WORLDBENCH: A CHALLENGING AND VISUALLY DIVERSE MULTIMODAL REASONING BENCHMARK

Anonymous authors
Paper under double-blind review

ABSTRACT

In real-world applications, models are expected to perform reliably across diverse settings. Yet, many existing multimodal benchmarks expand task types without capturing the visual diversity needed to handle open-ended visual inputs. We present WorldBench, a challenging and visually diverse reasoning benchmark to evaluate Multimodal Large Language Models (MLLMs). We build a taxonomy of thousands of visual concepts across multiple domains (*e.g.*, living things). Guided by this taxonomy, we curate a broad collection of images from search engines and existing datasets to comprehensively represent the visual world. Through structured trial-and-error, we manually design challenging questions that frontier MLLMs fail to answer. Quantitative and human evaluations provide complementary evidence that WorldBench is among the most visually diverse benchmarks we evaluate, while each evaluation has distinct limitations. Evaluating 15 MLLMs on WorldBench reveals weaknesses in visual understanding: even the strongest model reaches only 64.0% accuracy, while some models perform marginally above chance-level. We hope our work highlights the importance of visual diversity in building multimodal benchmarks.

1 INTRODUCTION

Comprehensively evaluating multimodal Large Language Models (MLLMs) requires benchmarks that capture the richness and diversity of the visual world. In recent years, many benchmarks (Yue et al., 2024; Fu et al., 2025a; Liu et al., 2024; Li et al., 2024) have been introduced as *standalone test suites* to evaluate models across different capabilities. Most of them follow a *task-centric* construction paradigm: they first design a set of task categories (*e.g.*, object recognition, OCR) and then build questions to cover each category.

However, this task-oriented approach overlooks the visual diversity of the images. The variety of the images determines what types of questions can be asked and, therefore, which model abilities can be evaluated. We introduce WorldBench, a reasoning benchmark that assesses multimodal understanding using a diverse image set paired with questions that are intuitive for humans yet challenging for models. As shown in Figure 1, WorldBench stands apart from existing diverse benchmarks by challenging models to *see the whole world*.

To comprehensively capture the visual world, we first construct a large-scale taxonomy containing thousands of fine-grained visual concepts across 7 visual domains. This process is semi-automated with a Large Language Model (LLM) and involves light human effort. This taxonomy guides the curation of high-quality, visually diverse images from web search engines and existing datasets. Following Lin et al. (2014), we prioritize non-iconic images that depict rich contextual scenes rather than object-centric views. For each image, we design a question through a structured trial-and-error process, iteratively refining it until at least one frontier MLLM fails to answer. Overall, WorldBench contains 2,000 carefully curated questions.

To assess the visual diversity of WorldBench relative to existing benchmarks, we perform both quantitative and human evaluations. Specifically, we measure visual diversity using the effective rank and participation ratio of the feature covariance matrix computed from image embeddings extracted by pre-trained vision encoders. Across the evaluated encoders and metrics, WorldBench ranks highly, although the precise ordering depends on the representation and metric. We further conduct a user study in which participants compare two image panels sampled from different benchmarks and vote

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

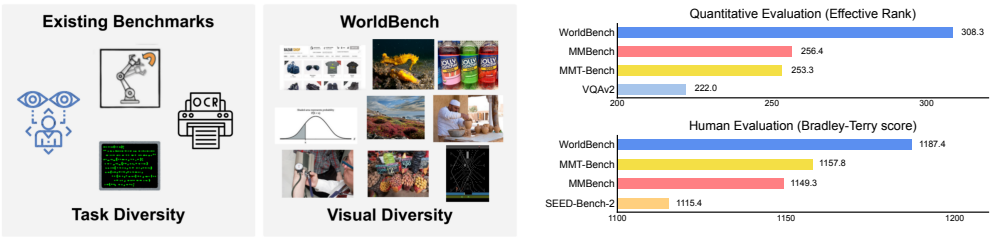


Figure 1. From task diversity to visual diversity. While existing multimodal benchmarks emphasize task diversity, WorldBench focuses on visual diversity. By covering images from different visual domains, WorldBench ranks highly across complementary quantitative and human evaluations of visual diversity, while remaining highly challenging for current MLLMs.

for the one they perceive as more diverse. We aggregate these pairwise votes into a global ranking using the Bradley–Terry model (Bradley & Terry, 1952), where WorldBench obtains the highest overall human-rated diversity score in this study.

We evaluate 15 MLLMs on WorldBench and find it very challenging for current models. Among all models, Gemini-3.1-Pro (Google, 2026) achieves the highest accuracy of 64.0% averaged across all domains, while the best-performing open-source model, Qwen3.5-VL-27B (Qwen, 2026), attains 56.6%. Several models remain only moderately above chance-level guess.

Moreover, we find that no model attains over 75% accuracy on questions in any domain. Current models fail on many simple cases. They often struggle with fine-grained perception (*e.g.*, counting) and make ungrounded inferences without closely looking at the image. We also find that while reasoning models outperform non-reasoning variants, increasing the number of reasoning tokens does not always improve performance on WorldBench, and this trend is observed across all domains.

2 WorldBench

Existing multimodal benchmarks (Fu et al., 2025a; Ying et al., 2024; Chen et al., 2025) often emphasize task diversity, but rarely consider the visual diversity of their images during construction. We present WorldBench, a new reasoning benchmark to challenge multimodal Large Language Models (MLLMs) to *see the whole world* by evaluating them on a broad set of visually diverse images across 7 domains, including **Living Things**, **Objects**, **Scenes**, **Digital World**, **Academics**, **Documents, Charts, & Tables**, and **Agents**. WorldBench comprises 2,000 questions that are challenging for frontier MLLMs but natural and intuitive for humans.

Figure 2 illustrates our three-step procedure to build WorldBench: (1) to capture the diversity of the visual world, we construct a large-scale taxonomy covering a spectrum of visual concepts with the help of a Large Language Model (LLM); (2) we curate a diverse set of images by selecting a high-quality sample for each concept; and (3) we manually propose difficult questions that frontier MLLMs answer incorrectly.

2.1 Building a Comprehensive Taxonomy

In this section, we discuss why existing taxonomies are insufficient for curating diverse images and introduce our taxonomy.

Limitations in existing taxonomies. Taxonomies have long been used to curate image datasets (Griffin et al., 2007; Deng et al., 2009; Everingham et al., 2010; Lin et al., 2014), but they are often centered on tangible, real-world entities and underrepresent web-native visual content such as webpages, screenshots, games, charts, and agent environments.

LLM-guided taxonomy construction. Manually constructing a diverse taxonomy is time-consuming, so we construct ours with LLM assistance and light human-in-the-loop effort. Starting from seven high-level domains—**Living Things**, **Objects**, **Scenes**, **Digital World**, **Academics**, **Documents, Charts, & Tables**, and **Agents**—we prompt an LLM to propose candidate subdomains, manually prune redundant or ambiguous entries, and merge semantically similar ones. We repeat this generate-

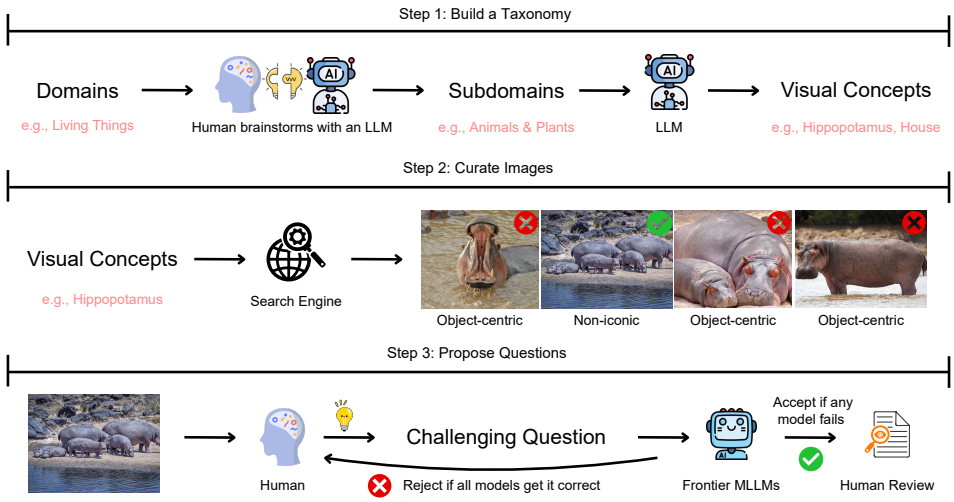


Figure 2. Three-step procedure to build WorldBench. Top: We construct a comprehensive taxonomy by defining a few high-level domains, expanding each into multiple subdomains, and then further into fine-grained concepts. **Middle:** We curate a high-quality image for each concept from search engines, prioritizing non-iconic and less object-centric images. **Bottom:** For each image, we design a challenging question that at least one frontier model answers incorrectly.

refinement cycle until no major new subdomains appear, then ask the LLM to expand each subdomain into fine-grained visual concepts.

For **Agents**, we additionally define robotics, games, and web-agent subdomains. Games and web agents follow the same LLM-assisted procedure, while robotics requires a different source because first-person images of robotic environments are difficult to find through web search. We therefore derive robotics concepts and images from AgentVQA (Anonymous, 2025), using an MLLM to identify leaf concepts that cover its major robotic scenarios followed by human verification. Figure 3 visualizes the resulting 2,000-concept taxonomy.

The taxonomy spans concrete real-world entities and more abstract visual content: concepts range from animals and plants to events, digital interfaces, scientific diagrams, charts, documents, games, and web-agent tasks. This breadth lets us sample images from many visual domains rather than repeatedly drawing from a small set of object-centric categories.

2.2 Curating Images from Taxonomy

Since the taxonomy defines the target visual concepts for our benchmark, we then collect high-quality images corresponding to these concepts from reliable sources.

Collecting diverse images from search engines. Following classic vision datasets (Deng et al., 2009; Lin et al., 2014), we use search engines as the primary image source. For each visual concept, we query Google and Bing to obtain a pool of candidate images, then select one high-quality image that best matches the concept. If the initial candidates do not contain a suitable image, we continue querying until a good match is found.

We observe that search engines tend to return object-centric, close-up images, particularly for real-world domains (*e.g.*, **Living Things** and **Objects**). We avoid overusing such images, as they reduce visual diversity. Similar to Lin et al. (2014), we prioritize non-iconic images (or non-canonical perspectives (Palmer, 1981)) with richer contexts and multiple objects.

For robotics concepts in **Agents**, we instead select representative frames from AgentVQA with MLLM assistance and human verification. This preserves the visual diversity of robot-centric images in the original benchmark while avoiding many near-duplicate frames from the same source.

Comparing to images in other diverse benchmarks. In total, WorldBench consists of 2,000 images. Figure 4 shows uncurated random samples from WorldBench along with those from existing diverse

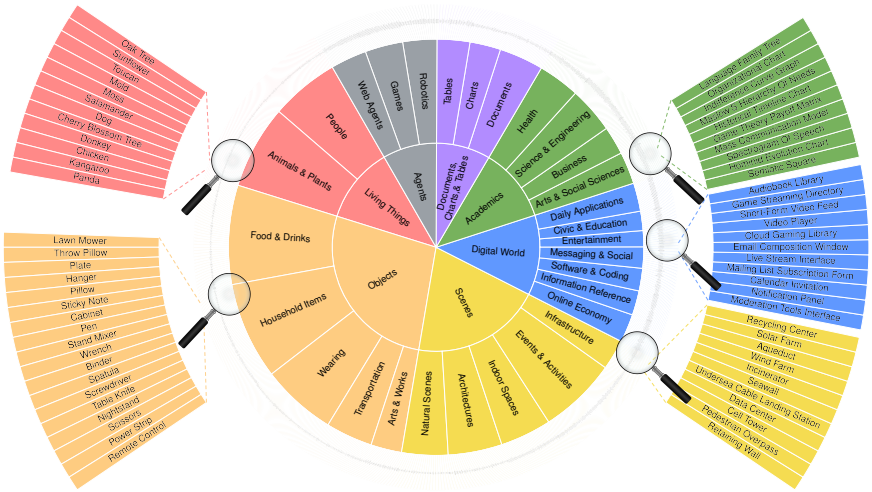


Figure 3. Taxonomy of WorldBench. We construct a large-scale taxonomy covering 2,000 visual concepts across 7 domains: **Living Things**, **Objects**, **Scenes**, **Digital World**, **Academics**, **Documents, Charts, & Tables**, and **Agents**.

benchmarks. MMMU (Yue et al., 2024) and MEGA-Bench (Chen et al., 2025) feature textbook-style charts and diagrams, while images in MMBench (Liu et al., 2024) often depict a single centered object with minimal context. In contrast, WorldBench offers broad visual coverage and a more balanced distribution across concepts. We quantitatively compare its visual diversity with existing diverse benchmarks in Section 3.

2.3 Proposing Challenging Questions

Now that we have curated a diverse set of images across domains, we describe how we design difficult questions for frontier MLLMs and perform human review to control annotation error. We also highlight the key features that set WorldBench apart from other challenging multimodal benchmarks.

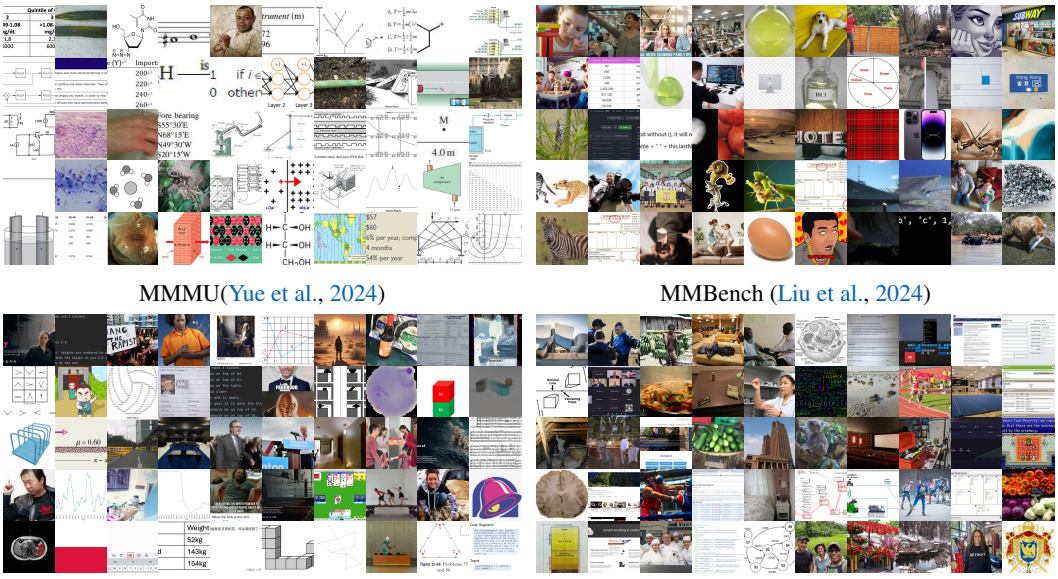
Designing challenging questions against models. We adopt a structured trial-and-error procedure to design difficult questions that require models to reason carefully before providing answers. We draft an initial question for each image and evaluate it on several frontier models (*i.e.*, GPT-5.4-Thinking (low) (OpenAI, 2026), Gemini-3.1-Pro (Google, 2026), Qwen3.5-VL-Plus-Instruct (Qwen, 2026), and Kimi-K2.5 (Kimi, 2026)). The question is iteratively revised until at least one model answers incorrectly. During benchmark construction, newer and more capable models are released, so different questions are evaluated against different sets of models; Table 2 reports these details.

We place only one constraint in question design: each question must be a four-option multiple-choice question. Note that the question is not required to be tied to the visual concept used to retrieve the image. We also ask annotators to optionally provide a concise explanation of the correct answer for questions they consider challenging and require substantial reasoning to arrive at the final answer. Appendix A illustrates our question annotation interface, along with the tools provided to annotators in designing questions.

Controlling annotation error. Frontier models may fail to answer certain questions not because the questions are difficult, but due to low-quality annotation (*e.g.*, ambiguous wording, multiple valid answers, or incorrect image descriptions). To mitigate this, we conduct several rounds of manual review of all questions and their explanations. We refine unclear language to eliminate alternative interpretations and verify that each question admits only one valid correct answer.

We also use Claude Code to assist question-quality review with light human oversight. Based on issues identified from volunteer audits, we ask Claude Code to flag and propose revisions for problematic questions, adopting changes only after manual verification.

Sample questions in WorldBench. Figure 5 shows 6 representative question samples in WorldBench along with the answers chosen by frontier models. While these questions are *intuitive* for humans,



MMMU(Yue et al., 2024) MMBench (Liu et al., 2024)

MEGA-Bench (Chen et al., 2025) WorldBench

Figure 4. Uncurated random image samples. Existing benchmarks often emphasize object-centric images or charts, while WorldBench spans real-world scenes, web content, games, and other visual domains.

frontier models often answer them incorrectly. Our questions are very different from those in other challenging reasoning benchmarks. For example, ZeroBench (Roberts et al., 2025) constructs its questions by chaining several independent sub-questions into a single, complex one (*e.g.*, count the chairs, count the letters, then multiply the two). In contrast, our questions are designed to be *more natural* and *closer to real-world settings*.

3 Measuring Visual Diversity of Images

We evaluate the visual diversity of the image set in WorldBench and compare it against existing benchmarks using quantitative and human evaluations. Quantitatively, we measure visual diversity through the covariance structure of image embeddings extracted by pre-trained vision encoders. We also conduct a user study to assess how humans perceive visual diversity across benchmarks.

For comparison, we include widely used multimodal benchmarks that emphasize diversity, including VQAv2 (Goyal et al., 2017), MME (Fu et al., 2025a), MMStar (Chen et al., 2024), MMBench (Yu et al., 2024a), MMMU (Yue et al., 2024), SEED-Bench-2 (Li et al., 2023a), MMT-Bench (Ying et al., 2024), and MEGA-Bench (Chen et al., 2025). Appendix D.2 details how we process images from each benchmark.

Vision encoders (Radford et al., 2021; Zhai et al., 2023; Oquab et al., 2024) are pre-trained on millions or even billions of images (Thomee et al., 2016; Chen et al., 2023; Gadre et al., 2023). As a result, they learn rich and generalizable visual representations that capture the visual world and we use them to measure visual diversity.

3.1 Quantitative Evaluation

Methods. We measure visual diversity from the covariance spectrum of image embeddings. For each benchmark, we randomly sample $N = 1000$ images, extract ℓ_2-normalized embeddings using a pre-trained vision encoder, and compute two spectrum-based diversity metrics: effective rank (Roy & Vetterli, 2007), equivalent to the Vendi Score (Friedman & Dieng, 2023) with cosine similarity, and participation ratio (Wiedmann et al., 2025). Intuitively, both metrics are higher when image embeddings spread across many independent directions rather than collapsing onto a few dominant modes. Full definitions are provided in Appendix D.1.

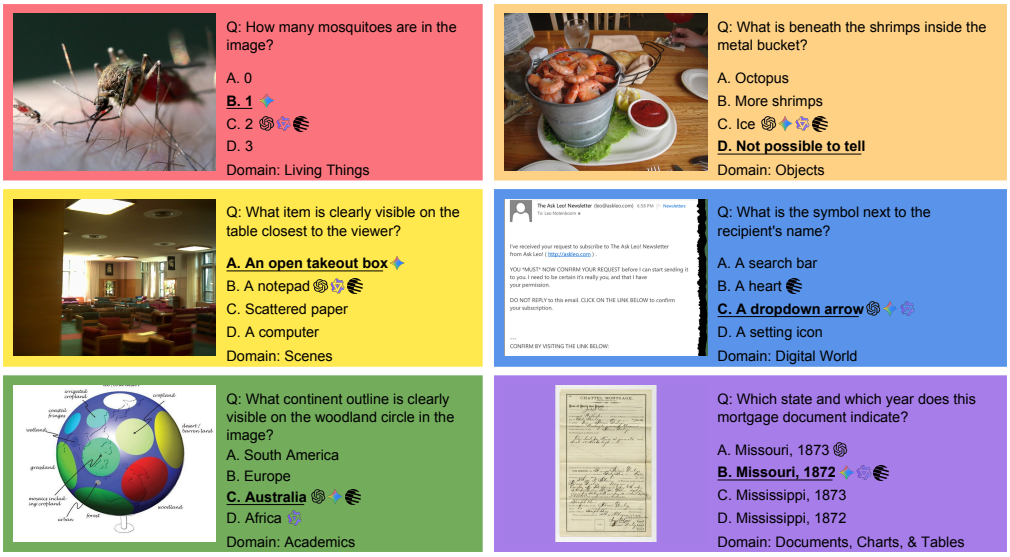


Figure 5. Question samples from WorldBench. We show representative questions with answers selected by frontier MLLMs. Ground-truth answers are highlighted in **bold and underline**.

Vision encoders. For a comprehensive evaluation, we use multiple pre-trained vision encoders—SigLIP 2 (Tschannen et al., 2025), Perception Encoder (Bolya et al., 2025), and DINO v3 (Siméoni et al., 2025). Each model is trained on distinct data distributions (*i.e.*, WebLI (Chen et al., 2023), MetaCLIP (Xu et al., 2024a), and Instagram images + ImageNet (Deng et al., 2009)), therefore offering different views of the visual world.

Results. Figure 6 shows the results. WorldBench often achieves the highest or second-highest effective rank and participation ratio across three vision encoders. These results place WorldBench among the most visually diverse benchmarks under the evaluated representations and metrics, while the precise ordering remains encoder- and metric-dependent.

We also observe that diversity rankings vary considerably across models. For example, MMBench obtains the highest value in participation ratio among all benchmarks under SigLIP 2, yet drops to the fourth place under Perception Encoder. This suggests that different encoders capture distinctive visual representations. Appendix B discusses the limitations of using a pre-trained vision encoder as a general visual diversity metric.

3.2 Human Evaluation

We conduct a human evaluation to compare benchmarks in visual diversity. We design a user interface that displays two image panels side by side, each containing 100 images randomly sampled from a benchmark. Users are asked to vote for the panel that *is visually more diverse*. Below, we describe how we collect user responses and how we aggregate them to produce an absolute ranking of visual diversity across benchmarks. Our diversity evaluation is conceptually similar to Chatbot Arena (Chiang et al., 2024): instead of choosing a more preferred model response, users select the image panel they perceive as visually more diverse.

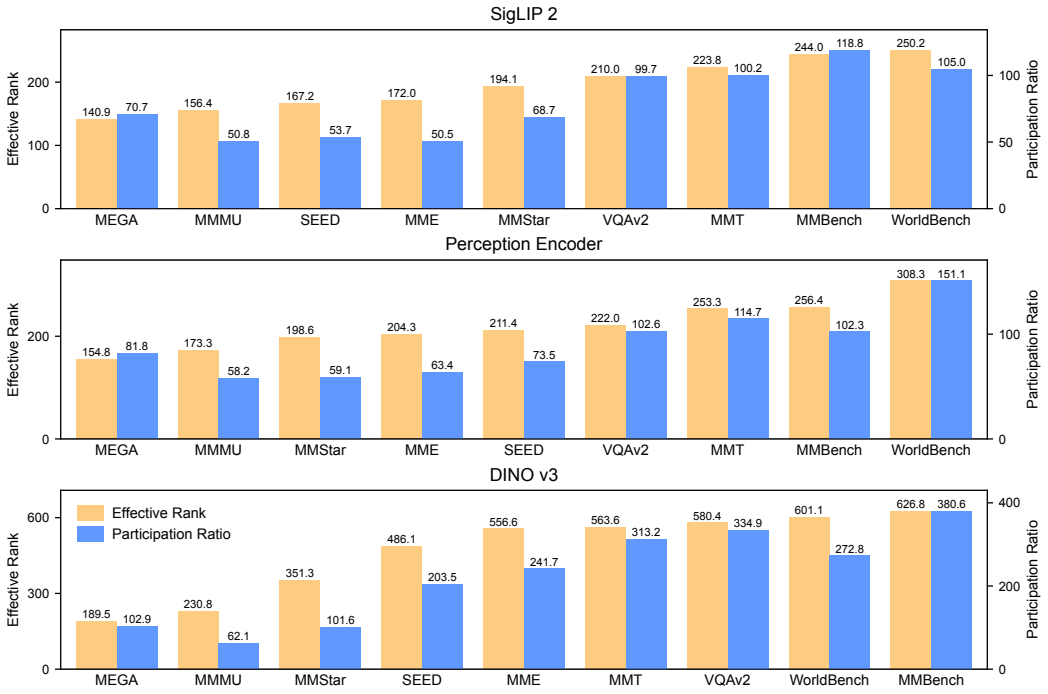


Figure 6. Quantitative diversity evaluation. WorldBench often ranks the first or second in effective rank (Roy & Vetterli, 2007) and participation ratio with embeddings extracted by different pre-trained vision encoders. This indicates better diversity under varying visual representations (*i.e.*, the distinct visual world captured by each encoder).

Aggregation. Following (Chiang et al., 2024), we aggregate pairwise comparisons with a Bradley-Terry model (Bradley & Terry, 1952) and estimate 95% confidence intervals with bootstrap resampling (DiCiccio & Efron, 1996). Additional details are provided in Appendix C.1.

Users. We collected 360 pairwise comparison results from 12 volunteers participating in our study. None of them were involved in image curation or had any prior knowledge of our benchmark before the test. Each user completes 30 rounds of comparisons. We do not impose any time limits for a round or the entire test, and only recommend spending at least one minute per round.

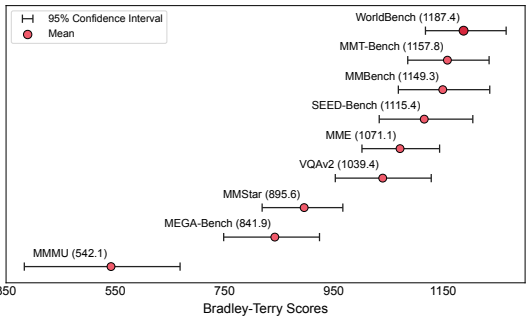


Figure 7. Human evaluation on visual diversity. Bradley–Terry scores with 95% confidence intervals estimated from pairwise human comparisons. WorldBench receives the highest human-rated diversity score.

Results. Figure 7 visualizes Bradley–Terry scores estimated from pairwise human comparisons. WorldBench achieves the highest score in this study, indicating that these participants judge its image set to be more visually diverse than the comparison benchmarks.

4 Model Performance on WorldBench

Models. We evaluate the performance of 15 MLLMs on WorldBench. For proprietary models, we consider GPT-5.4-Thinking (OpenAI, 2026), Gemini-3.1-Pro (Google, 2026), Gemini-3-Flash (Google, 2025), Claude-Opus-4.7 (Anthropic, 2026), Grok-4.2 (xAI, 2026), and Qwen3.5-VL-Plus-Thinking/Instruct (Qwen, 2026). These models are accessed through their official APIs. For

Table 1. Performance of MLLMs on WorldBench. Bold marks the highest value and underline the second-highest in each column. *: Models used during question proposal.

Models	Living	Objects	Scenes	Digital	Academics	DCT	Agents	Average
Proprietary Models								
Grok-4.2	49.8	53.6	49.2	59.7	53.9	57.0	54.2	53.3
Claude-Opus-4.7	46.1	52.8	51.3	59.3	57.0	58.7	56.0	53.7
GPT-5.4-Thinking (low)*	51.8	48.4	50.3	58.1	53.9	61.6	58.4	53.0
GPT-5.4-Thinking (high)	55.5	53.8	54.3	67.3	60.1	64.5	**63.9**	58.2
Qwen3.5-VL-Plus-Instruct*	45.7	49.4	45.1	55.2	47.8	49.1	51.2	48.7
Qwen3.5-VL-Plus-Thinking	56.3	**62.8**	55.0	58.9	60.1	60.2	63.3	59.3
Gemini-3-Flash	56.7	60.7	57.5	66.9	63.2	72.7	63.3	61.8
Gemini-3.1-Pro*	**62.0**	60.7	**61.4**	**71.4**	**64.0**	**73.8**	62.0	**64.0**
Open-source Models								
Gemma-4-E4B	34.3	33.6	29.7	38.7	36.8	39.0	37.3	34.6
Gemma-4-31B	49.4	47.4	46.0	51.2	54.4	52.9	54.8	49.7
InternVL-3.5	38.4	42.5	44.6	39.1	38.6	42.4	38.6	41.2
GLM-4.6V	47.8	41.1	43.4	38.7	40.4	42.4	44.6	42.5
Kimi-K2.5*	49.0	52.0	49.7	56.9	51.3	59.3	54.8	52.5
Qwen3.5-VL-35B-A3B	48.6	52.6	52.2	50.4	57.0	59.3	54.2	52.9
Qwen3.5-VL-27B	50.6	57.5	53.3	58.5	61.8	57.6	60.2	56.6

open-source models, we consider Qwen3.5-VL-35B-A3B (Qwen, 2026), Qwen3.5-VL-27B (Qwen, 2026), Kimi-K2.5 (1T) (Kimi, 2026), GLM-4.6V (106B) (GLM, 2025a), Gemma-4 (31B/E4B) (DeepMind, 2026), and InternVL3.5 (241B) (Wang et al., 2025). We run offline inference with vLLM (Kwon et al., 2023) for all open-source models except Kimi-K2.5, evaluated with API due to its size.

Evaluation setup. We use a unified prompt across all models to format the multiple choice questions, encouraging each model to provide a brief explanation before outputting its final answer choice. Final answers are extracted using regular expressions based on a fixed set of parsing rules. For generation, we use the default sampling parameters (e.g., temperature, top-k, and top-p) specified in each model's official configuration file or API. Additional evaluation details are provided in Appendix D.3.

4.1 Results

Table 1 reports the performances of all evaluated models on WorldBench. The results show that our benchmark presents a great challenge for both proprietary and open-source models. The strongest proprietary model (*i.e.*, Gemini-3.1-Pro) achieves an average accuracy of 64.0% across all domains, while the best open-source model (*i.e.*, Qwen3.5-VL-27B) attains 56.6%. Interestingly, a few open-source models (*e.g.*, Qwen3.5-VL-27B) outperform proprietary models (*e.g.*, Claude-Opus-4.7 and Grok-4.2) that are often considered more capable.

No single model achieves accuracy >75% on any domain. The performance gap between models within each domain is also substantial. For instance, Gemini-3.1-Pro reaches 73.8% accuracy on **Documents, Charts, & Tables**, while Gemma-4-E4B reaches only 39.0% on the same domain, only moderately above chance-level guess.

4.2 Analysis

Benefits of reasoning. Recent MLLMs (Bai et al., 2025; GLM, 2025b; Kimi, 2025) have adopted Reinforcement Learning in post training (Shao et al., 2024) to enable the generation of intermediate *Chain-of-Thought* reasoning steps (Wei et al., 2022) before producing final answers. To understand whether longer reasoning improves performance on WorldBench, we vary the reasoning budget of GPT-5.4 from none to high and measure accuracy across all domains. As shown in Figure 8, accuracy consistently improves from no reasoning to a low budget across all domains. Beyond a low budget, the trends diverge: **Digital World**, **Objects**, **Scenes**, and **Agents** continue to climb monotonically through the high budget, while **Documents, Charts, & Tables**, **Academics**, and **Living Things** saturate or regress at higher budgets. We find that generating more reasoning tokens can result in the model getting stuck in a reasoning loop.

Correlation with other benchmarks. To understand how the capabilities assessed by WorldBench differ from other benchmarks, we compute the Pearson correlation between model accuracies on WorldBench and other benchmarks (using the local models in Table 1). We select three benchmarks

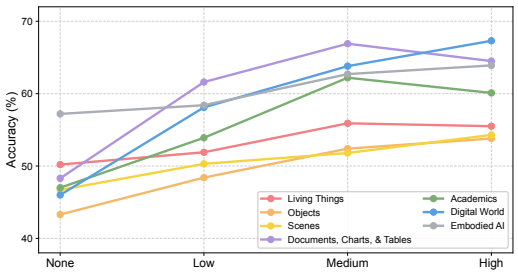


Figure 8. Performance with longer reasoning. Increasing reasoning budget does not always increase performance on WorldBench, with even slight decrease on some domains.

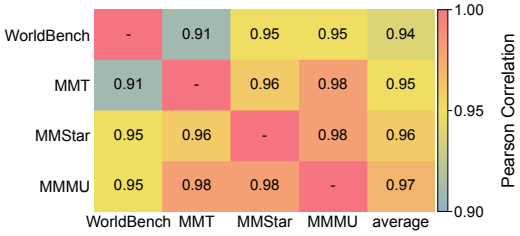


Figure 9. WorldBench shows the lowest average correlation with other diverse benchmarks. We compute the pairwise correlation of model accuracies between each pair of benchmarks.

from Section 3—MMMU, MMStar, and MMT-Bench—that have different levels of visual diversity according to human evaluation (Figure 7) and are not saturated while remaining manageable in size. In Figure 9, WorldBench shows the lowest average correlation of 0.94 with others. This suggests that WorldBench's emphasis on visual diversity offers a different dimension to evaluate MLLMs.

5 Related Work

Multimodal large language models. The capabilities of multimodal Large Language Models (MLLMs) have drastically improved over the past years. Proprietary models (OpenAI, 2026; Google, 2026; Anthropic, 2026; xAI, 2026) now achieve unprecedented performance on a wide range of visual understanding benchmarks. On the open-source side, progress has been equally rapid from early pioneers (Dai et al., 2023; Awadalla et al., 2023; Liu et al., 2023) to more recent models (Qwen, 2026; GLM, 2025a; Wang et al., 2025) that are increasingly competitive with frontier proprietary systems. Despite these advances, existing benchmarks (Yue et al., 2024; Tong et al., 2024) often evaluate models on image sets with limited visual diversity. Our work addresses this gap by evaluating MLLMs under a much broader and visually more diverse collection of images.

Multimodal benchmarks. Many multimodal benchmarks target specific image understanding capabilities of MLLMs, such as visual-centric perception tasks (Paiss et al., 2023; Tong et al., 2024; Fu et al., 2024; Tong et al., 2025), OCR (Wang et al., 2024; Ma et al., 2024; Fu et al., 2025b), object hallucination (Li et al., 2023b; Guan et al., 2024), GUI navigation (Li et al., 2025; Cheng et al., 2024; Xie et al., 2025; Rawles et al., 2024), and coding (Si et al., 2025; Yang et al., 2025). Beyond such task-specific benchmarks, various benchmarks (Xu et al., 2024b; Li et al., 2024; Yu et al., 2024b; Liu et al., 2024; Ying et al., 2024; Fu et al., 2025a) aim to provide broader task coverage by comprehensively evaluating MLLMs across a diverse set of tasks. In contrast to all prior work, WorldBench departs from a task-centric design and is instead built around *visual diversity*.
Image inspection and high-resolution reasoning. ZoomEye (Shen et al., 2024) uses training-free, tree-based image exploration to help MLLMs inspect fine-grained regions in high-resolution images. Such mechanisms may mitigate some failures on visually detailed questions, complementing our evaluation of current end-to-end systems.

6 Conclusion

This work introduces WorldBench, a visually diverse reasoning benchmark to challenge MLLMs by seeing the whole world. We build a taxonomy with thousands of visual concepts, curate a diverse image set from it, and design questions through a structured trial-and-error process that targets failure modes of frontier MLLMs. Quantitative and human evaluations provide complementary evidence that WorldBench is among the most visually diverse benchmarks we evaluate, while each evaluation has distinct limitations. Furthermore, our model evaluations reveal that WorldBench is very challenging for current models. We hope our work motivates building datasets that prioritize visual diversity.

Reproducibility Statement

The anonymized WorldBench dataset and evaluation code are available for review. Section 2.1 describes taxonomy construction, and Appendix A documents the question annotation interface. Appendix D describes the diversity metrics, benchmark preprocessing, inference settings, prompt template, and answer parsing; Appendix C.1 gives the human-evaluation aggregation and bootstrap procedure.

Ethics Statement

WorldBench draws images from web search results and existing datasets. As described in Appendix D.2, web-search results are filtered for Creative Commons-licensed images, and images and questions are manually reviewed to remove unsafe, private, or otherwise inappropriate content before release. The benchmark may nevertheless reflect biases in its sources, and performance on it should not be treated as sufficient evidence of safe or fair deployment. The human diversity evaluation involved volunteers comparing image panels and did not collect sensitive personal information.

References

Anonymous. AgentVQA: A unified benchmark for agentic visual understanding. In *Submitted to ICLR*, 2025. URL https://openreview.net/forum?id=aZFjp5wqck.

Anthropic. Introducing claude opus 4.7. *https://www.anthropic.com/news/claude-opus-4-7*, 2026.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, et al. Qwen3-vl technical report. *arXiv preprint arXiv: 2511.21631*, 2025.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. In *NeurIPS*, 2025.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 1952.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021.

Jiacheng Chen, Tianhao Liang, Sherman Siu, Zhengqing Wang, Kai Wang, Yubo Wang, Yuansheng Ni, Ziyan Jiang, Wang Zhu, Bohan Lyu, Dongfu Jiang, Xuan He, Yuan Liu, Hexiang Hu, Xiang Yue, and Wenhu Chen. MEGA-bench: Scaling multimodal evaluation to over 500 real-world tasks. In *ICLR*, 2025.

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? In *NeurIPS*, 2024.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model. In *ICLR*, 2023.

Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Li YanTao, Jianbing Zhang, and Zhiyong Wu. Seeclick: Harnessing gui grounding for advanced visual gui agents. In *ACL*, 2024.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *ICML*, 2024.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023.

Google DeepMind. Gemma 4: Our most intelligent open models, built from gemini 3 research and technology to maximize intelligence-per-parameter. https://deepmind.google/models/gemma/gemma-4/, 2026.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

Thomas J DiCiccio and Bradley Efron. Bootstrap confidence intervals. *Statistical science*, 1996.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *TMLR*, 2023.

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 *NeurIPS Datasets and Benchmarks Track*, 2025a.

Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, Jingqun Tang, Wei Chen, Lianwen Jin, Yuliang Liu, and Xiang Bai. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. 2025b.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *ECCV*, 2024.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. In *NeurIPS Datasets and Benchmarks Track*, 2023.

GLM. Glm-4.6v: Open source multimodal models with native tool use. https://z.ai/blog/glm-4.6v, 2025a.

GLM. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv: 2507.01006*, 2025b.

Google. Gemini 3 pro best for complex tasks and bringing creative concepts to life. https://deepmind.google/models/gemini/pro/, 2025.

Google. Gemini 3.1 pro: A smarter model for your most complex tasks. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/, 2026.

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 *CVPR*, 2017.

Gregory Griffin, Alex Holub, Pietro Perona, et al. Caltech-256 object category dataset. *Technical Report*, 2007.

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 *CVPR*, 2024.

Kimi. Kimi-vl technical report. *arXiv preprint arXiv: 2504.07491*, 2025.

Kimi. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench-2: Benchmarking multimodal large language models. *arXiv preprint arXiv:2311.17092*, 2023a.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. In *CVPR*, 2024.

Kaixin Li, Ziyang Meng, Hongzhan Lin, Ziyang Luo, Yuchen Tian, Jing Ma, Zhiyong Huang, and Tat-Seng Chua. Screenspot-pro: Gui grounding for professional high-resolution computer use. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

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 *ECCV*, 2024.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. In *NeurIPS*, 2024.

OpenAI. Introducing gpt-5.4. https://openai.com/index/introducing-gpt-5-4/, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *TMLR*, 2024.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Stephen E Palmer. Canonical perspective and the perception of objects. *Attention and performance*, 1981.

Qwen. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5, 2026.

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 *ICML*, 2021.

Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William Bishop, Wei Li, Folawiyo Campbell-Ajala, et al. Androidworld: A dynamic benchmarking environment for autonomous agents. *arXiv preprint arXiv:2405.14573*, 2024.

Jonathan Roberts, Mohammad Reza Taesiri, Ansh Sharma, Akash Gupta, Samuel Roberts, Ioana Croitoru, Simion-Vlad Bogolin, Jialu Tang, Florian Langer, Vyas Raina, et al. Zerobench: An impossible visual benchmark for contemporary large multimodal models. *arXiv preprint arXiv:2502.09696*, 2025.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *European Signal Processing Conference*, 2007.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS Datasets and Benchmarks Track*, 2022.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Haozhan Shen, Kangjia Zhao, Tiancheng Zhao, Ruochen Xu, Zilun Zhang, Mingwei Zhu, and Jianwei Yin. Zoomeye: Enhancing multimodal llms with human-like zooming capabilities through tree-based image exploration. *arXiv preprint arXiv:2411.16044*, 2024.

Chenglei Si, Yanzhe Zhang, Ryan Li, Zhengyuan Yang, Ruibo Liu, and Diyi Yang. Design2code: Benchmarking multimodal code generation for automated front-end engineering. In *NAACL*, 2025.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 2016.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, 2024.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *NeurIPS*, 2025.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022.

Luis Wiedmann, Orr Zohar, Amir Mahla, Xiaohan Wang, Rui Li, Thibaud Frere, Leandro von Werra, Aritra Roy Gosthipaty, and Andrés Marafioti. Finevision: Open data is all you need. *arXiv preprint arXiv:2510.17269*, 2025.

xAI. Grok 4.2. https://docs.x.ai/developers/models, 2026.

Tianbao Xie, Jiaqi Deng, Xiaochuan Li, Junlin Yang, Haoyuan Wu, Jixuan Chen, Wenjing Hu, Xinyuan Wang, Yuhui Xu, Zekun Wang, et al. Scaling computer-use grounding via user interface decomposition and synthesis. *arXiv preprint arXiv:2505.13227*, 2025.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. In *ICLR*, 2024a.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *PAML*, 2024b.

Cheng Yang, Chufan Shi, Yaxin Liu, Bo Shui, Junjie Wang, Mohan Jing, Linran Xu, Xinyu Zhu, Siheng Li, Yuxiang Zhang, Gongye Liu, Xiaomei Nie, Deng Cai, and Yujiu Yang. Chartmimic: Evaluating lmm's cross-modal reasoning capability via chart-to-code generation. In *ICLR*, 2025.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, Jiayi Lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. Mmt-bench: a comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. In *ICML*, 2024.

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 *ICML*, 2024a.

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 *ICML*, 2024b.

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 *CVPR*, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023.

A Question Annotation Interface

Figure 10 illustrates our interface to design challenging questions. To help annotators in proposing questions, we provide a "Generate Image Facts" button that queries an LLM (*e.g.*, GPT-5) to extract facts from an image. Annotators can also come up with questions in their own way. For each question, annotators are required to use the "Run Model Evaluation (4 Models)" button to check that at least one frontier model answers the question incorrectly. For questions that may confuse others or be mistakenly judged as incorrect, annotators need to provide a brief explanation clarifying the correct answer. The interface also includes an "Improve via GPT-5" button that automatically refines the wording of the question, answer choices, and explanation using an LLM (*e.g.*, GPT-5).

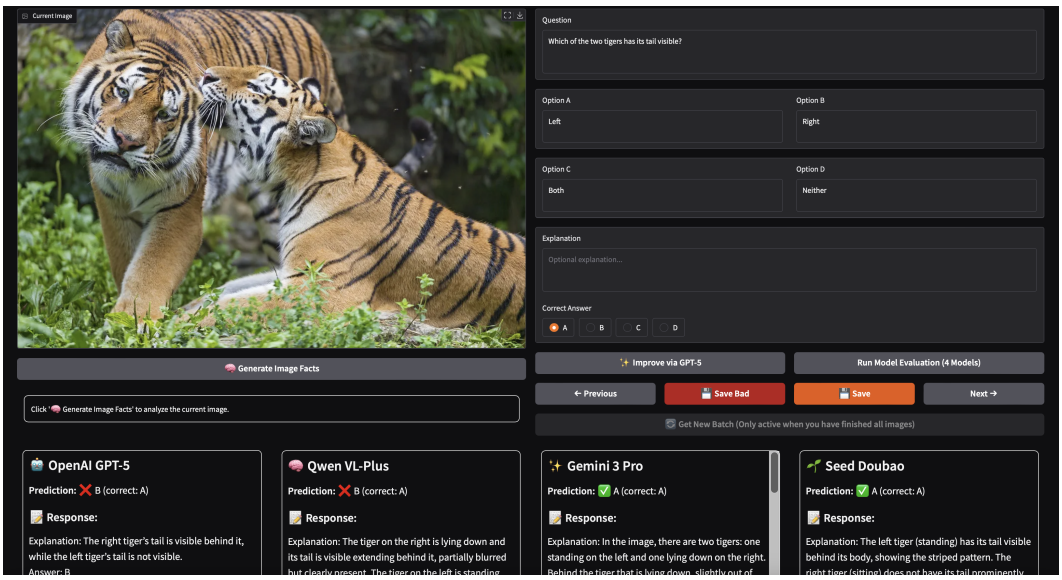


Figure 10. Question annotation interface. Annotators use the interface to examine each image, propose a question, evaluate it against frontier models, and refine the question wording as well as explanation with an LLM.

Table 2. Number of questions evaluated against each set of frontier models. Since newer and more capable models are released during benchmark construction, different questions are evaluated against different sets of models.

# of questions	models
321	GPT-5-Thinking, Gemini-2.5-Flash, Qwen3-VL-Plus-Instruct
352	GPT-5-Thinking, Gemini-2.5-Flash, Qwen3-VL-Plus-Instruct, Seed1.6-Vision
779	GPT-5-Thinking, Gemini-3-Pro, Qwen3-VL-Plus-Instruct, Seed1.6-Vision
542	GPT-5.4-Thinking, Gemini-3.1-Pro, Qwen3.5-VL-Plus-Instruct, Kimi K2.5

B Additional Results

In Section 3.1, we evaluate the visual diversity of images in existing multimodal benchmarks. We now extend this to several widely used pre-training datasets, including YFCC (Thomee et al., 2016), CC (Changpinyo et al., 2021), DataComp (Gadre et al., 2023), WIT (Srinivasan et al., 2021), LAION (Schuhmann et al., 2022), and ImageNet (Deng et al., 2009), to examine how different vision encoders assess their visual diversity. For each dataset, we randomly sample 1,000 images and compute visual diversity using effective rank and participation ratio across different pre-trained vision encoders. Table 3 reports the results.

We observe that different vision encoders rank the same datasets differently. SigLIP2 considers DataComp the most diverse, whereas DINO v3 rates ImageNet highest. This is likely influenced by differences in pre-training data distributions of these encoders. SigLIP2 is trained on WebLI, a proprietary dataset sourced from the Internet. This may lead its learned representations to be more

Table 3. Quantitative evaluation on pre-training datasets. We observe that vision encoders yield higher diversity scores on datasets that are similar to their pre-training distribution.

Dataset	SigLIP 2		Perception Encoder		DINO v3	
	Effective Rank	Participation Ratio	Effective Rank	Participation Ratio	Effective Rank	Participation Ratio
YFCC	17.9	103.9	27.7	133.7	572.1	540.5
CC	42.1	146.3	**50.9**	189.3	542.6	525.1
DataComp	**45.9**	**154.1**	49.7	**193.8**	516.8	511.9
WIT	19.8	78.0	29.3	108.2	453.2	293.1
LAION	48.5	146.8	49.3	170.7	452.1	393.1
ImageNet	24.9	107.1	30.2	120.3	**613.8**	**610.4**

aligned with other web-scale datasets, such as DataComp from the web archive data Common Crawl. In contrast, DINO v3 is trained on a mixture of Instagram images and ImageNet, with ImageNet comprising 1/10 of its pre-training data. As a result, this encoder assigns higher diversity scores to ImageNet. Overall, these results suggest that datasets more similar to an encoder's pre-training distribution tend to receive higher visual diversity scores under that encoder. More broadly, visual diversity is difficult to capture with any single metric or representation; our human evaluation provides a complementary signal, but future work should scale it to a larger and more demographically diverse set of judgments. WorldBench can help identify weaknesses in MLLMs across visually diverse settings, but because it is curated from web search results and existing datasets, it may reflect biases in those sources and should not be treated as sufficient evidence of safe or fair real-world deployment.

C User Study Interface for Human Evaluation

Figure 11 illustrates our interface to conduct human evaluation of visual diversity across benchmarks. For each comparison, we randomly sample 100 images from two benchmarks and display them side-by-side in a 20x5 image grid. Users can click any image to view it in full resolution. They are then asked to carefully examine all images on each side and choose the panel that is visually more diverse.

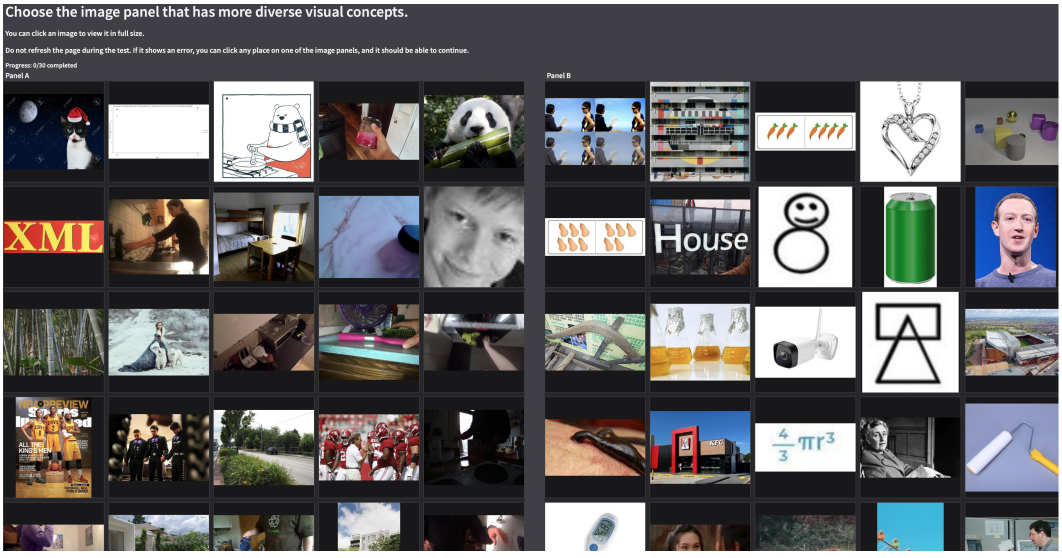


Figure 11. User study interface for human evaluation. We display two image panels side by side and ask users to choose the one that is visually more diverse.

C.1 Bradley-Terry Model in Human Evaluation

In Section 3.2, our human diversity evaluation collects pairwise comparisons between benchmarks. In each trial, a user is shown two image panels side by side, each containing 100 randomly sampled images from a benchmark, and is asked to select the panel that is visually more diverse.

Model formulation. Suppose there are K benchmarks indexed by $i \in \{1, \ldots, K\}$. The Bradley-Terry model (Bradley & Terry, 1952) assigns a score $\theta_i \in \mathbb{R}$ to each benchmark. The probability that benchmark i is preferred over benchmark j is

$$\Pr(i \succ j) = \frac{\exp(\theta_i)}{\exp(\theta_i) + \exp(\theta_j)} = \sigma(\theta_i - \theta_j), \tag{1}$$

where $\sigma(\cdot)$ is the sigmoid function.

Fitting model with maximum likelihood estimate. We collect N pairwise comparisons, where each comparison n involves a benchmark pair (i_n, j_n) and an outcome $y_n \in \{0, 1\}$. The outcome $y_n = 1$ indicates that benchmark i_n is considered visually more diverse than benchmark j_n, and $y_n = 0$ otherwise. Under the Bradley-Terry model, we estimate $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_K)$ by maximizing the log-likelihood

$$\mathcal{L}(\theta) = \sum_{n=1}^{N} \Big[y_n \log \Pr(i_n \succ j_n) + (1 - y_n) \log \Pr(j_n \succ i_n) \Big]. \tag{2}$$

Following (Chiang et al., 2024), the resulting scores are linearly rescaled to Elo-style ratings for readability.

Bootstrap confidence intervals. To quantify uncertainty, we compute bootstrap confidence intervals (DiCiccio & Efron, 1996). We resample pairwise comparisons with replacement, refit the Bradley-Terry model on each bootstrap sample, and record the resulting rating for each benchmark. During bootstrapping, we uniformly resample every benchmark pair to mitigate imbalance across comparisons in the original data. We perform 10,000 bootstrap rounds and report the mean rating with a 95% confidence interval defined by the 2.5th and 97.5th percentiles.

D Implementation Details

D.1 Diversity Metrics

Concretely, we randomly sample $N = 1000$ images from each benchmark and extract their ℓ_2-normalized embeddings $e_1, \ldots, e_N \in \mathbb{R}^d$ using a pre-trained vision encoder. Let $\bar{e}$ denote their empirical mean, and define the zero-centered feature matrix: $X = [\, e_1 - \bar{e},\ \ldots,\ e_N - \bar{e}\,] \in \mathbb{R}^{d \times N}$. We then compute the sample covariance matrix $C = \frac{1}{N} X X^\top$, and denote its eigenvalues by $\lambda_1, \ldots, \lambda_d \geq 0$.

To measure visual diversity, we use *effective rank* (Roy & Vetterli, 2007):

$$\mathrm{ER}(C) = \exp\left(-\sum_{i=1}^{d} p_i \log p_i \right), \quad p_i = \frac{\lambda_i}{\sum_{j=1}^{d} \lambda_j}. \tag{3}$$

This is equivalent to the Vendi Score (Friedman & Dieng, 2023) with cosine similarity. We also use *participation ratio*:

$$\mathrm{PR}(C) = \frac{\left(\sum_{i=1}^{d} \lambda_i \right)^2}{\sum_{i=1}^{d} \lambda_i^2}, \tag{4}$$

following (Wiedmann et al., 2025).

Intuitively, higher visual diversity leads to embeddings whose variance is distributed across many directions in feature space, rather than collapsing onto a few dominant ones. When this happens, the eigenvalues of C have a more uniform distribution, leading to higher effective rank and participation ratio.

D.2 Benchmarks Used in Diversity Evaluation

Removing similar images. Some benchmarks we evaluate in Section 3 contain identical images reused across different questions or multiple similar images within a single question. Using all

images in such cases would reduce their visual diversity and place them at a disadvantage relative to WorldBench.

Instead, we randomly sample only one image per question for each benchmark and then remove any *exact duplicate* within this set. For MEGA-Bench, we do not use any images from their in-context examples; we only sample from the images that models are actually evaluated on.

Benchmarks. We provide details on benchmarks we use in Section 3. We use the validation split of VQAv2 (Goyal et al., 2017), the full set of MME (Fu et al., 2025a), the full set of MMStar (Chen et al., 2024), the test split of MMBench (Yu et al., 2024a), the test split of MMMU (Yue et al., 2024), the full set of SEED-Bench-2 (Li et al., 2023a), the test split of MMT-Bench (Ying et al., 2024), and the full set of MEGA-Bench (Chen et al., 2025).

During WorldBench curation, we filter web-search results for Creative Commons-licensed images and manually review images and questions to remove unsafe, private, or otherwise inappropriate content before release.

D.3 Model Performance

Evaluation details. We report evaluation settings for proprietary models and open-source models. For proprietary models, we evaluate GPT-5.4-Thinking (OpenAI, 2026), Gemini-3.1-Pro (Google, 2026), Gemini-3-Flash (Google, 2025), Claude-Opus-4.7 (Anthropic, 2026), Grok-4.2 (xAI, 2026), and Qwen3.5-VL-Plus-Thinking/Instruct (Qwen, 2026). We use their default sampling settings to generate responses, except where a model exposes a reasoning-effort setting reported in the model name.

For open-source models, we run offline inference locally with vLLM (Kwon et al., 2023) for efficiency, except for Kimi-K2.5, which we evaluate through its API due to its size. Each local model evaluation uses approximately 4–8 NVIDIA A100 GPUs and takes roughly one hour, depending on model size. We use each local model's default or officially recommended generation parameters (*e.g.*, temperature, top-p, top-k). For reproducibility, we report the parameters used for local inference in Table 4.

Table 4. Sampling parameters for local open-source model inference. Kimi-K2.5 is excluded because it is evaluated through its API.

models	temperature	top-p	top-k
Qwen3.5-VL-35B-A3B (Qwen, 2026)	1.0	0.95	20
Qwen3.5-VL-27B (Qwen, 2026)	1.0	0.95	20
GLM-4.6V (GLM, 2025a)	0.8	0.6	2
Gemma-4-31B (DeepMind, 2026)	1.0	0.95	64
Gemma-4-E4B (DeepMind, 2026)	1.0	0.95	64
InternVL3.5 (Wang et al., 2025)	0.6	1.0	50

Prompt template and answer parsing. Figure 12 illustrates our prompt to format each question prompt and its answer choices. The prompt asks the model first to provide a brief explanation *before*

```
Question: {question_prompt}

Options:
A: {option_A}
B: {option_B}
C: {option_C}
D: {option_D}

Please answer the question and select the most appropriate answer from the given options.
First provide a brief explanation of your reasoning, then state your answer.

Format your response as:
Explanation: [Brief explanation of your reasoning]
Answer: [LETTER]
```

Figure 12. Prompt to format question and answer options. We encourage the model to produce a short explanation before giving the final answer. This yields slightly higher accuracy than asking for the answer before an explanation.

outputting the final answer. We find that this ordering slightly improves accuracy compared to asking for the final answer before an explanation. In the latter case, models sometimes change to a different answer during the explanation.

To robustly extract the model's chosen answer option, we apply three regular expression rules one by one. We first normalize the response to uppercase and search for an explicit "Answer: X" pattern, where X $\in$ {A, B, C, D}. If this fails, we fall back to parsing a single option letter at the beginning of the response, and finally to detecting a standalone option letter, anywhere in the output. If none of them match, we consider the response incorrect. Our parsing method achieves an average success rate of >99% across all questions in WorldBench. As a result, we do not rely on more advanced approaches (*e.g.*, LLM-as-a-judge).

E AI USE STATEMENT

LLMs assisted taxonomy expansion and refinement, extraction of image facts for annotators, and refinement of question, answer-choice, and explanation wording, as described in Section 2.1 and Appendix A. Human annotators curated concepts and images and verified the questions. Claude Code also assisted question-quality review by flagging and proposing revisions, which were adopted only after manual verification (Section 2.3). Models were also queried to identify challenging questions and to evaluate benchmark performance. Codex assisted manuscript editing, LaTeX formatting and build-script preparation, and layout checks. Beyond these uses, we did not use generative AI to develop hypotheses or research methodology, implement research methods, analyze data, interpret results, or produce theoretical claims, proofs, or translations. The authors take responsibility for the final content of this work, including AI-assisted text, claims, and artifacts.

F MODEL RESPONSES

We present example responses to WorldBench questions from all 15 models evaluated in Section 4. Figures 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, and 27 each illustrate one representative output from the corresponding model. We include exposed reasoning traces when they are available; some thinking models do not return their private reasoning traces through the API.



Q: What is the angle between the dump truck's bucket and the excavator's cab?

A. 30
B. 60
C. 90
D. 120

Domain: Objects

GPT-5.4-Thinking (high):

Explanation: The dump truck's bucket/bed appears roughly horizontal, while the excavator's cab is mostly vertical, so the angle between them is closest to a right angle.
Answer: C

Figure 13. Example response from GPT-5.4-Thinking (high) (OpenAI, 2026) on an Objects question.



Q: Where is this image likely taken?

A. Dessert store
B. Small supermarket
C. Coffee shop
D. Candy store

Domain: Digital World

GPT-5.4-Thinking (low):

Explanation: The image shows a checkout counter with Alipay and WeChat Pay signs, and behind it there appears to be café-style equipment such as a coffee/soft-serve machine and a service counter. This looks more like a café than a supermarket, candy store, or dedicated dessert shop.
Answer: C

Figure 14. Example response from GPT-5.4-Thinking (low) (OpenAI, 2026) on a Digital World question.



Q: Which color of hats do most people wear?

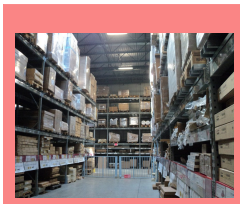
A. Gray
B. Black
C. Red
D. White

Domain: Objects

Gemini-3.1-Pro:

Explanation: Based on the image, the most prominent hat is worn by the person in the foreground, which is black. Another clearly visible person to the left is also wearing a black hat. While there are some lighter, straw-colored hats, among the given options (Gray, Black, Red, White), black is the clearly identifiable color worn by multiple people.
Answer: B

Figure 15. Example response from Gemini-3.1-Pro (Google, 2026) on an Objects question.



Q: What does the red sign in the far distance of the image indicate?

A. Firefighting equipment
B. Restricted area
C. Exit direction
D. Hazardous materials

Domain: Scenes

Gemini-3-Flash:

Explanation: In the far distance, above the blue metal gate at the end of the aisle, there is a small red sign. This sign is a standard "EXIT" sign used to indicate the direction of an emergency exit in a warehouse or public building.
Answer: C

Figure 16. Example response from Gemini-3-Flash (Google, 2025) on a Scenes question.



Q: What is the orientation of the longest scratch on the crocs?

A. Horizontal
B. Vertical
C. Diagonal
D. There is no scratch

Domain: Objects

Claude-Opus-4.7:

Explanation: Looking at the crocs in the image, I don't see any clear scratches. The patterns visible on the red/blue crocs are part of the marbled color design, not scratches.
Answer: D

Figure 17. Example response from Claude-Opus-4.7 (Anthropic, 2026) on an Objects question.



Q: Where is the ball positioned on the lane?

A. It is stuck on the gutter.
B. It is slightly to the right.
C. It is slightly to the left.
D. It is right in the middle.

Domain: Scenes

Grok-4.2:

Explanation: The image shows a blue bowling ball on the polished wooden surface of the lane, well clear of either side channel/gutter. It sits near the center line of the lane (between the visible side gutters and with the pin setup directly ahead), so it is not offset noticeably left or right.
Answer: D

Figure 18. Example response from Grok-4.2 (xAI, 2026) on a Scenes question.



Q: How was this image taken?

A. It was taken by a CCTV camera
B. It was taken from inside the
C. It was taken through a glass
D. It is a selfie

Domain: Scenes

Qwen3.5-VL-Plus-Thinking:

Explanation: The image shows a view of the interior of a convenience store (specifically a 7-Eleven, indicated by the uniform and signage). The camera angle is at eye level, positioned behind some foreground shelves (snacks on the left, hanging bags on the right) and looking towards an employee and further shelves. There are no reflections, frames, or distortions typical of shooting through a window, nor is it a high-angle grainy shot typical of CCTV. The photographer is clearly standing within the aisles of the store.
Answer: B

Figure 19. Example response from Qwen3.5-VL-Plus-Thinking (Qwen, 2026) on a Scenes question.



Q: Which item has a 5-star rating?

A. The top one
B. The bottom one
C. Both
D. Neither

Domain: Digital World

Qwen3.5-VL-Plus-Instruct:

Explanation: Both reviews shown in the image display five filled yellow stars, indicating a 5-star rating for each item. The top review is for Bluetooth headphones and the bottom one is for a vanity mirror — both are rated 5 stars by Jessica D.
Answer: C

Figure 20. Example response from Qwen3.5-VL-Plus-Instruct (Qwen, 2026) on a Digital World question.



Q: What are the fish in the water doing?

A. They are thrashing
B. They are dead
C. They are swimming freely
D. They are trying to dive deeper

Domain: Scenes

Qwen3.5-VL-35B-A3B:

<thinking>
The user wants me to identify the behavior of the fish in the water based on the image.

1. Analyze the image:
* I see three men in a body of water (likely a pond or fish farm).
* They are holding a large fishing net.
* Inside the net, there are many fish.
* The fish in the net are splashing and moving vigorously. Their bodies are twisted and contorted, indicating a struggle.
* There is also a blue barrel in the foreground filled with caught fish. These fish look relatively still, but the question asks about the fish *in the water* (specifically those being caught or just released, though the context implies catching). Wait, let's look closer. The fish *in the water* are the ones trapped in the net.
* The fish in the net are clearly agitated. They are splashing water and moving their bodies rapidly. This behavior is best described as "thrashing".

2. Evaluate the options:
* A: They are thrashing: This matches the visual evidence. The fish in the net are splashing and struggling against the net.
* B: They are dead: The fish in the net are moving vigorously, so they are alive. The fish in the barrel might be dying or dead, but the ones in the water (in the net) are definitely alive and moving. Even if the question refers to the fish in the barrel, "thrashing" is the most dynamic action happening in the water scene. However, the question asks "What are the fish in the water doing?". The fish *in the water* are the ones in the net. They are thrashing.
* C: They are swimming freely: They are trapped in a net, so they are not swimming freely.
* D: They are trying to dive deeper: While they might be trying to escape, "thrashing" is a more accurate description of the chaotic movement seen. They are splashi
...

Explanation: The image shows several men pulling a large net out of the water. Inside the net, numerous fish are visible. These fish are splashing vigorously, twisting their bodies, and creating a lot of movement and water disturbance. This chaotic and violent movement is best described as thrashing, indicating they are alive and struggling against the net. They are not swimming freely because they are trapped, nor do they appear dead given the amount of movement.

Answer: A

Figure 21. Example response from Qwen3.5-VL-35B-A3B (Qwen, 2026) on a Scenes question.



Q: From which side of the image does the light source originate?

A. Left
B. Right
C. Both
D. Neither

Domain: Scenes

Qwen3.5-VL-27B:

<thinking>
The user wants me to determine the direction of the light source based on the image.
1. Analyze shadows: I need to look at the shadows cast by the people and objects in the scene.
* Look at the girl in the foreground left (grey sweater, jeans). Her shadow is cast slightly behind her and to the right. Wait, let me look closer. Actually, looking at the girl in the grey sweater, there isn't a very distinct long shadow directly next to her feet, but the lighting seems fairly diffuse.
* Look at the girl in the black jacket next to her. Her shadow is faint but seems to fall slightly to the right.
* Look at the girl in the white shirt and blue shorts further back on the path. She is holding binoculars. Her shadow is clearly visible on the ground to her right side. It stretches towards the bottom right corner of the image.
* Look at the girl in the pink/white shirt and jeans next to the path. Her shadow is also cast to the right.
* Look at the girl on the far right holding a camera. Her shadow is cast to her right (towards the right edge of the frame). Wait, let me re-evaluate.
* Let's look at the bushes. The bushes on the left are lit up on their left/top side, but they are dense.
* Let's look at the general illumination. The sky is cloudy, which often creates diffuse light. However, there are directional shadows.
* Let's re-examine the shadows carefully.
* Girl in white t-shirt (middle right): Shadow is to her right. This means the sun is to her left.
* Girl in pink/white shirt (right of path): Shadow is to her right. Sun is to her left.
* Girl on far right: Shadow is to her right. Sun is to her left.
* Wait, let's look at the girl in the grey sweater again. If the sun is on t
...
</thinking>
</thinking>
</thinking>
</thinking>

Explanation: By observing the shadows cast by the people in the image, particularly the girls standing on and near the path, we can see that their shadows are falling to the right side of their bodies. For instance, the girl in the white t-shirt holding binoculars casts a shadow to her right. When shadows extend to the right, it indicates that the light source (the sun) is positioned to the left.
Answer: A

Figure 22. Example response from Qwen3.5-VL-27B (Qwen, 2026) on a Scenes question.



Q: What is the color of the bowling ball visible in the image?

A. Green
B. Blue
C. Black
D. Red

Domain: Scenes

Kimi-K2.5:

Explanation: Looking at the image, I can identify a bowling ball visible near the seating area on the right side of the image, close to where a person in a red shirt is sitting. The bowling ball appears to be black in color.
Answer: C

Figure 23. Example response from Kimi-K2.5 (Kimi, 2026) on a Scenes question.



Q: Which of the following best describes the text in the image?

A. Only the first letter of each sentence is capitalized
B. There is a shadow effect applied to the text
C. The text border is blue
D. The text fill is white

Domain: Digital World

GLM-4.6V:

<thinking>
Got it, let's analyze the options. First, look at the text in the image. The lyrics are written with each sentence starting with a capital letter, and the rest lowercase? Wait, no, let's check. The first option: "Only the first letter of each sentence is capitalized"—let's see the text. The lines like "I COUNT THE YEARS ONE HAND," "I HAVE FIVE BEEN TOGETHER," etc. Each sentence (or line) starts with a capital I, so yes, only the first letter of each sentence is capitalized. Now check other options: B: shadow effect? The text doesn't look like it has a shadow. C: text border blue? No, the text is black or dark, no blue border. D: text fill white? The text is dark, not white. So the correct answer is A.
</thinking>

Explanation: The text in the image (lyrics) shows that each sentence (or line) begins with a capital letter, and the rest of the letters in the sentence are lowercase, which matches option A. The other options (shadow effect, blue border, white fill) do not apply to the text as observed.
Answer: A

Figure 24. Example response from GLM-4.6V (GLM, 2025a) on a Digital World question.



Q: What is the shape of the throw pillow that has three different colors?

A. Triangle
B. Square
C. Rectangle
D. Circle

Domain: Objects

Gemma-4-31B:

<thinking>
The user wants to identify the shape of the throw pillow with three different colors in the image.
1. Identify the throw pillow with three different colors: There is a pillow resting on the purple blanket. It has three distinct colored sections: pink/peach, orange, and blue.
2. Determine its shape: Looking at the pillow, it has three sides and three corners. This is the definition of a triangle.
3. Compare with options:
* A: Triangle-Matches.
* B: Square-Does not match (it doesn't have four equal sides and right angles).
* C: Rectangle-Does not match (it doesn't have four sides).
* D: Circle-Does not match (it's not round).
Conclusion: The pillow is triangular.
</thinking>
</thinking>
</thinking>
</thinking>
</thinking>

Explanation: The throw pillow resting on the purple blanket features three distinct colored sections (pink, orange, and blue) and has a three-sided geometric form.
Answer: A

Figure 25. Example response from Gemma-4-31B (DeepMind, 2026) on an Objects question.



Q: What is the image most likely to be?

A. A glossary opened in a document editor
B. A PDF of a glossary
C. A scan of a paper glossary
D. A glossary on a webpage

Domain: Documents, Charts, & Tables

Gemma-4-E4B:

<thinking>
1. Analyze the image: The image displays a list of technical terms (301 Redirect, 302 Redirect, 403 Server Code, 404 Server Code, A.J.A.X, Algorithm, Alt Text, B2B, B2C) along with their definitions. The title is "Fundamentals Glossary".

2. Analyze the visual presentation:
* The text is clearly typeset, clean, and uses standard digital fonts.
* There are no signs of physical paper texture, curves, or artifacts typical of scanning (like ink bleed, scanner lines, uneven edges).
* The formatting (bullet points/list structure, clear headings) is consistent with digital document presentation.
* The colors are uniform (black text on a white background).

3. Evaluate the options based on the visual evidence:

* A: A glossary opened in a document editor (e.g., Word, Google Docs). This is highly plausible. The clean, formatted text fits this description.
* B: A PDF of a glossary. This is also highly plausible. PDFs maintain the exact digital formatting seen here.
* C: A scan of a paper glossary. Highly unlikely. A scan would almost certainly show imperfections (speckles, shadows, slight distortion, paper grain). This image is too crisp and clean for a typical scan.
* D: A glossary on a webpage. Highly plausible. Web content also has this clean, digital appearance.

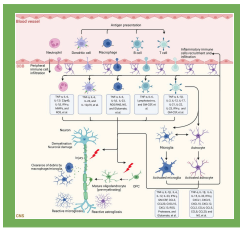
4. Distinguish between A, B, and D: While A, B, and D are all *possible* digital formats, the presentation style (simple text, clear hierarchy, centered title) is generic. However, when comparing the options, a document editor (A) or a webpage (D) are the most likely immediate sources.

5. Refining the choice (Digital Display vs. Document): The image looks like a screenshot or a digitally rendered document *from* an application (like a website or a word
...

Explanation: The image displays highly crisp, uniform text with a professional digital layout (clean fonts, structured headings, and definitions). There are no artifacts associated with physical objects, such as paper texture, ink bleed, shadows, or scanner distortions. This strongly suggests the content was created or captured digitally. While it could be a PDF or a webpage, the presentation is characteristic of a structured document that would be compiled using a document editor (like Microsoft Word or Google Docs) and then captured as a screenshot or image. This is the most accurate description of the visual nature of the file presented.

Answer: A

Figure 26. Example response from Gemma-4-E4B (DeepMind, 2026) on a **Documents, Charts, & Tables** question.



Q: What type of cell is the lightest blue in this image?

A. Macrophage
B. B Cell
C. Dendritic Cell
D. T Cell

Domain: Academics

InternVL-3.5:

<thinking>
Okay, let's see. The question is asking which cell type is the lightest blue in the image. The options are Macrophage, B Cell, Dendritic Cell, or T Cell. I need to look at the image and check the colors of each of these cells.

Looking at the image, there's a row of cells under the blood vessel and BBB. The cells shown from left to right are Neutrophil (pink), Dendritic cell (a bit darker blue), Macrophage (darker blue), B Cell (lighter blue?), T Cell (maybe even lighter?). Wait, the B Cell and T Cell might be similar in color. Let me check the labels. The B Cell is shown as a blue circle with receptors, and the T Cell is next to it. The B Cell might be a medium blue, while the T Cell is a lighter blue. Alternatively, maybe the T Cell is the lightest. Alternatively, looking at the colors in the image, the T Cell might be the lightest blue compared to the others. The Macrophage is a darker blue, Dendritic cell is also a darker blue. The B Cell and T Cell are both in the antigen presentation part. Wait, the B Cell is in the middle, maybe a medium blue, and the T Cell is even lighter. Alternatively, maybe the T Cell is the same as B Cell. Alternatively, looking at the image's color scheme, the B Cell is a bit darker than the T Cell. Let me check the image again. The T Cell is depicted with a very light blue, almost cyan, whereas the B Cell is a bit more blue. So the lightest blue would be the T Cell. So the answer is D.
</thinking>

Explanation: In the image, the T Cell is depicted in the lightest blue shade compared to the other cell types listed. The B Cell and Macrophage are shown in darker blue tones, while the T Cell has a paler blue coloration, making it the correct choice.

Answer: D

Figure 27. Example response from InternVL-3.5 (Wang et al., 2025) on an **Academics** question.