WorldVQA: Measuring Atomic World Knowledge in Multimodal Large Language Models

Runjie Zhou[1*] **Youbo Shao**[1*] **Haoyu Lu**[1*‡] **Bowei Xing**[1] **Tongtong Bai**[1] **Yujie Chen**[1]
Jie Zhao[1] **Lin Sui**[1] **Haotian Yao**[1] **Zijia Zhao**[1] **Hao Yang**[1] **Haoning Wu**[1] **Zaida Zhou**[1]
Zhiqi Huang[1] **Yiping Bao**[1] **Yangyang Liu**[1] **Y.Charles**[1] **Xinyu Zhou**[1]

[1] Moonshot AI

Abstract

We introduce WorldVQA, a benchmark designed to evaluate the atomic visual world knowledge of Multimodal Large Language Models (MLLMs). Unlike current evaluations, which often conflate visual knowledge retrieval with reasoning, WorldVQA decouples these capabilities to strictly measure "what the model memorizes." The benchmark assesses the atomic capability of grounding and naming visual entities across a stratified taxonomy, spanning from common head-class objects to long-tail rarities. We expect WorldVQA to serve as a rigorous test for visual factuality, thereby establishing a standard for assessing the encyclopedic breadth and hallucination rates of current and next-generation frontier models. The dataset can be found in WorldVQA Homepage.

1 Introduction

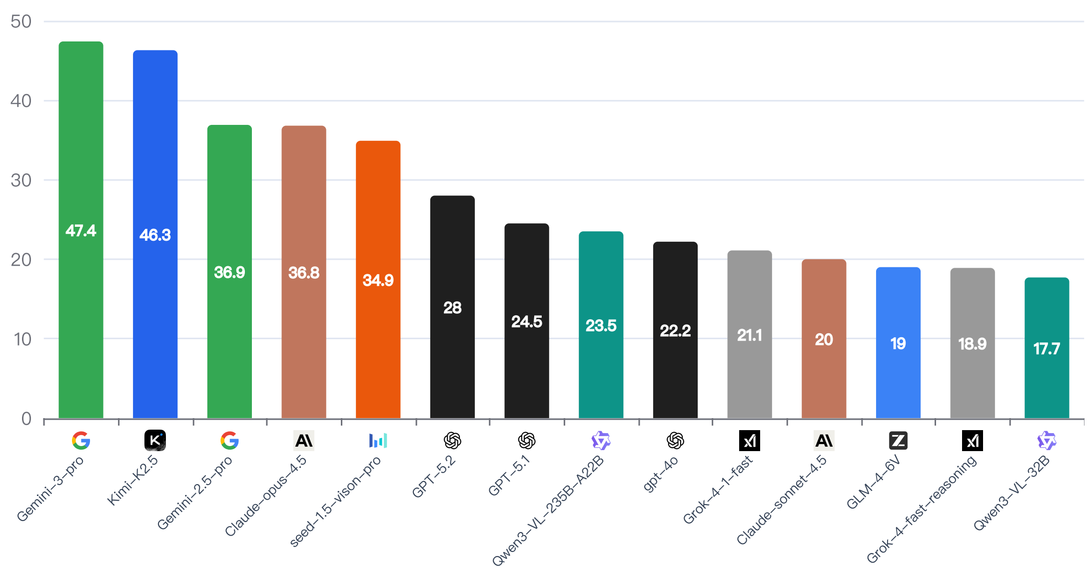


Figure 1: **Overall Model Accuracy on WorldVQA**. While the Gemini-3-pro (47.4%) and Kimi K2.5 (46.3%) currently lead the field, no evaluated model surpasses the 50% accuracy threshold, underscoring the significant challenge of grounding atomic visual knowledge.

---

[*] Equal contribution. [‡] Project lead.

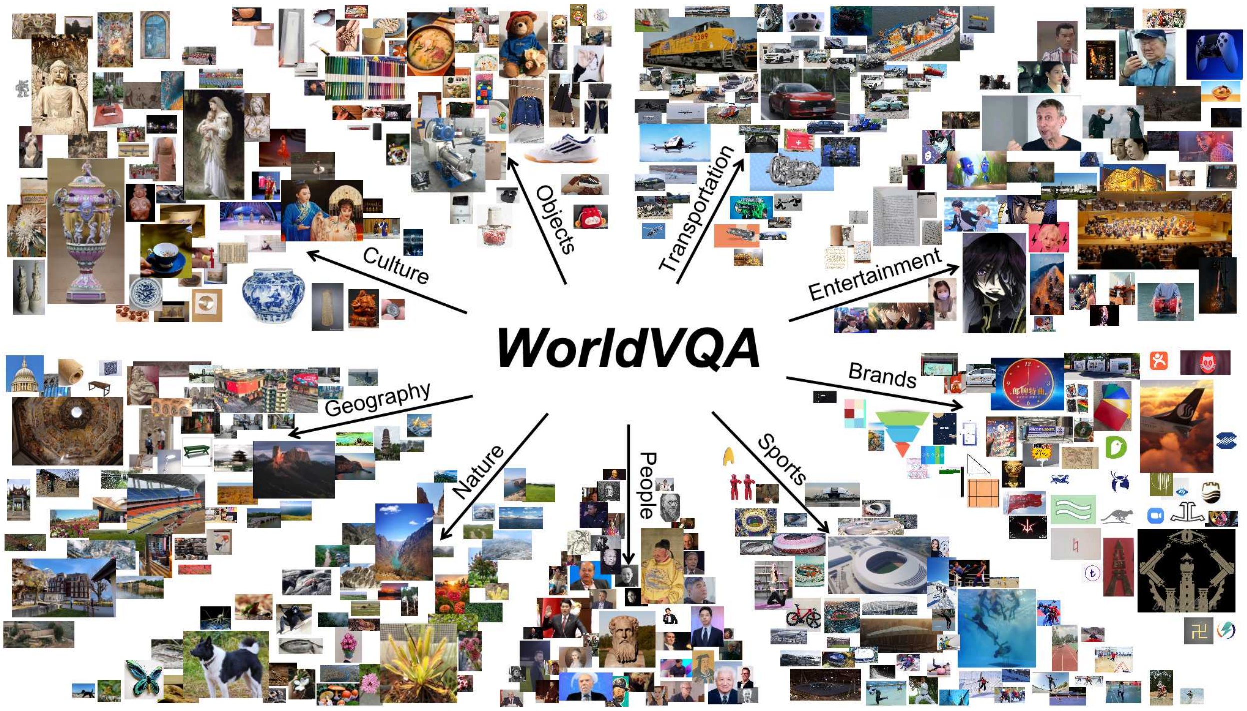


Figure 2: A visual overview of the **WorldVQA** dataset. The benchmark is organized into nine categories: Nature & Environment (Nature); Locations & Architecture (Geography); Culture, Arts & Crafts (Culture); Objects & Products (Objects); Vehicles, Craft & Transportation (Transportation); Entertainment, Media & Gaming (Entertainment); Brands, Logos & Graphic Design (Brands); Sports, Gear & Venues (Sports); Notable People & Public Figures (People). The visual entities curated to evaluate atomic world knowledge range from globally recognized "head-class" landmarks and logos to specific "long-tail" biological species and artisanal artifacts. To maintain **atomic isolation**, each image serves as an unambiguous visual stimulus for entity naming, strictly decoupled from complex reasoning or OCR dependencies.

The advancement of MLLMs is contingent upon two distinct capabilities: reasoning (processing logic and relationships) and knowledge grounding (mapping sensory inputs to factual reality). While recent frontier models have demonstrated encyclopedic mastery in the textual domain, their ability to maintain factual reliability in the visual domain remains inconsistent. True multimodal intelligence requires a robust visual encyclopedia, an internal representation that accurately maps raw pixel data to specific entity identities. Without this precise image grounding, MLLMs function as descriptive engines rather than knowledgeable observers, resulting in a propensity for visual hallucinations where models fabricate plausible but incorrect details.

Accurately quantifying such internal visual knowledge presents a methodological challenge. Current Visual-Question-Answering (VQA) benchmarks, such as MMMU (Yue, Ni, et al. 2024) and MMStar (Lin Chen et al. 2024), prioritize complex, multi-step reasoning. This approach conflates visual knowledge with logical deduction, making it difficult to isolate the source of error. Similarly, VQA benchmarks like SimpleVQA (Cheng et al. 2025) often couple visual recognition with secondary dependencies, such as language knowledge or optical character recognition (OCR). For instance, a failure to answer a question about a company's founding date may stem from a lack of textual historical knowledge rather than a failure to visually identify the company's logo. This entanglement prevents researchers from determining whether a deficit lies in the model's visual perception (the "eyes") or its semantic memory (the "brain").

To address this gap, we introduce WorldVQA, a targeted evaluation suite comprising 3,500 VQA pairs across 9 semantic categories (see Figure 2). Unlike general-purpose benchmarks, WorldVQA is engineered to assess **atomic world knowledge**, the direct, unassisted association between a visual stimulus and its specific proper noun or taxonomic name. To provide a clearer contrast, Table 1 summarizes the key differences between WorldVQA and these representative benchmarks across multiple dimensions, highlighting our focus on isolating atomic vision concepts.

WorldVQA follows four technical design principles:

1. Atomic Isolation. To eliminate confounding variables, we strictly isolate the target task of entity naming. The benchmark evaluates the most fundamental unit of visual knowledge by decoupling visual identification from external logic. We systematically exclude queries requiring OCR, arithmetic, or multi-hop knowledge retrieval. By focusing exclusively on visual grounding (e.g., *"What is the specific scientific name of this species?"*), we isolate the model's visual recognition capabilities from its reasoning engine.

2. Taxonomic Diversity. To rigorously test knowledge breadth, we synthesized a classification framework covering 9 primary categories (as visualized in Figure 2): *Nature & Environment (Nature); Locations & Architecture (Geography); Culture, Arts & Crafts (Culture); Objects & Products (Objects); Vehicles, Craft & Transportation (Transportation); Entertainment, Media & Gaming (Entertainment); Brands, Logos & Graphic Design (Brands); Sports, Gear & Venues (Sports); Notable People & Public Figures (People).* This taxonomy moves beyond unstructured web-crawling, ensuring a balanced distribution that tests both high-frequency "head" entities and the "long-tail" of rare instances. This structure allows us to profile the "encyclopedic boundary" of a model's visual memory.

3. Data Integrity & Verification. We implement a multi-stage curation pipeline to ensure WorldVQA serves as a gold standard oracle. To prevent contamination, we perform rigorous deduplication against common large-scale pre-training corpora. Verification utilizes a dual-gate mechanism combining automated consistency checks via high-performance MLLMs with human-in-the-loop validation, minimizing label noise and ensuring high-fidelity evaluation.

4. High Saturation Ceiling. The benchmark is calibrated to challenge current frontier models. As illustrated in Figure 1, even advanced systems exhibit performance ceilings, often failing to exceed 50% accuracy across all categories. The radar charts (see Figure 3) reveal distinct knowledge pits, particularly in *Nature* and *Culture*, where model performance lags significantly behind text-only equivalents. This demonstrates that WorldVQA provides high-resolution visibility into the limitations of current visual pre-training.

By isolating atomic knowledge retrieval from reasoning, WorldVQA provides a precise metric for visual hallucination and knowledge grounding. We release this benchmark to the community as a standard for assessing the factual reliability of next-generation MLLMs.

Benchmark	Data Size	Language	Question Type	Target Domain	Detail
MMMU-Pro	5.2k	English	Multiple-Choice	Academic Understanding	Evaluates expert-level academic knowledge, often conflating factual recall with complex logical reasoning.
MMBench	2.4k	CN & EN	Multiple-Choice	General Multi-modal Ability	Assesses various multimodal abilities by using perception and reasoning as its primary evaluation pillars.
RealWorldQA	765	English	Multiple-Choice	Spatial & Physical Perception	Measures understanding of physical environments and spatial relationships through situational queries.
SimpleVQA	2.0k	CN & EN	Generation	Vision Knowledge & Reasoning	Probes the factuality ability of MLLMs to answer natural language short questions.
WorldVQA	3.5k	CN & EN	Generation	Atomic Vision Knowledge	Isolates atomic world knowledge with stratified, encyclopedic taxonomy.

Table 1: Comparison of WorldVQA with existing Multimodal Benchmarks. While existing suites conflate factual recall with reasoning or secondary dependencies , WorldVQA stands apart by strictly isolating atomic visual knowledge through the principle of decoupling.

2 Data Collection and Verification

The construction of WorldVQA follows a rigorous pipeline designed to ensure atomic factuality and taxonomic breadth. The process comprises two primary phases. First, 10 expert annotators with over one year of experience in MLLMs evaluation curated the atomic entities and VQA pairs according to strict taxonomic and granularity standards. Second, these samples underwent a dual-verification process, including automated fact-checking by frontier MLLMs to assess visual clarity and independent blind validation by expert annotators to ensure absolute ground-truth reliability. Only samples passing all verification stages were included in the final benchmark.

2.1 Design Principles and Criteria

WorldVQA is strictly anchored by the principle of Atomic Isolation. Unlike benchmarks that conflate visual recognition with multi-hop reasoning, WorldVQA isolates the model's ability to ground specific visual entities. Beyond this core tenet, we prioritize Encyclopedic Diversity and Data Fidelity to guarantee data reliability. To achieve these, we enforce four strict criteria:

Atomic Isolation To evaluate world knowledge ability, we strictly decouple knowledge retrieval from multi-step reasoning. Questions are engineered to be single-hop, requiring only the direct identification of a visual entity or its specific attributes. Tasks involving OCR, arithmetic, or external logical deduction are excluded to ensure the evaluation reflects the model's internal parametric memory.

Encyclopedic Knowledge Coverage To ensure the benchmark serves as a global standard, data composition is governed by rigorous distribution rules rather than random sampling. First, each category maintains a sufficient volume of data to ensure statistical significance. Second, we prioritize cultural diversity by capping region-specific entities; specifically, the proportion of entities unique to the Chinese context is limited to under 50% for each individual category, resulting in a final aggregate of 36% Chinese-specific entities across the entire benchmark. Third, rather than selecting entities at random, annotators were instructed to deliberately sample entities across a broad spectrum of real-world prevalence. These distribution rules ensure that WorldVQAcaptures the natural distribution of encyclopedic knowledge, spanning from globally ubiquitous head-class concepts to highly specialized long-tail rarities, as empirically validated by our difficulty alignment analysis in Section 3.3.

Granularity Alignment To guarantee the validity of knowledge, we enforce a strict alignment between the specificity of the question and the granularity of the answer. Correctness is defined by taxonomic precision: for example, if an image depicts a Bichon Frise, the answer must identify the specific breed, whereas generic hypernyms such as dog are considered incorrect. This constraint serves two critical functions: it prevents models from bypassing the knowledge requirement through safe but vague generalizations, and it aligns the difficulty level with the true complexity of the visual signal.

Visual Reliability Images must serve as authentic, definitive evidence for the atomic fact being tested. We strictly enforce two standards for visual selection:

- **Sanitization**: Images must be devoid of textual leakage (e.g., labels, watermarks, overlay text) to preclude the model from using OCR-based shortcuts to "read" the answer.
- **Unambiguity**: The visual features must be distinct and strictly correspond to the target entity. The image must firmly support the ground truth while ruling out reasonable alternatives or confusing distractors. If an entity cannot be uniquely identified from the visual features alone, it is discarded.

2.2 Data Quality

The integrity of a benchmark is determined by the precision of its data. To ensure WorldVQA serves as a definitive measure of atomic world knowledge, we implemented a rigorous curation and verification pipeline.

2.2.1 Data Curation Pipeline

We employed a three-step pipeline to collect raw taxonomic entities and convert entities into high-quality VQA triplets:

Step 1: Seed Entity Collection. 10 annotators gathered initial data based on the taxonomy. Seed entities were sourced from internal lexicons. Adhering to the *Encyclopedic Knowledge Coverage* criteria, annotators selected appropriate entities and retrieved corresponding images from trusted web sources (following the *Visual Reliability* criteria), finally formulating QAs according to the *Granularity Alignment* criteria.

Step 2: Distributional Balancing and Global Expansion. To ensure the benchmark's international generalizability, we perform contextual labeling to identify and partition region-specific entities. We enforce a 50% per-category cap on entities unique to the Chinese context. For categories falling below our global representation threshold, we utilize an LLM-in-the-loop expansion strategy, leveraging GPT and Kimi for association search, to identify supplemental global entities, which are then integrated following the protocol in Step 1.

Step 3: Visual Deduplication. To eliminate redundancy and prevent data leakage from common pre-training corpora, we employed the Instance-level Semantic Content (ISC, Yokoo 2021) descriptor. We calculated the cosine similarity of ISC embeddings for each candidate image against massive open-source datasets, specifically LAION (Schuhmann et al. 2022) and Common Crawl. Applying a strict threshold of 0.95, any images identified as duplicates or leaked from these sources were discarded. To maintain data volume without duplication, we performed targeted re-collection for these entities by capturing new visual assets from video screenshots, which minimizes the likelihood of the model relying on memorized image-answer pairs. This rigorous protocol ensures that correct responses reflect the model's genuine internal encyclopedia rather than simple pattern retrieval from its training history.

2.3 Model-Based Difficulty Stratification

To ensure high discriminative capacity and mitigate the ceiling effect observed in existing benchmarks, we applied a Model-Performance-Based Stratification strategy.

We evaluated all candidate samples using an ensemble of five frontier MLLMs. To maximize the benchmark's utility and discriminative power, we discarded trivial samples correctly answered by all five models. The remaining samples were stratified into three difficulty levels based on model performance: Easy (>3 models correct), Medium (1–2 models correct), and Hard (0 models correct). To prevent the benchmark from being dominated by simpler entities and to maintain a focus on challenging long-tail knowledge, we performed random downsampling on the Easy category. We report the final proportions of each difficulty tier in Table 2.

Note on Bias. This stratification is not intended to "trap" specific models, but rather to counteract the ceiling effect prevalent in current benchmarks. By intentionally downsampling from the model-defined Easy tier, we ensure WorldVQA remains a challenging probe for next-generation frontier systems. Meanwhile, all Hard samples underwent a mandatory secondary human review to confirm that the difficulty stems from the rarity of the knowledge, not from visual ambiguity or annotation error.

2.4 Dual-Verification Mechanism

To ensure maximum ground-truth fidelity, we implement a rigorous dual-gate verification protocol consisting of automated model-based auditing and independent human validation. This multi-stage process is designed to filter out semantic noise and visual ambiguity.

Model Based Visual Auditing. We utilize few-shot prompted Gemini-3-Pro as automated fact-checker to evaluate each VQA triplet (see Visual Audit Prompt in Appendix A). The models enforce three non-negotiable requirements for data integrity:

- **Visual Clarity**: The image must provide sufficient resolution to permit unambiguous entity identification.
- **Semantic Exclusivity**: The image content must uniquely support the ground-truth label while actively ruling out reasonable alternative interpretations or distractors.
- **Contextual Completeness**: The visual context must encompass all necessary information required to resolve the question.

Human Blind Validation. Parallel to automated checks, we conducted independent human validation. An annotator, unaware of the ground truth, was required to answer the question. Any sample where the human prediction diverged from the ground truth was flagged for manual audit. Cases involving factual errors or visual ambiguity were permanently purged from the dataset.

2.5 Grading and Metrics

To facilitate a standardized comparison, we utilize Accuracy as our primary single-number metric to measure overall factual reliability. For a more granular experimental analysis, we also report Correct Given Attempted (CGA) and F-score. CGA isolates the precision of the model's internal knowledge by evaluating only the samples it chose to answer, effectively measuring its susceptibility to hallucination when it commits to a response. The F-score synthesizes coverage (attempt rate) and precision (CGA) into a single harmonic mean, penalizing both over-conservative refusal and over-aggressive guessing.

2.6 Benchmark Statistics

Statistics	Number	Statistics	Number
Data	3500	- Entertainment, Media & Gaming (Entertainment)	14.60%
- Chinese (CN)	1260 (36%)	- Brands, Logos & Graphic Design (Brands)	7.43%
- English (EN)	2240 (64%)	- Sports, Gear & Venues (Sports)	4.06%
Category Categories	9	- Notable People & Public Figures (People)	14.29%
- Nature & Environment (Nature)	9.31%	**Difficulty**	
- Locations & Architecture (Geography)	14.63%	- Easy	31.16%
- Culture, Arts & Crafts (Culture)	14.46%	- Medium	40.77%
- Objects & Products (Objects)	12.49%	- Hard	28.07%
- Vehicles, Craft & Transportation (Transportation)	8.74%		

Table 2: WorldVQA statistics across nine semantic categories and three difficulty tiers. Note that Easy samples (31.16%) are downsampled to maintain a high saturation ceiling for evaluating frontier MLLMs.

Following the curation protocols described above, we present a high-level overview of the WorldVQA dataset. As shown in Table 2, the benchmark consists of 3,500 pairs with a balanced linguistic and categorical spread. Importantly, to establish a global evaluation standard, we maintain a 1:1.78 Chinese-to-English ratio.

3 Experiments

3.1 Settings

To ensure a rigorous and fair evaluation across the diverse landscape of MLLMs, we maintained strict consistency in our experimental protocols. All models evaluated with unified prompts and official inference parameters. For the grading process, we employed GPT-oss-120b (OpenAI et al. 2025) as our primary judge model (see Appendix A for the judge prompt). To validate this automated grading, a manual audit of 160 random samples reveals a 98.1% alignment rate with human expertise (only 3 disagreements).

3.2 Main Results

Models	**Overall results**				**F-score on 9 task categories**								
	Accuracy	Not Attempted	Correct Given Attempted	F-score	Nature	Geo-graphy	Culture	Objects	Trans-portation	Entertain-ment	Brands	Sports	People
Closed-source MLLMs													
Gemini-3-pro	47.4	0.6	47.7	47.5	45.1	44.7	47.2	48.1	45.1	47.6	52.4	59.4	-
Gemini-2.5-pro	36.9	0.1	36.9	36.9	37.1	33.8	32.6	39.6	39.9	34.2	38.8	54.2	-
Seed-1.5-vision-pro	34.9	1.6	35.5	35.2	41.4	36.1	33.4	32.8	35.0	33.6	32.3	43.7	-
Claude-opus-4.5	36.8	3.4	38.1	37.5	32.5	36.5	34.1	39.6	43.5	29.0	47.6	54.9	-
Claude-sonnet-4.5	20.0	8.0	21.8	20.9	19.4	21.0	17.4	22.9	24.8	11.6	32.2	31.0	-
GPT-5.2	28.0	5.4	29.5	28.7	24.3	29.1	26.7	26.6	30.7	24.8	39.1	40.8	-
GPT-5.1	24.5	16.3	29.3	26.7	27.3	25.1	22.5	26.6	31.6	18.5	36.0	45.4	-
GPT-4o	22.2	9.1	24.4	23.3	25.6	20.6	17.8	19.1	26.2	19.1	35.2	44.5	-
Grok-4-1-fast-reasoning	21.1	0.1	21.1	21.1	18.4	23.6	20.2	25.2	23.5	11.4	25.8	30.3	-
Grok-4-fast-reasoning	18.9	0.2	19.0	18.9	17.8	19.0	18.6	22.0	20.3	8.3	26.6	34.5	-
Open-source MLLMs													
Kimi K2.5	46.3	2.1	47.3	46.8	40.6	46.8	43.0	44.7	47.4	48.1	52.6	64.8	50.9
Kimi-VL-16B-A3B	12.0	3.3	12.4	12.2	11.2	13.9	10.1	10.8	13.5	7.9	20.8	17.7	7.4
Qwen3-VL-235B-A22B-Instruct	23.5	0.0	23.5	23.5	26.1	24.8	22.9	26.1	28.8	15.5	22.3	26.1	26.2
Qwen3-VL-32B-Instruct	17.7	0.0	17.7	17.7	18.1	18.0	16.8	19.0	19.0	12.1	23.8	20.4	13.1
GLM-4.6V	19.0	0.0	19.0	19.0	24.5	21.5	17.8	19.2	18.6	12.5	20.4	23.2	10.7
GLM-4.6V-Flash	14.8	0.1	14.8	14.8	16.0	16.3	13.2	14.9	19.0	7.8	18.8	20.4	8.2

Table 3: Performance of frontier MLLMs on WorldVQA. Hyphen entries (-) denote scores omitted due to excessive refusal rates. Overall Results aggregate the first eight categories. "Notable People & Public Figures" (People) is excluded from the overall average to ensure a fair comparison, as systematic refusals in closed-source models, driven by privacy and safety guardrails, do not necessarily reflect underlying knowledge deficits.

Evaluations on WorldVQA reveal a substantial gap between frontier MLLMs and true encyclopedic proficiency. As detailed in Table 3, Gemini-3-pro leads with an F-score of 47.5%, followed closely by Kimi K2.5 (46.8%, the top-performing open-source model). Notably, no model surpasses the 50% threshold, underscoring the challenge of grounding the long-tail entities in our benchmark.

Category-wise analysis (see Figure 3) indicates higher proficiency in *Brands* and *Sports*, likely due to their over-representation in web-scale pre-training data. For instance, Gemini-3-pro achieves an F-score of 59.4 in *Sports*. Conversely, *Nature* and *Culture* emerge as significant weaknesses. In these domains, models frequently revert to generic hypernyms (e.g., "flower" instead of specific species), which are penalized under our Granularity Alignment criteria. This suggests that while MLLMs are "pop-culture savvy," their grasp of the natural world and diverse human heritage remains shallow, necessitating more diverse data sourcing.

The discrepancy between Correct Given Attempted (CGA) and F-score serves as a probe for model honesty. GPT-5.1 exhibits a high CGA (29.3%) but a low F-score (26.7%), indicating a conservative strategy where the model answers only when certain. In contrast, many smaller models show low CGA, reflecting a tendency to hallucinate names for obscure entities rather than admitting ignorance. This misalignment between a model's propensity to attempt an answer and its actual accuracy suggests that current MLLMs lack a reliable internal barometer of their own knowledge boundaries, a calibration deficit we analyze in depth in Section 3.4.

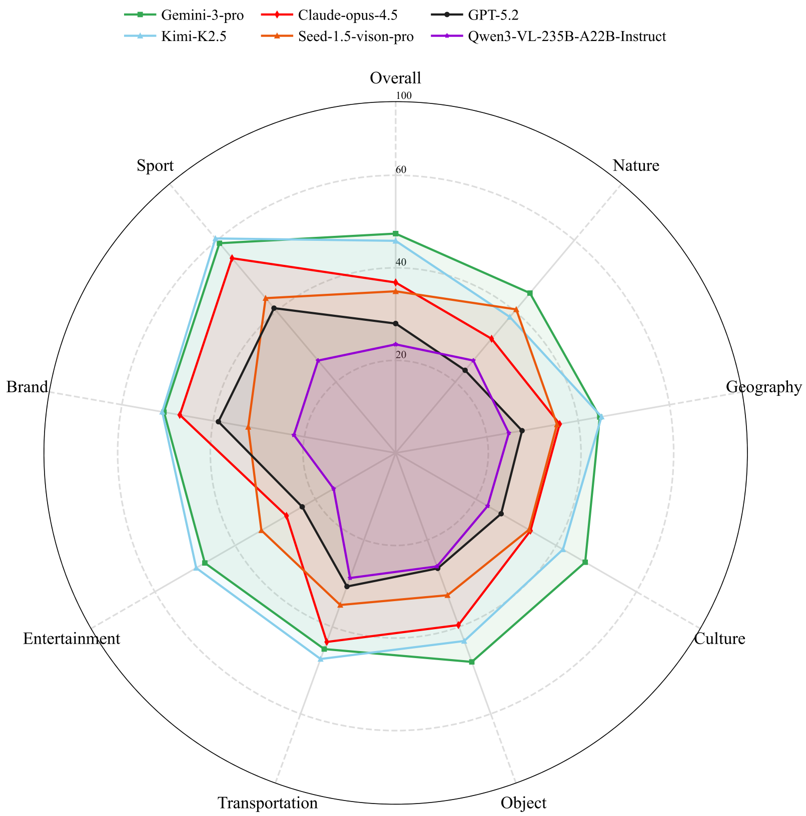


Figure 3: **Category-wise F-score comparison on WorldVQA.** This radar chart illustrates the performance profiles of frontier close-source and open-source MLLMs across the 8 semantic categories. The visualization highlights the relative proficiency in high-frequency domains like *Sports* and *Brands*, while revealing significant performance troughs in specialized domains such as *Nature* and *Culture*.

3.3 Validation of Difficulty Stratification

To validate whether our Model-Performance-Based Stratification accurately reflects real-world knowledge distribution, we utilize the rank frequency of entity terms in the MetaCLIP vocabulary (Hu Xu et al. 2025, Chuang et al. 2025) as a proxy for real-world prevalence. Figure 4 illustrates the density distributions of entity difficulties mapped against their MetaCLIP rank percentile.

The quantitative results demonstrate a distinct positive correlation between real-world rarity and benchmark difficulty. As shown by the fitted curves, Trivial and Easy samples concentrate heavily near the zeroth percentile, indicating that current MLLMs primarily master high-frequency head entities. As difficulty escalates to Medium and Hard, the distribution peaks progressively shift rightward toward higher rank percentiles. This systematic migration confirms that the difficulty in WorldVQA stems from genuine knowledge scarcity (long-tail entities) rather than confounding factors like visual ambiguity or annotation artifacts.

Furthermore, the analysis highlights WorldVQA's effective coverage of the knowledge spectrum. While Easy samples probe the high-frequency head, Medium and Hard categories successfully extend into critical long-tail regions often underrepresented in standard evaluations. Although minor variations exist, e.g., *Brands* and *People* skew slightly towards higher frequencies due to dense web coverage, the overarching trend of difficulty correlating with rarity remains robust. Thus, WorldVQA provides a stratified assessment that accurately mirrors the structural complexity of real-world encyclopedic knowledge.

3.4 Calibration Analysis

To evaluate whether MLLMs possess a reliable sense of their own knowledge boundaries, we adopt the calibration methodology. We prompt the models to provide their best guess for each question accompanied by a confidence score

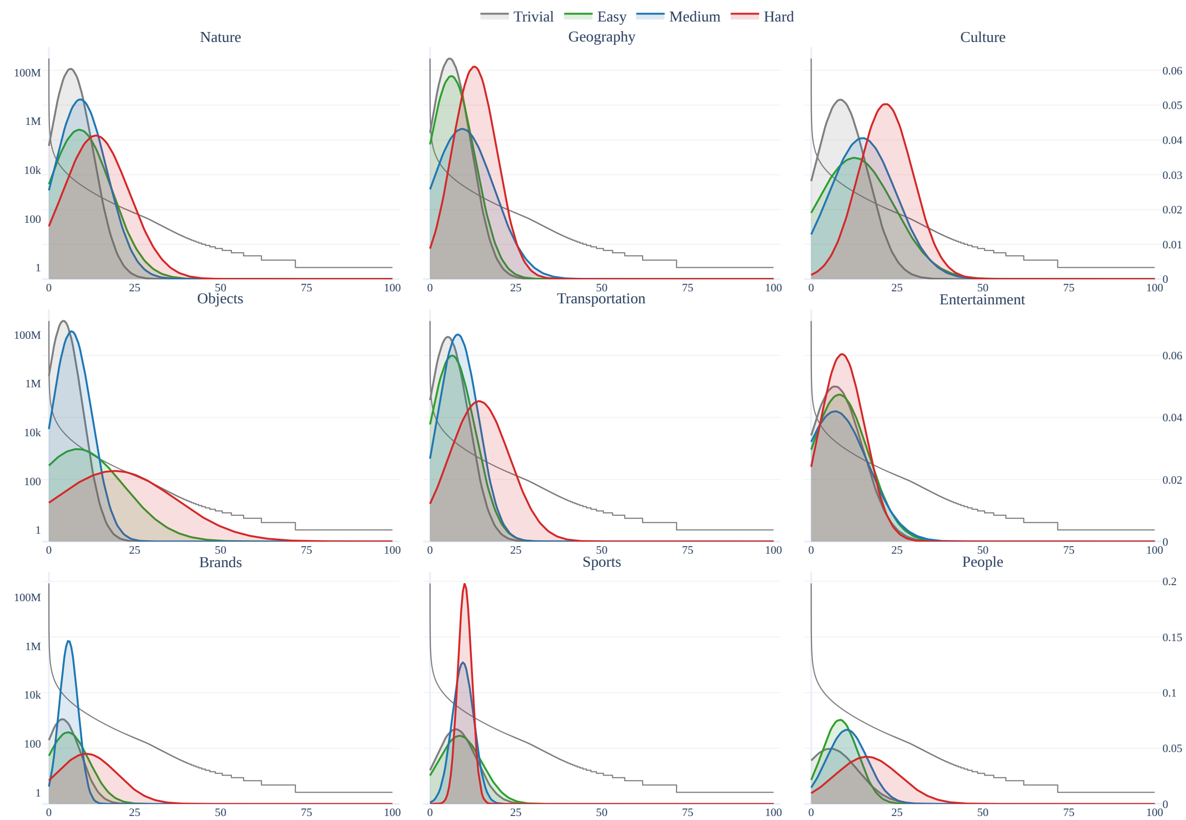


Figure 4: Entity Difficulty Distribution vs. MetaCLIP Frequency Rank Percentile across Categories. These plots illustrate the relationship between real-world entity frequency (proxied by MetaCLIP vocabulary rank percentile) and their assigned difficulty in WorldVQA.

on a scale of 0 to 100 (see Prompt *Question with Confidence* in Appendix A for prompt details). We evaluate calibration performance using two primary metrics calculated over M confidence-ordered bins:

- **Expected Calibration Error (ECE)**: Measures the alignment between subjective certainty and objective accuracy (optimal ECE is 0). It is formulated as: $ECE = \sum_{m=1}^{M} \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)|$
- **Weighted Average Slope (Slope)**: Assesses the correlation between accuracy and confidence. An optimal Slope is 1.0; values significantly below 1.0 indicate systemic overconfidence.

As shown in Figure 5, all models exhibit severe overconfidence. Kimi K2.5 achieves the best calibration (ECE: 37.9%, Slope: 0.550), maintaining the strongest alignment between internal confidence and actual performance, outperforming both GPT and Gemini series.

Regarding confidence distribution (Figure 5 Right), most models lack self-awareness. Gemini-3-pro shows binary behavior, assigning $\geq 95\%$ confidence in over 85% of cases regardless of accuracy. GPT-5.1 is the only model distinguishing low confidence, offering more honest uncertainty estimates despite a slightly higher ECE. This pervasive overconfidence likely stems from a lack of uncertainty samples in training data and alignment strategies favoring assertiveness.

4 Related Work and Discussion

In this paper, we propose a specialized benchmark for measuring atomic visual factuality. WorldVQA sits within an extensive evaluation landscape for MLLMs. A detailed comparison between WorldVQA and other representative benchmarks is provided in Table 1. Comprehensive suites such as MME Fu et al. 2025, MMBench Yuan Liu et al. 2024, SEED-Bench Bohao Li et al. 2023, and MMStar Lin Chen et al. 2024 assess holistic competence, where world

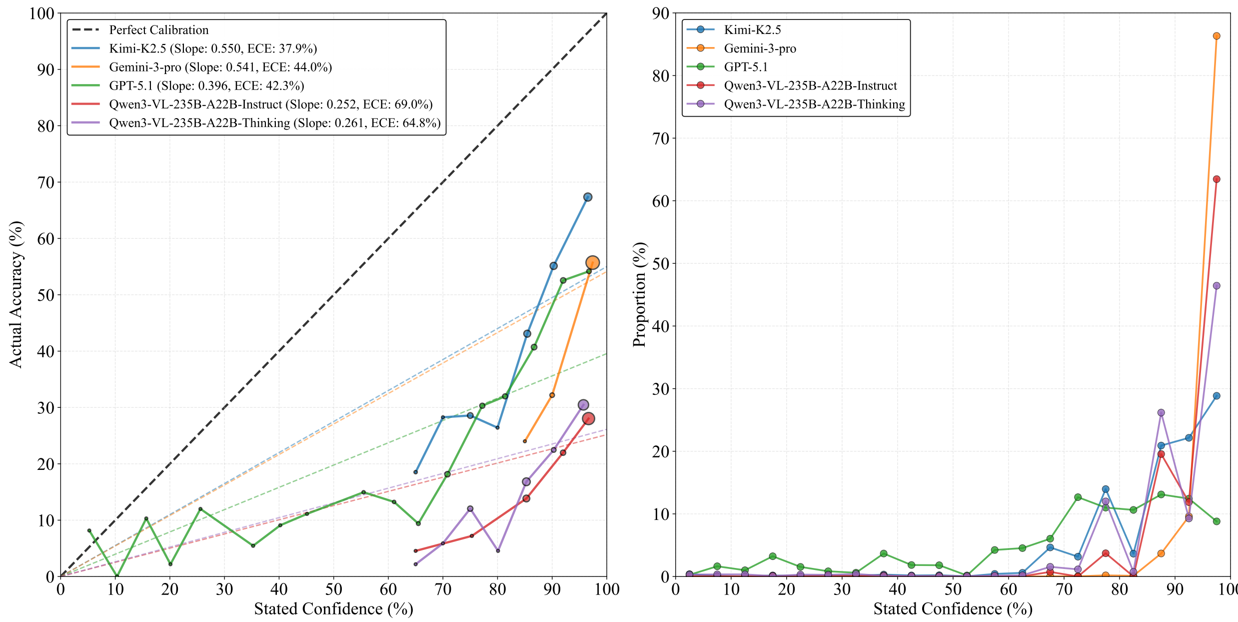


Figure 5: **Calibration and Confidence Distribution Analysis. Left:** Reliability diagrams plotting Actual Accuracy against Stated Confidence. To ensure statistical significance, only bins containing more than 20 samples are visualized. The size of each data point is proportional to the number of samples in that bin. The black dashed diagonal (y=x) represents perfect calibration, while colored dashed lines indicate the weighted average slope for each model. **Right:** The distribution of stated confidence scores across the full dataset (without sample thresholding). The plots reveal a severe overconfidence trend, with most models concentrating their predictions in the 90-100% confidence range.

knowledge is often an implicit prerequisite for tasks ranging from mathematical reasoning Lu et al. 2023 and text recognition Yuliang Liu et al. 2024 to diagram interpretation Hiippala et al. 2021 and spatial localization L. Yu et al. 2016. While expert-level benchmarks like MMMU Yue, Ni, et al. 2024 and MMMU-Pro Yue, T. Zheng, et al. 2025 test deep disciplinary knowledge, their emphasis on complex reasoning chains often obscures purely factual deficits. Closer to our aim are recent factuality probes like SimpleVQA Cheng et al. 2025 and VisualSimpleQA Yanling Wang et al. 2025; however, WorldVQA differentiates itself by focusing on the atomic recognition of entities across a stratified taxonomy—akin to ImageNet Deng et al. 2009 or LVIS Gupta et al. 2019—rather than composite retrieval tasks.

We also examine reliability through the lens of calibration and hallucination, drawing on methodologies established in the language domain Joshi et al. 2017; Kwiatkowski et al. 2019; S. Lin et al. 2022b; Junyi Li et al. 2023; Wei et al. 2024. Notably, prior studies demonstrate that while pre-trained models possess latent self-knowledge Kadavath et al. 2022; S. Lin et al. 2022a, this signal is likely distorted by post-training alignment Achiam et al. 2023. In the multimodal setting, evaluation has largely focused on existential or perceptual hallucination—checking for object presence Rohrbach et al. 2018; Y. Li et al. 2023 or attribute consistency T. Guan et al. 2024; Z. Sun et al. 2024; Junyang Wang et al. 2023. By isolating encyclopedic hallucination, we aim to disentangle the complex interplay between visual perception and parametric knowledge Hanchao Liu et al. 2024, offering a granular view of model trustworthiness distinct from previous polling-based metrics.

A main limitation of WorldVQA is that it measures factuality in a highly atomic setting. While this isolation allows for precise diagnosis of recognition failures, it remains an open research question whether the ability to correctly name specific entities correlates strongly with performance on complex, downstream multimodal tasks. Furthermore, we have not yet fully quantified how different Reinforcement Learning (RL) strategies specifically impact this atomic visual calibration. We hope that open-sourcing WorldVQA provides the community with a rigorous baseline to investigate these dynamics and develop alignment techniques that enhance factuality without compromising uncertainty estimation.

References

Achiam, Josh et al. “Gpt-4 technical report”. In: *arXiv preprint arXiv:2303.08774* (2023).

Chen, Lin et al. “Are we on the right way for evaluating large vision-language models?” In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 27056–27087.

Cheng, Xianfu et al. "Simplevqa: Multimodal factuality evaluation for multimodal large language models". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2025, pp. 4637–4646.

Chuang, Yung-Sung et al. *Meta CLIP 2: A Worldwide Scaling Recipe*. 2025. arXiv: 2507.22062 [cs.CV]. URL: https://arxiv.org/abs/2507.22062.

Deng, Jia et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

Fu, Chaoyou 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.

Guan, Tianrui 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*. 2024, pp. 14375–14385.

Gupta, Agrim, Piotr Dollar, and Ross Girshick. "Lvis: A dataset for large vocabulary instance segmentation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 5356–5364.

Hiippala, Tuomo et al. "AI2D-RST: a multimodal corpus of 1000 primary school science diagrams". In: *Language Resources and Evaluation* 55.3 (2021), pp. 661–688.

Joshi, Mandar et al. "Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension". In: *arXiv preprint arXiv:1705.03551* (2017).

Kadavath, Saurav et al. "Language models (mostly) know what they know". In: *arXiv preprint arXiv:2207.05221* (2022).

Kwiatkowski, Tom et al. "Natural questions: a benchmark for question answering research". In: *Transactions of the Association for Computational Linguistics* 7 (2019), pp. 453–466.

Li, Bohao et al. "Seed-bench: Benchmarking multimodal llms with generative comprehension". In: *arXiv preprint arXiv:2307.16125* (2023).

Li, Junyi et al. "Halueval: A large-scale hallucination evaluation benchmark for large language models". In: *arXiv preprint arXiv:2305.11747* (2023).

Li, Yifan et al. "Evaluating object hallucination in large vision-language models". In: *arXiv preprint arXiv:2305.10355* (2023).

Lin, Stephanie, Jacob Hilton, and Owain Evans. "Teaching models to express their uncertainty in words". In: *arXiv preprint arXiv:2205.14334* (2022).

– "Truthfulqa: Measuring how models mimic human falsehoods". In: *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*. 2022, pp. 3214–3252.

Liu, Hanchao et al. "A survey on hallucination in large vision-language models". In: *arXiv preprint arXiv:2402.00253* (2024).

Liu, Yuan et al. "Mmbench: Is your multi-modal model an all-around player?" In: *European conference on computer vision*. Springer. 2024, pp. 216–233.

Liu, Yuliang et al. "Ocrbench: on the hidden mystery of ocr in large multimodal models". In: *Science China Information Sciences* 67.12 (2024), p. 220102.

Lu, Pan et al. "Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts". In: *arXiv preprint arXiv:2310.02255* (2023).

OpenAI et al. *gpt-oss-120b & gpt-oss-20b Model Card*. 2025. arXiv: 2508.10925 [cs.CL]. URL: https://arxiv.org/abs/2508.10925.

Rohrbach, Anna et al. "Object hallucination in image captioning". In: *arXiv preprint arXiv:1809.02156* (2018).

Schuhmann, Christoph et al. *LAION-5B: An open large-scale dataset for training next generation image-text models*. 2022. arXiv: 2210.08402 [cs.CV]. URL: https://arxiv.org/abs/2210.08402.

Sun, Zhiqing et al. "Aligning large multimodal models with factually augmented rlhf". In: *Findings of the Association for Computational Linguistics: ACL 2024*. 2024, pp. 13088–13110.

Wang, Junyang et al. "Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation". In: *arXiv preprint arXiv:2311.07397* (2023).

Wang, Yanling et al. "VisualSimpleQA: A Benchmark for Decoupled Evaluation of Large Vision-Language Models in Fact-Seeking Question Answering". In: *arXiv preprint arXiv:2503.06492* (2025).

Wei, Jason et al. "Measuring short-form factuality in large language models". In: *arXiv preprint arXiv:2411.04368* (2024).

Xu, Hu et al. *Demystifying CLIP Data*. 2025. arXiv: 2309.16671 [cs.CV]. URL: https://arxiv.org/abs/2309.16671.

Yokoo, Shuhei. *Contrastive Learning with Large Memory Bank and Negative Embedding Subtraction for Accurate Copy Detection*. 2021. arXiv: 2112.04323 [cs.CV]. URL: https://arxiv.org/abs/2112.04323.

Yu, Licheng et al. "Modeling context in referring expressions". In: *European conference on computer vision*. Springer. 2016, pp. 69–85.

Yue, Xiang, Yuansheng Ni, 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*. 2024, pp. 9556–9567.

Yue, Xiang, Tianyu Zheng, et al. "Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark". In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2025, pp. 15134–15186.

Appendix

A Prompts

Question Prompt

Please provide as much detail as possible in your answer. {question}

Question with Confidence Prompt

Please provide as much detail as possible in your answer and give your best guess. {question}
At the end, please provide a confidence score (0-100%).

Visual Audit Prompt

You are a strict "Visual Fact-Checker." Your task is to determine whether the provided original image can serve as the sole, conclusive evidence to support the given answer to a question. Your core principle is: "Believe only what is seen; reject all speculation." You must judge if the image provides unique, conclusive, and exclusive evidence to validate the answer.
The input format is:
"""
Question: {question}
Ground Truth Answer: {ground_truth_answer}
"""
Evaluation Steps and Guidelines:
1. Analyze Question and Answer: Carefully read the Question (Q) and Answer (A) to understand the core knowledge or facts involved.
2. Verify Visual Evidence: Closely examine the image. Search for direct visual evidence that confirms the claims made in Answer (A).

- Point 1 - Clarity: Is the information in the image clear enough to identify Answer (A)?
- Point 2 - Uniqueness/Exclusivity: Does the visual information support *only* Answer (A) for Question (Q), or could it support other reasonable alternatives?
- Point 3 - Completeness: Does the image contain all the necessary information required to answer Question (Q)?

3. Write Reasoning (Mandatory): Detail how you arrived at your conclusion based on specific image details. Clearly list the key visual evidence that supports or refutes the answer.
Final Conclusion Format Requirements:
judge_reason: [Your detailed reasoning]
judge_result: [A, B, or C]
Based on your reasoning, you must select exactly one of the following three options as your judge_result:
A. Determinable (The image fully supports the answer)
B. Inconclusive (Cannot be clearly determined; requires more detail)
C. Incorrect (The image contradicts the answer)

Judge Prompt

Role
You are an expert judge specialized in evaluating the correctness of answers. Your task is to assess whether a model-generated answer is correct based on a given question, the model's response, and the ground truth answer.
Task: Evaluate Answer Correctness
Please classify the model's response into one of the following three categories. Ignore differences in formatting, punctuation, language (Chinese vs. English), or abbreviations/full names. Focus strictly on the core semantics and the level of detail (granularity):

1. **Correct**:
   - The model answer contains the core information of the ground truth.
   - The model answer is semantically consistent with the ground truth and contains no contradictions.
   - The granularity of the model answer is equal to or finer than the ground truth.
   - Extra irrelevant information is allowed as long as it does not conflict with the ground truth.
2. **Incorrect**:
   - The model answer provides information that contradicts the ground truth.
   - The model answer provides the wrong specific entity, value, or description.
   - The granularity of the model answer is coarser than the ground truth, leading to incomplete or insufficiently specific information.
   - Even if the model expresses uncertainty but follows up with a wrong answer (e.g., "I'm not sure, maybe it's B" when the truth is A), it is considered Incorrect.
3. **Unattempted**:
   - The model explicitly states it does not know the answer (e.g., "I don't know," "I cannot answer this question").
   - The model suggests the user search elsewhere (e.g., "Please search the internet").
   - The model answer contains no information from the ground truth but provides no incorrect or contradictory information.

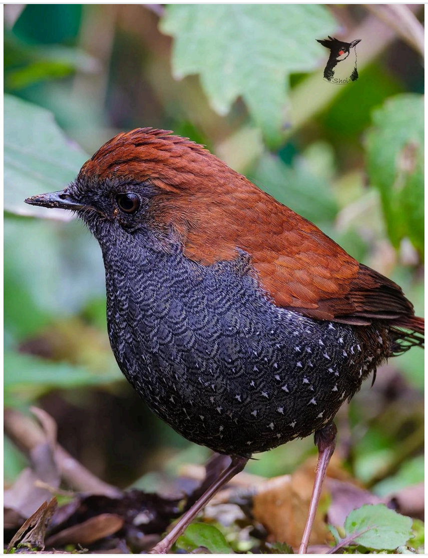
Output Format
Please strictly follow this two-line format for your output:

1. Evaluation: [A brief explanation of your reasoning]
2. Label: [Final classification: "Correct", "Incorrect", or "Unattempted"]

Examples
Input:
Example 1 (Correct - Finer Granularity)
Input:
"""
Question: What weather phenomenon is in the image?
Model Answer: Based on the visual evidence in the image, the weather phenomenon shown is a severe storm with extremely high winds, most likely a tornado or a very powerful hurricane/typhoon.
Ground Truth Answer: High winds
"""
Evaluation: The ground truth is "high winds," and a "tornado" is a more specific and granular type of high wind. The semantics are correct and the detail is finer.
Label: Correct
... (cases of incorrect and unattempted)
Current Task
Input:
"""
Question: {question}
Model Answer: {model_answer}
Ground Truth Answer: {ground_truth_answer}
"""
Evaluation:

B WorldVQA Showcases

B.1 Nature & Environment



What bird is in the picture?

Answer: **Chestnut Shortwing**

What's the name of the flower in the picture?

Answer: **Freesia**

B.2 Locations & Architecture

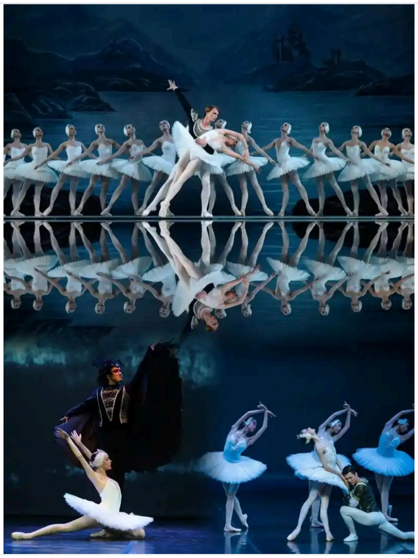
图中出现的内容/文物是/属于哪个遗址?

Answer: **善化寺**

What is the name of the natural landmark shown in the image?

Answer: **Cape of Good Hope**

B.3 Culture, Arts & Crafts



What is the title of the dance performance shown in the picture?

Answer: **Swan Lake**

这个图片是什么珍品

Answer: **战国水晶杯**

B.4 Objects & Products

What style of bag is shown in the picture?

Answer: **Shell bag**

What electronic consumer product is shown in the image? Provide the exact name and model number.

Answer: **iPhone 17 Pro**

B.5 Vehicles, Craft & Transportation

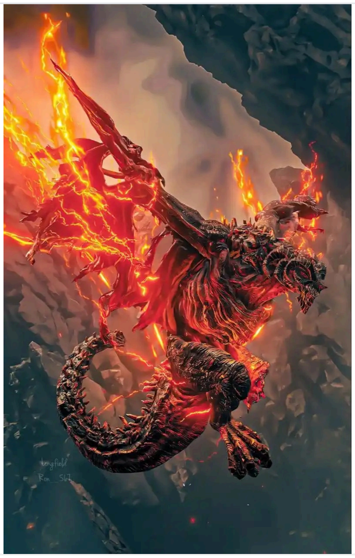
图中的飞行器是什么型号?

Answer: 中国歼 - 20战斗机

What specific attachment or accessory is this for the vehicle?

Answer: Roll cage

B.6 Entertainment, Media & Gaming



What is the name of the character in the picture?

Answer: Bayle the Dread

Which film or TV series is this image from?

Answer: Your Name

B.7 Brands, Logos & Graphic Design

What is the medium (carrier) of the advertisement in this image?

Answer: **Direct-mail advertisement**

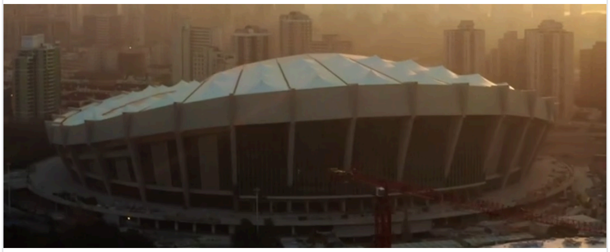
What is the name of the trademark or logo shown in the image?

Answer: **EgyptAir**

B.8 Sports, Gear & Venues

What track-and-field or gymnastics event is shown in the picture? Please be as specific as possible.

Answer: **Floor exercise**



图片中的建筑是哪座体育场馆?

Answer: **上海体育场**