MCSBench: Probing Multimodal Conceptual Structure of Multimodal LLMs

Anonymous CVPR submission

Paper ID 31

Abstract

Do multimodal LLMs (MLLMs) understand concepts structurally like humans? Inspired by cognitive principles, we formalize multimodal conceptual structure *(MCS) as the relational organization grounded in concept-attribute bindings, inter-concept relations, and hypothetical transformations across vision-language space. We introduce* MCSBENCH*, a diagnostic benchmark of 661 questions across seven tasks spanning four cognitive levels (from perceptual grounding to metacognitive verification), paired with 518 structural-integrity probes. We further propose the Structural Alignment Index (SAI), an integrity-aware metric that awards credit only when both the answer is correct and the underlying reasoning is structurally sound. Evaluating* $\sim$*80 MLLMs, we find that performance degrades sharply with cognitive depth, and SAI exposes structural brittleness that accuracy alone conceals. Notably, in-context golden evidence yields substantially smaller integrity gains than accuracy gains on difficult/all questions, suggesting that retrieval augmentation may inflate correctness without facilitating genuine structural reasoning. Additional analyses, such as item-response modeling, scaling analysis, and reasoning-graph diagnosis, validate that* MCSBENCH *provides a reliable, fine-grained diagnostic lens into conceptual-structure failures that existing benchmarks overlook. We will release our dataset and artifacts upon acceptance.*

1. Introduction

Cognitive science holds that concepts are not isolated labels but relationally organized systems: defined by discriminative attributes, structured by similarity and contrast with confusable neighbors, and stress-tested by counterfactual transformation [1, 4, 14]. A system that genuinely understands a concept should maintain this relational structure consistently across modalities, and its correctness should be auditable against evidence, not merely coincidental [11]. Whether multimodal large language models acquire such structure remains unknown because no existing benchmark jointly probes attribute binding, relational inference, counterfactual reasoning, and reasoning integrity.

We formalize multimodal conceptual structure (MCS) as this relational organization and introduce MCSBENCH (including 661 base questions across 7 tasks spanning 4 cognitive levels, overlaid with 518 integrity-checked questions), together with our proposed Structural Alignment Index (SAI), which requires correctness to be structurally justified.

Evaluating $\sim$ 80 MLLMs, we find that SAI can reorder model rankings relative to accuracy, exposing structurally unjustified correctness that current evaluation conceals. Most critically, providing models with explicit evidence substantially improves answer accuracy but yields near-zero integrity gains on difficult items—implying that retrieval-augmented pipelines may inflate correctness without instilling genuine structural reasoning. These results suggest that both current evaluation practices and evidence-grounding strategies require integrity-aware redesign.

Our contributions: `(1)` MCS formalized as a cognitively grounded evaluation construct, instantiated in MCSBENCH across four cognitive levels; `(2)` SAI, validated via IRT calibration, separating correctness from justified correctness; `(3)` We provide a diagnostic analysis suite that turns the benchmark into actionable insight through systematic analysis across different dimensions.

2. MCSBench

2.1. A Unified Relational Perspective

Human does not grasp concepts as isolated atoms. Cognitive science argues that concepts are internally structured: what defines a concept is not a checklist of surface features/properties but a relational organization—encoding which attributes are diagnostic and discriminative for a concept, how concepts neighbor and contrast with confusable alternatives, and how these relations are preserved or transformed under interventions and counterfactual edits [1, 31, 45]. We define **multimodal conceptual structure (MCS)** as this internal relational organization expressed consistently across vision and language: specifically, (i) *concept–attribute role knowledge*—the structured binding of discriminative attributes to concepts grounded in visual evi-

dence [16, 43], and (ii) *inter-concept relational organization*—the coherent arrangement of concepts into confusability neighborhoods and the relational geometry that governs similarity, contrast, and structural transformation across modalities [14, 40, 44]. Critically, MCS is a property of representation: a model possesses it only if its concept–attribute bindings and inter-concept relations are mutually consistent across visual and linguistic inputs, not merely correlated with label statistics. An equally important epistemic principle governs our evaluation design: answer correctness is necessary but insufficient for evidencing MCS, since a model may reach the correct option through surface shortcuts [13] without any structurally valid reasoning. We therefore require correct answers to be supported by verifiable evidence and a consistent reasoning structure—the core motivation behind our integrity-aware overlay, which decouples *correctness* from *justified correctness*.

Guided by these principles, MCSBENCH organizes its tasks into four cognitive levels that probe MCS with increasing operational depth: L1 *Perceptual Grounding*, L2 *Relational Inference*, L3 *Hypothetical Reasoning*, and L4 *Metacognitive Verification*. Each level is implemented as one or more task variants that hold the underlying relational structure fixed while varying the operation demanded of the model (see Table 1). Additional interdisciplinary discussion is provided in Appendix.

2.1.1. MCSBench Task Taxonomy

We define seven base tasks spanning four conceptual reasoning categories and four cognitive levels. Let q denote the query task, c a concept, I_c its object-centric image, R the relation set, A the attribute set, and $T(\cdot)$ a templated prompt.

C-A Align. predicts the correct concept–attribute relation: $\Pr(A \mid T(q),\, I_c)$. *CA-Discr.* picks a^+ over distractor a^-: $\Pr(A^+ \succ A^- \mid T(q),\, I_c)$, distinguishing the concept from its nearest confusable neighbor [44]. *CC-Align.* follows a given relation R to select the correct compositional answer: $\Pr(A^+ \succ A^- \mid T(q),\, R,\, I_c)$; success requires tracking the *relational role* of R, not category-level similarity [14]. *CC-Discr.* infers both the relation type and its filler without a relation label: $\Pr(R,\, c' \mid T(q),\, I_c)$.

Hypo. Reas. (Interventional) applies an explicit do-intervention [33]: $\Pr\big(c' \mid T(q,\, \mathrm{do}(A \leftarrow A')),\, I_c\big)$. *Hypo. Reas. (Counterfactual)* reasons over a fully observed instance under a counterfactual attribute edit: $\Pr\big(c_{A\leftarrow A'} = c^{\mathrm{cf}} \mid T(q),\, R,\, I_c\big)$, with $c_{A\leftarrow A'}$ the potential-outcome variable from Pearl's abduction–action–prediction.

MI-Discr. selects images satisfying a textual attribute/relation description from a candidate set. *MI-Match.* produces a globally consistent assignment between multiple images and multiple descriptions.

Reasoning-Chain Evaluation checks whether a model's explanation is structurally valid w.r.t. question and evidence, measuring the SAI overlay gap between answer correctness and justified correctness.

3. Evaluation Protocol

3.1. Structural Alignment Index (SAI)

We aim to evaluate the structural alignment of conceptual representations in MLLMs. Traditional accuracy metrics are insufficient: models frequently produce correct answers through flawed reasoning [10], exploit spurious shortcuts that intensify with scale [42], and in multimodality, over 70% of widely-used VQA samples can be solved by language priors alone without visual input [19]. Accuracy thus disguises whether a model has genuinely and faithfully learned the underlying conceptual structure or merely exploited statistical regularities. To address this, we propose the **Structural Alignment Index (SAI)**, which jointly requires both answer correctness and structural reasoning fidelity. Here, we introduce the probabilistic and empirical form of SAI as:

$$\mathrm{SAI} = \mathbb{E}_{q\sim\mathcal{Q},\,\tau\sim\mathcal{T}_q}\Big[\Pr_{\pi}\big(\hat{y} = y_q \mid q\big)\ \Pr_{\pi}\big(\hat{t} = t_{q,\tau} \mid q, \tau\big)\Big], \tag{1}$$

$$\widehat{\mathrm{SAI}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{K}\sum_{k=1}^{K}\mathbb{I}(\hat{y}_i = y_i)\ \mathbb{I}(\hat{t}_{ik} = t_{ik}), \tag{2}$$

where $\mathcal{T}_q = \{\tau_{q,1}, \ldots, \tau_{q,K}\}$ denotes K reasoning probes paired with each base question q, each targeting certain structural relation(s) that a correct solution must preserve, with a gold answer $t_{q,k}$. The product structure ensures that a model receives nonzero credit only when it answers q correctly *and* demonstrates faithful structural reasoning on the associated probes. Empirically, SAI reduces to the multiplication of base accuracy and mean probe accuracy per question.

3.2. Experimental Setup

Our evaluation spans 82 multimodal LLMs across three groups: proprietary frontier models (*e.g.*, GPT-5.2, Claude-Opus-4.6, Gemini-3), a broad open-source family (*e.g.*, InternVL3.5/3, Qwen-VL3/2.5, GLM, MiniCPM), and their respective thinking variants. All models are evaluated under two conditions. In the standard *closed-book* setting as default, the model receives only the task image and question prompt, with no access to external knowledge. In the *open-book* setting, the model is additionally provided with in-context evidence of: (i) web-retrieved evidence via search tool (web), or (ii) ground-truth evidence from our curated evidence (evd). We enforce greedy decoding for all models unless specified.

4. Experiments and Analyses

We organize our analyses as: starting from the capability landscape under accuracy and integrity-aware measurement

Table 1. Overview of task schema design in MCSBENCH with eight concrete task variants. Cognitive levels (L0–L4) indicate increasing reasoning complexity. L1=Perceptual Grounding, L2=Relational Inference, L3=Hypothetical Reasoning, and L4=Metacognition.

Task Group	Task Variants	Cog. Level	Task Target
Concept-Attribute	Discrimination, Alignment	L1	*Visual-Attribute Binding*: Align linguistic attribute semantics with visual evidence and discriminate confusing candidates.
Concept-Concept	Discrimination, Alignment	L2	*Cross-Modal Relational Inference*: Align inter-concept (multi-hop) relations across modalities with linguistic specifications and discriminate against confusing concepts or spurious relations.
Hypothetical Reasoning	Interventional/Counterfactual	L3	*Causal Relation Prediction*: Predict how conceptual relations shift under population-level interventions or instance-level counterfactual edits across modalities.
Multi-Image Concept	Discrimination, Matching	L2	*Multi-Image Relational Matching*: Infer many-to-many correspondences between multiple concept images and linguistic descriptions specifying concept-attribute or concept-concept relations.
Reasoning-Chain	Evaluation	L4	*Reasoning Chain Verification*: Evaluate explanatory reasoning chains for evidence faithfulness, logical correctness, and soundness.

(§5.1), ruling out semantic confounders (§5.2), and establishing reliability via IRT (§5.3); then diagnosing what helps, in-context analysis (§5.4) and scaling trend (§5.5), before testing external consistency through cross-benchmark correlations (§5.6).

5. Main Evaluation

5.1. Evaluation Results

Overview. Table 3 reports closed-book performance across all models. The consistent performance gaps across task axes confirm that `MCSBench` captures distinct structural capabilities beyond standard VQA, providing a fine-grained diagnostic that general benchmarks cannot offer.

Proprietary models lead with a clear gap. GPT-5 achieves the highest average (78.37%), and the best open-source models (InternVL3-78B, Qwen2.5VL-72B; 60.06%) trail the proprietary frontier by $\sim$ 14 pp, while many models fall below 30%.

A consistent cognitive gradient: L0 > L1 > L2. Performance decreases monotonically along the cognitive ladder across nearly all models (*e.g.*, InternVL3-78B: 80.46 $\rightarrow$ 66.37 $\rightarrow$ 25.19), with the L1$\rightarrow$L2 drop sharpest for non-frontier models, indicating a persistent inability in inter-concept and hypothetical relations.

Multi-image processing $\neq$ MCS. Several models achieve competitive multi-image scores while remaining weak on L2 (*e.g.*, MiniCPM-V-4.5: MI-Discr. 75.78%, L2 21.37%), showing the bottleneck is higher-order structured conceptual inference, not multi-image processing *per se*.

CoT $\not\Rightarrow$ MCS. Thinking variants do not consistently outperform base counterparts (GLM-4.1V-Thinking 52.04% vs. GLM-4.5V 58.09%), confirming MCS competence is not reducible to chain-of-thought prompting.

Table 4 reports performance on the split with Reasoning-Chain overlay:

SAI exposes structural brittleness hidden by Base accuracy. Across all models and settings, SAI is strictly lower than Base with systematically varied gap magnitude. *CC-Align* and *Hyp. Reas.* show the steepest Base$\rightarrow$SAI drops across all tiers, indicating unfaithful reasoning or alignment in higher-order relations.

Proprietary–open-source gap widens under SAI. The top proprietary model (Gemini-3-Pro-Preview) leads the best open-source counterpart (InternVL3-38B) by $\sim$ 24 SAI pp—wider than the Base gap alone, highlighting the discriminative power of SAI.

Open-book evidence narrows but does not close the SAI gap. Even with ground-truth evidence, strong open-source models (*e.g.*, InternVL3-78B: 57.59 SAI) still trail top proprietary closed-book scores, demonstrating the ability to integrate evidence into faithful reasoning matters.

6. Mechanistic Diagnostics

6.1. Recognition $\neq$ Structure

6.1.1. Semantic Controls Setup

We run two controls to test whether MCSBench difficulty reduces to superficial factors: sensitivity to semantic label cues in the prompt, and performance dependence on category recognition. We augment instructions with label/descriptor cues (semantic-label intervention) across N=12 models on 661 base-task items, and separately filter evaluation to each model's top-k% recognized categories ($k \in \{10, 20, 30, 40, 50\}$). Implementation details see Appendix.

Semantic label cues have negligible effect. The overall accuracy change is small (+0.87 pp, 95% CI: $[0.32, 1.46]$), and an equivalence test (TOST, $\pm$2 pp margin) confirms practical negligibility (p_{TOST}=0.0016). Item-level prediction patterns are stable ($0/12$ models show significant change under McNemar's test), and task-level effects are sparse—only C-A Discr. (+5.88 pp, p=0.019) and C-C Align. (+4.28 pp,

p=0.028) reach significance (Table 5).

Strong recognition does not rescue performance. Recognition filtering improves recognition-adjacent tasks at small k (C-A Align. +19.4 pp at k=10%), confirming the filter is selective, yet relation- and reasoning-heavy tasks do not consistently benefit and can degrade (C-C Discr. −9.8 pp at k=10%–20%, significant after BH-FDR; Fig. 4).

Implication. MCSBench difficulty is not attributable to semantic cues or recognition limitations; it genuinely probes underlying multimodal structure beyond superficial recognition.

6.2. In-Context Diagnostics

We ask whether explicit evidence can close the remaining gap under open-book evaluation, providing k in-context evidence of three types (*Fact*, *Observation*, *Mixed*) and measuring performance gains over a no-evidence baseline on ALL and MEDIUM-HARD subsets. We define Evidence-Utilization Efficiency $\mathrm{EUE}(k) = \frac{\sum_m \Delta\mathrm{SAI}_m(k)}{\sum_m \Delta\mathrm{Acc}_m(k)}$ to quantify structural integrity gains relative to accuracy gains.(1) *Evidence contribution.* Mixed evidence yields the largest mean ΔAcc, yet all types produce substantially lower ΔSAI gains (Fig. 5a): evidence improves surface-level correctness far more than structural reasoning. (2) *Diminishing returns.* Accuracy gains follow a saturation curve ($\Delta(k) = 0.314(1 - e^{-k/4.6})$, $R^2 = 0.981$; $k_{90} \approx 11$), while ΔSAI saturates at a markedly lower ceiling (Fig. 5b). Beyond $k \approx$ 8–12, additional evidence yields negligible benefit for correctness and even less for integrity. (3) *Integrity gap.* EUE < 1 across all k, with efficiency collapsing on MEDIUM-HARD items (Fig. 5c): evidence acts as decision support, not *structural* integrity instiller. This bottleneck motivates our structural diagnosis below.

6.3. Reasoning-Graph Diagnosis of Structural Reasoning

We extract typed reasoning graphs from model rollouts (N=2310; 9 models) and compute WL-based *Stability* and *Faithfulness* scores alongside evidence-anchoring edge features, diagnosing how models reason beyond answer correctness.

Three findings emerge. First, only items that are both stable and faithful achieve high accuracy (mean 0.63 for Robust vs. 0.35 for Disorganized; Fig. 6), showing that structural alignment—not consistency alone, drives success. Second, grounding edges are the strongest OLS predictor of accuracy (std. β=0.099, p<.001), exceeding Stability (β=0.049) and Faithfulness (β=0.042) (Table 6b), implicating explicit evidence-to-structure binding as the key bottleneck. Third, under difficult settings, Stability is an unreliable proxy: its correlation with accuracy flips from ρ=+0.29 on easy items to ρ=−0.14 on hard items (Fisher z=8.59, p<.05; Table 6a), revealing a stable-but-wrong regime that motivates integrity-aware scoring over self-consistency.

7. Scaling Trend

We provide an interpretable and testable decomposition of scaling behavior on MCSBench: how much variance is explained by size, how much by training recipe/lineage, and which structural relation scale the most. We regress per-task accuracy on $\ln P$ and decompose the explained variance (Fig. 7, Tab. 7).

Scaling exists but explains only moderate variance. Most tasks show positive slopes (Avg. $\beta = 7.4$), yet scale alone accounts for only a moderate fraction of variance (Avg. $R^2_{\mathrm{sc}} \approx 0.32$), confirming that MCSBench is not reducible to a parameter-count leaderboard.

Training recipe dominates residual variance. Adding family/lineage intercepts yields the largest incremental gain (Avg. $\Delta R^2_{\mathrm{fam}} \approx 0.50$, p<.05), raising the average fit to $R^2_{\mathrm{full}} \approx 0.87$. Performance gaps thus manifest primarily as recipe-level offsets rather than different scaling rates; family×scale interactions contribute negligibly (see Appendix).

Scaling is task- and depth-dependent. Slopes vary substantially across task groups (Fig. 7): multi-image tasks scale most strongly (MI-Match $\beta = 13.5$), while hypothetical reasoning and deeper cognitive levels are persistent bottlenecks (Hypo. Reas. $\beta = 2.1$; Cog-L3 $\beta = 2.1$), with weaker family gains (Tab. 7), indicating that higher-order structural reasoning is not reliably acquired through capacity scaling alone.

MCSBench exhibits clear but moderate scaling, with training recipe accounting for the majority of residual variation. Critically, scaling gains are non-uniform across conceptual relations, making MCSBench a diagnostic tool for identifying where scale helps and where targeted training are needed.

8. Conclusion

MCSBench formalizes multimodal conceptual structure as a measurable evaluation dimension and reveals, via SAI, that correct answers frequently lack structural justification, a gap unresolved by evidence provision or scale alone. These findings argue for integrity-aware evaluation and structurally grounded system design as field priorities.

References

[1] Lawrence W Barsalou. Perceptual symbol systems. *Behavioral and brain sciences*, 22(4):577–660, 1999. 1, 7

[2] R Darrell Bock and Murray Aitkin. Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, 46(4):443–459, 1981. 12

[3] David Byrne. Complexity, configurations and cases. *Theory, culture & society*, 22(5):95–111, 2005. 7, 8

[4] Ruth MJ Byrne. Precis of the rational imagination: How people create alternatives to reality. *Behavioral and brain sciences*, 30(5-6):439–453, 2007. 1

[5] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024. 7

[6] Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can pre-trained vision and language models answer visual information-seeking questions? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14948–14968, 2023. 7, 8

[7] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2818–2829, 2023. 10

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 10, 19

[9] Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, et al. Agent ai: Surveying the horizons of multimodal interaction. *arXiv preprint arXiv:2401.03568*, 2024. 7

[10] Don Fallis and Peter J Lewis. Right for the wrong reasons: common bad arguments for the correct answer to the monty hall problem. *Synthese*, 207(1):53, 2026. 2

[11] John H Flavell. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, 34(10):906, 1979. 1, 7, 9

[12] Gregor Geigle, Radu Timofte, and Goran Glavaš. African or european swallow? benchmarking large vision-language models for fine-grained object classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2653–2669, 2024. 7

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

[14] Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2):155–170, 1983. 1, 2, 7, 9

[15] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14375–14385, 2024. 7, 8

[16] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990. 2

[17] Ray S Jackendoff. *Languages of the mind: Essays on mental representation*. mit Press, 1995. 7

[18] Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, et al. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. *arXiv preprint arXiv:2502.09621*, 2025. 7, 8

[19] Siddharth Joshi, Haoli Yin, Rishabh Adiga, Ricardo Monti, Aldo Carranza, Alex Fang, Alvin Deng, Amro Abbas, Brett Larsen, Cody Blakeney, et al. Datbench: Discriminative, faithful, and efficient vlm evaluations. *arXiv preprint arXiv:2601.02316*, 2026. 2

[20] Jihyung Kil, Zheda Mai, Justin Lee, Arpita Chowdhury, Zihe Wang, Kerrie Cheng, Lemeng Wang, Ye Liu, and Wei-Lun Harry Chao. Mllm-compbench: A comparative reasoning benchmark for multimodal llms. *Advances in Neural Information Processing Systems*, 37:28798–28827, 2024. 7

[21] Jeonghwan Kim and Heng Ji. Finer: Investigating and enhancing fine-grained visual concept recognition in large vision language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6187–6207, 2024. 7

[22] Michael J Kolen and Robert L Brennan. Linking. In *Test equating, scaling, and Linking: methods and Practices*, pages 487–536. Springer, 2014. 14

[23] Asher Koriat. How do we know that we know? the accessibility model of the feeling of knowing. *Psychological review*, 100(4):609, 1993. 7

[24] Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. Naturalbench: Evaluating vision-language models on natural adversarial samples. *Advances in Neural Information Processing Systems*, 37:17044–17068, 2024. 7

[25] Yijiang Li, Qingying Gao, Tianwei Zhao, Bingyang Wang, Haoran Sun, Haiyun Lyu, Robert D Hawkins, Nuno Vasconcelos, Tal Golan, Dezhi Luo, et al. Core knowledge deficits in multi-modal language models. *arXiv preprint arXiv:2410.10855*, 2024. 7, 8

[26] Yian Li, Wentao Tian, Yang Jiao, Tianwen Qian, Na Zhao, Bin Zhu, Jingjing Chen, and Yu-Gang Jiang. Look before you decide: Prompting active deduction of mllms for assumptive reasoning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 2713–2722, 2025. 8

[27] Zhiyuan Li, Heng Wang, Dongnan Liu, Chaoyi Zhang, Ao Ma, Jieting Long, and Weidong Cai. Multimodal causal reasoning benchmark: Challenging vision large language models to discern causal links across modalities. *arXiv preprint arXiv:2408.08105*, 2024. 8

[28] John M Linacre et al. What do infit and outfit, mean-square and standardized mean. *Rasch measurement transactions*, 16 (2):878, 2002. 14

[29] Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, et al. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. *arXiv preprint arXiv:2408.02718*, 2024. 7

[30] Thomas Mensink, Jasper Uijlings, Lluis Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, and Vittorio Ferrari. Encyclopedic vqa: Visual questions about detailed properties of fine-grained categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3113–3124, 2023. 7, 8

[31] Gregory Murphy. *The big book of concepts*. MIT press, 2004. 1, 7

[32] Gregory L Murphy and Douglas L Medin. The role of theories in conceptual coherence. *Psychological review*, 92(3):289, 1985. 7

[33] Judea Pearl. *Causality*. Cambridge university press, 2009. 2, 7, 8, 9

[34] Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples. *arXiv preprint arXiv:2402.14992*, 2024. 14

[35] Chenhui Qiang, Zhaoyang Wei, Xumeng Han, Zipeng Wang, Siyao Li, Xiangyuan Lan, Jianbin Jiao, and Zhenjun Han. Ver-bench: Evaluating mllms on reasoning with fine-grained visual evidence. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 12698–12705, 2025. 7, 8

[36] Georg Rasch. *Probabilistic models for some intelligence and attainment tests.* ERIC, 1993. 12

[37] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021. 10, 19

[38] Eleanor Rosch. Principles of categorization. In *Cognition and categorization*, pages 27–48. Routledge, 2024. 7

[39] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024. 7, 8

[40] Roger N Shepard. Toward a universal law of generalization for psychological science. *Science*, 237(4820):1317–1323, 1987. 2

[41] Nino Shervashidze, Pascal Schweitzer, Erik Jan Van Leeuwen, Kurt Mehlhorn, and Karsten M Borgwardt. Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research*, 12(9), 2011. 15

[42] Zechen Sun, Yisheng Xiao, Juntao Li, Yixin Ji, Wenliang Chen, and Min Zhang. Exploring and mitigating shortcut learning for generative large language models. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)*, pages 6883–6893, 2024. 2

[43] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980. 2, 7

[44] Amos Tversky. Features of similarity. *Psychological review*, 84(4):327, 1977. 2, 7, 9

[45] Lorraine K Tyler and Helen E Moss. Towards a distributed account of conceptual knowledge. *Trends in cognitive sciences*, 5(6):244–252, 2001. 1, 7

[46] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset-supplementary material. *Reptilia*, 32(400): 1–3, 2017. 10

[47] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024. 7

[48] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, 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)*, pages 15134–15186, 2025. 7

[49] Zhiwei Zha, Xiangru Zhu, Yuanyi Xu, Chenghua Huang, Jingping Liu, Zhixu Li, Xuwu Wang, Yanghua Xiao, Bei Yang, and Xiaoxiao Xu. Constructure: Benchmarking concept structure reasoning for multimodal large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4954–4968, 2024. 7

[50] Jusheng Zhang, Kaitong Cai, Xiaoyang Guo, Sidi Liu, Qinhan Lv, Ruiqi Chen, Jing Yang, Yijia Fan, Xiaofei Sun, Jian Wang, et al. Mm-cot: a benchmark for probing visual chain-of-thought reasoning in multimodal models. *arXiv preprint arXiv:2512.08228*, 2025. 8

[51] Letian Zhang, Xiaotong Zhai, Zhongkai Zhao, Yongshuo Zong, Xin Wen, and Bingchen Zhao. What if the tv was off? examining counterfactual reasoning abilities of multimodal language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21853–21862, 2024. 8

A. Towards A Unified Relational Framework

The pursuit of increasingly general artificial intelligence foregrounds a foundational question: how should foundational models or agents abstract and organize experience into reusable conceptual knowledge [9]? Cognitive science grounds this intuition: experience-distilled conceptual components underpin basic cognitive operations and, in turn, scaffold the higher-level capabilities required for real-world tasks. Across decades of debate, cognitive science has converged on one foundational principle: **conceptual knowledge is internally organized and relational** [14, 32, 38]. In this paper, we formalize *multimodal conceptual structure (MCS)* in multimodal LLMs, motivated by this convergent view, as the grounded relational organization that determines how a concept is internally constituted, how it is situated among interconnected concepts, and how it can be causally explained [17, 32].

From Flat Features to Structured Concepts. Classical definitional theories treated concepts as sets of necessary-and-sufficient features: an object either satisfies the checklist or it does not. Rosch's *prototype theory* overturned this view by demonstrating that category membership is graded, *i.e.*, some members are more typical than others, establishing that concepts possess internal organization beyond binary inclusion [38]. *Structure-mapping theory* shifted the focus from the features themselves to the relations among them: conceptual competence requires preserving structured correspondences that support comparison, contrast, and analogical transfer [14]. *Theory-theory* pushed further still, arguing that even relational similarity is insufficient without explanatory constraints: what lends a concept its coherence is an underlying explanatory theory that specifies why features co-occur and how they are causally linked [32]. On this view, knowing that birds have wings and fly is not merely a statistical regularity but reflects an implicit causal model in which wings enable flight.

B. Extended Related Work

As summarized in Table 2, recent MLLM benchmarks have become increasingly comprehensive, spanning vision-centric evaluation [5, 24, 48], fine-grained recognition [12, 21], web/knowledge base (KB) grounding [6, 30], comparative and multi-image reasoning [20, 29, 49], and reasoning-trace diagnostics [15, 18, 25, 35, 39]. However, these efforts are rarely organized around a cognitively motivated ability ladder that explicitly distinguishes and diagnoses (i) perceptual grounding and attribute binding [1, 43], (ii) relational inference over conceptual structure [14, 44], (iii) counterfactual and interventional predictions [3, 33], and (iv) metacognitive verification of reasoning integrity [11, 23]. This diagnostic need is practically consequential to achieve reliable and robust multimodal intelligent systems [47]. MCSBench is designed to fill this gap by proposing a unified relational benchmark to diagnose concept–attribute, concept–concept, multi-image, and hypothetical reasoning in cognitive conceptual space, paired with integrity-aware evaluation. Complete related works are in Appendix.

We expand the main-paper positioning (Table 2) along three axes that jointly define the gap MCSBench is designed to fill.

Recent MLLM benchmarks have made substantial progress along complementary directions: vision-centric evaluation, fine-grained recognition, web/knowledge base (KB) grounding, comparative and multi-image reasoning, and reasoning-trace diagnostics [5, 6, 12, 15, 18, 20, 21, 24, 25, 29, 30, 35, 39, 48, 49]. Two structural gaps, however, persist across this landscape. First, existing benchmarks are rarely organized around a **cognitively motivated** framework that explicitly distinguishes and diagnoses perceptual grounding, relational inference, hypothetical reasoning, and metacognitive verification as progressively deeper human cognitive operations [1, 3, 11, 44]. Second, much concept/category-centric evaluation remains largely *entity- or label-centric*, *i.e.*, testing whether a model can align a concept name to visual evidence, retrieve associated facts, or classify among semantic labels, rather than adopting a **relational view** in which a concept is characterized by its discriminative attributes, its conceptually confusing neighbors, and the hypothetical transformations that changes its identity/population [14, 31, 45]. Below, we discuss how each gap emerges across three evaluation families.

B.1. From Entity-Centric to Relational Evaluation

Fine-grained benchmarks such as FOCI [12] and Finer [21] aim to evaluate frontier MLLMs on discriminative semantic cues among confusing categories, but their evaluation unit remains the individual concept-label pair success means aligning the correct label to visual evidence, without systematically probing the relational structure (attribute bindings, neighborhood contrasts, hypothetical transformations) that characterizes human conceptual organization [1, 31]. Comparative benchmarks extend the scope toward relational reasoning, but along narrower axes: Constructure [49] evaluates concept abstraction and specialization along taxonomic hierarchies (*i.e.*, superordinate and subordinate), yet does not probe lateral relations among confusers at arbitrary semantic levels; MLLM-CompBench [20] and MMIU [29] benchmark pairwise image relativity and multi-image understanding at scale, yet neither controls for concept-level confusability nor grounds its comparisons in the discriminative attribute structure of a concept family. Noticeably, our primary focus is probing internal conceptual alignment; we additionally include an open-book retrieval setting as a diagnostic lens on evidence utilization.

	Task Design						Cognitive Abilities				
Benchmark	Fine-gr.	Vision-centric	Ext-KB/Web	Comp./Relat.	Ctf./Interv.	Integrity eval.	Percep.	Concept	Relat.	Ctf.	Meta-verify
General vision benchmarks											
MMStar	✗	✓	✗	✗	✗	∼	✓	✗	✗	✗	✗
MMMU-Pro	✗	✓	✗	∼	✗	✗	✓	✓	∼	✗	✗
NaturalBench	∼	✓	✗	✓	✗	✗	✓	∼	✓	✗	✗
Fine-grained concept benchmarks											
FOCI	✓	∼	✗	✗	✗	✗	✓	✓	✗	✗	✗
Finer	✓	∼	✗	✗	✗	✗	✓	✓	✗	✗	✗
Knowledge-intensive / web-grounded VQA											
Encyclopedic-VQA	✓	∼	✓	✗	✗	∼	✓	✓	∼	✗	✗
InfoSeek	✗	∼	✓	✗	✗	✗	✓	✓	∼	✗	✗
Concept-structure / relational reasoning											
CONSTRUCTURE	∼	∼	✗	✓	✗	✗	✓	✓	✓	✗	✗
MLLM-CompBench	✗	∼	✗	✓	✗	✗	✓	∼	✓	✗	✗
MMIU	✗	✓	✗	✓	✗	∼	✓	∼	✓	✗	✗
Evidence / reasoning-trace / integrity diagnostics											
Visual CoT	∼	✓	✗	✗	✗	∼	✓	∼	∼	✗	∼
VER-Bench	✓	✓	✗	∼	✗	∼	✓	∼	✓	✗	∼
MME-CoT	✗	∼	✗	∼	✗	✓	∼	∼	∼	✗	✓
HallusionBench	✗	✓	✗	∼	✗	✓	✓	∼	∼	✗	∼
CoreCognition	✗	∼	✗	∼	∼	✓	∼	✓	∼	∼	✓
MCSBench (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 2. Comparison of MCSBench with representative benchmarks across task-design axes (fine-grained, vision-centric, external knowledge, comparative/relational, counterfactual, integrity evaluation) and cognitive ability dimensions (perception, concept knowledge, relational and counterfactual reasoning, metacognitive verification). ✓: explicitly present as evaluation objective; ∼: implicitly present; ✗: not targeted.

MCSBench adopts the relational cognitive view explicitly: every task is constructed around a local conceptual-space topology, *i.e.*, a target concept situated among semantically close confusers sharing controlled attribute overlap, and requires specific relational operations within that topology, progressing from attribute–concept binding (L1) through inter-concept contrast and cross-image matching (L2) to hypothetical transformation under attribute edits (L3) and metacognitive verification of reasoning integrity (L4). Success therefore requires reasoning over the structure that systematically defines a concept in the complex network, not merely recognizing or labeling it in isolation.

B.2. Hypothetical Reasoning in MLLMs

Hypothetical and counterfactual reasoning has received growing attention as a diagnostic dimension beyond recognition. C-VQA [51] augments VQA questions with counterfactual presuppositions, revealing sharp performance drops, yet its counterfactuals operate at the scene level and are distinct from conceptual structure; CFMM [26] and MuCR [27] probe broad causal or counterfactual competence through binary classification and siamese image pairs, yet neither tests how attribute-level interventions shift concept identity within a controlled confuser set; MM-CoT [50] evaluates chain-of-thought under adversarial distractors but remains verification-centric rather than structure-centric. MCSBench introduces hypothetical reasoning over the conceptual topology, an advanced prediction capability of human [3, 33].

B.3. Evidence and Reasoning-Trace Diagnostics

Prior work evaluates reasoning traces and evidence utilization along complementary but structurally unanchored axes: Visual CoT [39], VER-Bench [35], and MME-CoT [18] assess chain-of-thought quality across perception and cognition subtasks; HallusionBench [15] and CoreCognition [25] diagnose hallucination-style failures; Encyclopedic-VQA [30] and InfoSeek [6] measure evidence utilization yet equate success with answer correctness alone. Since none is organized around a controlled concept-centric topology, attributing failures to specific conceptual structural deficits remains difficult.

C. Additional Details on Dataset Profile

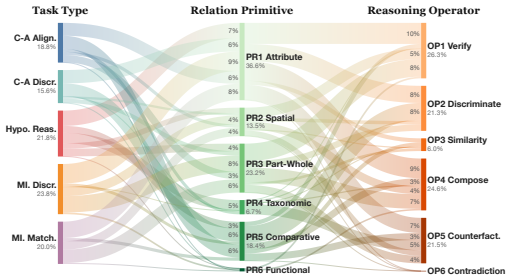

Figure 1. Alluvial diagram of `MCSBench` composition across three axes: task type (left), relation primitive (center), and reasoning operator (right). Edge width encodes the proportion of items linking each pair; percentages denote the marginal share of each node.

MCSBench Task Taxonomy We define seven base tasks spanning four conceptual reasoning categories and four cognitive levels. Let q denote the query task, c a concept, I_c the object-centric image of c, R the relation set, A the attribute set, and $T(\cdot)$ a templated prompt.

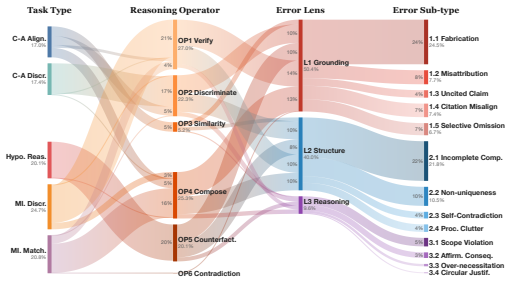

Figure 2. Error distribution of MCSBench overlay across task type, reasoning operator, error lens, and error sub-type.

Concept-Attribute Alignment (C-A Align.) aims to correctly align the attribute relations with the query concept encoded in the image. Given an image and task query, the model predicts the correct concept relation $\Pr(A \mid T(q), I_c)$.

Concept-Attribute Discrimination (CA-Discr.) asks the model to pick the correct attribution a^+ over a distractor a^- for the same question, i.e. $\Pr(A^+ \succ A^- \mid T(q), I_c)$, distinguishing the concept from its nearest confusable neighbor [44].

Concept-Concept Alignment (CC-Align.) requires following a specified relation R to identify the correct compositional answer. Given $T(q)$, R, and I_c, the model prefers the correct candidate: $\Pr(A^+ \succ A^- \mid T(q), R, I_c)$. Because candidates come from the same taxonomic neighborhood, success depends on tracking the *relational role* of R, not category-level similarity [14].

Concept-Concept Discrimination (CC-Discr.) checks whether the model can jointly recover the structural ordering σ, relation R, and concept c' from the query and image: $\Pr(R, c' \mid T(q), I_c)$. Unlike CC-Align, no relation label is given, so the model must infer both the relation type and its filler while rejecting spurious same-neighborhood pairings.

Hypo. Reas. (Interventional) evaluates how the model responds to an explicit intervention on an attribute. Following Pearl's do-calculus [33], given $T(q, \mathrm{do}(A \leftarrow A'))$ and I_c, the model predicts the concept $\Pr\big(c' \mid T(q, \mathrm{do}(A \leftarrow A')), I_c\big)$. Failure suggests the model relies on co-occurrence rather than relational structure.

Hypo. Reas. (Counterfactual) focuses on counterfactual reasoning for a fully observed factual instance. Given evidence $e \coloneqq (T(q), \tau, R, I_c)$ that fixes the realized instance, the model predicts the counterfactual concept under the alternative attribute value A': $\Pr\big(c_{A \leftarrow A'} = c^{\mathrm{cf}} \,\big|\, T(q), R, I_c\big)$, where $c_{A \leftarrow A'}$ is the potential-outcome variable defined by Pearl's abduction–action–prediction.

Multi-Image Discrimination (MI-Discr.) asks the model to identify which images in a candidate set satisfy a textual description of attributes or relations. The output is a selection (or ranking) of images, evaluated by whether true positives are preferred over visually similar distractors, so the task requires grounding relational specs across multiple images at once.

Multi-Image Matching (MI-Match.) tests whether a model can produce a globally consistent correspondence between multiple image instances and multiple textual descriptions. Given a set of images and a set of descriptions, the model outputs a multi-to-multi (or constrained) assignment between them.

Reasoning-Chain Evaluation assesses whether a model's proposed explanation is structurally valid w.r.t. the question and the provided evidence. We construct candidates where the final answer can be correct even when intermediate steps are unsupported or logically inconsistent, thereby measuring the SAI overlay [11] gap between answer correctness and justified correctness.

Fine-grained Perspective on Structure. Fig. 3(c) decomposes each item along two axes: the *relation primitive* it engages—spanning attribute/morphological (PR1), spatial (PR2), part-whole (PR3), taxonomic (PR4), and comparative (PR5) relations—and the *reasoning operator* it demands, ranging from verification (OP1) and discrimination (OP2) through compositional constraint satisfaction (OP4) to counterfactual intervention (OP5). The distribution is intentionally weighted toward PR1 (34.3%) and PR3 (26.4%), emphasizing perceptually grounded relational distinctions over coarse taxonomic knowledge. Crucially, the many-to-many flow confirms that no task type reduces to a single relation–operator pair: Concept–Attribute Alignment routes primarily through OP1 and OP4, while Hypothetical Reasoning concentrates on OP5 despite drawing on equally diverse primitives. This ensures that cross-task performance differences reflect the demanded cognitive operation rather than surface content shifts.

Relation-Operator Decomposition. Fig. 1: We thematically annotate each reasoning-chain question along two axes: relation primitives (PR1–PR6, multi-label), *i.e.*, the semantic relation engaged by the question body and options, and a primary reasoning operator (OP1–OP6), *i.e.*, the cognitive move required to answer correctly. The distribution of MCSBench is intentionally weighted toward perceptually grounded relations with PR1 Attribute (36.6%), PR3 Part-Whole (23.2%), and PR5 Comparative (18.4%) over coarse taxonomic knowledge (PR4, 6.7%). Operators spread more evenly across verification (OP1, 26.3%), composition (OP4, 24.6%), counterfactual (OP5, 21.5%), and discrimination (OP2, 21.3%). Crucially, routing is many-to-many: each task type draws on distinct primitive-operator combinations, confirming that cross-task performance gaps reflect the demanded cognitive operation rather than surface content overlap.

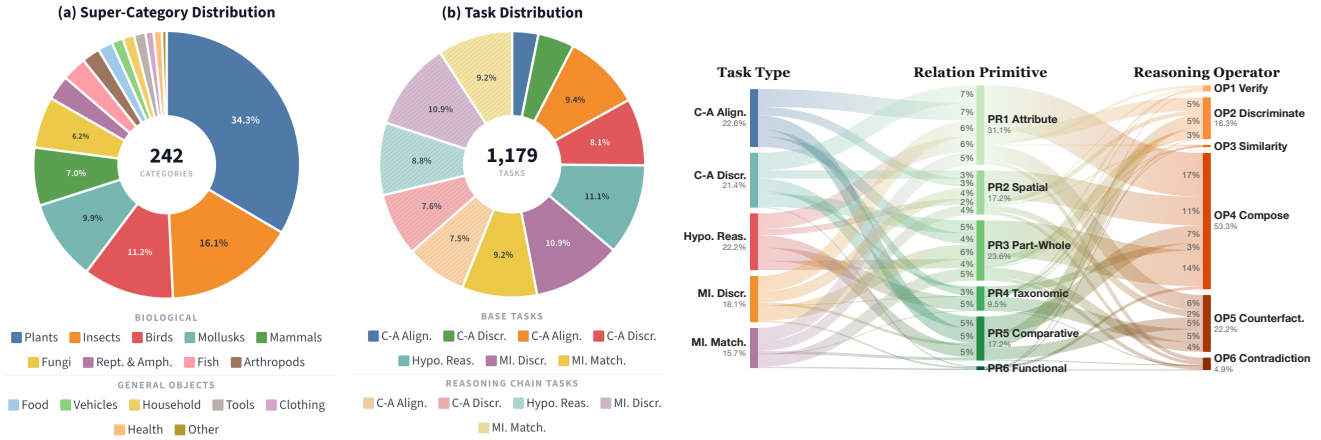

Figure 3. Our MCSBenchis profiled with **(a)** Supercategory distribution of 242 categories. **(b)** 661 questions in 7 base tasks, where 518 are paired with reasoning tasks. **(c)** Fine-grained structural primitives.

Structural Error Taxonomy. We annotate each distractor in the reasoning-chain overlay with a primary error drawn from a three-lens hierarchy applied bottom-up (Fig. 2). L1 Grounding errors (50.4%) concern misalignment between claims and the provided evidence or observations: fabrication of unsupported attributes (24.5%), misattribution of a correct trait to the wrong entity (7.7%), citation misalignment where a cited source does not support the claim (7.4%), and selective omission of required constraints (6.7%). L2 Structure errors (40.0%) concern argument completeness: incomplete comparison, *i.e.*, confirming one option without eliminating plausible confusers (21.8%), and non-uniqueness, where the stated cues remain compatible with multiple candidates (10.5%). L3 Reasoning errors (9.6%) involve formal logical faults such as scope violations (*e.g.*, using out-of-scope criteria (*e.g.*, habitat) as decisive evidence (5.0%)) and affirming the consequent (3.0%). The steep concentration in L1/L2 reveals that the primary bottleneck is evidentiary grounding and contrastive elimination which are precisely the structural competencies our task targets.

D. Details on Dataset Construction & Qualit Control

This section specifies our integral implementations on constructing MCSBench with quality control.

D.1. Dataset Construction

Building Concept Bank. MCSBench evaluates conceptual structure centered on multimodal concepts sourced from a general-purpose concept bank construction pipeline. Each concept is documented with its linguistic name, image set, definition, meronymic attributes, and a set of visually confusing distractors. We source concept names from three visually rich taxonomies: iNat2021 [46] (10,000 fine-grained species across 11 supercategories), ImageNet [8] (1,000 object categories spanning everyday and abstract concepts), and ImageNet-21K [37] (the *small* subset of ∼10,000 synsets sampled from the full 21,841-class hierarchy). Noticeably, we only adopt ImageNet images in the benchmark: our manual inspection reveals that a substantial fraction of iNat2021 images fail to unambiguously depict the target species, often proving indistinguishable among visually confusing distractors or even absent. We therefore retain iNat2021 solely as a concept-name source. Concept names are canonicalized and deduplicated. Intangible concepts (15 rejection super-categories: abstract concepts, emotions, processes, events, relationships, etc.) are also filtered. The curated list contains 15,343 unique and concrete visual concepts. *Next*, we construct a meronymic hierarchy (part-whole attributes) for each concept by prompting a web-augmented agent (o4-mini) based on authoritative sources with Prompt (details see Section I). The obtained hierarchical attributes act as contextual references for the task generation prompts. *Lastly*, we derive visually confusing concepts from a foundational vision encoder (Open-CLIP [7]) and linguistically confusing concepts from web-augmented agents. An optional expansion stage further enriches diversity: additional confusing concepts are iteratively generated and added to the bank, with corresponding images crawled from public sources (*e.g.*, WikiCommons). Noisy images are automatically filtered by subsequent pipeline stages.

Generating Base Tasks. Base items are formatted as 5-way multiple-choice VQA (one correct option). The implementation uses task-specific prompt conditioned generators that collectively instantiate the seven base tasks in Table 2.

Prompting and distractor control. For each sampled (concept, image(s)) pair, the generator is conditioned on concept name/definition, meronymic attribute hierarchy, and (when applicable) the confusable concept set. Prompts of

CVPR #31

each generation is additionally contextualized with the task-specific distractor strategy (*e.g.*, visual similarity, linguistic proximity), enabling both controlled hardness and avoiding shortcuts. In our runs, a web-augmented LLM produces 5 questions per image, each with 5 options; outputs are format-validated (key presence, option format correctness, answer index validity), with automatic retries on malformed outputs; unanswerable images are discarded.

Sampling scale (generation pool). To balance diversity and cost, we sample 100 concepts per concept source partition for confusing-category tasks and 50 concepts per partition for single-image and multi-image tasks, with up to 3 images per concept. This yields a raw pool of thousands of candidate questions before filtering, which is progressively curated by the quality-control pipeline into the final benchmark.

Committee-Based Refinement (Auditor–Challenger–Testee). We implement the committee in Fig. 2 as an iterative refinement loop that targets three failure modes:

- **Challenger (hardening):** refines or regenerates items to prevent trivial elimination strategies (*e.g.*, distractors that are syntactically inconsistent, or apparently solvable without visual evidence), and to increase distractor plausibility while preserving the target test relation.
- **Auditor (faithfulness):** checks overall quality of the question in four aspects: Factual Correctness (answer verifiable), Distractor Correctness (distractor demonstrably incorrect), Image Dependency (visual information involved), and Self-Containedness (no missing context or ambiguity); it then filters likely hallucinated questions based on the quality score, minimal reference number (≥ 3), and target attribute presence.
- **Testee (difficulty sanity):** probes whether items are degenerate (too easy via priors or has modality shortcut); trivial items are discarded.

In our implementation, this committee behavior is instantiated through (i) optional challenge refinement iterations (notably for single-image tasks), and (ii) staged automated auditing used as gating signals: quality scoring, hallucination/grounding checks, answer/distractor verification, and justification sufficiency (detailed in D.2).

Constructing Reasoning Chains (Integrity Overlay). For base tasks with complex relations ($\geq$L2), we construct reasoning-chain evaluation items that probe whether correctness is supported by structurally valid reasoning (measured by SAI). For each retained base item, an Explainer produces a gold reasoning chain (with controlled quality) grounded in (i) explicit visual observations and (ii) atomic textual claims referencing authoritative sources. A red-team attacker then generates structurally perturbed distractors, plausible but incorrect chains with diverse error types (see Fig. 2). One base task is paired with three RC questions. Each RC question is assembled as a multiple-choice selection over candidate chains by sampling 4 combinations from 5 generated distractors and inserting the gold chain at a random position. In total, the benchmark contains 661 base items spanning 7 tasks. Of these, 518 contain the integrity overlay ($\geq$L2), producing 1,554 reasoning-chain evaluation items (3 per base item).

D.2. Quality Control

Overview. Our quality control is a four-stage pipeline: `(1)` shortcut filtering removes items solvable from language alone; `(2)` LLM judgment applies automated multi-criteria adjudication for correctness/grounding/plausibility; `(3)` human curation removes residual ambiguity and validates optimality; `(4)` optional psychometric filtering calibrates item difficulty and screens potentially biased items via measurement modeling.

Shortcut Filtering. We detect language-only shortcuts by evaluating each candidate item under text-only settings using a model ensemble (Gemini-2.5-Pro, Gemini-2.5-Flash, o4mini as Testee) and discarding items that remain solvable without images. Specifically, we run a shortcut detector that flags an item if text-only accuracy exceeds a fixed threshold, indicating likely reliance on linguistic priors rather than image grounding.

LLM Judgment for multi-criteria check. Surviving items are filtered by an automated LLM adjudication stack that evaluates complementary failure modes:

1. **Quality check** (Auditor): assigns a 1–5 quality score and sub-scores over factual correctness, distractor validity, image dependency, and self-containedness.
2. **Hallucination check**: removes items whose claimed evidence is ungrounded, inconsistent, or not aligned with the image and references, using stricter criteria for settings prone to hidden priors.
3. **Option verification**: verifies that only the answer option is correct and optimal, and that each distractor is plausible yet incorrect; items failing answerability or with high leakage are discarded or refined.
4. **Justification** (for RC overlay): evaluates whether the gold reasoning chain with referenced evidence is sufficient, factually correct, and logically sound and coherent, and filters out items where the reasoning cannot be reliably justified given the evidence.

Human Curation After automated filtering, we conduct two rounds of human curation, including annotation and consistency check, on Prolific platform with increased stringency to confirm that `(i)` the gold option is correct and optimal, `(ii)` distractors are plausible and incorrect, and `(iii)` each question is answerable and clear. Among all participants, all native English speakers residing in the United States (60.8%) or the United Kingdom (39.2%) and holding postgraduate degrees (81.0% Master's, 19.0% Doctoral), ensuring the reading comprehension and domain knowledge required for nuanced visual-reasoning judgments. The pool

was demographically balanced (59.5% female; mean age 37.2, σ=8.1, range 23–50; five self-reported ethnic groups) and experienced, with a median of 1,457 previously approved Prolific tasks, reducing annotator bias and strengthening ground-truth reliability.

Psychometric Filtering. As the final quality-control stage, we optionally calibrate items with a Rasch 1PL model so that the released benchmark functions as a stable measurement instrument. We retain items that fall within a target difficulty range ($[-1.5, 1.5]$ logits) and satisfy item-fit and discrimination screens, and then stratify the retained set into easy, medium, and hard tiers for controlled evaluation. Full model specification, filtering criteria, sampling procedure, and bootstrap estimation details are provided in App. F.4.

E. Details on Evaluation Protocol

Evaluation protocol. All models are evaluated through a unified inference pipeline with greedy decoding (`do_sample=False`, t=0.0, `max_new_tokens` = 4096) with fixed random seed to ensure deterministic generation. The default instruction asks the model to select from options A–E (or F–J for reasoning-chain evaluation) and end with the phrase without enforcing or suppressing chain-of-thought; this lets each model respond in its natural style while ensuring fairness. Images are inserted before the text prompt following each model's native multi-image template, resized to a maximum dimension of 448 px with the same aspect-ratio. Answers are parsed by a multi-stage rule-based parser along with GPT LLM with the highest scores retained. For API-served models, we use t=0.0 (whenever feasible) with asynchronous inference (concurrency = 30) and exponential-backoff retries; all other settings are identical. Non-API models are inferred locally with 8×A6000 NVIDIA GPUs. The evaluation scripts and prompts will be released with the benchmark.

F. Details on Experiments & Analysis

F.1. Main Results

F.2. Psychometrics Benchmarking via IRT

A diagnostic benchmark should function as a reliable measurement instrument. We apply Rasch 1PL to a task-balanced variance-sampled short form (149 items) to validate MCSBench under both ACC and SAI scoring, and to quantify how SAI reshapes model ordering. Model abilities are estimated with 95% bootstrap CIs ($B = 200$); SAI abilities are linked to the ACC scale via mean–sigma linking for comparison. Full details are in the Appendix.

Reliable measurement. Both scoring schemes achieve high person-separation reliability (ACC: 0.973; SAI: 0.980), with all items passing infit acceptance (Table 8), confirming MCSBench stably separates model abilities.

Table 5. **Per-task sensitivity to semantic intervention.** For each task, we report the mean accuracy change Δ (pp) across $n = 12$ models, paired t-test statistic t, and effect size d (Cohen's d). $^*p < 0.05$.

Task	Δ (pp)	t	p	d
C-A Align.	-1.16	-0.77	0.459	-0.22
C-A Discr.	+5.88	2.74	0.019*	0.79
C-C Align.	+4.28	2.53	0.028*	0.73
C-C Discr.	+1.22	1.43	0.180	0.41
Hypo. Reas.	+0.89	1.26	0.235	0.36
MI-Discr.	-2.99	-2.01	0.070	-0.58
MI-Match.	-0.08	-0.08	0.934	-0.02
Overall	+0.87	2.88	0.015*	0.83

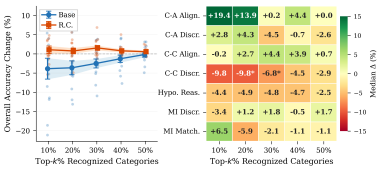

Figure 4. **Recognition-conditioned evaluation.** MCSBench is re-evaluated on the subset of the top-k% recognized categories ($k \in \{10, 20, 30, 40, 50\}$). **Left:** Overall accuracy trend for Base and reasoning-chain (R.C.) tasks across models. **Right:** Per-task median accuracy change Δ (pp) under the same filtering; *significant under Wilcoxon signed-rank test.

SAI imposes stricter difficulty. Mean item difficulty shifts from $b = 0.038$ (ACC) to $b = 0.477$ (SAI), as SAI penalizes "lucky-correct" responses lacking structural justification.

Preserved global order, non-trivial local re-ranking. Model ordering is largely maintained (Kendall $\tau = 0.870$) with no systematic directional bias ($p = 0.525$, n.s.), yet re-rankings are substantial (mean $|\Delta\text{Rank}| = 2.80$, max = 10). Table 9 shows that models such as Kimi-VL-A3B drop 10 positions under SAI while Gemma-3-12B and MiniCPM-V-4 each rise by 7, exposing structural gaps that ACC obscures. SAI thus serves as a *conservative structural refinement*: preserving broad rank structure while revealing where model correctness is structurally unjustified.

F.3. Recognition-Conditioned Evaluation

Setup. Since open-ended generation is difficult to reliably evaluate across MLLMs, we operationalize recognition as a multiple-choice classification task. For each category, we randomly sample 10 images and present the model with a forced-choice question identifying the depicted concept. Distractors are constructed by retrieving the nearest neighbors in OpenCLIP embedding space, ensuring semantically proximate options that prevent trivial elimination.

Additional Results. We present additional performance sensitivity for base and reasoning tasks.

F.4. Item-Response Modeling

Implementation details. To characterize model behavioral patterns and calibrate measurement, we leverage a Rasch 1PL model [36], which places model ability θ_i and item difficulty b_j on a common logit scale: $P(X_{ij}{=}1 \mid \theta_i, b_j) = \sigma(\theta_i - b_j)$. Difficulties are estimated by marginal maximum likelihood [2] (`girth`, 61-point Gauss–Legendre quadrature, $\mathcal{N}(0, 1)$ prior) and abilities by expected *a posteriori* inference. We calibrate under two views: binary accuracy and SAI for both correctness and integrity. Starting from 518

CVPR
#31

Table 3. Performance evaluation on MCSBench base tasks under the closed-book setup. *Align.=Alignment; Discr.=Discrimination; Match.=Matching; Cog.=Cognitive; Con.=Concept; Attr.=Attribute; Hypo.=Hypothetical.*

Model	Con.-Attr.		Con.-Con.		Hypo.	Multi-Image		Cog. Ladder			Avg.
	Align.	Discr.	Align.	Discr.	Reason.	Discr.	Match.	L0	L1	L2	
Proprietary Models											
GPT-5	91.67	84.31	89.58	61.07	64.86	85.16	87.96	87.36	81.72	61.07	78.37
Claude-Opus-4.6	91.67	82.35	83.33	53.44	64.86	84.38	85.19	86.21	79.46	53.44	75.19
Gemini-2.5-Pro	88.89	82.35	81.25	50.38	61.26	87.50	85.19	85.06	79.01	50.38	74.13
Gemini-2.5-Flash	77.78	76.47	73.96	37.40	55.86	78.12	84.26	77.01	73.14	37.40	66.57
O4-Mini	77.78	70.59	81.25	35.11	43.24	85.94	85.19	73.56	74.04	35.11	66.26
GPT-5-Mini	69.44	68.63	76.04	34.35	47.75	81.25	79.63	68.97	71.33	34.35	63.69
Claude-4.5-Sonnet	91.67	78.43	75.00	32.06	37.84	72.66	62.96	83.91	62.08	32.06	59.00
Open-Sourced Models											
InternVL3-78B-Instruct	83.33	78.43	77.08	25.19	38.74	72.66	77.78	80.46	66.37	25.19	60.06
Qwen2.5VL-72B-Instruct	80.56	74.51	72.92	28.24	39.64	77.34	74.07	77.01	66.14	28.24	60.06
Qwen3VL-30B-A3B-Instruct	77.78	58.82	69.79	27.48	43.24	71.88	81.48	66.67	66.59	27.48	58.85
GLM-4.5V	80.56	70.59	72.92	22.90	36.94	74.22	76.85	55.17	60.95	20.61	58.09
InternVL3.5-38B-Instruct	80.56	64.71	69.79	24.43	31.53	70.31	68.52	71.26	60.05	24.43	54.46
MiniCPM-V-4.5	63.89	37.25	57.29	21.37	31.53	75.78	64.81	48.28	58.01	21.37	49.47
Ovis2.5-9B	77.78	54.90	61.46	21.37	31.53	57.81	63.89	64.37	53.50	21.37	48.56
Gemma-3-27B-it	69.44	47.06	58.33	16.79	27.03	64.84	60.19	56.32	52.82	16.79	46.14
Kimi-VL-A3B-Instruct	75.00	47.06	67.71	16.79	26.13	57.81	57.41	58.62	51.92	16.79	45.84
NVILA-15B	77.78	60.78	69.79	24.43	33.33	43.75	35.19	67.82	44.70	24.43	43.72
DeepSeek-VL2	44.44	37.25	60.42	26.72	24.32	59.38	49.07	40.23	48.31	26.72	42.97
Phi-4-Multimodal-Instruct	63.89	37.25	59.38	19.08	14.41	50.78	41.67	48.28	41.31	19.08	37.82
LLaVA-OV-1.5-8B-Instruct	88.89	64.71	64.58	18.32	25.23	27.34	27.78	74.71	34.99	18.32	36.91
R1-Onevision-7B	52.78	33.33	48.96	20.61	23.42	46.88	35.19	41.38	38.60	20.61	35.40
Llama3.2-11B-Vision-Instruct	61.11	41.18	64.58	17.56	23.42	14.84	15.74	49.43	27.99	17.56	28.74
Idefics3-8B-LLama3	52.78	43.14	55.21	17.56	18.02	19.53	21.30	47.13	27.31	17.56	27.99
LongVILA-R1-7B	55.56	47.06	51.04	17.56	25.23	8.59	25.93	50.57	26.19	17.56	27.69
Thinking Models											
Qwen3VL-30B-A3B-Thinking	83.33	66.67	64.58	32.06	32.43	81.25	76.85	73.56	64.33	32.06	59.15
GLM-4.1V-9B-Thinking	75.00	58.82	67.71	25.95	23.42	68.75	68.52	65.52	57.11	25.95	52.04
Kimi-VL-A3B-Thinking-2506	69.44	50.98	53.12	18.32	26.13	53.91	59.26	47.13	41.53	17.56	43.57

Table 4. Performance on MCSBENCH with Reasoning-Chain evaluation overlaid across five base tasks. We report base accuracy and SAI per task, averaged across all questions. Open-book settings provide models with web search (*web*) or ground-truth evidence (*evd*) as context. Models are ranked by descending average SAI within each group.

Model	C-C Discr.		C-C Align.		Hyp. Reas.		MI-Discr.		MI-Match.		Avg.	
	Base	SAI	Base	SAI	Base	SAI	Base	SAI	Base	SAI	Base	SAI
Open-Book Eval												
Gemini-3-Flash-Preview (web)	84.44	82.59	76.14	68.94	74.04	70.83	80.47	78.91	84.26	80.25	79.92	76.51
Gemini-3-Pro-Preview (web)	85.56	79.26	79.55	71.21	79.81	72.44	86.72	82.55	87.04	75.00	83.98	76.45
GPT-5.2-2025-12-11 (web)	80.00	74.81	61.36	50.38	39.42	35.26	86.72	78.12	83.33	68.83	71.04	62.29
GPT-5-2025-08-07 (web)	87.78	70.00	68.18	53.41	58.65	41.99	89.06	75.52	84.26	64.20	78.19	61.71
InternVL3-78B-Instruct (evd)	77.78	77.04	40.91	38.26	24.04	22.44	72.66	71.35	77.78	74.69	59.46	57.59
Qwen2.5-VL-72B-Instruct (evd)	74.44	68.52	42.05	35.23	27.88	24.04	77.34	76.56	74.07	70.99	60.23	56.44
Qwen3-VL-30B-A3B-Instruct (evd)	67.78	60.37	44.32	39.39	27.88	23.72	71.88	67.19	81.48	71.60	59.65	53.47
MiniCPM-V-4.5 (evd)	56.67	52.96	30.68	23.86	21.15	16.67	75.78	71.88	64.81	52.47	51.54	45.30
Proprietary Models—Closed-Book Eval												
Gemini-3-Pro-Preview	84.44	77.41	73.86	65.53	72.12	66.35	90.62	85.42	89.81	75.31	82.82	74.71
Claude-Opus-4.6	82.22	75.93	68.18	56.44	51.92	49.04	84.38	79.69	85.19	88.58	75.93	68.15
O4-Mini	80.00	68.15	45.45	37.12	36.54	30.45	85.94	75.00	85.19	70.68	67.95	57.53
GPT-5-2025-08-07	88.89	62.22	64.77	48.11	58.65	39.10	85.16	66.41	87.96	61.11	77.61	55.98
GPT-5.2-2025-12-11	70.00	63.70	42.05	34.47	39.42	31.41	68.75	59.90	53.70	45.68	55.41	47.55
Open-Sourced Models—Closed-Book Eval												
Internvl3-38B-Instruct	78.12	75.00	33.33	26.13	25.19	19.85	65.62	59.38	79.63	62.96	58.55	50.23
Qwen2.5-VL-72B-Instruct	74.44	61.85	42.05	35.98	27.88	24.36	77.34	64.84	74.07	55.56	60.23	49.36
Qwen3-VL-30B-A3B-Instruct	67.78	47.04	44.32	31.06	27.88	22.44	71.88	52.60	81.48	59.26	59.65	43.31
Gemma-3-27B-it	55.56	33.70	25.00	18.56	19.23	14.42	64.84	49.48	60.19	37.96	46.33	32.05
Kimi-VL-A3B-Thinking-2506	45.56	18.52	30.68	14.02	21.15	11.54	38.28	16.67	55.56	10.80	38.42	14.29
LLaVA-OneVision-1.5-8B-Instruct	62.22	23.33	26.14	11.74	21.15	5.13	27.34	10.94	27.78	10.49	32.05	11.97

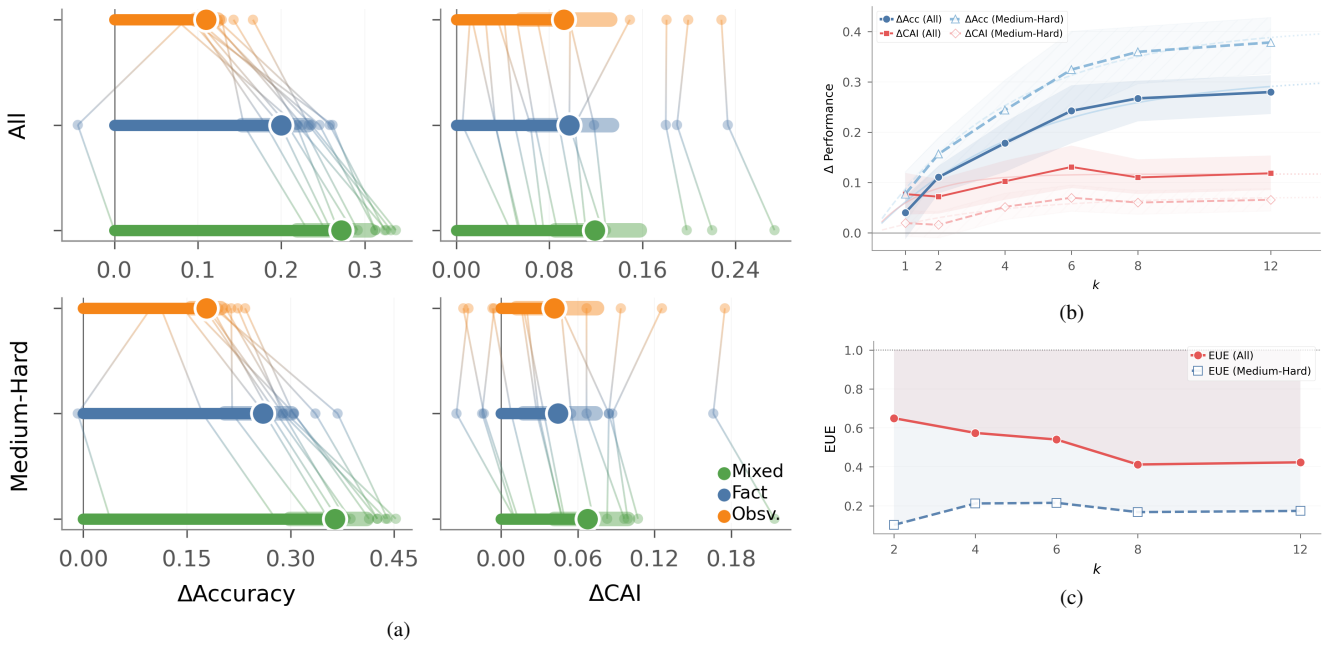

Figure 5. **In-context diagnostics: evidence type, scaling, and utility.** **(a)** Model-wise ΔAccuracy and ΔSAI gains over the no-evidence baseline across three evidence types, for ALL and MEDIUM-HARD subsets. **(b)** ΔAccuracy and ΔSAI trends w.r.t. random evidence count k; Acc gains saturate; SAI gains minorly. **(c)** Evidence-Utilization Efficiency (EUE) across k; EUE < 1 indicates low evidence gains on reasoning integrity.

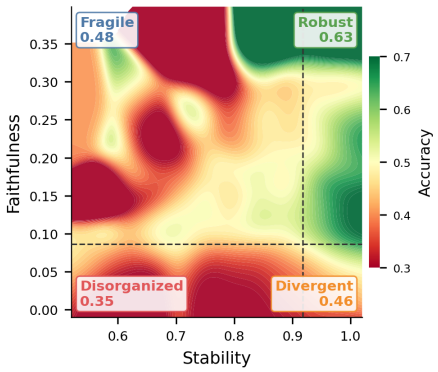

Figure 6. Reasoning Stability–Faithfulness landscape. Contours show binned mean accuracy (w/ thinking rollout); dashed lines mark median thresholds partitioning items into four regimes (Robust, Fragile, Divergent, Disorganized) with inset mean accuracies. $N = 2310$ model×question pairs.

Table 6. Graph-structure diagnostics for `MCSBench`. **(a)** Spearman ρ between Stability and accuracy across Easy/Hard split (Fisher's z: $p < .05$). **(b)** Top OLS predictors of accuracy on reasoning graph features.

(a) Difficulty interaction (Spearman ρ)			
Metric	ρ_{Easy}	ρ_{Hard}	Fisher z
Stability	+0.29	−0.14	8.59*
(b) Top structural predictors (OLS, std. β)			
Feature	β	p	Rank
Edge Type	+0.099	<.001	1
Stability	+0.049	<.001	2
Faithfulness	+0.042	<.001	3

items across 56 models, to satisfy Rasch model's assumption, a two-stage screen [28] removes 180 misfitting items (infit/outfit MNSQ $\notin [0.5, 1.5]$) and 21 low-discriminating items (point-biserial $r < 0.15$); per fit iteration, variance-based proportional sampling [34] then yields a 149-item short form. SAI abilities are linked to the ACC scale via mean–sigma equating [22] (A=0.7497, B=0.0974), with uncertainty from bootstrap resampling (N=200, percentile 95% CIs). Both views achieve person-separation reliability ≥ 0.97, confirming stable measurement while exposing ranking shifts that accuracy alone conceals.

Additional Results. We present entire model ability ranking and delta rank in Fig. 12 9. The distribution of capability and local rank adjustment validate our claims in the main paper.

F.5. Reasoning Graph Analysis

For each model×question pair we sample k=5 stochastic chain-of-thought rollouts (temperature $t \in \{0.7, 0.8\}$) and extract typed reasoning graphs from the resulting rationales using Gemini-3-Pro-Preview with a fixed structured-JSON

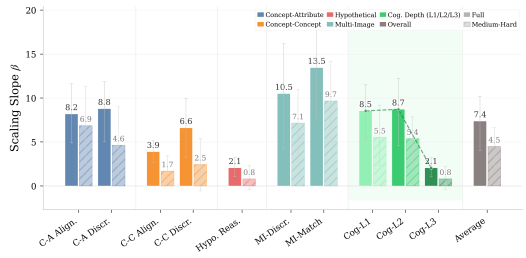

Figure 7. **Scaling slopes β per task** (OLS on $\ln P$) with 95% bootstrap CIs, coloured by task group. Scale alone accounts for moderate variance ($R^2_{\text{sc}} \approx .32$); family intercepts contribute the largest incremental gain ($\Delta R^2_{\text{fam}} \approx .50$), indicating that recipe-level shifts dominate over scaling rates.

Table 7. Variance decomposition of task-level accuracy via nested OLS. R^2_{scale}: $\ln P$ alone; ΔR^2_{fam}: family intercepts; R^2_{full}: full model. $^*p < .05$; $^\dagger p < .10$.

Task	R^2_{scale}	ΔR^2_{fam}	R^2_{full}
C-A Align.	.209	.418*	.691
C-A Discri.	.320	.287*	.747
C-C Align.	.393	.320*	.820
C-C Discri.	.252	.288*	.661
Hypo-Reas.	.179	.285†	.642
MI-Discri.	.200	.618*	.866
MI-Match.	.300	.513*	.867
Avg.	.321	.500*	.870

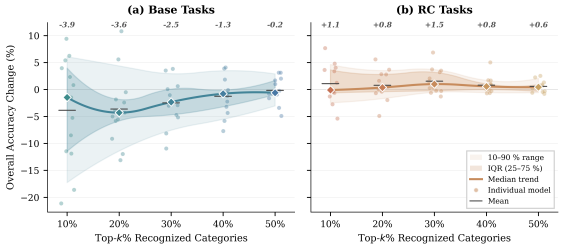

Figure 8. Performance sensitivity to removing top-k% most recognized categories. Base accuracy even minorly degrades, while RC accuracy remains flat, indicating that recognition does not aid correctness and reasoning integrity. Bands: IQR and 10–90 % range across models.

schema. Each graph comprises six node types (`concept`, `img`, `obs`, `fact`, `claim`, `answer`) and five edge types (`grounds`, `supports`, `refutes`, `compares`, `concludes`), assigned deterministically via keyword-based rules. An edge is *evidence-anchored* if it is a `grounds` edge originating from a visual-evidence node (`img`/`obs`). The full extraction pipeline is released with the benchmark.

Graph similarity kernel. We quantify structural similarity with a Weisfeiler–Leman (WL) subtree kernel [41] using h=3 iterations, node-type label initialization, directed neighborhoods encoded with OUT/IN markers and edge-type signatures. Two summary scores are derived from this kernel:

- *Stability*: mean pairwise WL similarity across all $\binom{5}{2}$=10 rollout pairs per question, capturing whether a model converges on a single reasoning strategy or vacillates across stochastic samples.
- *Faithfulness*: mean WL similarity between each rollout graph and the ground-truth reference graph extracted from the released golden reasoning chains (`rc_breakdown`, available for 518/661 items), measuring how closely the model's spontaneous reasoning structure mirrors the intended solution path, independent of answer correctness.

F.6. Scaling Trend

Implementation details. We evaluate 63 proprietary models or open-source VLMs spanning 1.98B–107.71B parameters, representing scale as $x = \log_{10}(\text{params}_\text{B})$. Each model is scored on MCSBENCH under two splits: FULL (661 items) and MEDIUM+HARD (319 items), reporting both ACC and SAI at the task and cognitive-depth level.

Scaling estimation. Scaling slopes are fitted via ordinary least squares (OLS) on the log-linear form $y = A + \beta x + \epsilon$ for each task and cognitive depth group, with 95% confidence intervals obtained by bootstrap resampling over models (N_{boot}=2000, α=0.05).

Scaling trend visualization. We visualize the per-task scaling trend in Fig. 10. We can observe that the variance is larger among the medium-sized models, which further suggests that other confounders, *e.g.*, training recipe, have larger effects, supporting our main claims.

G. Additional Results

H. External Consistency

To establish external consistency, we correlate `MCSBench` overall accuracy with three established benchmarks, *i.e.*, MME, HallusionBench, and SEEDBench, on overlapping model sets (setup in Appendix). Table **??** reports statistically significant rank correlations across all three suites, confirming that `MCSBench` tracks the core visual–language competence axis in the community. Fig. **??** further reveals that MCSBench tasks provide a diagnostically richer decomposition. Most notably, **SAI** uniquely correlates with Hallucination Robustness, confirming that structural-awareness recovers robustness invisible to raw accuracy. We also notice that *Hypo. Reas.* has a lower correlation, suggesting that it may capture missing capabilities.

Table 8. IRT calibration (Rasch 1PL) comparing ACC and SAI scoring on 149 variance-sampled items (95% bootstrap CIs, $B = 200$). Reliability: Rasch person-separation reliability; mean b: item difficulty (logits). Rank agreement between scoring schemes is evaluated after linking SAI to ACC scale.

Key Stats	ACC	SAI
Reliability	$0.973^{+.002}_{-.001}$	$0.980^{+.001}_{-.001}$
Infit acceptable	100%	100%
Mean b (logits)	0.038	0.477
Kendall τ	$0.870^{+.020}_{-.033}$	
Mean $\lvert\Delta\text{Rank}\rvert$	$2.80^{+0.66}_{-0.37}$ (Max = 10)	
Wilcoxon (Rank)	$p = 0.525$	(*n.s.*)

Table 9. Largest rank movers when switching from ACC to structure-aware SAI, sorted by $|\Delta R|$. $\Delta\theta^* = \theta^*_{\text{SAI}} - \theta_{\text{ACC}}$ is the linked IRT ability shift. ▲ rank improves; ▼ rank worsens. Full ranking in Appendix.

Model	ΔR	$\Delta\theta^*$
Kimi-VL-A3B	▼ 10	−0.481
Gemma-3-12B	▲ 7	+0.388
MiniCPM-V-4	▲ 7	+0.396
InternVL3.5-20B-A4B	▲ 6	+0.384
DeepSeek-VL2-Small	▼ 5	−0.298
LLaVA-OV-Qwen2-7B	▼ 5	−0.641
GPT-5	▼ 3	−0.698
Qwen2.5-VL-72B	▲ 3	+0.158
Gemini-3-Flash	▲ 2	+0.014

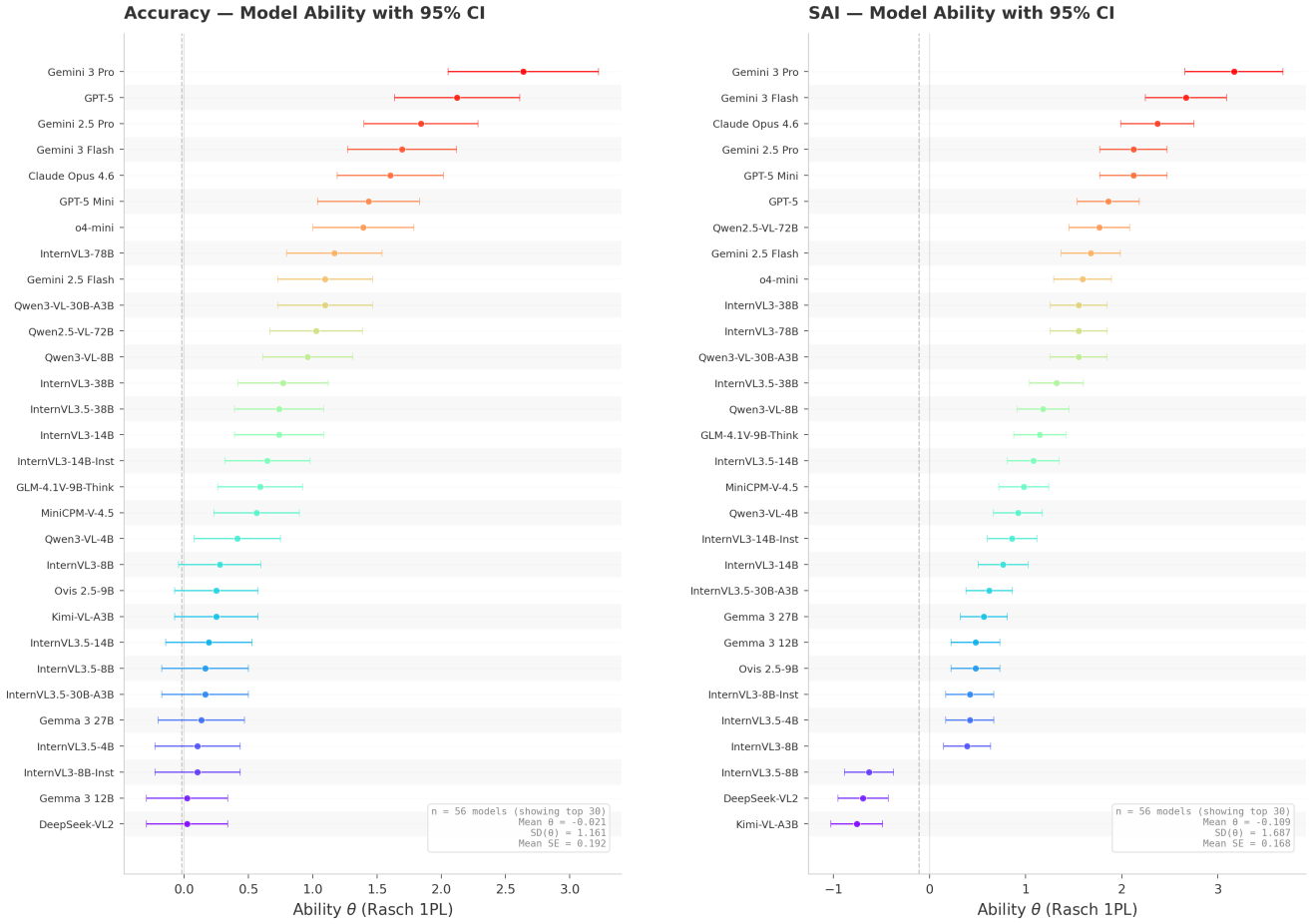

Figure 9. Rasch model ability estimates for top-30 models ranked by accuracy (left) and SAI (right).

H.1. Prompt Sensitivity Analysis

Five prompting variants, Direct, Chain-of-Thought (CoT), Justification, Expert Role, and Contrast, are compared across 9 models with complete prompt grids (Fig. 11). Four conclusions emerge: **(1)** overall accuracy is only mildly sensitive to prompt wording; **(2)** CoT prompting does not improve performance over simple direct instruction; **(3)** Contrast prompting is consistently harmful, suggesting that current models lack robust discriminative comparison ability; and **(4)** within each

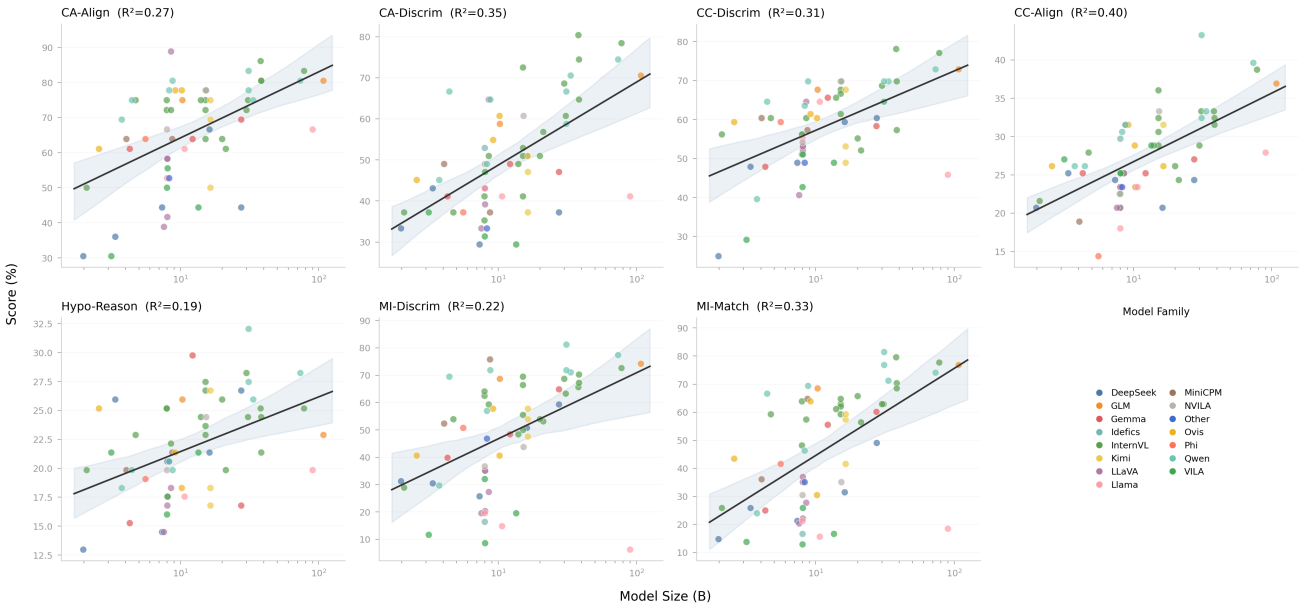

Figure 10. Per-task scaling trend visualization.

cognitive level, the difficulty ordering is preserved across all prompts—prompt engineering modulates absolute scores but does not reshape the benchmark's difficulty profile.

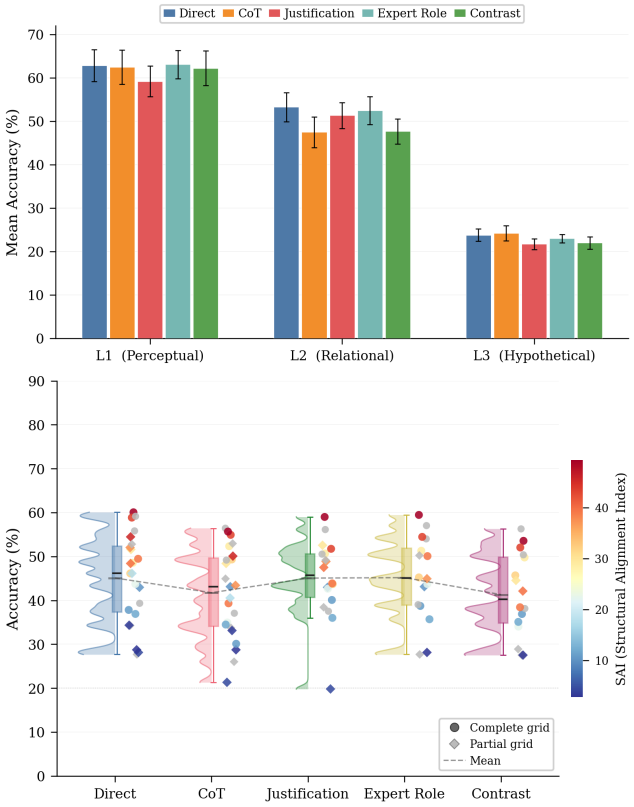

Figure 11. **Prompt sensitivity across cognitive levels and prompt type.** *Left:* Mean accuracy under five prompt variants aggregated over 9 models, with error bars showing between-model variability. Prompt effects are largest at L2 and negligible at L3. *Right:* Per-prompt accuracy distributions; marker color encodes SAI and shape denotes grid completeness. No prompt yields a uniform gain.

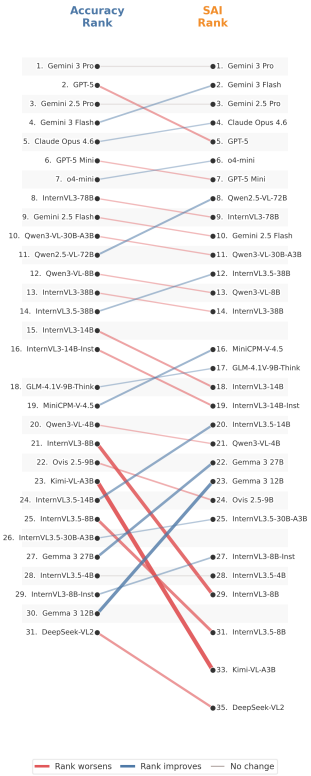

Figure 12. IRT-calibrated rank shift from accuracy to SAI. Rasch ability estimates reorder the leaderboard substantially, confirming that correctness alone does not reflect reasoning integrity.

I. Prompts

I.1. Definition: Relation Primitives

We annotate each question with one or more *relation primitives* that characterize the semantic relation(s) explicitly invoked by the question stem and answer options. Tags are **multi-label**: a single item may involve multiple primitives (e.g., an attribute comparison over a part of an object).

- **PR1 Attribute / Morphological.** Visual attributes and intrinsic properties such as color, shape, texture, size, material, or state (`has-attribute`, `has-value`).
- **PR2 Spatial Configuration.** Relative position and layout relations such as above/below, inside/outside, next-to, orientation, or arrangement (`spatially-related-to`, `relative-position`).
- **PR3 Mereological / Part–Whole.** Part structure and compositional relations including subcomponents, containment, and component membership (`has-part`, `part-of`, `component-of`).
- **PR4 Taxonomic / Definitional.** Category membership and definitional relations such as class/subclass, instance-of, or definitional inclusion (`is-a`, `defined-as`, `instance-of`).
- **PR5 Comparative / Contrastive.** Explicit comparisons or contrastive judgments (e.g., more/less, same/different, similarity/contrast) over concepts or attributes (`differs-from`, `more/less-than`).
- **PR6 Functional / Contextual.** Purpose, causal, or contextual/ecological relations (e.g., used for, caused by, occurs in a situation or environment) (`used-for`, `caused-by`, `occurs-in`).
- **PR0 Miscellaneous.** The relation is ambiguous, underspecified, or does not fit PR1–PR6.

Annotation rule. Tagging is determined solely by the relations *explicitly* engaged by the question and options; annotators do *not* infer additional relations from background knowledge.

I.2. Definition: Reasoning Operators

Each question is assigned exactly one primary operator and up to one secondary operator. Selection follows a strict priority ordering: OP6 > OP5 > OP4 > OP2 > OP3 > OP1.

- **OP1 Verify / Ground**: verify that a concept or image satisfies a described attribute profile.
- **OP2 Discriminate**: identify the discriminative attribute(s) separating a target from confusers.
- **OP3 Similarity / Align**: identify the closest concept under shared-attribute constraints.
- **OP4 Compose / Bind**: satisfy multiple constraints simultaneously; conjunction or multi-image matching.
- **OP5 Counterfactual / Intervene**: hypothetical attribute edit mapped to the nearest concept.
- **OP6 Contradiction / Consistency**: locate which clause is contradicted by evidence or image.

I.3. Relation Classification Prompt

The following prompt is used to classify each question along both dimensions via an LLM-based annotation pipeline.

```
ROLE: You are an expert conceptual-structure
annotator for a multimodal benchmark (MCSBench).
Given a question stem and its answer options,
classify the question along two dimensions: Relation
Primitives and Reasoning Operator.

INPUT:
Question: {ori_question}
Answer options: {ori_options}
Correct answer: {ori_answer}

OUTPUT FORMAT: Return only valid JSON:
{
"rationale": "<1--2 sentence analysis>",
"pr": ["PR#", ...],
"op_primary": "OP#",
"op_secondary": "OP#" | null
}
```

I.4. Prompts for Dataset Generation

J. Discussion and Limitations

K. Data Licensing

MCSBench contains source images from Wikimedia Commons (WikiCM)[1] and ImageNet-21K (IN21K) [8, 37]. All WikiCM images are dominantly verified to carry permissive Creative Commons licenses (CC0, CC BY, or CC BY-SA) at the time of collection. No images with *Non-Commercial* or *NoDerivatives* restrictions were included. Per-image license metadata and author attribution are provided in the released dataset files. IN21K images are governed by a research-only license that prohibits redistribution;[2]. Accordingly, we release only filenames and synset IDs so that authorised users may reconstruct the image set on their own. All MCSBench annotations will be released under CC BY-SA 4.0. which is intended for only non-commercial academic research.

[1] https://commons.wikimedia.org
[2] https://image-net.org/accessagreement

CVPR
#31

Quality Scoring (conf./single variant shown; MI variant analogous)

Role: Domain expert of {category_name} and meticulous evaluator. *Has web search access.*
Goal: Evaluate the quality of existing MC-VQA tasks by analyzing hallucination errors. Score each task on a 1–5 scale.

Core Criteria:
A. **Factual Correctness** — answer must be verifiably correct based on: (1) image evidence, (2) attribute list, (3) authoritative web sources, (4) logical deduction.
B. **Distractor Correctness** — each distractor must contain at least one verifiable inaccuracy or be provably less correct than the labeled answer.

Input: Image of {category_name}, attribute list, existing Q&A tasks.
Output: JSON with per-task ``score'' (1–5), ``explanation'', and ``issues''.

Hallucination Detection

Role: Domain expert in {category_name} and task-auditor for MC-VQA benchmarks.
Goal: Detect hallucination errors in MC-VQA tasks. *Focus solely on the question, correct answer, and image. Disregard distractor errors.*

Procedure:
1. **Review** image and attribute list (silent).
2. **Web search** for verification of biological/visual claims.
3. **Evaluate** each task: check whether the question + correct answer are consistent with the image.

Input: Image, attribute list (§4), VQA tasks (§5).
Output: JSON with ``hallucination_detected'' (bool), ``explanation'', per task.

Accuracy Check

Role: Domain expert in {category_name} and impartial MC-VQA auditor.
Goal: Detect errors in MC-VQA tasks — analyze *every* option (answer + all distractors).

Procedure:
1. **Review** image and task (silent).
2. **Web search** for relevant authoritative information.
3. **Evaluate** — analyze each option *independently* of the original label:
   A. Analyze each option independently for factual accuracy against image.
   B. Determine the correct answer from evidence alone.
   C. Compare with labeled answer; flag discrepancies.

Output: Per-option accuracy verdict, overall task verdict, corrected answer if needed.

Figure 13. Multi-round quality control pipeline. Each part has conf./single-image/MI variants sharing the same structure.

Evidence-Based Explanation (conf. variant shown; single/MI analogous)

Role: Domain expert in {category_name} and evidence-first adjudicator.
Goal:
1. Extract every piece of nomenclature from the task body (question + options).
2. Explain each term in plain language; list common synonyms.
3. Provide authoritative web references for definitions and answer correctness.
4. Construct step-by-step justification grounded only in collected evidence.

Input: Image, confusing categories {confusing_categories_list}, VQA task (§4).
Output: JSON with ``attributes'' (nomenclature terms), ``references'' (web sources), ``evidence'' (atomic facts from image/web), ``reasoning'' (step-by-step chain).

Justification Quality Filter (conf. variant shown; MI analogous)

Role: Domain expert in {category_name} and impartial dataset auditor.
Goal:
1. Evaluate the quality of the reasoning chain produced by *Evidence-Based Explanation*.
2. If incorrect, *minimally tweak* the chain to make it correct.

Input: Image, confusing categories, original VQA task (§4), reasoning chain context (§5): Observations (O#), Evidence (E#), Reasoning steps.
Output: Quality verdict + corrected chain if needed.

Reasoning-Chain (RC) Adversarial Attack (conf. variant shown; single/MI analogous)

Role: Reasoning-chain adjudicator and *adversarial task designer*. *Has web search access.*
Goal: Given a VQA task + its reasoning chain for the correct answer, generate a *new* MC-VQA task where each option is an **incorrect reasoning chain for the same answer**.

Inputs:
- Image of {category_name}; {category_def}
- Confusing categories: {confusing_categories_list}
- Original VQA task with answer (§7)
- REASON_CHAIN: JSON with attributes, references, evidence, reasoning (§8)

Procedure:
1. **Evaluate** the reasoning chain: factual accuracy, logical soundness, image grounding.
2. **Generate distractors:** 4 incorrect reasoning chains (options A–D) that are plausible but flawed.
3. Each distractor must contain a *subtle error* detectable only by an expert:
   - ▷ Fabricated observation, misattributed evidence, uncited claim
   - ▷ Incomplete comparison, self-contradiction, procedural clutter
   - ▷ Scope violation, affirming the consequent, over-necessitation, circular justification

Output: New MC task — question asks "Which reasoning chain correctly justifies the answer?"
Options: 1 correct chain (original) + 4 adversarial chains. JSON with ``error_type'' per distractor.

Figure 14. Evidence-based explanation (5a) and justification quality filter (5b) produce and verify reasoning chains. RC adversarial attack generates MC tasks where distractors are *plausible but subtly flawed* reasoning chains.

Task Def – Identifying Discriminating Features (C-A Align., conf.)

Goal: Pinpoint the specific attribute(s) that best describes the *visual associations* (similarity, difference) between the image category T and confusing categories $[C]$.

Question Variations:
- ▷ "Which single attribute best distinguishes T from C [and $C_2, C_3 \ldots$]?"
- ▷ "What is the *minimal set* of attributes needed to distinguish T from C_1 and C_2?"

Distractor Strategies:
1. Non-Minimal Set — includes redundant features
2. Insufficient Set — fails to distinguish from *all* specified Cs
3. Non-Distinguishing (Shared) — T and all Cs are visually similar on these attributes
4. Partially Discriminating — distinguishes from *some* but not all Cs
5. Subtle/Quantitative — differs only in expert-level quantitative manner

Task Def – Mapping Attributes to Categories (C-A Discr., conf.)

Goal: Identify which confusing category C matches specified criteria for similarity/difference with target T on given attribute(s) $[A]$, $[B] \ldots$

Question Variations:
- ▷ "Which category is most visually similar to T regarding attribute A?"
- ▷ "Which C is similar to T on A *but* different on B?"

Distractor Strategies:
1. Opposite Relationship — opposite relation on the specified attribute
2. Mismatched Criteria — matches on different, unasked-for criteria
3. Partial Match — satisfies some but not all required criteria
4. Target Inclusion — T itself as an option (self-reference error)
5. Subtle Match — nuanced similarity that is hard to verify visually
6. Satisfies Canonically, Not Visually — correct for the species generally, wrong for *this image*

Task Def – Hypothetical Reasoning (conf.)

Goal: Predict which confusing category C a target T would most resemble if specific visual attributes were hypothetically modified (changed, added, removed).

Question Variations:
- ▷ "If T's [Attr. A] were changed to [Value V], which C would the modified T most resemble?"
- ▷ "Which hypothetical modification would make T most similar to C?"

Distractor Strategies:
1. Target as Distractor (T) — when the modification is significant
2. Similarity Decrease — C becomes *less* similar after modification
3. Persistent Difference — C remains clearly distinct even after modification
4. Second-Best Match — second most similar to the modified T
5. Close Relative of Correct — taxonomically close to correct C but subtly different

Figure 15. Task type definitions for **confusing-category** tasks (C-A Alignment, C-A Discrimination, Hypothetical Reasoning). Each definition specifies the goal, question variations, and task-specific distractor strategies injected into the generation prompt.

Task Def – Image → Visual Description Matching (single-image)

Goal: Present candidate attribute–description pairs; choose the one that most accurately describes a feature visible in the image.

Distractor Strategies:
1. Correct Attribute, Incorrect Description
2. Incorrect Attribute, Correct Description
3. Generalization Mismatch — plausible for the category but inaccurate for *this* image
4. Less Significant/Precise Match — less prominent feature than the correct answer
5. Conditional State Mismatch — correct if subject were in a different state/angle
6. Minor Feature Match — real but much less defining feature

Task Def – Detecting Discrepancies (single-image)

Goal: Present a plausible but incorrect statement describing the image. Identify which specific detail is contradicted by visible evidence.

Distractor Strategies:
1. Correct Detail Identification — detail that is actually *correct*
2. Non-Visual/Unverifiable Error — cannot be confirmed from the image alone
3. Minor Error Focus — trivial inaccuracy when a more significant error exists
4. Imprecise Error Description — incomplete description of the error

⋮ *+ 3 more strategies (unrelated attribute, hypothetical context, less critical)*

Task Def – Image Selection by Attributes (multi-image)

Goal: Identify a specific visual instance from a set of 4 closely related images, based on a textual description of 1–2 key diagnostic visual attributes. Category names are *not* revealed; options reference images directly (Image-1, . . . , Image-4).

Distractor Strategies:
1. Confusing Attribute — shared attribute visually confusing among multiple images
2. Partial Match — matches one attribute but fails on another
3. Qualifier Distractor — shared attribute with different visual qualifiers/values
4. Indirect Reference — uses subtle visual descriptions instead of direct image references
5. Set Confusion — adds/removes one borderline image from the correct set

Figure 16. Task type definitions for **single-image** tasks (Image Matching, Discrepancy Detection) and **multi-image** tasks (Image Selection by Attributes). Together with Figure 15, these 7 task types cover the complete MCSBench taxonomy.

Phase 0a: Supercategory Assignment

Task: For each fine-grained category listed below, provide a list of relevant supercategory names.

Requirements:
- ▷ Supercategories should be common and not overly broad or vague.
- ▷ Provide at least 4 supercategory names; favor colloquial names.
- ▷ Rank from most specific (less coarse) to most general (coarser).

Input: {fine-grained categories}
Output: JSON — {``[concept]'': ['supcat_1'', ``supcat_2'', ...]}

Phase 0b: Concept Tangibility Filter

Role: Filter WordNet synset names from ImageNet21K for tangible, visible objects.

Accept if: concrete physical entity with defined boundaries, clearly photographable, consistent visual appearance, can be physically touched.

Reject if falls into meta-categories: abstract concept, emotion, motion/action, scene without boundaries, temporal concept, process/phenomenon, event, relationship, property, social construct, theory, natural phenomenon, activity, . . .

Input: Batch of {n_batch} synset names with definitions.
Output: JSON — [{``[synset]'': ``ACCEPT/REJECT'', ``meta-category'': ..., ``explanation'': ...}]

Figure 17. Phase 0: Concept Preparation. (a) Supercategory assignment builds the taxonomy hierarchy. (b) Tangibility filter removes abstract/non-visual concepts from the candidate pool.

Phase 1: MC-VQA Task Generation (confusing-category variant shown; single-image & multi-image variants analogous)

Role: Domain expert for {category_name}.
Goal: Design challenging, domain-specific MC questions testing fine-grained visual knowledge of {category_name}.

§1 Provided Information:
- Image of {category_name}; {category_def}
- Textual list of professional visual attributes (§9)
- List of visually confusing categories: {confusing_categories_list}

§2 Your Task:
1. Analyze image, attribute list, and confusing categories (internal only).
2. Generate {num_questions} MC tasks; craft an internal plan for difficulty.
3. Web search for additional information about confusing categories as needed.

§3 Question Content: Visual focus, evidence-based, domain attributes only. *Task Definition:* {task} (one of 7 task types from Figure 15)

§4 Choice Format: 5 choices (A–E); exactly 1 correct + 4 distractors.

§5 Strict Constraints:
- *Visual Verification* — all options must reference visible/verifiable features; image takes priority over text.
- *Professional Nomenclature* — formal domain-specific terms only; no colloquial hints.
- *Content Restrictions* — avoid revealing {category_name} or broader classification; no quantitative measurements.
- *Anti-Shortcut* — cannot be answered by elimination or text knowledge alone.

§6 Distractor Requirements: Visual similarity, terminology proximity, expert-level challenge, plausibility, category integration, task-specific challenges, knowledge–image conflict.

§7 Output: JSON list — `[{''question'', ''options'': [A...E], ''answer'', ''strategy''}]`

§8 Exception Handling: Return error JSON if image blurred/irrelevant, attributes corrupted, etc.

§9 Attribute list for {category_name}: {attribute_list}

Figure 18. Phase 1: Main MC-VQA task generation prompt (confusing-category variant). The prompt injects a task-type definition (§3) from one of 7 task types. Single-image and multi-image variants share the same structure, differing only in §1 inputs and category references.

Phase 1 (Challenge): Iterative Difficulty Refinement (2 iterations; single/conf/MI variants)

Role: Expert biologist, evaluator and refiner of AI-generated visual assessments. *Has access to web search.*
Goal: Enhance difficulty of existing MC-VQA tasks so they are significantly more challenging for both expert humans and advanced MLLMs.

§1 Provided Information: Image of {category_name}, attribute list (§7), existing Q&A tasks (§8), web search, original task definition: {task}.

§2 Your Task:
1. Review image, attributes, and existing tasks; create internal mental description.
2. Plan to strategically increase difficulty following all requirements (§3–5).
3. Web search to connect with confusing categories, incorporate subtle visual info, ensure factual correctness.
4. Refine tasks: maximize difficulty by revising questions and options (especially distractors).

§3 Core Refinement Criteria:
A. Maximize distractor plausibility & confusion.
B. Apply task-specific strategies + broader misleading principles: {challenge}
C. Prevent short-cuts: minimal redundancy, consistent format across options, require visual inspection.
D. Maintain original task constraints and definition alignment.

§4 Constraints: *(same as generation: visual verification, nomenclature, content restrictions)*

Meta-Strategy Injected via {challenge}: *Subtle detail inaccuracies; attribute misattributions; plausible but absent features; incorrect spatial/relational descriptions; near-synonym substitutions; phylogenetic neighbor confusion; trait value shifts; contextual re-framing.*

Output: JSON with `''refinement_notes''` explaining changes (iteration 0: 4 choices; iteration 1: 5 choices).

Figure 19. Phase 1 (Challenge): Iterative difficulty refinement. Applied twice (k=0, 1) after initial generation. The meta-strategy list is injected via the {challenge} variable. Each iteration receives the output of the previous round as §8.

Phase 0a: Supercategory Assignment

Task: For each fine-grained category listed below, provide a list of relevant supercategory names.

Requirements:
- ▷ Supercategories should be common and not overly broad or vague.
- ▷ Provide at least 4 supercategory names; favor colloquial names.
- ▷ Rank from most specific (less coarse) to most general (coarser).

Input: {fine-grained categories}
Output: JSON — {``[concept]'': [``supcat_1'', ``supcat_2'', ...]}

Phase 0b: Concept Tangibility Filter

Role: Filter WordNet synset names from ImageNet21K for tangible, visible objects.

Accept if: concrete physical entity with defined boundaries, clearly photographable, consistent visual appearance, can be physically touched.

Reject if falls into meta-categories: abstract concept, emotion, motion/action, scene without boundaries, temporal concept, process/phenomenon, event, relationship, property, social construct, theory, natural phenomenon, activity, ...

Input: Batch of {n_batch} synset names with definitions.
Output: JSON — [{``[synset]'': ``ACCEPT/REJECT'', ``meta-category'': ..., ``explanation'': ...}]

Figure 20. Phase 0: Concept Preparation. (a) Supercategory assignment builds the taxonomy hierarchy. (b) Tangibility filter removes abstract/non-visual concepts from the candidate pool.

Phase 1: MC-VQA Task Generation (confusing-category variant shown; single-image & multi-image variants analogous)

Role: Domain expert for {category_name}.
Goal: Design challenging, domain-specific MC questions testing fine-grained visual knowledge of {category_name}.

§1 Provided Information:
- Image of {category_name}; {category_def}
- Textual list of professional visual attributes (§9)
- List of visually confusing categories: {confusing_categories_list}

§2 Your Task:
1. Analyze image, attribute list, and confusing categories (internal only).
2. Generate {num_questions} MC tasks; craft an internal plan for difficulty.
3. Web search for additional information about confusing categories as needed.

§3 Question Content: Visual focus, evidence-based, domain attributes only. *Task Definition:* {task} (one of 7 task types from Figure 15)

§4 Choice Format: 5 choices (A–E); exactly 1 correct + 4 distractors.

§5 Strict Constraints:
- *Visual Verification* — all options must reference visible/verifiable features; image takes priority over text.
- *Professional Nomenclature* — formal domain-specific terms only; no colloquial hints.
- *Content Restrictions* — avoid revealing {category_name} or broader classification; no quantitative measurements.
- *Anti-Shortcut* — cannot be answered by elimination or text knowledge alone.

§6 Distractor Requirements: Visual similarity, terminology proximity, expert-level challenge, plausibility, category integration, task-specific challenges, knowledge–image conflict.

§7 Output: JSON list — [{``question'', ``options'': [A...E], ``answer'', ``strategy''}]

§8 Exception Handling: Return error JSON if image blurred/irrelevant, attributes corrupted, etc.
§9 Attribute list for {category_name}: {attribute_list}

Figure 21. Phase 1: Main MC-VQA task generation prompt (confusing-category variant). The prompt injects a task-type definition (§3) from one of 7 task types. Single-image and multi-image variants share the same structure, differing only in §1 inputs and category references.

Phase 1 (Challenge): Iterative Difficulty Refinement (2 iterations; single/conf/MI variants)

Role: Expert biologist, evaluator and refiner of AI-generated visual assessments. *Has access to web search.*
Goal: Enhance difficulty of existing MC-VQA tasks so they are significantly more challenging for both expert humans and advanced MLLMs.

§1 Provided Information: Image of `{category_name}`, attribute list (§7), existing Q&A tasks (§8), web search, original task definition: `{task}`.

§2 Your Task:

1. Review image, attributes, and existing tasks; create internal mental description.
2. Plan to strategically increase difficulty following all requirements (§3–5).
3. Web search to connect with confusing categories, incorporate subtle visual info, ensure factual correctness.
4. Refine tasks: maximize difficulty by revising questions and options (especially distractors).

§3 Core Refinement Criteria:

A. Maximize distractor plausibility & confusion.
B. Apply task-specific strategies + broader misleading principles: `{challenge}`
C. Prevent short-cuts: minimal redundancy, consistent format across options, require visual inspection.
D. Maintain original task constraints and definition alignment.

§4 Constraints: *(same as generation: visual verification, nomenclature, content restrictions)*

Meta-Strategy Injected via `{challenge}`: *Subtle detail inaccuracies; attribute misattributions; plausible but absent features; incorrect spatial/relational descriptions; near-synonym substitutions; phylogenetic neighbor confusion; trait value shifts; contextual re-framing.*

Output: JSON with ``refinement_notes'' explaining changes (iteration 0: 4 choices; iteration 1: 5 choices).

Figure 22. Phase 1 (Challenge): Iterative difficulty refinement. Applied twice (k=0, 1) after initial generation. The meta-strategy list is injected via the `{challenge}` variable. Each iteration receives the output of the previous round as §8.