MCSBench: Behavioral Probes of Multimodal Conceptual Structure in MLLMs

Anonymous ACL submission

Abstract

Cognitive accounts treat concepts not as isolated labels but as relationally organized structures: attributes bind to concepts, concepts contrast with confusers, and hypothetical edits reshape these relations. This exposes a gap in multimodal evaluation: a model may pick the correct answer while violating the constraints that justify it. We introduce **MCSBench**, a diagnostic benchmark for multimodal conceptual-structure consistency: 661 base questions across seven task types and 518 reasoning-chain (RC) verification overlays, spanning concept–attribute binding, inter-concept relations, hypothetical reasoning, and multi-image relations. We further propose the **S**tructural **A**lignment **I**ndex (**SAI**), a conservative joint score crediting a prediction only when the answer is correct *and* passes its RC verification probe. Across ∼70 MLLMs, performance drops sharply under higher structural demand, unexplained by label priors. The accuracy–SAI gap widens on inter-concept and hypothetical tasks, where added evidence saturates accuracy but barely narrows the gap; scale helps unevenly, with family and training recipe outweighing parameter count; and self-consistency does not close it. MCSBench reframes MLLM evaluation from picking the right option to satisfying the constraints that justify it. Artifacts: Huggingface/MCSBench.

1 Introduction

Human concept knowledge is grounded in multimodal experience, yet cognitive science argues that concepts are not isolated feature lists but relationally organized systems defined by attribute binding, structured by contrast with connected neighbors, and stress-tested under hypothetical transformation (Barsalou, 1999; Gentner, 1983; Byrne, 2007). This view exposes a gap in current multimodal evaluation: a model may predict the correct answer through shortcuts or priors while failing to preserve the local structural constraints as justified (Geirhos et al., 2020; Flavell, 1979), which has been under-investigated by existing MLLM benchmark protocols (Chen et al., 2024; Yue et al., 2025; Kim and Ji, 2024; Mensink et al., 2023; Zhang et al., 2023).

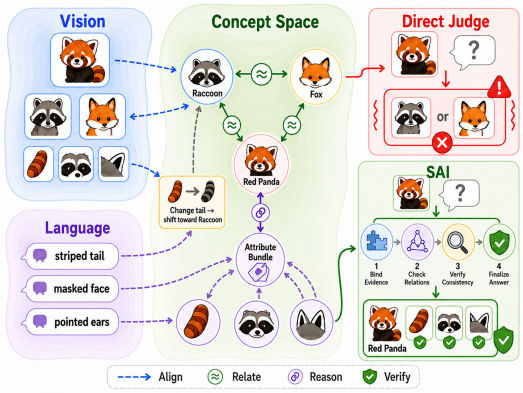


Figure 1: Illustration of multimodal conceptual structure.

We operationalize *multimodal conceptual structure* (MCS) as the behavioral signature of this relational organization across vision and language, and instantiate it in MCSBench: 661 base questions across seven tasks spanning four cognitive levels, paired with 518 integrity overlays. Tasks cover concept–attribute binding, inter-concept relations, hypothetical reasoning, and multi-image relations, each built around visually or semantically confusing concepts so that success requires identifying discriminative attributes, local neighborhood relations, or how hypothetical transformations reshape the structure (Murphy, 2004; Tversky, 1977). To separate correctness from *structurally justified* correctness, we further propose the Structural Alignment Index (SAI), a reasoning integrity-aware score that credits a prediction only when both the answer and its associated structural probe are satisfied.

Evaluating ∼70 MLLMs across closed- and open-book settings, we find a consistent failure

Benchmark Family	Representative Benchmarks	Conceptual & Evidential Design (L0-L3)							Verification & Scoring (L4)	
		Fine-gr./Ent.	Vision-dep.	Confusers	Attr./Rel.	Hypoth.	Comp./MI	Ext-KB	Rat. Verif.	Joint
General VQA	MMStar (Chen et al., 2024); MMMU-Pro (Yue et al., 2025); NaturalBench (Li et al., 2024a)	✗	✓	∼	∼	✗	✗	✗	✗	✗
Fine-Grained Concept Recognition	FOCI (Geigle et al., 2024); Finer (Kim and Ji, 2024)	✓	✓	✓	∼	✗	✗	✗	✗	✗
V-L Compositionality	ARO (Yuksekgonul et al., 2022); VALSE (Parcalabescu et al., 2022); SugarCrepe (Hsieh et al., 2023); Winoground (Thrush et al., 2022)	✗	✓	✓	✓	✗	∼	✗	✗	✗
Knowledge / Web-Grounded VQA	Encyclopedic-VQA (Mensink et al., 2023); InfoSeek (Chen et al., 2023)	✓	∼	✗	∼	✗	✗	✓	∼	✗
Concept / Relational / Multi-Image Reasoning	CONSTRUCTURE (Zha et al., 2024); MLLM-CompBench (Kil et al., 2024); MMIU (Meng et al., 2024)	∼	✓	∼	✓	✗	✓	✗	✗	✗
Hypothetical / Causal VQA	C-VQA (Zhang et al., 2024); MuCR (Li et al., 2024c)	✗	✓	✗	∼	✓	∼	✗	∼	✗
Rationale / Evidence / Cognitive Diagnostics	Visual CoT (Shao et al., 2024); VER-Bench (Qiang et al., 2025); MME-CoT (Jiang et al., 2025); HallusionBench (Guan et al., 2024); CoreCognition (Li et al., 2024b)	∼	✓	∼	∼	∼	✗	∼	✓	✗
MCSBench (Ours)		✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: **Related representative MLLM benchmarks** on target design dimensions relevant to behavioral evaluation of MCS with scope comparison. ✓ denotes an explicit evaluation objective or protocol component, ∼ partial coverage, and ✗ not targeted or featured. *Fine-gr./Ent.*: fine-grained / entity-specific concepts; *Vision-dep.*: vision-dependent design (controls against text-only solving); *Confusers*: controlled semantic confusers; *Attr./Rel.*: attribute binding / relational; *Hypoth.*: hypothetical (attribute/relation-edit) transformation; *Comp./MI*: comparative / multi-image relation; *Ext-KB*: external evidence / KB setting; *Rat. Verif.*: rationale, evidence, CoT, or hallucination/consistency verification; *Joint*: item-level joint answer+verification scoring.

pattern. Accuracy degrades sharply from attribute grounding to inter-concept and hypothetical reasoning, and this difficulty is not reducible to shallow shortcuts: injecting semantic label cues yields negligible accuracy gains. SAI further exposes a gap hidden by accuracy, widest where relational structure is most demanded. Standard remedies fall short: open-book evidence improves accuracy but yields much smaller SAI gains with quick saturation; scaling helps unevenly, and reasoning consistency alone across rollouts does not imply structural integrity. These results argue that current evaluation practices and evidence-grounding strategies require integrity-aware reposition.

Our contributions include: (1) MCSBench, a concept-centric diagnostic benchmark that operationalizes multimodal conceptual-structure alignment via four cognitive tiers; (2) SAI, an integrity-aware metric that separates answer correctness from structurally justified correctness; (3) Broad evaluation and analysis of MLLMs show that current models remain brittle on conceptual-structure reasoning.

2 A Unified Relational Perspective

Cognitive science converges on a relational view of concepts: across prototype theory (Rosch, 2024), structure-mapping (Gentner, 1983), and theory-theory (Murphy and Medin, 1985), concepts are not feature sets but coherently organized relational systems. A concept is defined by which attributes are diagnostic (Tversky, 1977), how it contrasts with confusing neighbors/other concepts (Shepard, 1987), and how these relations transform under hypothetical edits (Byrne, 2007). We use *multimodal conceptual structure* (MCS) as an operational evaluation construct that probes whether the model behavior across vision and language is consistent with this relational organization (yet not about latent representations). Specifically, we operationalize MCS on three behavioral dimensions across vision-language space (Tyler and Moss, 2001; Barsalou, 1999): (i) *attribute-concept binding*: identifying which attributes are diagnostic for a concept and binding them to visual evidence (Tversky, 1977; Treisman and Gelade, 1980; Harnad, 1990); (ii) *inter-concept relational inference*: discriminating a target concept from its confusing neighbors through relational structure rather than surface similarity (Gentner, 1983; Shepard, 1987); and (iii) *hypothetical structural transformation*: predicting how a concept identity or neighborhood relations shift under counterfactual attribute or relation edits (Byrne, 2007; Pearl, 2009).

A second epistemic principle governs evaluation. Classical epistemology holds that a correct belief constitutes knowledge only when supported by appropriate justification (Gettier, 1963). The same empirically applies to MLLMs, which can reach the correct option through shortcuts or post-hoc rationalization (Geirhos et al., 2020; Turpin et al., 2023; Fallis and Lewis, 2026). We therefore require correct answers to also satisfy structural probes—the motivation for our integrity-aware overlay, which decouples *correctness* from *structurally justified correctness*.

Guided by epistemological principles above, MCSBench organizes its tasks into four cognitive levels that probe MCS with advancing cognitive levels: Perceptual Grounding (L1) (Harnad, 1990;

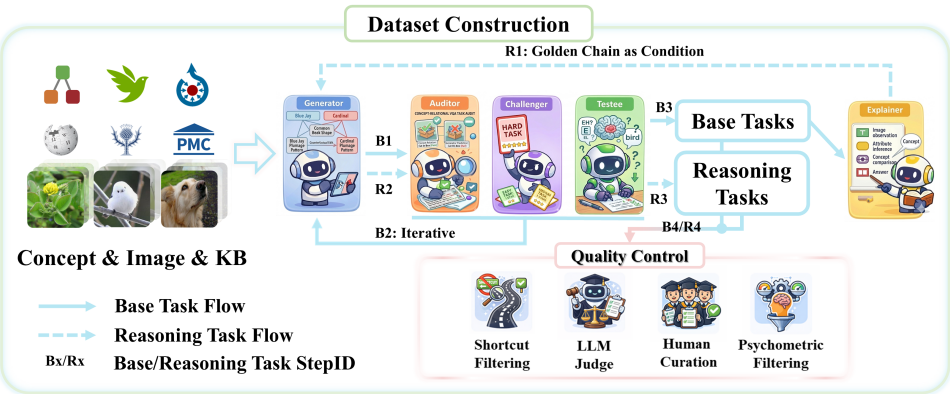


Figure 2: **Dataset construction workflow of MCSBench.** Initially, a generator constructs base tasks fed with collected concept bank, category images, and KB attributes. Generated tasks are iteratively refined and filtered by the Auditor–Challenger–Testee committee; an Explainer then derives golden reasoning chains that condition a second generation pass to yield reasoning tasks. All generated tasks pass through our quality control pipeline comprising shortcut filtering, LLM judgment, human curation, and psychometric filtering for IRT analysis.

Barsalou, 1999), Relational Inference (L2) (Gentner, 1983), Hypothetical Reasoning (L3) (Byrne, 2007), and Metacognitive Verification (L4) (Gettier, 1963; Flavell, 1979; Jacovi and Goldberg, 2020). Each level is implemented as one or more task variants (see Table 2). We denote L0 as the naive recognition capability baseline (Rosch et al., 1976; Palmeri and Gauthier, 2004). Extended theoretical discussions are in the Appendix.

3 Related Work on MLLM benchmarks

As summarized in Table 1, recent MLLM benchmarks advance along complementary but usually separate parts of this design space: general capability VQA (Chen et al., 2024; Yue et al., 2025; Li et al., 2024a), fine-grained recognition (Geigle et al., 2024; Kim and Ji, 2024), vision-language compositionality with hard negatives (Yuksekgonul et al., 2022; Parcalabescu et al., 2022; Hsieh et al., 2023; Thrush et al., 2022), knowledge-grounded VQA (Mensink et al., 2023; Chen et al., 2023), concept-structure, comparative, and multi-image reasoning (Zha et al., 2024; Kil et al., 2024; Meng et al., 2024), counterfactual / hypothetical VQA (Zhang et al., 2024; Li et al., 2024c), and rationale-trace or hallucination diagnostics (Shao et al., 2024; Qiang et al., 2025; Jiang et al., 2025; Guan et al., 2024; Li et al., 2024b). However, these dimensions are typically evaluated in isolation. MCSBench draws these aspects together around controlled concept neighborhoods and provides behavioral evidence about whether models satisfy conceptual-structure constraints. We provide extended related work in Appendix B.

4 MCSBench: Multimodal Conceptual Structure Benchmark

4.1 Dataset Profile and Taxonomy

Dataset profile. MCSBench comprises 661 base questions across seven task types; 518 carry 3 paired reasoning-chain (RC) questions, yielding 1,554 RC items, and span 242 fine-grained categories from established vision taxonomies (Fig. 3). Each item is annotated with relation primitives (PR1–PR6) and reasoning operators (OP1–OP6; Fig. 3c), disentangling cognitive operation from content via many-to-many routing rather than fixed content–operation pairings. Every item is built over a local concept neighborhood with a constraint set Γ_x; the gold answer satisfies Γ_x while each distractor, drawn from the same neighborhood, violates at least one constraint, so MCSBench probes behavioral consistency with explicit conceptual-structure constraints.

L1 — Perceptual Grounding probes *attribute–concept binding*: CA-Align. aligns attributes with the depicted concept, $\Pr(A \mid T(q), I_c)$; CA-Discr. prefers the correct attribute over its nearest confuser, $\Pr(A^+ \succ A^- \mid T(q), I_c)$ (Tversky, 1977).

L2 — Relational Inference probes *inter-concept relations*: CC-Align. follows a relation set R to the correct attribute, $\Pr(A^+ \succ A^- \mid T(q), R, I_c)$, tracking the role of R over category similarity (Gentner, 1983); CC-Discr. jointly recovers R and filler c', $\Pr(R, c' \mid T(q), I_c)$; MI-Discr. and MI-Match. extend this to a candidate image set and a many-to-many correspondence.

L3 — Hypothetical Reasoning probes *coun-*

terfactual structural transformation: given a hypothetical attribute edit, the model predicts the nearest resulting concept, over *interventional* and *counterfactual* families—hypothetical, not causal-identification, reasoning (Byrne, 2007).

L4 — Metacognitive Verification probes whether a chosen answer is supported by structurally valid reasoning: the RC overlay pairs ≥L2 items with candidate explanations, and the model selects the valid chain from distractors given the question, answer, and evidence. RC is the integrity overlay underpinning SAI (Sec. 4.3).

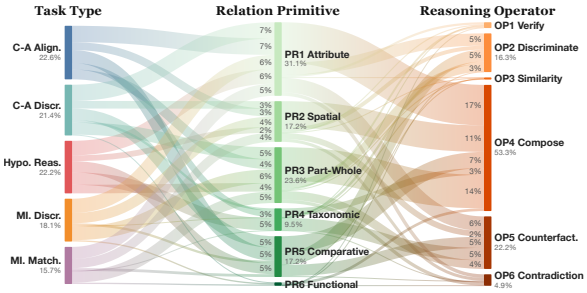


Figure 3: Our MCSBench is profiled with **(a)** Supercategory distribution of 242 categories. **(b)** 661 questions in 7 base tasks, where 518 are reasoning tasks. **(c)** Fine-grained structural primitives.

4.2 Data Construction & Quality Control

Dataset construction. The pipeline employs five specialized agents, *i.e.*, Generator, Auditor, Challenger, Testee, and Explainer, operating in two phases. The Generator produces base tasks conditioned on over 30K concepts paired with clean images, web-retrieved attributes, and visually confusing distractors. To prevent potential model biases, three agents form an inline review committee with cross-model checking: the Challenger forces harder reformulations, the Auditor performs turn-by-turn review, and the Testee monitors difficulty; each is a different model backbone. Curated base tasks are forwarded to the Explainer, which produces a gold reasoning chain with grounding evidence; these chains are fed back to the Generator to produce structurally perturbed distractors under the same committee.

Quality control. All tasks pass a four-stage pipeline ordered by increasing cost: (i) *shortcut filtering* discards items solvable by a text-only model-zoo ensemble; (ii) *LLM-panel judgment* screens answerability, correctness, distractor plausibility, and logical soundness; (iii) *human quality assurance* flags residual ambiguity, visual-grounding, and option-mapping issues, with redundant review on a calibration subset; (iv) optional *psychometric filtering* fits a 1-PL Rasch model and stratifies retained items into easy, medium, and hard tiers. As a release-time independent audit, four frontier LLM judges further evaluate the released benchmark on six quality dimensions, with mean pass rates $> 94\%$ on every dimension (see Appendix).

4.3 Evaluation Metric: Structural Alignment Index (SAI)

Accuracy alone is insufficient as a measure of conceptual-structure consistency: models frequently reach correct answers through flawed reasoning (Fallis and Lewis, 2026), exploit spurious shortcuts that intensify with scale (Sun et al., 2024), and over 70% of widely-used VQA samples are solvable from language priors alone (Joshi et al., 2026). Echoing the epistemic principle of Sec. 2, *i.e.*, a correct answer is not knowledge unless it satisfies the constraints that justify it, we propose the **Structural Alignment Index (SAI)**, a conservative joint score that credits a prediction only when the base answer is correct *and* the model selects a reasoning chain consistent with the item's annotated conceptual-structure constraints. SAI is a behavioral measure under the RC verification protocol.

Each base question q is paired with K reasoning-chain probes $\mathcal{T}_q = \{\tau_{q,1}, \ldots, \tau_{q,K}\}$. Treating (q, τ) with $\tau \sim \mathrm{Uniform}(\mathcal{T}_q)$ as a joint sampling unit, SAI is the expected joint correctness of the final answer and the selected reasoning chain, $\mathbb{I}(\hat{y} = y_q \wedge \hat{t} = t_{q,\tau}) = \mathbb{I}(\hat{y} = y_q) \cdot \mathbb{I}(\hat{t} = t_{q,\tau})$:

$$\begin{aligned} \mathrm{SAI} &= \mathbb{E}_{q,\tau}\Big[\, \mathbb{I}(\hat{y} = y_q) \cdot \mathbb{I}(\hat{t} = t_{q,\tau}) \,\Big] \\ &= \mathbb{E}_{q}\Big[\, \mathbb{I}(\hat{y} = y_q) \cdot \mathbb{E}_{\tau|q}\big[\mathbb{I}(\hat{t} = t_{q,\tau})\big] \,\Big], \end{aligned} \tag{1}$$

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

where y_q and $t_{q,\tau}$ denote the gold answer and the valid reasoning chain for question q. The factorization holds because $\mathbb{I}(\hat{y} = y_q)$ does not depend on τ, yielding base accuracy weighted by mean RC-verification accuracy per question. By design, SAI is a conservative lower bound on the accuracy satisfying the structural constraints.

Table 2: **Overview of task schema design** in MCSBench with eight concrete task variants. Cognitive levels (L0–L4) are operationally defined tiers of cognitive demand that organize tasks by the primary structural operation required. L0=Recognition Baseline, not included in MCSBench. L1=Perceptual Grounding, L2=Relational Inference, L3=Hypothetical Reasoning, and L4=Metacognitive Verification. The BASE and RC columns present the number of base and reasoning-chain overlay questions per group. RC overlays measured only for L2–L3 base tasks.

Task Group	Task Variants	Cog. Level	# Base	# RC	Task Target
Concept-Attribute	Discrimination, Alignment	L1	87 (7.4%)	–	*Visual-Attribute Binding*: Align linguistic attribute semantics with visual evidence and discriminate confusing candidates.
Concept-Concept	Discrimination, Alignment	L2	207 (17.6%)	178 (15.1%)	*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	131 (11.1%)	104 (8.8%)	*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	236 (20.0%)	236 (20.0%)	*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	–	518 (43.9%)	*Reasoning Chain Verification*: Evaluate explanatory reasoning chains for evidence faithfulness, logical correctness, and soundness.
Total	7 Base Tasks + RC Overlay		661 (56.1%)	518 (43.9%)	1,179 questions

Table 3: Performance evaluation on MCSBench base tasks under the closed-book setup. Only selected models are present. *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
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
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
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
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
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
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-235B-A22B-Thinking	72.22	74.51	78.12	25.19	43.24	82.81	81.48	73.56	71.56	25.19	62.63
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
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

5 Experiments and Analyses

5.1 Experimental Setup

We evaluate ~ 70 multimodal LLMs spanning 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 reasoning-tuned ("thinking") variants of both. Each model is evaluated under two conditions. In the default *closed-book* setting, the model receives only the task image and prompt. In the *open-book* setting, the model additionally receives in-context evidence from one of two sources: WEB—retrieved on demand via the model's search tool or EVD—extracted from the gold reasoning chain of each item. All inference uses greedy decoding under a unified pipeline; full implementation details are in Appendix.

5.2 Main Evaluation

Tables 3 and 4 report closed-book base-task performance and RC evaluation, which support:

The bottleneck is relational, not perceptual. Performance is highest when an item is solvable by attribute grounding or instance selection, and drops sharply once the model must separate a target concept from its confusers. The clearest bottleneck is CC-Discr.: GPT-5 falls 23% from CA-Discr. to CC-Discr., InternVL3-78B 53%, and GLM-4.5V 48%. This is not a multi-image-throughput effect: GLM-4.5V scores $74.2/76.9$ on the two multi-image tasks and Qwen3VL-30B-A3B-Instruct $71.9/81.5$, yet both collapse on CC-Discr. (22.9 and 27.5). The failure is on the relation that distinguishes a

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. Only selected models are present.

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-5-2025-08-07 (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
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
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
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

concept from nearby confusers.

Accuracy and SAI diverge most where structure is demanded. SAI is stricter than accuracy by design; the informative signal is the size and location of the gap. On the RC-overlay subset, base-to-SAI gaps exceed 25% on relation-heavy and matching tasks (*e.g.*, GPT-5: CC-Discr. 88.9→62.2, MI-Match 88.0→61.1). High base accuracy on these items is therefore often obtained without satisfying the conceptual-structure constraints the probe checks, so answer accuracy overestimates conceptual-structure consistency exactly where relational demand is highest.

Open-book evidence narrows answer gaps more than verification gaps. Open-book settings help several models, but evidence does not turn weaker models into ones that pass conceptual-structure verification. Even with gold evidence in context, InternVL3-78B reaches only 57.6 average SAI, trailing the best closed-book proprietary model (Gemini-3-Pro-Preview, 74.7) by 17 pp. This motivates the evidence diagnostic in Sec. 6: the question is not whether evidence helps, but whether it helps answer selection and RC verification to the same degree.

SAI re-ranks models locally without replacing accuracy. On a task-balanced 149-item IRT short form (see App.), ACC and SAI share a strong dominant ability axis, yet SAI still induces local re-rankings (mean absolute rank shift 2.80, max 10). SAI is therefore not a redundant score which surfaces model-specific gaps that accuracy ignores.

6 Mechanistic Diagnostics

6.1 Semantic Controls Setup

To test whether MCSBench difficulty is reducible to superficial factors, we run two controls across n=12 models on 661 base items: a *semantic-label intervention* augmenting instructions with label/descriptor cues, and a *recognition filter* re-evaluating each model on its top-k% recognized categories ($k \in \{10, 20, 30, 40, 50\}$); see Appendix for details.

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

Semantic label cues have negligible effect. Augmenting prompts with label cues shifts overall accuracy by only $+0.87$ pp (95% CI $[0.32, 1.46]$), with a TOST equivalence test confirming practical negligibility within ± 2 pp (p_{TOST}=0.0016) and $0/12$ models showing item-level shifts under McNemar. Two tasks reach significance—C-A Discr. ($+5.88$ pp) and C-C Align. ($+4.28$ pp)—but neither alters

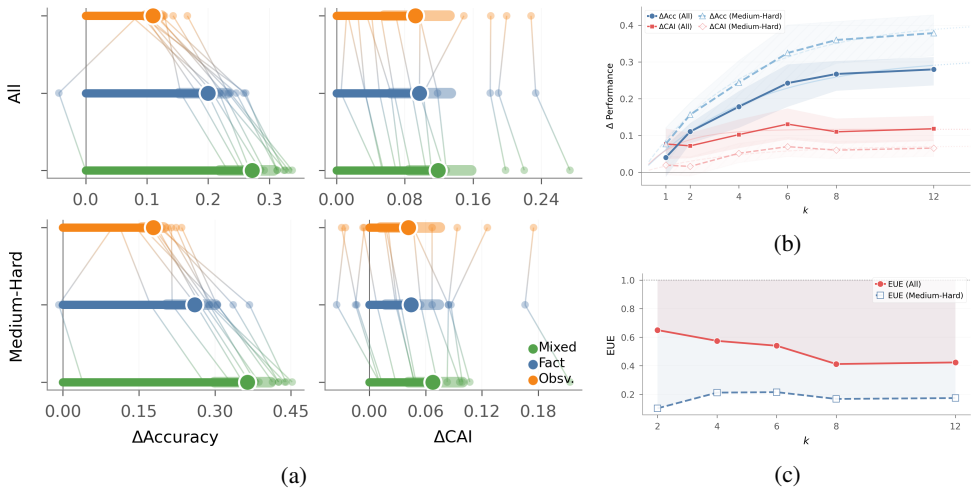


Figure 4: **In-context diagnostics: evidence type, scaling, and utility. (a)** Model-wise ΔAccuracy and ΔSAI gains over the no-evidence baseline across three evidence types, for ALL and MEDIUM-HARD subsets. **(b)** ΔAccuracy and ΔSAI trends *w.r.t.*random evidence count k; Acc gains saturate; SAI gains minorly. **(c)** Evidence-Utilization Efficiency (EUE) across k; EUE < 1 indicates low evidence gains on reasoning integrity.

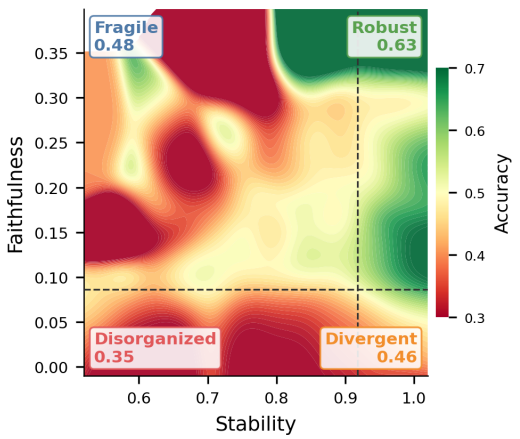


Figure 5: Reasoning Stability–Faithfulness landscape. Contours show binned mean accuracy (w/ thinking rollout); dashed lines mark median thresholds partitioning items into four regimes (Robust, Fragile, Divergent, Disorganized) with inset mean accuracies. $N = 2310$ model×question pairs.

the overall task-difficulty ordering.

Recognition does not rescue relational tasks. 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. 11). MCSBench difficulty is not attributable to semantic cues or recognition limitations; it genuinely probes underlying multimodal structure beyond superficial recognition.

6.2 Evidence Helps Accuracy More Than Verification

A natural hypothesis is that models fail MCSBench because they lack the relevant facts. We test this by supplying in-context evidence and measuring gains in both answer accuracy and SAI. If evidence primarily fills knowledge gaps, accuracy and RC-verified consistency should improve together. Instead, the gains are asymmetric: across evidence types and counts, ΔAccuracy consistently exceeds ΔSAI, with accuracy saturating quickly in k while SAI plateaus at a lower level (Fig. 4). We summarize this with Evidence-Utilization Efficiency,

$$\mathrm{EUE}(k) = \frac{\sum_m \Delta\mathrm{SAI}_m(k)}{\sum_m \Delta\mathrm{Acc}_m(k)},$$

the share of an evidence-driven accuracy gain that also carries through to RC verification. EUE stays below 1 at every tested k and falls further on Medium–Hard items: evidence helps models pick the right option more readily than pass the corresponding RC verification probe.

This asymmetry separates knowledge availability from conceptual-structure consistency. Supplying relevant facts can raise decision accuracy, yet the model may still fail to bind those facts to the right visual observation, rule out the right confuser, or preserve the hypothetical relation the item requires. Evidence therefore acts more as decision support than as a remedy for conceptual-structure consistency.

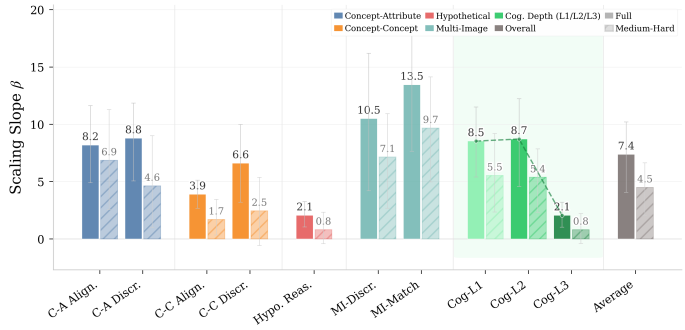


Figure 6: **Scaling slopes β per task** (OLS on $\ln P$) with 95% bootstrap CIs, coloured by task group. Scale alone accounts for moderate variance ($R^2_{\text{sc}} \approx .32$); family intercepts contribute the largest incremental gain, indicating that recipe-level shifts dominate over scaling rates.

Table 6: 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

6.3 Self-Consistency Is An Imcomplete Proxy

One might hypothesize common rollout self-consistency as an alternative to RC verification (Wang et al., 2022). We test this by extracting typed reasoning graphs from k=5 stochastic rollouts per item (N=2310 model×question pairs, 9 models) and comparing pairwise Weisfeiler–Leman *Stability* against accuracy and *Faithfulness* to the gold reference graph of RC (see Appendix).

Stability flips sign on hard items. The Stability–accuracy correlation reverses with difficulty: ρ=+0.29 on Easy vs. ρ=−0.14 on Hard items (Fisher z=8.59, p<.05; Appendix Tab. 9), indicating a *stable-but-wrong* regime in which models converge on the same incorrect reasoning across rollouts. The joint Stability–Faithfulness landscape (Fig. 5) makes the same point: high accuracy concentrates only where both axes are high (Robust, 0.63), not where stability alone is high (Fragile, 0.48). High rollout agreement therefore does not by itself indicate conceptual-structure consistency: it can equally reflect a stable shortcut. which is why This justifies our RC probe over self-consistency as a sufficient proxy.

6.4 Scale helps, but it is not the full explanation.

Across ∼60 models (1.98B–107.71B), log parameter count alone explains only moderate accuracy variance (R^2≈0.32); adding model-family intercepts raises the fit to R^2≈0.87 (ΔR^2_{fam}≈0.50, p<.05), suggesting hidden factors, *e.g.*, training recipe, data, and alignment matter beyond size. Slopes are also task-dependent: multi-image matching scales fastest (β=13.5), whereas hypothetical reasoning (β=2.1) remains a bottleneck (Fig. 6, Tab. 6). Larger models tend to perform better, but scale alone does not explain the MCSBench profile.

7 Conclusion

8 Conclusion

MCSBench operationalizes multimodal conceptual structure as a diagnostic evaluation construct, and SAI shows that correct answers frequently fail the RC verification probes that justify them. This accuracy–SAI gap is widest on inter-concept and hypothetical tasks, where structural demand is highest, and it is narrowed for answer accuracy but not for verification by in-context evidence, while scale, model family, and rollout self-consistency each close it only unevenly. These results suggest that answer accuracy alone overstates conceptual-structure consistency, and that conceptual-structure-aware evaluation is a concrete next step toward MLLMs whose correct answers rest on the constraints that justify them.

9 Limitations

Construct scope. MCS is used as an *operational behavioral construct*: MCSBench diagnoses whether model behavior across vision and language is consistent with the relational organization that cognitive theories attribute to concepts, rather than certifying latent representations. The four cognitive levels (L1–L4) are an evaluation scaffold, not a developmental claim about how MLLMs acquire structure.

Domain coverage. The release is weighted toward fine-grained visual categories where discriminative attributes and confusable neighborhoods are well-defined. We provide supercategory metadata so that users can restrict evaluation as appropriate; broader domain coverage is a natural extension.

References

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

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

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

Susan Carey. 1985. Conceptual change in childhood. *(No Title)*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023. 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.

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

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

Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, and 1 others. 2024. Agent ai: Surveying the horizons of multimodal interaction. *arXiv preprint arXiv:2401.03568*.

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

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

Gregor Geigle, Radu Timofte, and Goran Glavaš. 2024. 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.

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

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

Dedre Gentner and Arthur B Markman. 1997. Structure mapping in analogy and similarity. *American psychologist*, 52(1):45.

Edmund L. Gettier. 1963. Is justified true belief knowledge? *Analysis*, 23(6):121–123.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. 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.

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

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. 2023. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36:31096–31116.

Ray S Jackendoff. 1995. *Languages of the mind: Essays on mental representation*. mit Press.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4198–4205.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, and 1 others. 2025. Mmecot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. *arXiv preprint arXiv:2502.09621*.

Siddharth Joshi, Haoli Yin, Rishabh Adiga, Ricardo Monti, Aldo Carranza, Alex Fang, Alvin Deng, Amro Abbas, Brett Larsen, Cody Blakeney, and 1 others. 2026. Datbench: Discriminative, faithful, and efficient vlm evaluations. *arXiv preprint arXiv:2601.02316*.

Frank C Keil. 1992. *Concepts, kinds, and cognitive development*. mit Press.

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

Jeonghwan Kim and Heng Ji. 2024. 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.

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

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

Yian Li, Wentao Tian, Yang Jiao, Tianwen Qian, Na Zhao, Bin Zhu, Jingjing Chen, and Yu-Gang Jiang. 2025. 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.

Yijiang Li, Qingying Gao, Tianwei Zhao, Bingyang Wang, Haoran Sun, Haiyun Lyu, Robert D Hawkins, Nuno Vasconcelos, Tal Golan, Dezhi Luo, and 1 others. 2024b. Core knowledge deficits in multi-modal language models. *arXiv preprint arXiv:2410.10855*.

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

John M Linacre and 1 others. 2002. What do infit and outfit, mean-square and standardized mean. *Rasch measurement transactions*, 16(2):878.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, and 1 others. 2024. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. *arXiv preprint arXiv:2408.02718*.

Thomas Mensink, Jasper Uijlings, Lluis Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, and Vittorio Ferrari. 2023. 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.

Gregory Murphy. 2004. *The big book of concepts*. MIT press.

Gregory L. Murphy and Douglas L. Medin. 1985. The role of theories in conceptual coherence. *Psychological Review*, 92(3):289–316.

Thomas J Palmeri and Isabel Gauthier. 2004. Visual object understanding. *Nature Reviews Neuroscience*, 5(4):291–303.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280.

Judea Pearl. 2009. *Causality*. Cambridge university press.

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

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

Georg Rasch. 1993. *Probabilistic models for some intelligence and attainment tests.* ERIC.

Bob Rehder. 2003. A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(6):1141.

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

Eleanor Rosch. 1975. Cognitive representations of semantic categories. *Journal of experimental psychology: General*, 104(3):192.

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

Eleanor Rosch and Carolyn B Mervis. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, 7(4):573–605.

Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive psychology*, 8(3):382–439.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. 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.

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

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

Steven A Sloman, Bradley C Love, and Woo-Kyoung Ahn. 1998. Feature centrality and conceptual coherence. *Cognitive science*, 22(2):189–228.

Edward E Smith and Douglas L Medin. 1981. *Categories and concepts*. Harvard University Press.

Zechen Sun, Yisheng Xiao, Juntao Li, Yixin Ji, Wenliang Chen, and Min Zhang. 2024. 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.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

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

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36.

Amos Tversky. 1977. Features of similarity. *Psychological review*, 84(4):327.

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

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

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, and 1 others. 2025. 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.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2022. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*.

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

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

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Appendix Contents

A Towards A Unified Relational Framework **12**

B Extended Related Work **12**
- B.1 From Entity-Centric to Relational Evaluation 13
- B.2 Hypothetical Reasoning in MLLMs 13
- B.3 Evidence and Reasoning-Trace Diagnostics 13

C Additional Details on Dataset Profile **14**

D Details on Dataset Construction & Qualit Control **14**
- D.1 Dataset Construction 14
- D.2 Quality Control 16

E Details on Evaluation Protocol **16**

F Details on Experiments & Analysis **17**
- F.1 Recognition-Conditioned Evaluation 17
- F.2 Item-Response Modeling 17
- F.3 Reasoning Graph Analysis 17
- F.4 Scaling Trend 18
- F.5 External Consistency 19

G Additional Results **19**
- G.1 Prompt Sensitivity Analysis 19

H Qualitative Examples **22**

I Prompts **29**
- I.1 Definition: Relation Primitives . . 29
- I.2 Definition: Reasoning Operators . 29
- I.3 Relation Classification Prompt . . 29

J Discussion and Limitations **29**

K Dataset Quality Model-Checking **30**

L Data Licensing **30**

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 (Durante et al., 2024)? 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** (Rosch, 2024; Gentner, 1983; Murphy and Medin, 1985; Smith and Medin, 1981; Murphy, 2004). 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 (Murphy and Medin, 1985; Jackendoff, 1995; Carey, 1985; Rehder, 2003).

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 (Smith and Medin, 1981). 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 (Rosch, 2024; Rosch and Mervis, 1975; Rosch, 1975). *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 (Gentner, 1983; Gentner and Markman, 1997). *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 (Murphy and Medin, 1985; Carey, 1985; Keil, 1992). 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 (Rehder, 2003; Sloman et al., 1998).

B Extended Related Work

We expand the main-paper positioning (Table 1) along three axes that jointly define the gap MCS-Bench 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 (Chen et al., 2024; Yue et al., 2025; Li et al., 2024a; Geigle et al., 2024; Kim and Ji, 2024; Mensink et al., 2023; Chen et al., 2023; Zha et al., 2024; Kil et al., 2024; Meng et al., 2024; Shao et al., 2024; Qiang et al., 2025; Jiang et al., 2025; Guan et al., 2024; Li et al., 2024b). 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 (Barsalou, 1999; Tversky, 1977; Byrne, 2007; Flavell, 1979). 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 (Gentner, 1983; Murphy, 2004; Tyler and Moss, 2001). Below, we discuss how each gap emerges across three evaluation families.

B.1 From Entity-Centric to Relational Evaluation

Fine-grained benchmarks, exemplified by FOCI (Geigle et al., 2024) and Finer (Kim and Ji, 2024), 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 (Murphy, 2004; Barsalou, 1999). Comparative benchmarks extend the scope toward relational reasoning, but along narrower axes: Constructure (Zha et al., 2024) 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 (Kil et al., 2024) and MMIU (Meng et al., 2024) 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.

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 (Zhang et al., 2024) 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 (Li et al., 2025) and MuCR (Li et al., 2024c) 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 (Zhang et al., 2023) 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 (Pearl, 2009; Byrne, 2007).

B.3 Evidence and Reasoning-Trace Diagnostics

Prior work evaluates reasoning traces and evidence utilization along complementary but structurally unanchored axes: Visual CoT (Shao et al., 2024), VER-Bench (Qiang et al., 2025), and MME-CoT (Jiang et al., 2025) assess chain-of-thought quality across perception and cognition subtasks; HallusionBench (Guan et al., 2024) and CoreCognition (Li et al., 2024b) diagnose hallucination-

style failures; Encyclopedic-VQA (Mensink et al., 2023) and InfoSeek (Chen et al., 2023) 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

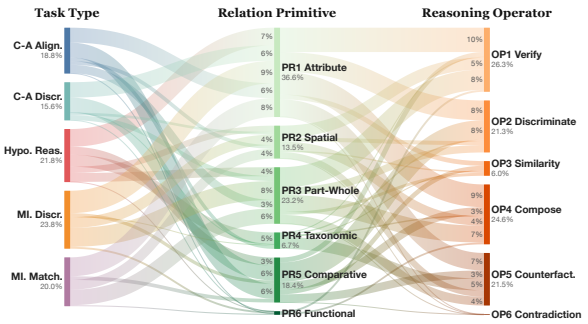


Figure 7: Alluvial diagram of MCSBench composition across three axes: task type (left), relation primitive (center), and reasoning operator (right). Edge width encodes the proportion of items linking each pair; percentages denote the marginal share of each node.

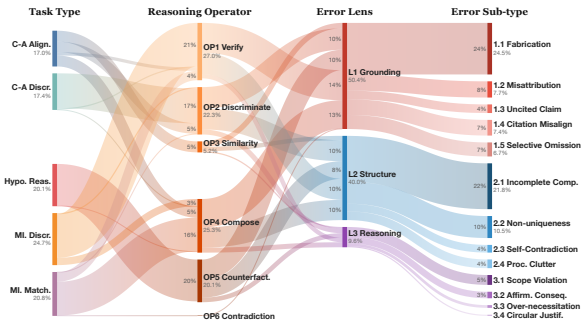


Figure 8: Error distribution of MCSBench overlay across task type, reasoning operator, error lens, and error sub-type.

Relation-Operator Decomposition. Fig. 7: 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.

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. 8). 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 (Van Horn et al., 2017) (10,000 fine-grained species across 11 supercategories), ImageNet (Deng et al., 2009) (1,000 object categories spanning everyday and abstract concepts), and

ImageNet-21K (Ridnik et al., 2021) (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 (Cherti et al., 2023)) 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 confusing concept set. Prompts of 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 (≥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. 8). 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 (≥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.2.

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 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.2 Item-Response Modeling

Implementation details. To characterize model behavioral patterns and calibrate measurement, we leverage a Rasch 1PL model (Rasch, 1993), 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 (Bock and Aitkin, 1981) (`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 items across 56 models, to satisfy Rasch model's assumption, a two-stage screen (Linacre et al., 2002) 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 (Polo et al., 2024) then yields a 149-item short form. SAI abilities are linked to the ACC scale via mean–sigma equating (Kolen and Brennan, 2014) (A=0.7497, B=0.0974), with uncertainty from bootstrap resampling (N=200, percentile 95% CIs). Both views achieve person-separation reliability $\geq$ 0.97, confirming stable measurement while exposing ranking shifts that accuracy alone conceals.

A diagnostic benchmark should function as a reliable measurement instrument. We fit a Rasch 1PL model on a task-balanced 149-item short form under both ACC and SAI scoring, with SAI abilities linked to the ACC scale via mean–sigma linking; full details are in Appendix.

Reliable measurement. Both scorings reach high person-separation reliability (ACC 0.973; SAI 0.980), with all items passing infit acceptance (Table 7). MCSBench thus supports stable ability estimation for both correctness- and integrity-aware evaluation.

SAI carries information beyond ACC. Global ordering is largely shared (Kendall $\tau = 0.870$; no systematic directional bias, $p = 0.525$), yet item-level SAI scoring induces substantial local re-rankings (mean $|\Delta\text{Rank}| = 2.80$, max $= 10$). Table 8 illustrates concrete cases: Kimi-VL-A3B drops 10 positions and LLaVA-OV-Qwen2-7B drops 5 under SAI, while Gemma-3-12B and MiniCPM-V-4 each rise 7. SAI therefore aligns with ACC on the dominant ability axis while surfacing model-specific gaps that accuracy alone absorbs.

Additional Results. We present entire model ability ranking and delta rank in Fig. 10. The distribution of capability and local rank adjustment validate our claims in the main paper.

F.3 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 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 (Shervashidze et al., 2011) 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

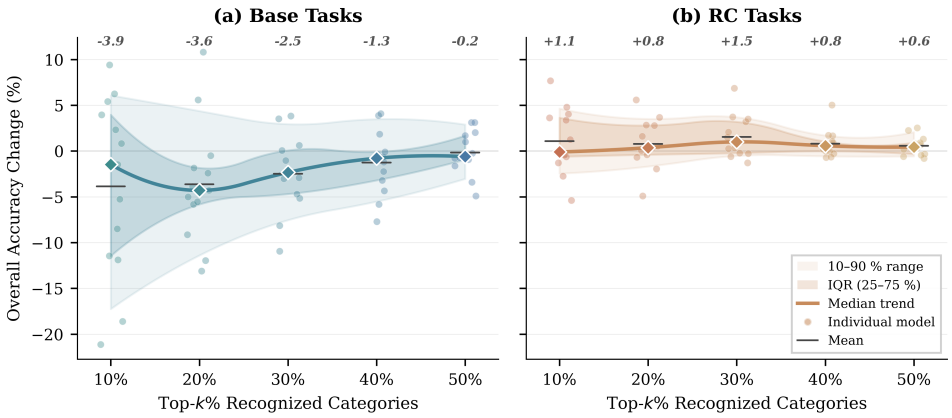


Figure 9: Performance sensitivity to removing top-$k\%$ most recognized categories. Base accuracy even minorly degrades, while RC accuracy remains flat, indicating that recognition does not aid correctness and reasoning integrity. Bands: IQR and 10–90 % range across models.

Table 7: 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 8: 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

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.

Additional Results Table 9 reports the full breakdown. (a) The Stability–accuracy correlation is positive on Easy and negative on Hard items, with a significant difficulty interaction (Fisher z=8.59). (b) An OLS fit on graph features ranks Edge-Type diversity, Stability, and Faithfulness as the top accuracy predictors (all p<.001).

Table 9: 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

F.4 Scaling Trend

Implementation details. We evaluate 63 proprietary models or open-source VLMs spanning 1.98B–107.71B parameters, representing scale as

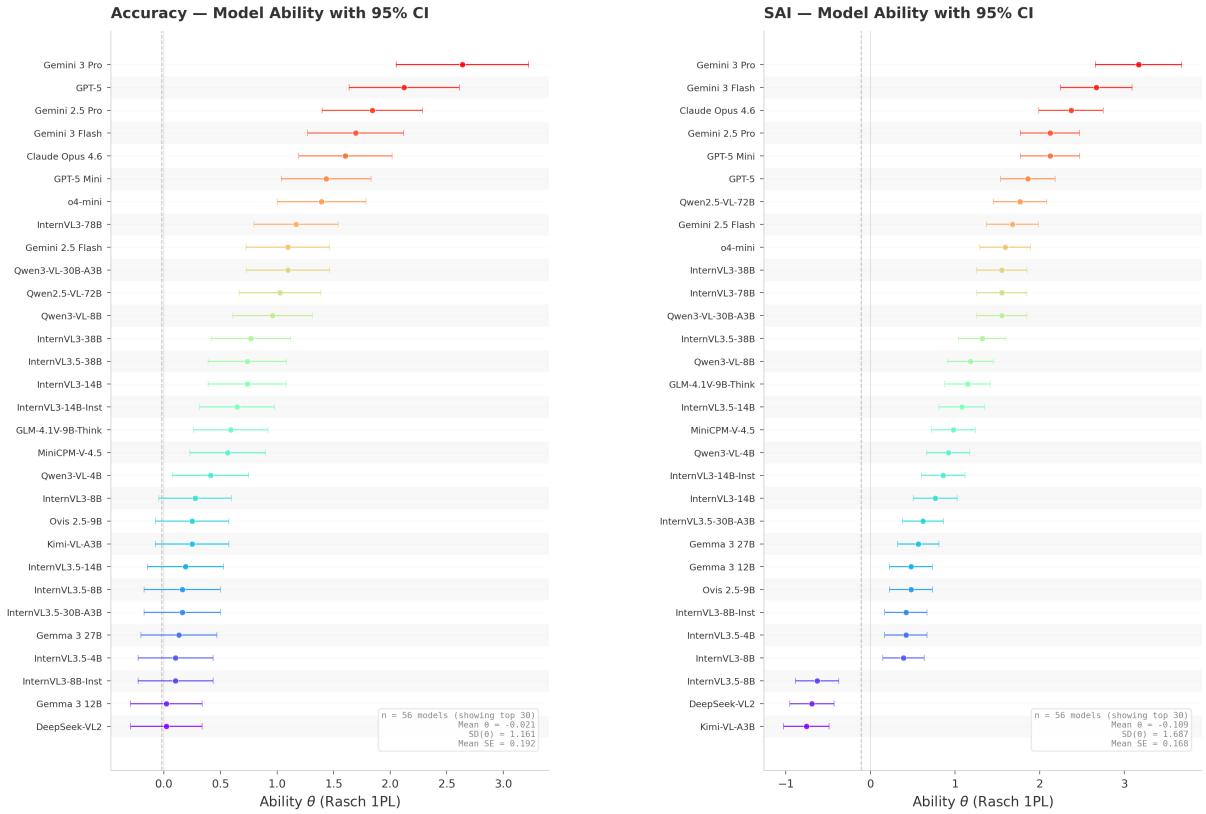


Figure 10: Rasch model ability estimates for top-30 models ranked by accuracy (left) and SAI (right).

$x = \log_{10}(\text{params}_\text{B})$. 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. 12. 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.

F.5 External Consistency

MCSBench accuracy correlates positively and significantly with MME, HallusionBench, and SEED-Bench on the exact set of overlapping models (Spearman ρ=0.61–0.86; Table 11). Among capability axes examined (Fig. 14), overall SAI exhibits the highest correlation with hallucination-robustness, consistent with SAI penalizing structurally unjustified correctness.

G Additional Results

G.1 Prompt Sensitivity Analysis

Five prompting strategies (*i.e.*, *Direct Prompting, Chain-of-Thought (CoT), Justification, Expert Role Playing, and Contrastive Prompting*) are compared across 9 models with various capability levels Sensitivity results are presented in Fig. 13, which support four main findings: (1) Overall accuracy is only mildly sensitive to prompting strategies; (2) Counterintuitively, CoT prompting does not improve performance over simple direct answering variant; (3) Contrastive prompting is consistently harmful, suggesting that most models lack robust structural or discriminative reasoning ability; and (4) Within each cognitive level, the difficulty ordering is mostly preserved across all prompting strategies. Overall, the choice of prompting strategy does not alter the benchmark difficulty profile.

Table 10: **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

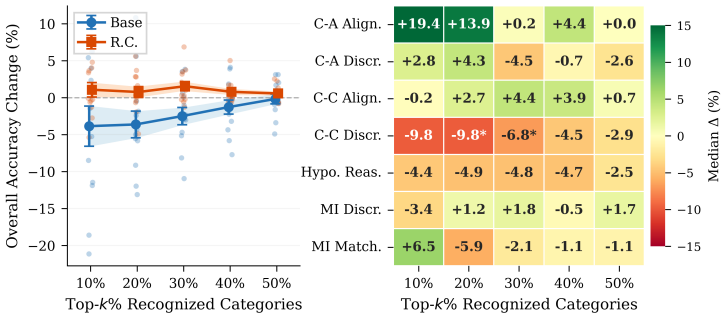


Figure 11: **Recognition-conditioned evaluation.** MCSBench is re-evaluated on the subset of the top-k% recognized categories ($k \in \{10, 20, 30, 40, 50\}$). **Left:** Overall accuracy trend for Base and reasoning-chain (R.C.) tasks across models. **Right:** Per-task median accuracy change Δ (pp) under the same filtering; *significant under Wilcoxon signed-rank test.

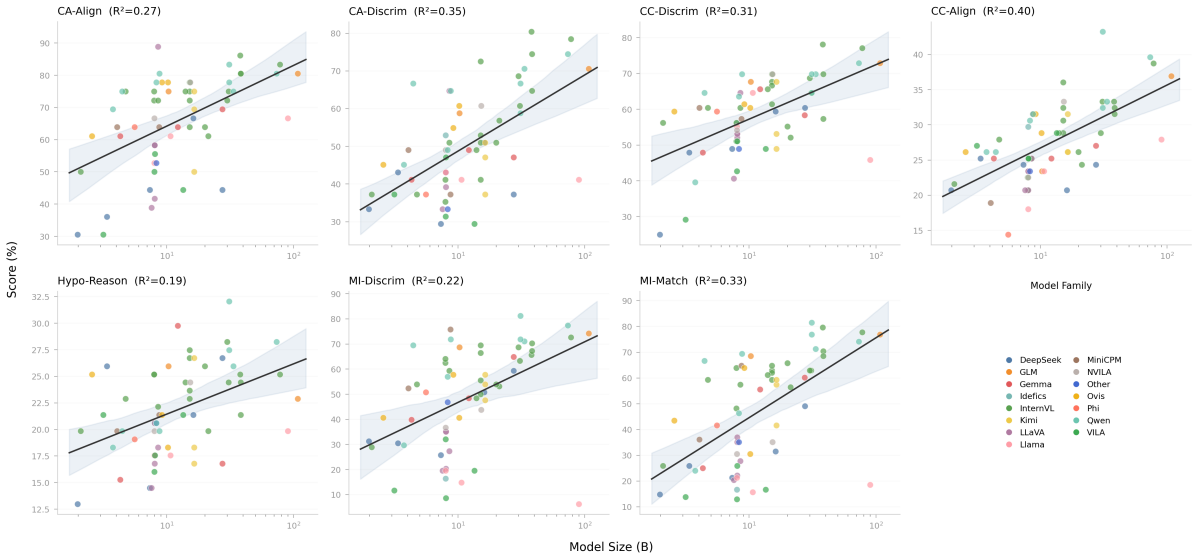


Figure 12: Per-task scaling trend visualization.

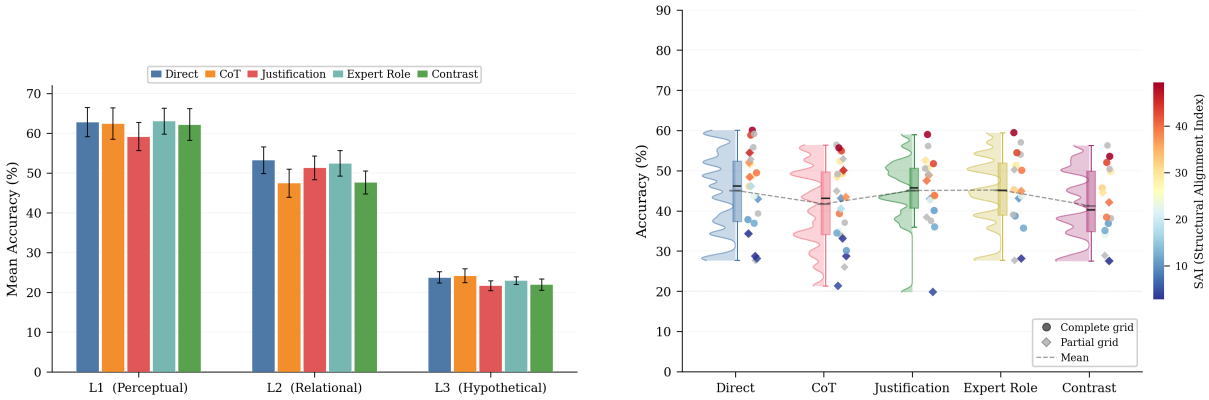


Figure 13: **Prompt sensitivity across cognitive levels and prompt type.** *Left:* Mean accuracy under five prompt variants aggregated over 9 models, with error bars showing between-model variability. Prompt effects are largest at L2 and negligible at L3. *Right:* Per-prompt accuracy distributions; marker color encodes SAI and shape denotes grid completeness. No prompt yields a uniform gain.

Table 11: **Rank agreement between MCSBench and established MLLM benchmarks.** Spearman and Kendall correlations (with p-value) between MCSBench and external benchmarks, computed on a randomly sampled overlapping model set.

Benchmark	**Spearman ρ**	**Kendall τ**
MME	0.858 (0.004)	0.718 (0.001)
HallusionBench	0.753 (0.004)	0.595 (0.007)
SEEDBench	0.612 (0.045)	0.574 (0.015)

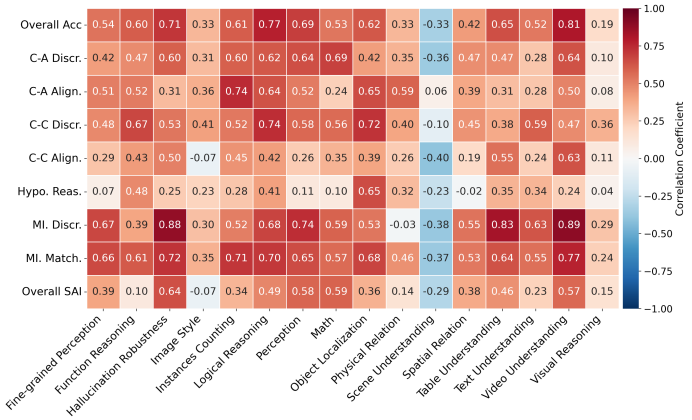


Figure 14: Task–capability Spearman correlations between MCSBench tasks and existing MLLM benchmark core capabilities. $*$ denotes $p < 0.05$.

H Qualitative Examples

Capitulum Arrangement Identification (task_base_192)

VQA Question

Identify the category that is **different from the specimen on capitulum arrangement**: the option with *multiple small heads tightly clustered near the stem apex* on *branched, spiny-winged stems*.

A *Cirsium palustre*
B *Centaurea debeauxii*
C *Cirsium acaule*
D *Cirsium dissectum*
E *Cirsium andersonii*

Correct Answer: A

Reasoning Chain

1 O1 The specimen shows a thistle-like plant with a **single prominent terminal flower head** on a mostly unbranched stem in this frame.

2 E1–E3 define *capitulum arrangement* and what it means for heads to be *multiple, tightly clustered at the stem apex* on *branched, spiny-winged stems*.

3 The question asks for the **option that has the clustered, branched-head arrangement** (the opposite of the single-head look in O1).

4 E4 describes *Cirsium palustre* as a tall thistle with **branched upper stems** bearing **many small flower heads densely clustered at branch tips** and with **spiny-winged stems** — matching the prompt's distinguishing arrangement.

5 Other options are inconsistent with the required arrangement: E5 (*Cirsium acaule*) is typically stemless with a solitary head, E6 (*Centaurea*) tends to have single heads per branch, and E7–E8 (other *Cirsium*) do not characteristically form dense, tight terminal clusters like *C. palustre*.

6 **Therefore** the correct category is A ***Cirsium palustre***, the option best matching a branched, tightly clustered capitulum arrangement.

Observations

O1 Field photo: a thistle-like plant with one large purple head visible near the top; the head is not shown as a dense cluster of many small heads.

Evidence List

E1 **Capitulum arrangement**: how the flower heads (capitula) are positioned on the plant (solitary terminal head vs. multiple heads clustered at branch tips, etc.).

E2 **Multiple small heads tightly clustered**: many small capitula produced close together, appearing as a tight bunch near the stem/branch apex.

E3 **Spiny-winged stems**: stems with flattened wing-like ridges bearing spines, often formed by decurrent leaf bases.

E4 ***Cirsium palustre*** (marsh thistle): commonly described as having branched upper stems with numerous small heads densely clustered near the ends, and distinctly spiny-winged stems.

E5 ***Cirsium acaule*** (stemless/dwarf thistle): typically a basal rosette with a solitary flower head (not a branched, clustered apex).

E6 ***Centaurea debeauxii*** (knapweed): branched stems may occur, but flower heads are generally borne singly at branch tips rather than as tight clusters of many small heads.

E7 ***Cirsium dissectum*** (meadow thistle): a slender thistle whose heads are usually solitary or few rather than dense terminal clusters.

E8 ***Cirsium andersonii*** (Anderson's thistle): may bear a small number of heads (e.g., 1–6) but not typically the dense tight apex clusters characteristic of *C. palustre*.

E9 **Original specimen (*Cirsium heterophyllum*)**: commonly characterized by a more solitary head presentation compared with the marsh thistle's dense clustered inflorescences.

Conclusion

From the observation **O1** (a specimen not showing a dense cluster of small heads), applying the capitulum/inflorescence definitions **E1**–**E3**, and matching the required "many small heads tightly clustered on branched, spiny-winged stems" pattern to the category description **E4** while excluding conflicting alternatives **E5**–**E8**, the option that fits the prompt is ***Cirsium palustre*** (**A**).

Inflorescence Modification (*Cercis*) (task_rc_65_0)

VQA Question

In the image, flowers are in tight fascicles directly on old wood with short pedicels and no leaves present. **If those fascicles were replaced by short axillary racemes bearing up to about a dozen flowers per raceme on darker gray-black bark, which category would the modified specimen most closely resemble?**

A *Cercis griffithii*
B *Cercis canadensis*
C *Cercis occidentalis*
D *Cercis glabra*
E *Cercis chinensis*

Correct Answer: D

Reasoning Chain

1. O1 The photo shows pink pea-type blossoms clustered tightly along leafless woody branches (flowers on old wood, short individual stalks).

2. E1–E4 clarify the inflorescence vocabulary: *fascicle* (tight bundle) vs. *raceme* (flowers on a short axis), and what *axillary* placement and *pedicels* mean.

3. The hypothetical change specifies **short axillary racemes** with up to **about a dozen flowers per raceme** and **darker gray-black bark**. This is a key combination of inflorescence structure + bark character.

4. E5 notes that *Cercis glabra* is characterized by flowers borne on **short axillary racemes** (commonly a few to around 10–12 flowers) and bark described as **grayish-black/dark** — matching the modified description.

5. In contrast, E6–E8 (*C. canadensis*, *C. occidentalis*, *C. griffithii*) are commonly described with flowers in **small tight clusters/fascicles on older wood** rather than the explicit axillary raceme structure, and E9 (*C. chinensis*) is more variable but not the best match for the raceme + dark bark combination.

6. **Therefore** the modified specimen most closely resembles D ***Cercis glabra***.

Observations

O1 Branches are leafless and carry many pink blossoms in tight clusters along older wood; individual flowers appear on short stalks.

Evidence List

E1 **Fascicle**: a very compact bundle/cluster of flowers with little or no visible central axis.

E2 **Raceme**: an unbranched inflorescence with multiple flowers attached by short pedicels along a central axis (rachis).

E3 **Axillary**: arising from a leaf axil (the angle between a leaf and stem). "Axillary racemes" originate at these nodes.

E4 **Pedicel**: the small stalk attaching an individual flower to an inflorescence.

E5 ***Cercis glabra*** (smooth/Yunnan redbud): described as having flowers on short axillary racemes (often a few to about 10–12 flowers per raceme) and darker gray-black bark — matching the modified condition.

E6 ***Cercis canadensis*** (eastern redbud): typically flowers in small tight clusters (fascicles) on older branches/trunk, often before leaves.

E7 ***Cercis occidentalis*** (western redbud): commonly presented with showy spring flowers in conspicuous clusters; not usually emphasized as short axillary racemes with many flowers.

E8 ***Cercis griffithii*** (Griffith's redbud): produces abundant pink flowers on branches and trunk; typically depicted as clustered blooms rather than the specific raceme structure.

E9 ***Cercis chinensis*** (Chinese redbud): inflorescences can be described as clustered or short racemes in different accounts, but the prompt's specific "short axillary racemes with up to a dozen flowers" plus dark bark is most directly aligned with E5.

Conclusion

Starting from the observed fasciculate flowering habit O1, applying the inflorescence definitions E1–E4, and matching the hypothetical replacement with axillary racemes and darker bark to the diagnostic description in E5 while excluding alternative redbuds E6–E9, the modified specimen most closely resembles ***Cercis glabra*** (D).

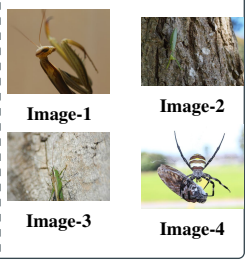
Multi-Image Wing Marker Detection

VQA Question

Select the image in which the **forewing** shows a distinct **triangular, pale stigma** rimmed by **darker pigment** and a **hyaline discoidal area**.

A Image-2
B Image-1
C Image-1 and Image-2
D Image-4
E Image-3

Correct Answer: A



Reasoning Chain

1 E1–E4 define the requested wing characters: *forewing*, *stigma* (triangular pale mark rimmed dark), *hyaline* (transparent), and the *discoidal area*.

2 O2 (Image-2) shows a mantis with a forewing bearing a **pale triangular stigma** near the leading edge and a **transparent/hyaline central area** consistent with a discoidal region.

3 E5 notes that in mantids such as *Hierodula patellifera*, a pale triangular stigma rimmed with darker pigment and a more transparent discoidal area are commonly used identification marks.

4 The other images do not satisfy the combined condition: O1 (Image-1) shows a mantis wing pattern without the same distinct pale triangular stigma; O3 (Image-3) is a grasshopper/katydid with different forewing structure; O4 (Image-4) is a spider with prey and does not present a mantis-type forewing stigma.

5 **Therefore** the only image matching the described **pale triangular stigma + hyaline discoidal area** is **Image-2**, i.e., A.

Observations

O1 **Image-1**: Close-up of a mantis with forewings visible; markings appear darker/mottled without an obvious pale triangular stigma.

O2 **Image-2**: Green mantis on bark; forewing shows a light triangular patch near the leading edge and a clearer translucent region.

O3 **Image-3**: Green grasshopper/katydid on bark; wings are long and narrow and do not show a mantis-like triangular stigma.

O4 **Image-4**: Spider on a web holding prey; no relevant forewing stigma/discoidal wing structure is the focus.

Evidence List

E1 **Forewing (tegmen)**: the front wing pair; in mantids these are leathery covers that may carry diagnostic marks.

E2 **Stigma**: a thickened or pigmented patch near the leading edge of the forewing; in many mantids it appears as a *triangular pale mark rimmed by darker pigment*.

E3 **Hyaline**: glass-like, transparent (used here for a clear wing region).

E4 **Discoidal area**: the central wing cell/area (often delimited by veins) that can appear more transparent in some insects.

E5 ***Hierodula patellifera* forewing character**: commonly illustrated with a pale triangular stigma on the tegmina and a comparatively hyaline/transparent discoidal region, matching the phrasing in the question.

E6 ***Stagmomantis californica* (Image-1)**: a mantis whose forewing markings can be darker or less distinctly pale-triangular in the stigma region compared with *Hierodula* in the prompt.

E7 **Conocephalus / grasshopper-type wings (Image-3)**: long narrow forewings without the mantis-specific triangular stigma + hyaline discoidal combination.

E8 **Cicada / spider scene (Image-4)**: does not present the target mantis forewing stigma structure.

Conclusion

Given the required wing-marker definitions E1–E4 and the per-image observations O1–O4, only O2 exhibits a forewing with a clearly **pale triangular stigma** rimmed dark and a **hyaline discoidal area**, consistent with the mantis wing character summarized in E5. Therefore the correct choice is **Image-2** (A).

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.

CONTEXT: <Enumerated relation primitives and reasoning
operators>

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
}
```

J Discussion and Limitations

Construct scope. MCS is used as an *operational behavioral construct*: MCSBench diagnoses whether model behavior across vision and language is consistent with the relational organization that cognitive theories attribute to concepts, rather than certifying latent representations. The four cognitive levels (L1–L4) are an evaluation scaffold, not a developmental claim about how MLLMs acquire structure.

Domain coverage. The release is weighted toward fine-grained visual categories where discriminative attributes and confusable neighborhoods are well-defined. We provide supercategory metadata (Fig. 1) so that users can stratify or restrict evaluation as appropriate; broader domain coverage (abstract, social, or thin-attribute concepts) is a natural extension, but these may reach performance saturation in practice due to greater data exposure.

SAI as a strict lower bound. SAI is the joint indicator of base-answer correctness and reasoning-probe correctness, by construction a *conservative* estimate of structurally justified correctness rather than an exhaustive certification. Items where multiple valid reasoning chains coexist are flagged during quality control and excluded from the scored RC subset, so SAI scores admit a clear interpretation: they reward answers that satisfy audited structural constraints, and abstain elsewhere.

Intended use. MCSBench is a diagnostic evaluation instrument for research on multimodal foundation models. It is not designed as a deployment-stage certification or as a sole criterion for high-stakes capability claims; we encourage users to combine it with domain-specific evaluations.

K Dataset Quality Model-Checking

To independently verify the released benchmark, we conduct a release-time audit using four frontier LLM judges from two provider families (GPT-5.5, GPT-5.4, GEMINI-3.1-PRO, GEMINI-3.1-FLASH). The audit is deliberately separated from the construction-time LLM-as-a-judge stage (Sec. D.2): it uses different model versions, runs on the final released items only, and evaluates six independent quality dimensions: overall quality (QC), gold-answer correctness (GA), gold-answer optimality (OA), reasoning-chain correctness (RC; RC-overlay items only), referenced-evidence correctness (RE; RC-overlay items only), and image-reference mapping (IM). For each item, a dimension is recorded as "pass" if all four judges return non-rejecting verdicts; we report the mean pass rate in Table 12.

The audit is intended as a transparency signal rather than a correctness certificate: high pass rates indicate that frontier LLM judges, prompted independently of the generation pipeline, do not flag systematic issues in the released items. It does not substitute for the human quality-assurance stage in Sec. D.2, nor for users' own validation on subsets they care about.

L Data Licensing

MCSBench contains source images from Wikimedia Commons (WikiCM)[1] and ImageNet-21K (IN21K) (Deng et al., 2009; Ridnik et al., 2021). 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.

Table 12: Quality check on our MCSBench, assessed by four independent frontier LLM judges (GPT-5.5, GPT-5.4, GEMINI-3.1-PRO, GEMINI-3.1-FLASH). All entries report the mean pass rate (%) across judges. **QC**: overall quality; **GA**: gold-answer correctness; **OA**: gold-answer optimality; **RC**: gold-reasoning-chain correctness; **RE**: referenced-evidence correctness; **IM**: image-reference mapping/observational correctness. Every dimension exceeds 94% across all tasks, indicating high item quality.

	Con.-Attr.		Con.-Con.		Hypo.	Multi-Image		
Metric	Align.	Discr.	Align.	Discr.	Reason.	Discr.	Match.	**Overall**
QC (%)	97.9	96.1	93.2	92.8	94.4	96.4	95.4	**94.8**
GA (%)	97.9	98.0	95.4	94.8	96.3	98.4	96.8	**96.6**
OA (%)	97.9	97.1	95.6	93.6	95.1	98.3	96.5	**96.1**
RC (%)	–	–	94.4	93.5	94.0	96.0	96.5	**95.0**
RE (%)	–	–	96.1	96.8	95.9	97.0	98.0	**96.9**
IM (%)	100.0	100.0	99.9	99.9	100.0	99.9	100.0	**99.9**

[1] https://commons.wikimedia.org

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

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 15: 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 16: 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 17: 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 18: Task type definitions for **single-image** tasks (Image Matching, Discrepancy Detection) and **multi-image** tasks (Image Selection by Attributes). Together with Figure 17, 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 19: 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 17)

§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 20: 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 21: 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.