CrossCheck-Bench: Diagnosing Compositional Failures in Multimodal Conflict Resolution

Baoliang Tian[1*], Yuxuan Si[1,2*], Jilong Wang[1,3*], LingYao Li[1], Zhongyuan Bao[1], Zineng Zhou[1], Tao Wang[1†], Sixu Li[1], Ziyao Xu[1], Mingze Wang[1], Zhouzhuo Zhang[1], Zhihao Wang[1], Yi Ke Yun[1], Ke Tian[1], Ning Yang[3†], Minghui Qiu[1]

[1]ByteDance
[2]Zhejiang University
[3]Institute of Automation, Chinese Academy of Sciences
walton.wang929@gmail.com [†], ning.yang@ia.ac.cn [†]

Abstract

Multimodal Large Language Models are primarily trained and evaluated on aligned image-text pairs, which leaves their ability to detect and resolve real-world inconsistencies largely unexplored. In open-domain applications visual and textual cues often conflict, requiring models to perform structured reasoning beyond surface-level alignment. We introduce CrossCheck-Bench, a diagnostic benchmark for evaluating contradiction detection in multimodal inputs. The benchmark adopts a hierarchical task framework covering three levels of reasoning complexity and defines seven atomic capabilities essential for resolving cross-modal inconsistencies. CrossCheck-Bench includes 15k question-answer pairs sourced from real-world artifacts with synthetically injected contradictions. The dataset is constructed through a multi-stage annotation pipeline involving more than 450 expert hours to ensure semantic validity and calibrated difficulty across perception, integration, and reasoning. We evaluate 13 state-of-the-art vision-language models and observe a consistent performance drop as tasks shift from perceptual matching to logical contradiction detection. Most models perform well on isolated entity recognition but fail when multiple clues must be synthesized for conflict reasoning. Capability-level analysis further reveals uneven skill acquisition, especially in tasks requiring multi-step inference or rule-based validation. Additional probing shows that conventional prompting strategies such as Chain-of-Thought and Set-of-Mark yield only marginal gains. By contrast, methods that interleave symbolic reasoning with grounded visual processing achieve more stable improvements. These results highlight a persistent bottleneck in multimodal reasoning and suggest new directions for building models capable of robust cross-modal verification.

Code — https://github.com/bytedance/CrossCheck-Bench

Introduction

Multimodal content in the open world frequently exhibits noisy, unreliable, and even deceptive characteristics. A product page may display a luxury brand logo with a suspiciously low price, or pair an image of a sports shoe with a textual description of formal wear. Humans can instinctively recognize when such visual and textual clues do not align, flagging potential fraud or misrepresentation. This ability to resolve cross-modal conflicts is fundamental for any robust multimodal reasoning system (Wang et al. 2023b; Wolf et al. 2023). However, vision-language models (VLMs), which now underpin many content understanding deployments, have not been thoroughly evaluated on their ability to detect and reject contradictions (Tan, Plummer, and Saenko 2020). Most existing VLMs (Li et al. 2023a; Alayrac et al. 2022; Liu et al. 2023a) are primarily trained and evaluated on aligned datasets, where vision and language describe the same semantic content. This alignment-centric paradigm encourages cross-modal consistency but overlooks a critical question: *can models verify whether multimodal signals are logically compatible?* The lack of this capability would pose a tangible risk: models may confidently affirm incompatible clues, producing outputs that are not only inaccurate but also logically inconsistent with the input evidence (Li et al. 2023b; Liu et al. 2023b).

Consequently, there is an urgent need for a benchmark that can effectively evaluate such an ability (Ma et al. 2023). While existing benchmarks effectively assess tasks like retrieval (Li et al. 2022; Wasserman et al. 2025), descrip-

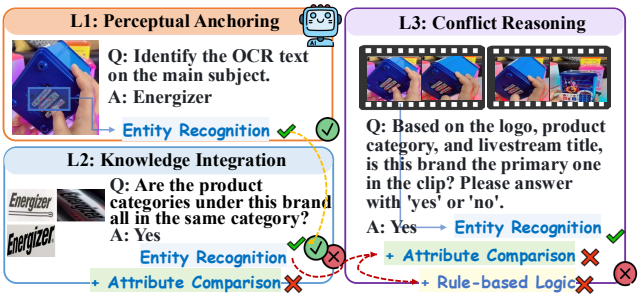

Figure 1: CROSSCHECK-BENCH cascade case: the model answers the Level-1 perception query correctly, yet fails the dependent Level-2 integration and Level-3 conflict-reasoning tasks.

*These authors contributed equally.
†Corresponding author.

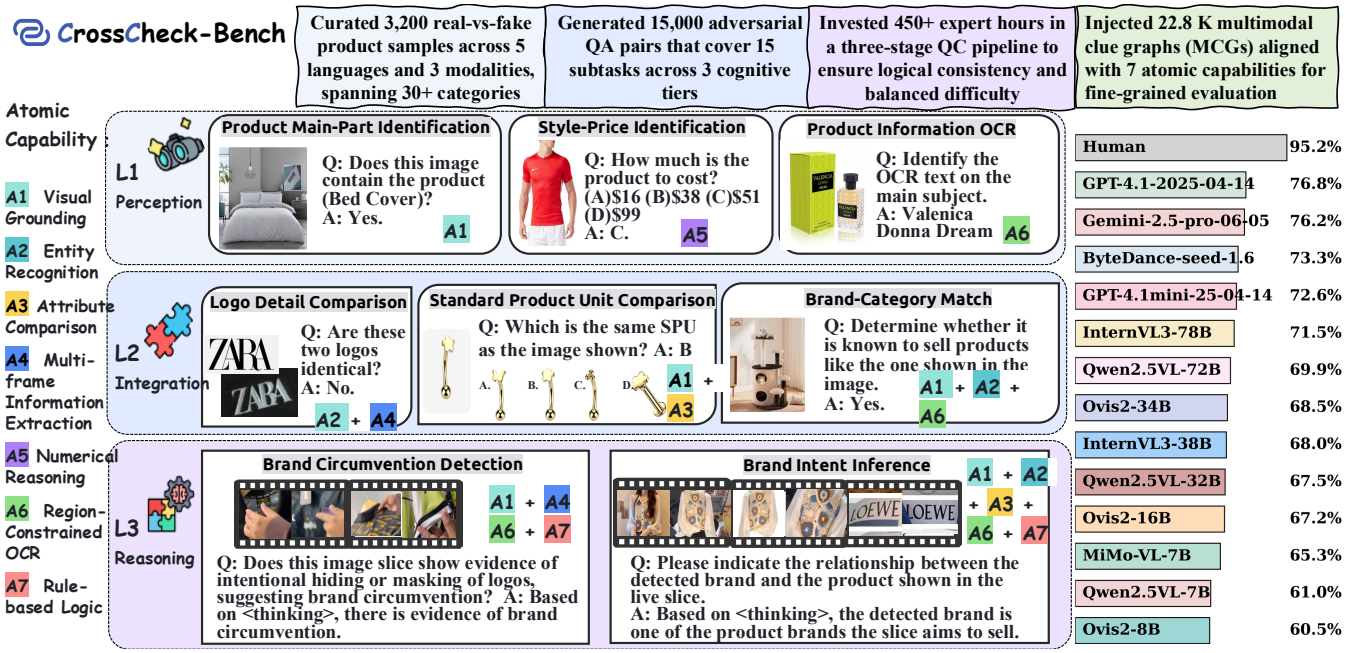

Figure 2: Overview of CROSSCHECK-BENCH. The benchmark spans three cognitive tiers—L1 Perception, L2 Integration, and L3 Reasoning—grounded in seven atomic capabilities (A1–A7) and eight representative tasks. It offers thousands product samples, 14,690 adversarial QA pairs, and 22.8 K multimodal clue graphs, curated through 450+ expert hours. The right-hand bar chart contrasts average model accuracy with the human upper bound.

tion (Maaz et al. 2023; Yue et al. 2024), and entailment (Xie et al. 2019), few explicitly test whether multimodal inputs are jointly compatible in a structured, diagnostic way. Unlike generation correctness evaluation, validating conflict resolution requires more than local grounding or factual recall—it demands compositional verification across modalities (Johnson et al. 2017). Specifically, this ability requires models to (1) localize corresponding semantic entities or attributes in each modality and (2) assess whether these aligned clues are logically compatible. However, constructing such data is highly challenging, as it depends on reliable data sources to identify contradictory entities across different modalities and create realistic conflict-based evaluation tasks (Johnson et al. 2017).

To address this gap, we introduce *CrossCheck-Bench*, the benchmark designed to evaluate and diagnose VLMs' capacity to resolve multimodal inconsistencies (Li et al. 2024; Bai et al. 2024; Gunjal, Yin, and Bas 2024). As illustrated in Figure 2, we structure the benchmark into a three-level hierarchy reflecting increasing reasoning complexity (Yue et al. 2024; Lu et al. 2023). L1 (Perceptual Anchoring) evaluates whether the model can extract atomic entities from each modality. L2 (Knowledge Integration) assesses whether the model can compare cross-modal attributes. L3 (Conflict Reasoning) requires the model to detect implicit contradictions that arise from multi-attribute combinations. Each level builds on prior capabilities: from recognizing entities, to comparing attributes, to applying rule-based logic under uncertainty. To enable fine-grained diagnosis, we further decompose the evaluation into seven atomic capabilities spanning Entity Recognition, Attribute Comparison, Numerical Reasoning, Multi-Frame Extraction, etc. By structuring both tasks and skills hierarchically, CrossCheck-Bench reveals how early-stage failures—such as misidentifying logos or misreading text—can propagate upward, leading to confident but flawed high-level inferences. As exemplified by the case study in Figure 1, a single input sample may trigger success at L1 while failing at L2 and L3, illustrating how surface-level perception often masks deeper reasoning collapse. Each QA sample is derived from real-world stimuli (e.g., e-commerce listings, ads, social posts) with injected inconsistencies requiring nontrivial inference.

We evaluated 13 current top performing VLMs, including GPT-4.1 (Achiam et al. 2023), Gemini-2.5 (Comanici et al. 2025), Qwen2.5-VL (Bai et al. 2025), InternVL3 (Zhu et al. 2025), and MiMo-VL (Xiaomi et al. 2025), on 15k conflict-oriented QA samples spanning all three task levels. Our results reveal a stark performance drop from L1 to L3: while most models succeed at local grounding, nearly all fail at multi-attribute contradiction detection, affirming implausible combinations with high confidence. We further examine whether prompting strategies and lightweight adaptation can enhance model performance on conflict-sensitive tasks. Our findings show that conventional methods such as Chain-of-Thought prompting and visual grounding through annotated regions provide limited benefit and may even reinforce superficial patterns. In contrast, approaches that support iterative reasoning across visual and textual inputs, including lightweight supervised fine-tuning, result in more consistent improvements. Overall, these results highlight a criti-

cal bottleneck in the ability of MLLMs to perform robust, integrated reasoning—a key challenge for future research. To summarize, we make the following contributions:

- We propose *CrossCheck-Bench*, a new benchmark to evaluate and diagnose VLMs' ability to resolve multimodal inconsistencies, structured around a three-level reasoning hierarchy and seven atomic capabilities.
- We construct a large-scale dataset of 15k QA samples derived from real-world artifacts, with programmatically injected contradictions that require compositional reasoning to resolve.
- We benchmark 13 leading VLMs, revealing their systematic failure on L3 conflict reasoning, identifying the limitations of current prompting methods, and highlighting iterative verification as a promising path forward.

Related Work

Multimodal Reasoning Benchmarks. Existing benchmarks primarily evaluate VLMs on compositional tasks where modalities *reinforce* each other. Datasets like VCR (Zellers et al. 2019), NLVR2 (Suhr et al. 2019), and SNLI-VE (Xie et al. 2019) assess textual entailment or reasoning with aligned visual support, assuming visual-textual consistency. Recent benchmarks such as MMMU (Yue et al. 2024) and MathVista (Lu et al. 2023) emphasize complex multimodal reasoning but remain confined to scenarios with concordant inputs. While these efforts demonstrate VLMs' growing proficiency in integrated understanding, they fail to assess models' resilience when modalities *conflict*—the core challenge addressed by our work.

Inconsistency Detection in Vision-Language. Prior research on multimodal inconsistency has primarily focused on specialized, narrow tasks (Elazar et al. 2021; Tahmasebi, Müller-Budack, and Ewerth 2024). The MMIR benchmark (Yan et al. 2025) evaluates inconsistency reasoning in layout-rich artifacts but restricts its scope to predefined error types and lacks granular diagnosis of underlying capability failures. Works like Beyond Appearance (Xu et al. 2025) study modality gaps in specific attributes (e.g., color, shape) but do not scale to compositional real-world conflicts. Crucially, VLM2-Bench (Zhang et al. 2025) addresses a distinct challenge: visually *linking matching cues* across different images (e.g., identifying the same person), which represents an orthogonal problem to resolving *contradictory multimodal evidence* within unified inputs. CrossCheck-Bench thus fills a critical void by introducing the first hierarchical evaluation framework (L1 $\rightarrow$ L3) to diagnose *why* models fail at cross-modal conflict resolution—requiring both structural alignment and logical comparison within unified multimodal inputs.

Diagnostic Evaluation of VLMs. Efforts to diagnose VLM failures typically dissect atomic capabilities in isolation. SpaCE-10 (Gong et al. 2025) decomposes spatial intelligence into 10 atomic skills but does not examine their interplay under *conflicting evidence*. BEiT-3 (Wang et al. 2023a) and LLaVA (Liu et al. 2023a) analyze modality biases via probing tasks, yet these remain decoupled from real-world inconsistency scenarios. Our benchmark uniquely integrates capability-centric diagnosis with adversarial multimodal conflict: we not only define 7 atomic cross-checking capabilities (e.g., entity grounding, attribute verification) but deliberately stress-test their composition in failure-prone contexts where cues contradict, thereby exposing cascading reasoning breakdowns unseen in prior benchmarks.

CrossCheck-Bench

To enable fine-grained diagnosis of multimodal inconsistency resolution in Vision-Language Models (VLMs), we introduce *CrossCheck-Bench*, a factually grounded benchmark comprising 7 atomic capabilities and 15 systematically constructed tasks. As shown in Figure 2, CrossCheck-Bench follows a hierarchical structure, categorizing tasks into Perception (L1), Integration (L2), and Reasoning (L3) levels, based on the combination of atomic capabilities and task complexity. To ensure the reliability and consistency, the construction of CrossCheck-Bench involves three key stages (see Figure 3): Data Collection via Multimodal Clue Graphs, Hierarchical Tasks Generation, and Quality Verification. In total, **over 450 expert-hours** were invested to curate 14.69k high-fidelity, semantically grounded QA tasks.

Data Collection

We curated a diverse e-commerce dataset from major platforms, comprising 22.8k listings with $\geq$ 5 verified attributes and high-resolution images. To represent factual signals, we define *Multimodal Cue Graphs* (MCGs) as quadruples: `(entity, modality, attribute, value)` (Kommineni, König-Ries, and Samuel 2024; Yao et al. 2025). MCG construction proceeds in three stages:

Entity Extraction: We employ an ensemble of YOLOv8-L (Yi et al. 2023), GroundingDINO (Liu et al. 2024), and visual embeddings for image-level detection, alongside fine-tuned Qwen3-8B (Yang et al. 2025) for textual entity recognition. Attribute Extraction: Visual (e.g., OCR, shape) and textual (e.g., brand, price) signals are extracted via rule-based templates and GPT-4o augmentation. Cross-Validation: GPT-4o identifies cross-modal inconsistencies. Discrepant pairs are corrected, with a 15% manual audit yielding 98.2% accuracy.

The final set includes 22.8k MCGs, each containing on average 12.7 verifiable clues, with a semantic consistency rate of 97.3% based on expert auditing.

Hierarchical QA Generation

To enable diagnostic evaluation across varying levels of reasoning complexity, we formulate a three-tiered task taxonomy grounded in atomic visual-language capabilities. Each task is derived from structured information encoded in the MCGs, allowing precise targeting of distinct skills and their combinations.

The Three-Tier Diagnostic Taxonomy. We define seven atomic capabilities for assessing multimodal reasoning: A1 (Visual Grounding), A2 (Entity Recognition), A3 (Attribute Comparison), A4 (Multi-frame Reasoning), A5 (Numerical

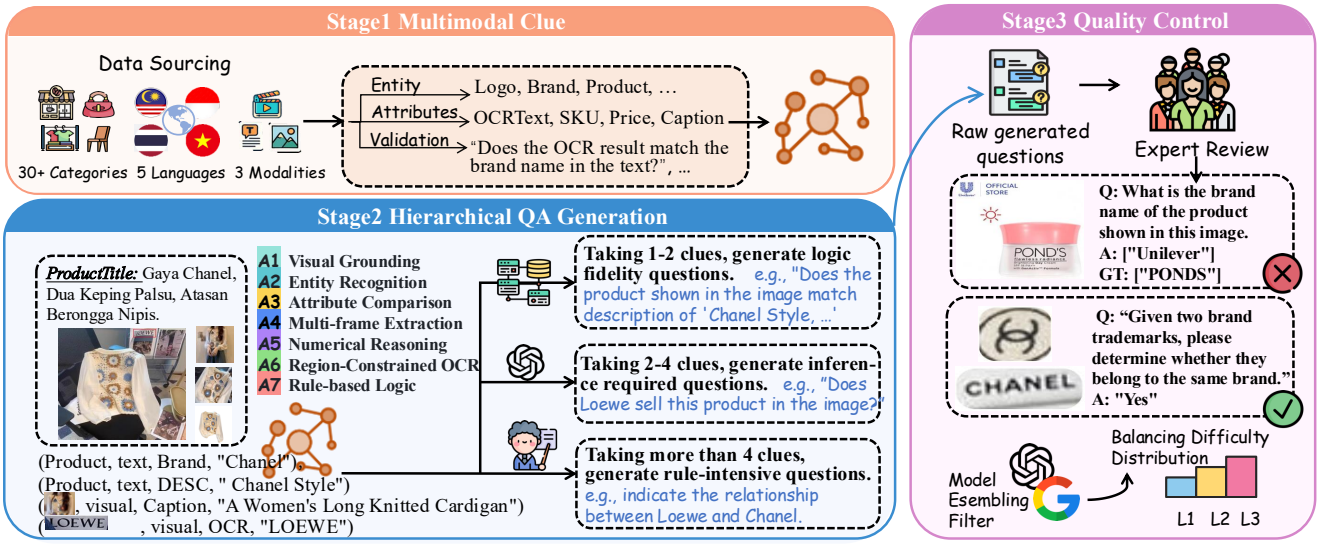

Figure 3: Dataset–construction pipeline for CROSSCHECK-BENCH. Stage 1: Clue Encoding. Aggregates multimodal data (30+ categories, 5 languages) into clue graphs binding entities with validated attributes. Stage 2: QA Composition. Samples $1 - n$ clues to generate hierarchical QA pairs targeting 7 capabilities across 3 cognitive tiers (L1–L3). Stage 3: Quality Control. Employs a three-step loop (expert review, model filtering, and difficulty balancing) to ensure correctness and task uniformity.

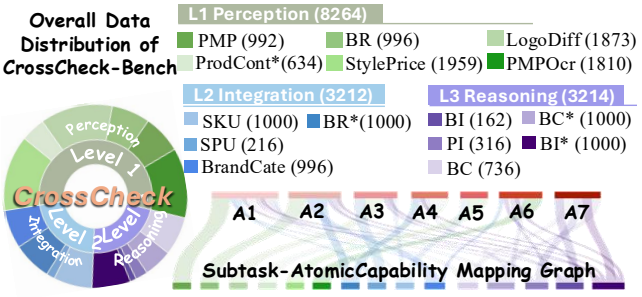

Figure 4: Dataset statistics for CROSSCHECK-BENCH. The benchmark contains $\sim$ 15,000 question–answer pairs distributed over 15 subtasks and three cognitive levels (left). Six subtasks probe a single atomic capability (A1–A6), while the remaining nine require compositions capabilities (right). Subtask names followed by "*" take multi-frame input.

Plausibility), A6 (Region-Constrained OCR), and A7 (Rule-based Logic). These skills represent fundamental cross-modal abilities required to detect real-world inconsistencies. Building on them, we construct a three-level hierarchy to measure increasing reasoning depth. Perception (L1) isolates a single atomic skill to evaluate basic cross-modal alignment. Integration (L2) combines two to three capabilities to test coordination across modalities. Reasoning (L3) requires synthesizing multiple clues, applying commonsense or domain knowledge, and resolving implicit contradictions. This taxonomy is both diagnostic and necessary, enabling fine-grained attribution of failures to perceptual, integrative, or reasoning limitations.

Question Generation.

To generate diverse questions aligned with our taxonomy, we adopt a hybrid framework combining template, model, and human strategies: (1) Template-based (L1). For atomic tasks, we design 45+ rule-based templates operating on MCGs to test grounding and attribute consistency. (2) Model-assisted (L2). For integration tasks, GPT-4o is prompted with structured instructions to generate questions involving cross-modal or temporal reasoning, refined by human reviewers. (3) Human-authored (L3). For reasoning tasks requiring multi-step inference, experts manually craft questions targeting compositional challenges like intention deception or rule violations. All questions undergo expert validation to ensure semantic rigor. The final benchmark contains 14.69k QA pairs balanced across three levels. A detailed composition breakdown is shown in Figure 4.

Quality Verification

In order to ensure the reliability of CrossCheck-Bench, we implement verification pipeline for addressing difficulty labeling, logic consistency, and robustness. Tasks are labeled as L1 (Perception), L2 (Integration), or L3 (Reasoning) via model consensus (76% agreement across GPT-4o, GPT-4.1, Gemini 2.5 Pro) and finalized by expert annotators under a clear rubric. Experts override model votes in 18% of cases.

We perform adversarial validation on 40% of samples: three reviewers assess each QA pair for ambiguity or shortcuts, with 12% flagged for revision. Inter-annotator agreement (IAA) on a separate 10% confirms annotation stability. These steps ensure that CrossCheck-Bench is both challenging and reliable for multimodal reasoning evaluation.

Models	Avg.	Perception						Integration				Reasoning				
		LD	BR	PMP	PMO	PD*	SP	BR*	SPU	SKU	BDC	BC	BC*	BI	PI	BI*
Open-Source MLLMs (Small Group)																
Qwen2.5VL-7B	61.0	59.1	61.8	78.8	56.6	53.2	42.9	62.5	78.4	59.8	64.9	58.9	50.0	80.9	51.6	55.3
Ovis2-8B	60.5	**77.9**	59.8	72.1	51.1	72.6	36.6	67.6	67.9	56.6	75.5	58.7	50.1	66.1	50.3	44.6
MiMo-VL-7B	65.3	66.8	**93.0**	73.7	59.5	62.9	52.1	66.1	76.2	56.4	91.5	60.7	52.8	59.3	61.4	46.7
Open-Source MLLMs (Medium Group)																
Ovis2-16B	67.2	77.2	75.4	74.1	61.4	73.5	52.0	71.3	78.0	69.8	89.9	58.3	52.6	75.3	57.6	42.0
Qwen2.5VL-32B	67.5	67.3	69.5	79.7	67.7	71.9	53.9	69.4	81.2	52.9	93.1	67.8	49.9	65.9	50.0	**72.6**
Ovis2-34B	68.5	67.8	71.6	72.6	70.2	72.4	51.0	69.0	85.3	66.2	89.8	58.6	50.7	63.0	82.7	55.9
InternVL3-38B	68.0	64.2	59.7	**92.4**	64.2	71.5	52.2	71.6	68.4	67.3	93.4	58.9	53.9	63.4	75.6	63.4
Open-Source MLLMs (Large Group)																
Qwen2.5VL-72B	69.9	67.3	69.5	79.7	75.1	72.1	51.9	69.1	86.7	68.3	90.8	58.6	50.6	63.6	72.6	69.2
InternVL3-78B	71.5	73.3	69.1	89.0	74.4	61.5	50.1	74.0	81.2	71.3	91.9	59.6	53.1	78.8	81.5	64.0
Closed-Source MLLMs																
GPT-4.1-mini-2025-04-14	72.6	65.8	85.7	78.1	79.1	**74.8**	60.4	74.5	86.7	**73.9**	91.8	67.4	54.0	74.7	68.4	54.0
ByteDance-seed-1.6	73.3	62.9	90.3	76.2	66.2	59.2	63.2	71.0	89.5	73.7	96.2	63.1	57.8	**85.5**	**81.0**	63.8
GPT-4.1-2025-04-14	**76.8**	68.2	89.9	78.1	**85.3**	71.1	70.2	80.1	**90.4**	72.1	**98.1**	67.8	65.4	74.7	75.7	65.4
Gemini-2.5-pro-p-06-05	76.2	69.6	92.0	76.8	80.9	70.7	**72.2**	**83.7**	85.8	48.2	97.7	**71.0**	**66.3**	77.2	**81.0**	70.2
Human	**95.2**	**94.5**	**98.1**	**96.7**	**93.2**	**88.5**	**85.6**	**92.4**	**97.8**	**89.1**	**99.5**	**85.2**	**82.1**	**94.3**	**92.8**	**88.0**

Table 1: Main results of the CROSSCHECK-BENCH. The best performance is highlighted by bold and underline. Human performance in bold as an upper-bound reference. * denotes multi-frame input.

Experiments and Analysis

Setup

We evaluate 13 vision-language models (VLMs), including both proprietary and open-source systems capable of processing interleaved image-text inputs. The proprietary models comprise Gemini 2.5 Pro(Comanici et al. 2025), GPT-4.1 (Achiam et al. 2023), GPT-4.1mini (Achiam et al. 2023), and ByteDance-seed-1.6 (Team 2025). The open-source group spans multiple families and scales: Qwen2.5-VL(7B, 32B, 72B)(Bai et al. 2025), InternVL3(38B, 78B)(Zhu et al. 2025) , Ovis2(8B, 16B, 34B)(Lu et al. 2024), MiMo-VL-7B(Xiaomi et al. 2025). All models are evaluated on CrossCheck-Bench using a unified zero-shot QA protocol with standardized prompts. Evaluation employs hybrid scoring: deterministic single-choice questions are assessed via exact match, while open-ended responses are semantically judged by GPT-4o. Open-source models are executed using official implementations on NVIDIA H100 GPUs. Human baseline performance is collected from seven expert annotators following the same QA protocols.

Overall Results

Table 1 summarizes the performance of 13 VLMs across 15 benchmark tasks, categorized into three task levels: perception, integration, and reasoning.

Human vs. MLLMs Human accuracy provides an upper bound across all task levels. The average score reaches 95.2%, surpassing the best proprietary model by over 18 points. Even on reasoning tasks, human performance remains above 88%, while most models fall short of 76%. This gap persists across categories, confirming that current models struggle with consistent multimodal alignment, particularly under conflicting or compositional input.

Open vs. Closed Models Proprietary models consistently outperform open baselines. GPT-4.1 and Gemini 2.5 Pro achieve average scores above 76%, while the strongest open-source model peaks at 71.5%. Although open models perform competitively on localized tasks, their advantage vanishes as task complexity increases. The gap widens on composition-intensive queries, where proprietary systems maintain a consistent lead.

Performance Declines from Anchoring to Reasoning Across all models, accuracy consistently declines as tasks progress from perceptual anchoring (L1) to knowledge integration (L2) and conflict reasoning (L3). Closed-source models show a clear downward trend, with GPT-4.1 dropping from 85.3% on L1 tasks to 75.7% on L3. Open-source models exhibit even sharper declines. InternVL3-78B falls from 71.5% at L1 to 64.0% at L3, and Qwen2.5-VL-72B drops from 69.9% to 63.9%. This pattern reveals that while most models can extract atomic entities and perform attribute-level comparisons, they struggle with reasoning over conflicting cues and enforcing logical consistency. The most severe degradation occurs in tasks that require multi-attribute fusion and rule-based contradiction detection, confirming that compositional reasoning remains the most fragile capability across architectures.

Model Scaling Yields Uneven Benefits Across Levels Scaling improves performance on low-level tasks but brings inconsistent or diminishing returns at higher levels. For L1 tasks, larger models show notable gains. InternVL improves by 3.5 points from 38B to 78B, and Qwen2.5-VL improves by nearly 9 points from 7B to 72B. However, gains on L2 tasks are unstable. Qwen2.5-VL-72B improves marginally over 32B on some tasks, while degrading on others. At L3, reasoning accuracy stagnates or declines, even in large-scale

Task	Capacity	GPT-4.1	InternVL3-78B	Qwen2.5-VL-72B	Ovis2-34B	MiMo-VL-7B	Overall-A(%)
BR	A2	89.9	73.9	69.5	71.6	93.0	75.9
BR*	**+ A4**	80.1 (↓9.8)	74.0 (↑0.1)	69.1 (↓0.4)	69.0 (↓2.6)	66.1 (↓26.9)	71.5 (↓4.4)
SPU	A1,3	90.4	81.2	86.7	85.3	76.2	81.2
SKU	**+ A5**	72.1 (↓18.3)	71.3 (↓9.9)	68.3 (↓18.4)	66.2 (↓19.1)	56.6 (↓19.6)	64.4 (↓16.8)
BDC	A1,2,6	98.1	91.9	90.8	89.8	91.5	89.6
PI	**+ A7**	75.7 (↓22.4)	81.5 (↓10.4)	72.6 (↓18.2)	82.7 (↓7.1)	61.4 (↓30.1)	73.8 (↓15.8)
BI	**+ A3,7**	74.7 (↓23.4)	78.8 (↓13.1)	63.6 (↓27.2)	63.0 (↓26.8)	59.3 (↓32.2)	66.0 (↓23.6)
BI*	**+ A3,4,7**	65.4 (↓32.7)	64.0 (↓27.9)	69.2 (↓21.6)	55.9 (↓33.9)	46.7 (↓44.8)	59.0 (↓30.6)

Table 2: Accuracy comparison on atomic vs. compositional tasks.Atomic variants and corresponding compositional variants are paired per task. Values in parentheses denote accuracy drops due to capability addition, measured relative to the baseline.

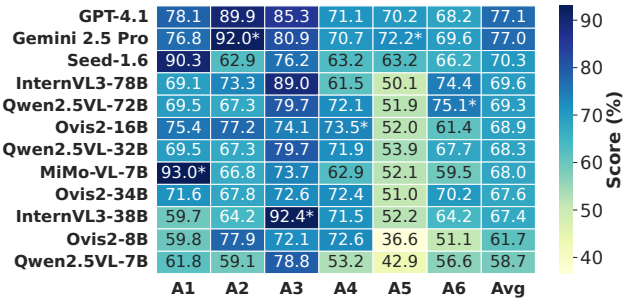

Figure 5: Model-wise Performance on Atomic Capabilities. Each cell indicates accuracy on one capability, with * denoting the best-performing model per capability.

models. These results indicate that increased capacity enhances surface perception but does not resolve the bottlenecks in cross-modal composition. The inability to align symbolic and visual information under constraint suggests that reasoning limitations persist independently of scale.

Capability Analysis

Atomic vs. Compositional Performance. Table 2 presents a controlled comparison between atomic and compositional tasks, revealing how capability integration impacts model reliability. While models handle isolated tasks (e.g., A1–A3) reasonably well, accuracy drops by 12%–35% when tasks involve numerical plausibility (A5), cross-frame reasoning (A4), or rule compliance (A7).

This compositional collapse is most evident when all capabilities (A1–A7) are required: top models often fail to exceed 50% accuracy, with smaller ones below 40%. Failures correlate strongly with tasks involving temporal (A4), numerical (A5), or rule-based (A7) reasoning, underscoring persistent architectural blind spots. CrossCheck-Bench disentangles these challenges, enabling precise attribution of integration failures.

Capability-Specific Trends and Scaling Effects. Figure 5 summarizes model accuracy on atomic capabilities A1 through A6. These span from perceptual anchoring to symbolic reasoning. A clear separation emerges between perceptual skills (A1–A3) and reasoning-oriented capabilities

Intervention	A5: Numeric	A6: OCR	A7: Logic
Base (Vanilla)	61.2	58.7	49.1
CoT Prompting	62.0	56.3	50.8 ↑
SoM Prompting	62.4	60.9 ↑	48.6
CoT + SoM	61.8	59.3	50.1
CSFT	63.5 ↑	60.2	49.5
MM-CoT	**65.3** ↑	**61.7** ↑	**53.5** ↑

Table 3: Accuracy (%) under prompting interventions on three capabilities. Gray cells mark best results. Bold and ↑ indicate top-2 gains over base.

(A4–A6). Models improve steadily on A1–A3 as scale increases. GPT-4.1 exceeds 85% on A2 and A3. These skills rely on spatial anchoring and shallow visual-textual mapping, which align well with current pretraining objectives. In contrast, A4 through A6 remain challenging across model families. Even top-tier models score below 75% on average across these three capabilities. Smaller models collapse entirely. Ovis2-8B drops to 36.6% on A5, and most open models perform below 55% on A6. These failures suggest persistent fragility in symbolic inference when it depends on temporal alignment, numerical estimation, or rule-based logic.

Model scaling does not resolve these weaknesses uniformly. InternVL3-38B outperforms GPT-4.1 on A3, and Qwen2.5-VL-72B achieves the best A6 result overall. This suggests that architectural tuning or supervision strategies may contribute more than parameter count for certain atomic skills. While perceptual capabilities scale smoothly, symbolic coordination remains an unsolved frontier in vision-language alignment.

Prompting-Based Diagnostic Interventions

We test whether prompting or light tuning can recover capability-specific weaknesses revealed by CrossCheck-Bench. We focus on three difficult skills: numeric reasoning (A5), region-based OCR (A6), and logic inference (A7), where most models show low accuracy.

Prompting and Fine-tuning Configurations. We evaluate vision-language models under four strategies: Chain-of-Thought (CoT) prompting(Wei et al. 2022), Set-of-Mark (SoM) visual guidance (Yang et al. 2023), a combined

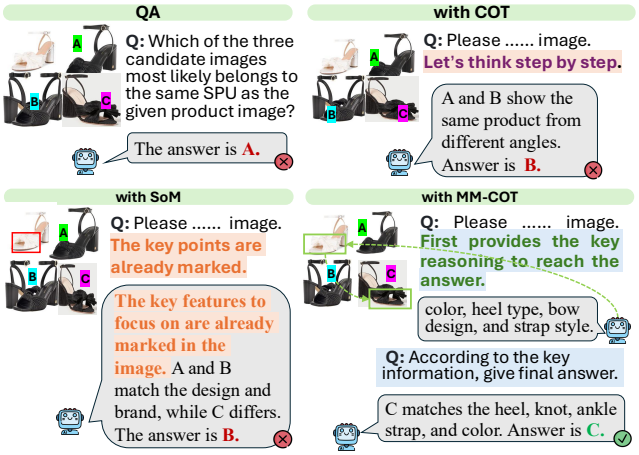

Figure 6: Prompting Strategies. We visualize answer rationales and visual focus under three prompting configurations. MM-CoT coordinates reasoning and grounding, leading to correct identification through structured inference.

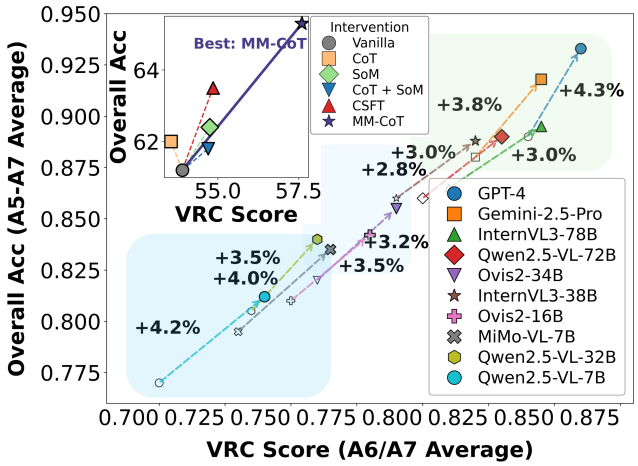

Figure 7: Effect of MM-CoT. The figure illustrates the performance shift across models with different prompting interventions. MM-CoT consistently enhances reasoning alignment and overall accuracy. Inset: comparison of intervention strategies on reasoning-intensive tasks (A5–A7).

CoT+SoM setting, and supervised fine-tuning (CSFT) with 500 curated QA pairs. CoT adds symbolic reasoning scaffolds. SoM uses bounding boxes to guide attention. The combined setup merges both. CSFT tests model adaptability with lightweight supervision.

As shown in Table 3, CoT improves A7 performance (+1.7% on average) but often harms perception-centric tasks due to hallucinated logic paths. SoM is particularly effective on A6 tasks for proprietary models with stronger visual pretraining, though results vary in open models. Combining CoT and SoM yields no consistent improvement and may introduce modality interference. CSFT helps moderately on A5 and A6 but fails to rectify A7 reasoning failures, suggesting tuning alone cannot repair logic abstraction gaps.

Multimodal Interleaved CoT (MM-CoT) To further address the failures, we propose *Multimodal Interleaved CoT (MM-CoT)*, a two-stage prompting protocol designed to weave together grounding and reasoning. In Stage 1, models generate candidate answers with free-form rationales, which are parsed to extract relevant visual elements and highlight them using bounding-box overlays. In Stage 2, the model is re-invoked with both the SoM-augmented input and its own previous reasoning trace. This procedure encourages iterative inference, linking visual localization and symbolic logic through an explicit feedback loop. Figure 6 shows a representative case under four prompting variants.

As illustrated in Figure 7, MM-CoT consistently shifts models toward higher reasoning accuracy and VRC scores, especially for tasks involving compositional numerical and rule-based inference. MM-CoT outperforms all prior strategies: GPT-4o sees a +4.4% gain over vanilla prompting, and open models average +2.1% improvement. Benefits are most significant in tasks demanding chained reasoning over values and constraints, such as those involving both A5 and A7. These findings demonstrate that tightly interleaved cross-modal prompting offers a viable path for addressing complex multimodal failures.

Overall, these results show that prompting strategies can selectively recover model deficits but struggle under compositional complexity. MM-CoT, by encouraging reasoning-grounding feedback loops, represents a promising direction for enhancing cross-modal inference—especially for tasks where visual and symbolic information must be reconciled through structured reasoning.

Conclusion

In this paper, we introduce CrossCheck-Bench, a diagnostic benchmark for evaluating vision-language models (VLMs) under multimodal inconsistency. The benchmark targets models' ability to detect conflicts and reason compositionally across visual, textual, and symbolic inputs. CrossCheck-Bench defines a three-tier capability structure and spans 15 tasks covering atomic and compositional reasoning. Through systematic evaluation, we uncover three key trends: (1) performance degrades consistently from perception to reasoning; (2) models prioritize internal priors over conflicting external signals; (3) tasks requiring cross-frame alignment and rule-grounded inference remain major bottlenecks. These failures are not random but traceable to structural capability gaps. By isolating where compositional failures emerge, CrossCheck-Bench offers a roadmap for improving alignment between symbolic inference and grounded perception. Additional probing shows that conventional prompting strategies such as Chain-of-Thought and Set-of-Mark yield only marginal gains. By contrast, methods that interleave symbolic reasoning with grounded visual processing achieve more stable improvements. We hope this work enables more robust, reliable, and cognitively grounded VLMs for real-world applications, and provides a foundation for future research on multimodal consistency.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alayrac, J.-B.; Donahue, J.; Lucic, P.; Miech, A.; Barr, I.; Gokalp, Y.; Smaira, L.; Laptev, I.; Lecun, Y.; and Sivic, J. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. *Advances in Neural Information Processing Systems*, 35: 30479–30493.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Bai, Z.; Wang, P.; Xiao, T.; He, T.; Han, Z.; Zhang, Z.; and Shou, M. Z. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Elazar, Y.; Kassner, N.; Ravfogel, S.; Ravichander, A.; Hovy, E.; Schütze, H.; and Goldberg, Y. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9: 1012–1031.

Gong, Z.; Li, W.; Ma, O.; Li, S.; Wang, Z.; Ji, J.; Yang, X.; Luo, G.; Yan, J.; and Ji, R. 2025. Space-10: A comprehensive benchmark for multimodal large language models in compositional spatial intelligence. *arXiv preprint arXiv:2506.07966*.

Gunjal, A.; Yin, J.; and Bas, E. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 18135–18143.

Johnson, J.; Hariharan, B.; Van Der Maaten, L.; Fei-Fei, L.; Lawrence Zitnick, C.; and Girshick, R. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2901–2910.

Kommineni, V. K.; König-Ries, B.; and Samuel, S. 2024. From human experts to machines: An LLM supported approach to ontology and knowledge graph construction. *arXiv preprint arXiv:2403.08345*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, J.; Lu, W.; Fei, H.; Luo, M.; Dai, M.; Xia, M.; Jin, Y.; Gan, Z.; Qi, D.; Fu, C.; et al. 2024. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*.

Li, J.; Wang, D.; Li, L.; Gan, Z.; and Hoi, S. C. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. *International Conference on Machine Learning*, 12888–12900.

Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; and Wen, J.-R. 2023b. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023a. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, 38–55. Springer.

Liu, Y.; Yao, Y.; Ton, J.-F.; Zhang, X.; Guo, R.; Cheng, H.; Klochkov, Y.; Taufiq, M. F.; and Li, H. 2023b. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*.

Lu, P.; Bansal, H.; Xia, T.; Liu, J.; Li, C.; Hajishirzi, H.; Cheng, H.; Chang, K.-W.; Galley, M.; and Gao, J. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Lu, S.; Li, Y.; Chen, Q.-G.; Xu, Z.; Luo, W.; Zhang, K.; and Ye, H.-J. 2024. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*.

Ma, Z.; Hong, J.; Gul, M. O.; Gandhi, M.; Gao, I.; and Krishna, R. 2023. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10910–10921.

Maaz, M.; Rasheed, H.; Khan, S.; and Khan, F. S. 2023. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. arXiv:2306.05424.

Suhr, A.; Zhou, S.; Zhang, A.; Zhang, I.; Bai, H.; and Artzi, Y. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, 6418–6428.

Tahmasebi, S.; Müller-Budack, E.; and Ewerth, R. 2024. Multimodal misinformation detection using large vision-language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2189–2199.

Tan, R.; Plummer, B.; and Saenko, K. 2020. Detecting cross-modal inconsistency to defend against neural fake news. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2081–2106.

Team, B. S. 2025. Seed1.5-VL Technical Report. *arXiv preprint arXiv:2505.07062*.

Wang, W.; Bao, H.; Dong, L.; Bjorck, J.; Peng, Z.; Liu, Q.; Aggarwal, K.; Mohammed, O. K.; Singhal, S.; Som, S.; et al. 2023a. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19175–19186.

Wang, Y.; Zhong, W.; Li, L.; Mi, F.; Zeng, X.; Huang, W.; Shang, L.; Jiang, X.; and Liu, Q. 2023b. Aligning large

language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

Wasserman, N.; Pony, R.; Naparstek, O.; Goldfarb, A. R.; Schwartz, E.; Barzelay, U.; and Karlinsky, L. 2025. REAL-MM-RAG: A Real-World Multi-Modal Retrieval Benchmark. *arXiv preprint arXiv:2502.12342*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wolf, Y.; Wies, N.; Avnery, O.; Levine, Y.; and Shashua, A. 2023. Fundamental limitations of alignment in large language models. *arXiv preprint arXiv:2304.11082*.

Xiaomi, L.; Xia, B.; Shen, B.; Zhu, D.; Zhang, D.; Wang, G.; Zhang, H.; Liu, H.; Xiao, J.; Dong, J.; et al. 2025. MiMo: Unlocking the Reasoning Potential of Language Model–From Pretraining to Posttraining. *arXiv preprint arXiv:2505.07608*.

Xie, N.; Lai, F.; Doran, D.; and Kadav, A. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*.

Xu, Z.; Tang, F.; Chen, Z.; Su, Y.; Zhao, Z.; Zhang, G.; Su, J.; and Ge, Z. 2025. Toward modality gap: Vision prototype learning for weakly-supervised semantic segmentation with clip. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 9023–9031.

Yan, Q.; Fan, Y.; Li, H.; Jiang, S.; Zhao, Y.; Guan, X.; Kuo, C.-C.; and Wang, X. E. 2025. Multimodal inconsistency reasoning (MMIR): A new benchmark for multimodal reasoning models. *arXiv preprint arXiv:2502.16033*.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yang, J.; Zhang, H.; Li, F.; Zou, X.; Li, C.; and Gao, J. 2023. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*.

Yao, L.; Peng, J.; Mao, C.; and Luo, Y. 2025. Exploring large language models for knowledge graph completion. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Yi, H.; Liu, B.; Zhao, B.; and Liu, E. 2023. Small object detection algorithm based on improved YOLOv8 for remote sensing. *IEEE journal of selected topics in applied earth observations and remote sensing*, 17: 1734–1747.

Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9556–9567.

Zellers, R.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6720–6731.

Zhang, J.; Yao, D.; Pi, R.; Liang, P. P.; and Fung, Y. R. 2025. VLM2-Bench: A Closer Look at How Well VLMs Implicitly Link Explicit Matching Visual Cues. *arXiv preprint arXiv:2502.12084*.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.