

VCU-Bridge: Hierarchical Visual Connotation Understanding via Semantic Bridging

Ming Zhong^{1,*} Yuanlei Wang^{3,*} Liuzhou Zhang² Ruichuan An^{2,♠} Renrui Zhang⁴

Hao Liang² Ming Lu² Ying Shen^{3,5,6,†} Wentao Zhang^{2,7,8,†}

¹Zhejiang University ²Peking University ³Sun Yat-sen University ⁴CUHK ⁵Peng Cheng Laboratory

⁶Guangdong Provincial Key Laboratory of Fire Science and Intelligent Emergency Technology

⁷Zhongguancun Academy ⁸Beijing Key Laboratory of Data Intelligence and Security

Project page: <https://vcu-bridge.github.io>

Abstract

While Multimodal Large Language Models (MLLMs) excel on benchmarks, their processing paradigm differs from the human ability to integrate visual information. Unlike humans who naturally bridge details and high-level concepts, models tend to treat these elements in isolation. Prevailing evaluation protocols often decouple low-level perception from high-level reasoning, overlooking their semantic and causal dependencies, which yields non-diagnostic results and obscures performance bottlenecks. We present **VCU-Bridge**, a framework that operationalizes a human-like hierarchy of visual connotation understanding: multi-level reasoning that advances from foundational perception through semantic bridging to abstract connotation, with an explicit evidence-to-inference trace from concrete cues to abstract conclusions. Building on this framework, we construct **HVCU-Bench**, a benchmark for hierarchical visual connotation understanding with explicit, level-wise diagnostics. Comprehensive experiments demonstrate a consistent decline in performance as reasoning progresses to higher levels. We further develop a data generation pipeline for instruction tuning guided by Monte Carlo Tree Search (MCTS) and show that strengthening low-level capabilities yields measurable gains at higher levels. Interestingly, it not only improves on HVCU-Bench but also brings benefits on general benchmarks (average +2.53%), especially with substantial gains on MMStar (+7.26%), demonstrating the significance of the hierarchical thinking pattern and its effectiveness in enhancing MLLM capabilities.

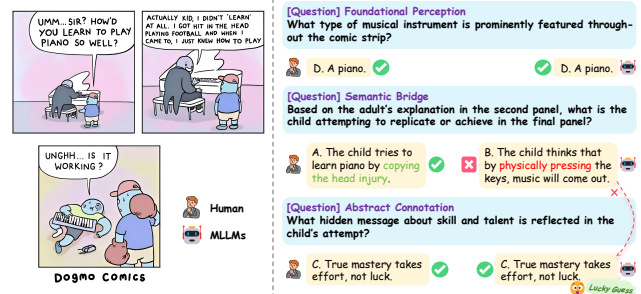


Figure 1. **Showcase of different pattern between human and model.** Models can appear capable by correctly answering both concrete and abstract questions while fundamentally failing at the reasoning that bridges them. Current evaluation may miss critical reasoning failures when models produce correct answers at both concrete and abstract levels.

1. Introduction

While Multimodal Large Language Models (MLLMs) excel on benchmarks, their processing paradigm differs from the human ability to integrate visual information. As illustrated in Figure 1, unlike humans who naturally bridge details and high-level concepts, models tend to treat these elements in isolation. This phenomenon misses a vital aspect of human vision: our capacity to form abstract, connotative meanings based on concrete perceptual evidence [30]. The inferential process that connects seeing to interpreting remains largely unevaluated.

Existing evaluation paradigms fall short in capturing this critical inferential process. Benchmarks focused on low-level perception [15, 29, 52] test the ability to identify what is in an image, but do not evaluate the ability to infer what is implied by it. Conversely, benchmarks for high-level cognition [33, 34, 38, 45] often assess the final abstract conclusion but do not require the model to ground its reasoning in specific visual details, leaving the inferential process a

* Equal contribution, Email: chime@zju.edu.cn

♠ Project Leader

† Corresponding authors

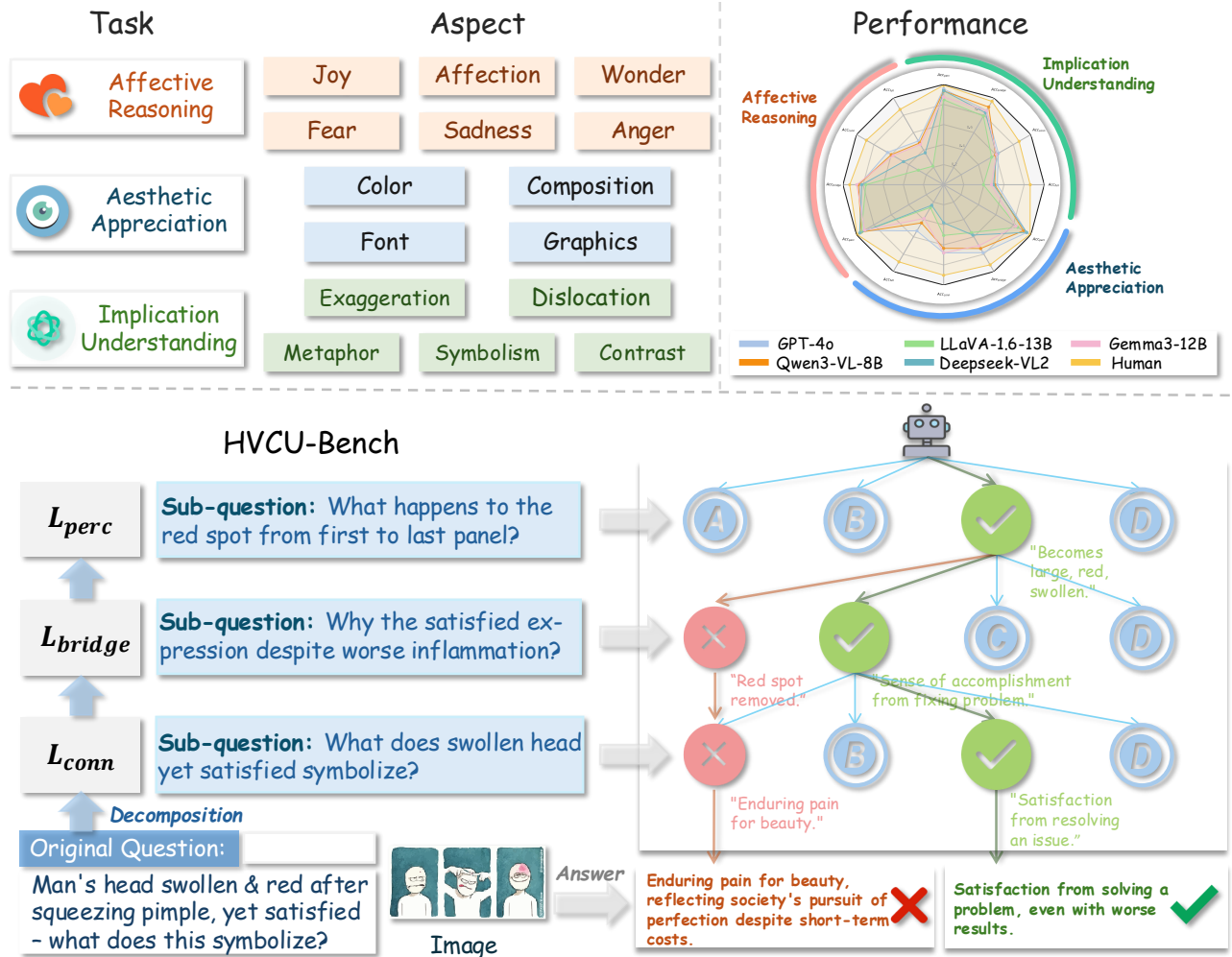


Figure 2. **Overview of HVCU-Bench.** We evaluate MLLMs across 3 task families spanning 15 diverse aspects (top left). Our benchmark employs hierarchical decomposition: each question is systematically broken down into sub-questions across three levels (L_{perc} , L_{bridge} , L_{conn}), with validation ensuring logical coherence. During evaluation, models progress from low to high levels, constructing inter-level reasoning chains that emulate human visual comprehension (bottom). While GPT-4o achieves top performance among MLLMs, it falls substantially short of human capability, exposing a significant gap (top right).

black box. What remains missing is an explicit focus on the bridging inference: the associative reasoning that connects concrete perceptual evidence to abstract connotative meaning (e.g., dark clouds and a hunched posture create a scene of a person alone in bad weather, which connotes a sense of gloom) [30, 39], and addressing this critical gap is the primary motivation of our work.

To address this, we introduce and formalize the concept of **VCU-Bridge**, which refers to hierarchical understanding of visual connotation. We propose that effective visual comprehension operates as a hierarchical, multi-level reasoning process. It begins at the **Foundational Perceptual Level** (L_{perc}), which involves identifying low-level, objective visual primitives like objects and their attributes. It culminates in the **Abstract Connotative Level** (L_{conn}), where high-level, subjective interpretations such as aesthet-

ics, emotion, or symbolic meaning are inferred. Critically, these two levels are connected by an intermediate **Semantic Bridge Level** (L_{bridge}), which provides factual, descriptive statements that explain how perceptual evidence gives rise to connotative meaning. It is this unmodeled “Bridge” level, which explicitly links the concrete and the abstract, that remains a key unsolved challenge for MLLMs. VCU-Bridge is the first to focus on the connotative hierarchy, demanding that the model articulate the associative link from L_{perc} , through L_{bridge} , to L_{conn} . The core of our contribution is thus the modeling of this inter-level association, which serves as the foundation for evaluation.

To systematically explore and evaluate the VCU-Bridge capabilities of MLLMs, we construct **HVCU-Bench**, a novel benchmark specifically designed to measure this hierarchical visual reasoning ability. HVCU-Bench explic-

Table 1. Comparison of various recent multimodal benchmarks about visual understanding.

Benchmark	VCU-Bridge				Metrics		Human Evaluation	Open-Source
	Foundational Perception	Semantic Bridge	Abstract Connotation	Inter-Level Association	Accuracy	Full-Chain Accuracy		
Low-level								
MME [13]	✓	×	×	×	✓	×	×	✓
MMBench [26]	✓	×	×	×	✓	×	✓	✓
MMStar [11]	✓	×	×	×	✓	×	×	✓
HallusionBench [16]	✓	×	×	×	✓	×	×	✓
HCA [32]	✓	×	×	×	✓	✓	×	✓
High-level								
MMM [46]	×	×	✓	×	✓	×	✓	✓
MME-Reasoning [45]	×	×	✓	×	✓	×	✓	✓
II-Bench [27]	×	×	✓	×	✓	×	✓	✓
EEmo-Bench [14]	×	×	✓	×	✓	×	×	✓
Multi-level								
Lens [43]	✓	×	✓	×	✓	×	×	✓
MVP-Bench [19]	✓	×	✓	×	✓	×	✓	×
HiBench [18]	✓	×	✓	×	✓	×	×	×
InsightVision [44]	✓	×	✓	×	✓	×	×	×
HVCU-Bench (Ours)	✓	✓	✓	✓	✓	✓	✓	✓

itly models the hierarchical process of visual understanding, requiring models to reason step-by-step from low-level perception to high-level abstraction, which forms tightly coupled inter-level reasoning chains. Through hierarchical evaluation, HVCU-Bench enables tracing failures at higher levels back to potential causes in lower levels, thereby facilitating a more fine-grained analysis of model capability bottlenecks. Figure 2 provides an overview of HVCU-Bench.

Subsequently, to enhance the hierarchical reasoning ability of models, we propose a targeted low-to-high-level data generation approach based on Monte Carlo Tree Search (MCTS) [5, 20, 28, 42, 51]. Specifically, we design a data generation pipeline that starts from low-level perceptual details, progressively builds up to high-level abstract questions, and validates each level to ensure the logical coherence and quality of the hierarchical reasoning chain. The MCTS strategy enables efficient exploration and generation of high-quality training data within the hierarchical reasoning space for instruction tuning.

In summary, our contributions are as follows:

- We novelly introduce and formalize **VCU-Bridge**, modeling visual connotation understanding as a three-level progressive process (L_{perc} / L_{bridge} / L_{conn}), and highlighting the importance of modeling inter-level associations.
- We construct **HVCU-Bench**, the first benchmark to explicitly evaluate VCU-Bridge capabilities. HVCU-Bench enables hierarchical evaluation and diagnosis of deep causes for high-level reasoning failures, providing fine-grained analysis of model capability bottlenecks.
- We propose a novel MCTS-based data generation pipeline that explores the hierarchical reasoning space, progressively constructing chains from low-level perceptual details to high-level abstract concepts with level-wise validation to ensure logical coherence and quality.

- Extensive experiments on HVCU-Bench reveal universal performance degradation from perception to connotation and substantial gains with lower-level context. Models instruction-tuned with data from our pipeline achieve significant improvements on HVCU-Bench (+6.17%) while maintaining strong performance on diverse general benchmarks (average +2.53%).

2. Related Work

Visual Understanding Evaluation. The evaluation of Multimodal Large Language Models (MLLMs) has advanced rapidly [4, 6, 21, 22, 49], but a significant gap persists between low-level perception and high-level connotation. Many existing benchmarks tend to isolate these capabilities, failing to address the inferential processes that link perception with connotation. **Low-level Perception.** Benchmarks such as VQA [7] and MMBench [26] test object recognition, spatial understanding, and factual grounding but do not evaluate the inferential leap from *what is in* an image to *what is implied* by it. Although HCA [32] measures full-chain accuracy across hierarchical categories, it focuses on taxonomic classification relying on pretrained knowledge rather than genuine hierarchical visual understanding. **High-level Connotation.** Benchmarks such as II-Bench [27] and EEmo-Bench [14] assess logical inference and metaphorical interpretation from *what is implied* but do not require grounding in *what is in* the image. **Multi-level.** Benchmarks such as Lens [43] and MVP-Bench [19] cover multiple levels but without inter-level associations, leaving the reasoning pathway from *what is in* to *what is implied* implicit. A comprehensive comparison between our benchmark and prior works is presented in Table 1. Additional discussion on data generation approaches for instruc-

tion tuning is provided in Appendix A.

3. Methods

3.1. Problem Formulation

We formalize **VCU-Bridge** as the structured inference of abstract scene interpretations grounded in perceptual evidence, with each intermediate semantic justification made explicit and auditable. Formally, VCU-Bridge is represented as a structured triple of statements $\mathcal{H} = (l_1, l_2, l_3)$, ordered by their level of abstraction $l_1 \rightarrow l_2 \rightarrow l_3$. A valid hierarchy must satisfy a pairwise support constraint $S(l_k, l_{k+1})$ for $k \in \{1, 2\}$, where S indicates that the higher-level statement l_{k+1} is sufficiently justified by the lower-level one l_k . The three levels are defined as follows:

- **Foundational Perceptual Level** ($L_{perc} = l_1$): Objective, low-level visual facts that are directly observable in the image.
- **Semantic Bridge Level** ($L_{bridge} = l_2$): Explanatory statements that causally link perceptual evidence to higher-level meaning.
- **Abstract Connotative Level** ($L_{conn} = l_3$): Subjective, high-level interpretations of the scene.

This formulation motivates two subsequent components:

- (1) **HVCU-Bench**, which operationalizes the three-level structure and evaluates both per-level and full-chain performance; and
- (2) a targeted data generation pipeline that supplies high-quality hierarchical supervision to systematically and significantly improve MLLMs on VCU-Bridge.

3.2. HVCU-Bench: A Benchmark for VCU-Bridge

To systematically evaluate VCU-Bridge, we introduce HVCU-Bench, a benchmark constructed through a model-driven generation-validation framework powered by Gemini-2.5-Pro. The core principle is a sequential generation process from abstract to concrete, where each level conditions on the previous one, with immediate validation applied after generating each subsequent level to ensure logical coherence and hierarchical dependencies. This interleaved generation-validation approach produces question-answer pairs that are both visually grounded and exhibit strong hierarchical relationships. Detailed statistics, quality assurance procedures, and representative samples are provided in Appendix B.

Task Design. HVCU-Bench comprises three task families covering fifteen fine-grained aspects: *Affective Reasoning* (joy, affection, wonder, anger, fear, sadness); *Aesthetic Appreciation* (color, composition, font, graphics); *Implication Understanding* (metaphor, symbolism, contrast, exaggeration, dislocation). All items follow a hierarchical multiple-choice QA format with four options per question.

For Affective and Implication tasks, we leverage images and high-level concepts from existing datasets II-

Bench [27] and EEmo-Bench [14] to anchor the L_{conn} level, then systematically construct answer options and the lower levels (L_{bridge} , L_{perc}) through our generation framework. For Aesthetic tasks, we source images from LayerD [31] and apply controlled perturbations for each aspect to isolate the targeted aesthetic factor while keeping the hierarchical reasoning structure consistent. For instance, color variants are created by adjusting hue, saturation, and brightness; composition variants by modifying cropping, reframing, and scale; font variants generated by changing size, weight, and spacing; diverse graphics variants generated by altering stroke width, texture, and density.

Sequential Generation with Interleaved Validation.

Each benchmark sample is generated through a three-stage process, where each level l_i is instantiated as a question-answer pair (q_i, a_i) . Given an input image I , the generation proceeds from abstract to concrete:

- **Stage 1 - Connotative Synthesis** (L_{conn}): An MLLM performs holistic image analysis to generate the highest-level question-answer pair (q_3, a_3) , capturing abstract interpretations such as aesthetic quality or emotional tone.
- **Stage 2 - Semantic Bridge Formulation** (L_{bridge}): Conditioned on (q_3, a_3) , the model generates an intermediate-level pair (q_2, a_2) that provides semantic justification for the connotative interpretation. Validation then checks whether (q_2, a_2) offers sufficient logical support for (q_3, a_3) . If validation fails, a refinement loop is triggered to regenerate (q_2, a_2) .
- **Stage 3 - Perceptual Grounding** (L_{perc}): Conditioned on both (q_3, a_3) and (q_2, a_2) , the model generates the most concrete pair (q_1, a_1) , grounding the reasoning chain in observable visual evidence. Similarly, validation ensures (q_1, a_1) provides foundational support for (q_2, a_2) , triggering a refinement loop if validation fails.

Validation Mechanism and Refinement. The validation enforces two critical criteria: (1) *logical dependency*—the lower-level answer must be directly useful for reasoning about the higher-level question, and (2) *difficulty progression*—the lower level must be significantly more objective, concrete, and observable compared to the higher level’s interpretative nature.

When validation fails, a refinement loop is triggered. The failed generation is discarded, and the stage re-executes with two adaptive strategies: (1) incorporating the specific validation failure reason as explicit guidance to more effectively avoid similar logical flaws, and (2) progressively increasing the sampling temperature to further enhance output diversity and escape potential local optima. Prompts for generation and validation are provided in Appendix G.

Evaluation Metrics. We evaluate model performance on HVCU-Bench using metrics that capture both level-specific

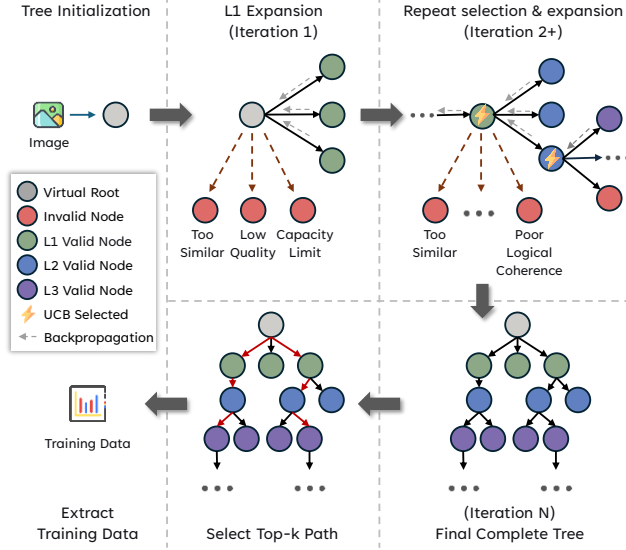


Figure 3. **Overview of our hierarchical data generation pipeline.** An MCTS-driven approach for generating high-quality hierarchical training data.

correctness and full-chain consistency. For each level i , let a_i denote the ground-truth answer and $\hat{a}_i$ the model’s prediction.

- **Per-Level Accuracy (Acc_i):** The proportion of correct predictions at each individual level, formally $Acc_i = \mathbb{P}[\hat{a}_i = a_i]$ for $i \in \{1, 2, 3\}$, providing insight into capability at different abstraction levels.
- **Full-Chain Accuracy (Acc_{full}):** The most stringent metric, defined as $Acc_{full} = \mathbb{P}[\bigwedge_{i=1}^3 (\hat{a}_i = a_i)]$, requires simultaneous correctness across all three levels. This evaluates the model’s ability to perform hierarchical reasoning from perception to connotation.

To provide a single aggregate performance indicator across all three tasks, we also report an Overall Score ($Score$), computed as the mean of the Acc_{full} scores across all these tasks, enabling a fair, robust, and direct model comparison on overall VCU-Bridge capability.

3.3. Hierarchical Data Generation for Instruction Tuning

To enhance the hierarchical reasoning capabilities of MLLMs, we propose a novel data generation pipeline for instruction tuning. In contrast to the benchmark’s top-down sequential generation of individual samples, our training data is constructed through a large-scale bottom-up search. Rather than refining a single reasoning chain per image, we employ MCTS to explore a tree of candidate paths, ultimately selecting diverse high-quality chains for training. Figure 3 illustrates the complete pipeline.

MCTS-driven Reasoning Path Discovery. The core of our generation pipeline is an MCTS procedure that iteratively

constructs a reasoning tree. Starting from a virtual root, the algorithm progressively expands question-answer pairs across the configured hierarchical levels, with each node representing a (q, a) pair at a specific reasoning level. The search balances exploration and exploitation via the Upper Confidence Bound (UCB) strategy, proceeding in three iterative phases:

- **Selection:** The algorithm selects nodes to expand based on UCB: $UCB_j = \bar{X}_j + c\sqrt{\frac{\ln N}{n_j}}$, where $\bar{X}_j$ denotes node j ’s average reward, n_j its visit count, N the parent’s visits, and c is the exploration constant controlling the exploration-exploitation trade-off. This effectively balances exploitation (high-reward nodes) and exploration (less-visited nodes), while prioritizing promising yet underexplored reasoning paths.
- **Expansion & Evaluation:** For each selected node, the pipeline generates a batch of candidate children through parallel MLLM calls, with batch size constrained by a configurable capacity limit per expansion. Each candidate undergoes immediate quality assessment to obtain a numerical score (ranging from 0 to 1), with rejections occurring when candidates exhibit poor logical coherence with the parent, excessive semantic similarity to existing siblings, exceed per-level capacity limits, or fall below a quality threshold (e.g., insufficient image alignment or inappropriate difficulty level). Only candidates successfully passing these checks are admitted to the tree with their evaluation scores, ensuring a more robust balance between exploration breadth and reasoning validity.
- **Backpropagation:** Evaluation scores propagate back up the tree, updating statistics for all ancestors. Specifically, for each ancestor node j on the path, we update $n_j \leftarrow n_j + 1$ and $\bar{X}_j \leftarrow \frac{(n_j-1) \cdot \bar{X}_j + r}{n_j}$, where r is the newly observed reward. This credit assignment ensures high-quality branches accumulate higher rewards, receiving more exploration in iterations.

The pipeline maintains a hierarchical tree with configurable depth and per-level node capacity limits. Through MCTS cycles, combined with quality filtering and diversity checks, the search efficiently explores the reasoning space at scale, constructing a diverse tree of promising, high-quality candidate paths with significantly improved exploration-exploitation trade-offs. Detailed prompts for node generation and evaluation are in Appendix G.

Top-K Reasoning Path Selection. After MCTS convergence, the pipeline extracts the top-K highest-rated complete reasoning paths from the search tree, where each path traverses the configured hierarchical levels from low-level perceptual grounding to high-level abstract interpretation. Path ranking is based on the average evaluation score across all nodes in the path, ensuring that selected paths exhibit both high individual node quality and strong inter-level co-

Table 2. **Overall results on HVCU-Bench.** The best-performing model is **in-bold**, and the second is underlined.

Model	Model Size	Implication Understanding				Aesthetic Appreciation				Affective Reasoning				Score
		<i>Acc_{perc}</i>	<i>Acc_{bridge}</i>	<i>Acc_{conn}</i>	<i>Acc_{full}</i>	<i>Acc_{perc}</i>	<i>Acc_{bridge}</i>	<i>Acc_{conn}</i>	<i>Acc_{full}</i>	<i>Acc_{perc}</i>	<i>Acc_{bridge}</i>	<i>Acc_{conn}</i>	<i>Acc_{full}</i>	
Basic Reference														
Human	-	99.25	96.00	86.50	86.00	99.14	92.29	90.29	88.86	99.33	93.33	88.67	86.67	87.18
GPT-4o	-	95.50	<u>85.25</u>	62.75	53.25	95.43	78.29	68.00	53.14	91.33	83.67	64.33	50.33	52.24
Open-Source MLLMs														
Qwen3-VL-Instruct	4B	86.75	82.75	58.00	43.25	90.57	70.86	60.00	41.14	90.33	82.67	56.67	39.33	41.24
Qwen3-VL-Instruct	8B	93.50	89.50	59.50	<u>50.75</u>	91.71	<u>73.43</u>	<u>63.43</u>	<u>44.00</u>	<u>94.33</u>	84.67	60.00	<u>48.00</u>	<u>47.58</u>
LLaVA-1.6	7B	81.75	58.00	40.25	18.75	79.14	36.86	33.14	9.43	92.00	58.00	19.33	12.00	13.39
LLaVA-1.6	13B	84.75	79.00	55.00	39.50	84.86	55.14	50.57	26.29	<u>94.33</u>	77.33	29.00	21.33	29.04
Deepseek-VL2-tiny	MoE 1B/3B	88.25	62.25	49.75	29.25	89.71	45.14	41.14	19.71	93.33	65.33	29.00	19.00	22.65
Deepseek-VL2	MoE 4.5B/27B	<u>93.75</u>	83.00	60.75	49.50	<u>95.14</u>	58.00	38.00	23.71	96.33	81.33	46.00	36.67	36.63
Gemma3	4B	76.50	72.00	49.75	30.75	68.86	62.86	68.00	29.14	87.00	76.00	51.00	36.00	31.96
Gemma3	12B	87.50	<u>85.25</u>	60.50	47.50	82.86	70.29	68.00	38.86	90.67	86.33	58.00	46.33	44.23
InternVL3.5	4B	82.50	83.75	58.50	42.00	82.86	64.57	40.00	23.43	91.00	81.67	<u>60.67</u>	47.67	37.70
InternVL3.5	8B	82.00	<u>85.25</u>	55.75	41.75	84.00	68.00	60.57	36.29	86.00	83.67	55.67	42.00	40.01
Phi-4-Multimodal-Instruct	6B	90.25	56.50	42.75	32.25	90.29	42.57	23.14	15.14	90.00	<u>85.00</u>	45.33	33.67	27.02
Phi-3.5-Vision-Instruct	4B	84.25	83.25	<u>61.25</u>	44.75	88.29	61.14	53.43	33.14	91.33	82.00	54.33	41.33	39.74

herence. These paths serve directly as instruction tuning data, enabling MLLMs to learn explicit hierarchical reasoning patterns through instruction tuning.

4. Experiments

4.1. Experimental Setup

Models. To comprehensively assess VCU-Bridge capabilities across diverse model architectures and scales, we evaluate 13 MLLMs on HVCU-Bench. Our evaluation includes the leading proprietary model GPT-4o [17], alongside a broad spectrum of open-source models spanning multiple architectural families and parameter scales: Qwen3-VL-Instruct (4B, 8B) [8, 9, 36, 41], LLaVA-1.6 (7B, 13B) [23–25], Gemma3 (4B, 12B) [35], InternVL3.5 (4B, 8B) [37], Phi-4-Multimodal-Instruct (6B) [2], and Phi-3.5-Vision-Instruct (4B) [1]. We further include Mixture-of-Experts architectures with Deepseek-VL2-tiny (MoE 1B/3B) and Deepseek-VL2 (MoE 4.5B/27B) [40] to examine whether sparse expert routing confers advantages for bridging perceptual and connotative levels. Finally, we evaluate Qwen3-VL-4B-Bridge, our instruction-tuned model based on Qwen3-VL-4B-Instruct, to validate the efficacy of our hierarchical data generation pipeline. Complete training configurations and hyperparameters are detailed in Appendix C. This diverse model selection enables systematic analysis of how architecture, scale, and hierarchical instruction data influence VCU-Bridge performance.

Evaluation Benchmark. We evaluate all models on HVCU-Bench, reporting results on its held-out test split. We conduct evaluations under two settings: a “base” setting where models answer each level independently, and a “context” setting where previous levels’ predictions are provided as context. This design allows us to isolate and quantify the contribution of hierarchical dependencies. To assess whether instruction tuning on hierarchical data improves general reasoning capabilities beyond patterns specific to HVCU-Bench, we evaluate Qwen3-VL-4B-Bridge

on four benchmarks: MMBench [26], MMMU [46], MMStar [11], and HallusionBench [16]. These evaluations follow standard protocols without task-specific tuning.

Human Evaluation. To establish a human performance baseline, we collected evaluations from undergraduate students. For each VCU-Bridge level (L_{perc} , L_{bridge} , and L_{conn}), we recruited three students to evaluate it and reported the average score for that level. We intentionally used different groups of annotators for each level to prevent information leakage between them. This ensures the evaluation mirrors the “base” setting for models, as using the same person to evaluate all levels would inadvertently replicate the “context” setting.

4.2. Main Results

Gap between Humans and MLLMs As shown in Table 2, humans nearly saturate all HVCU-Bench levels across all tasks, achieving an overall score of 87.18% with consistently strong performance at both L_{bridge} and L_{conn} . In contrast, top models fall far behind at higher abstraction levels. On Implication Understanding, humans achieve 99.25% at L_{perc} , 96.00% at L_{bridge} , and 86.50% at L_{conn} , while GPT-4o demonstrates near-human performance at L_{perc} (-3.75%) but exhibits a substantial disparity at L_{conn} (-23.75%). The strongest open-source baseline, Qwen3-VL-8B-Instruct, demonstrates a similar pattern with performance gaps of -5.75% at L_{perc} and -27.00% at L_{conn} . These level-specific disparities result in significantly lower Acc_{full} , with GPT-4o and Qwen3-VL-8B-Instruct underperforming humans by -32.75% and -35.25% respectively. This pattern holds consistently across all three tasks, with the overall score revealing a substantial gap of -34.94% for GPT-4o and -39.60% for Qwen3-VL-8B-Instruct, demonstrating that current MLLMs still lack a stable semantic bridge from concrete evidence to abstract meaning.

Universal Performance Degradation in VCU-Bridge. Beyond this human-model gap, Table 2 reveals a

Table 3. Comparison of “base” and “context” settings on HVCU-Bench. + Δ denotes gain over “base”.

Model	Task	Accuracy				Score
		Acc_{perc}	Acc_{bridge}	Acc_{conn}	Acc_{full}	
GPT-4o(Base)	Impl.	95.50	85.25	62.75	53.25	52.24
	Aesth.	95.43	78.29	68.00	53.14	
	Affect.	91.33	83.67	64.33	50.33	
GPT-4o(Context)	Impl.	95.50	89.75	76.50	65.00	68.18 +15.94
	Aesth.	95.43	82.29	87.71	72.86	
	Affect.	91.33	86.00	80.67	66.67	
Qwen3-VL-8B(Base)	Impl.	93.50	89.50	59.50	50.75	47.58
	Aesth.	91.71	73.43	63.43	44.00	
	Affect.	94.33	84.67	60.00	48.00	
Qwen3-VL-8B(Context)	Impl.	93.50	90.00	74.75	62.75	62.28 +14.70
	Aesth.	91.71	74.00	82.57	59.43	
	Affect.	94.33	89.00	76.00	64.67	
Gemma3-4B(Base)	Impl.	76.50	72.00	49.75	30.75	31.96
	Aesth.	68.86	62.86	68.00	29.14	
	Affect.	87.00	76.00	51.00	36.00	
Gemma3-4B(Context)	Impl.	76.50	78.25	63.50	40.75	36.39 +4.43
	Aesth.	68.86	65.14	82.57	35.43	
	Affect.	87.00	75.00	50.00	33.00	

widespread degradation pattern across evaluated MLLMs. Most models, regardless of scale or architecture, exhibit a sharp, cascading decline when moving from perception to connotation. While nearly all achieve high accuracy at L_{perc} , their performance systematically deteriorates at L_{bridge} and typically drops precipitously at L_{conn} . This pattern is pronounced in strong models: GPT-4o experiences a degradation of -32.75%, Qwen3-VL-8B-Instruct degrades by -34.00%, and Gemma3-12B suffers a -27.00% decline on Implication Understanding. Even models that deviate from this trend on specific tasks still struggle at L_{conn} , leading to consistently low Acc_{full} . Such cascading error propagation validates that a critical weakness exists in bridging perception to abstract reasoning.

Analysis of Model Scale and Architecture. Table 2 further reveals that while increasing model scale generally improves performance, it does not resolve the fundamental challenges of VCU-Bridge. For example, LLaVA-1.6-13B achieves an overall score of 29.04%, significantly outperforming its 7B counterpart at 13.39%, yet its performance at L_{conn} remains weak. This suggests that larger models possess stronger foundational capabilities but still lack the specialized knowledge required for connotative reasoning. We also observe distinct performance profiles across different model architectures. Certain architectures, such as Qwen3-VL exhibit more balanced performance across levels, whereas LLaVA-1.6 shows a particularly steep decline after L_{perc} , indicating a weaker ability to understand the semantic bridge. These patterns strongly suggest that the challenge of VCU-Bridge transcends mere model scale and architecture, pointing instead to a deeper, more fundamental gap in current multimodal understanding paradigms.

Hierarchical Dependency in HVCU-Bench. To verify that performance on higher levels depends on lower-level capabilities, we conducted a comparative analysis between the “base” and “context” settings. As shown in Table 3

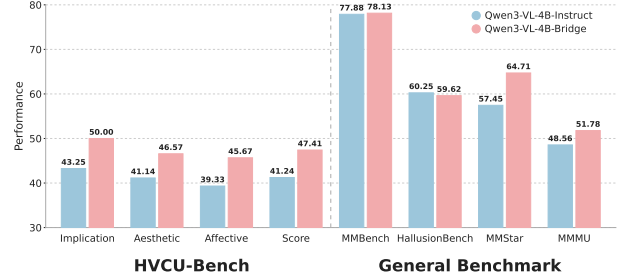


Figure 4. Performance of Qwen3-VL-4B-Bridge on (left) HVCU-Bench and (right) general benchmarks.

(detailed results in Appendix D.2), providing hierarchical context yields substantial performance gains across all evaluated models. Specifically, GPT-4o demonstrates an overall improvement of +15.94%, while Qwen3-VL-8B-Instruct achieves a gain of +14.70%. Examining the level-wise improvements, we observe that most models, particularly stronger ones, exhibit marked accuracy gains across L_{bridge} and L_{conn} , with GPT-4o improving by approximately +3.6% and +16.6% on average across the two levels, respectively. Even in cases where certain tasks show irregular patterns, such as Gemma3-4B on Affective Reasoning, the overall score consistently improves (+4.43%), a pattern that further aligns with our expectation that models with weaker lower-level capabilities would show more constrained higher-level improvements. These pervasive and significant uplifts demonstrate that HVCU-Bench indeed exhibits strong inter-level dependencies, with lower levels providing critical grounding for higher-level connotative reasoning. Beyond validating our hierarchical evaluation, this finding underscores that connotative inference is not a monolithic task but a hierarchical process that fundamentally relies on a coherent chain of reasoning from perception through semantic bridging to abstract interpretation.

4.3. Effectiveness of Hierarchical Data Generation

To validate the effectiveness of our data generation pipeline, we instruction-tune Qwen3-VL-4B-Instruct on approximately 10k hierarchical QA pairs generated from 1k images (Section 3.3), yielding Qwen3-VL-4B-Bridge.

Improvements and Generalization. As shown in Figure 4 (left), Qwen3-VL-4B-Bridge exhibits consistent improvements across all three tasks of HVCU-Bench. Although training supervision is provided exclusively for Implication Understanding, the model achieves substantial gains not only on this task (+6.75% in Acc_{full}) but also on Aesthetic Appreciation (+5.43%) and Affective Reasoning (+6.34%), where no direct training signals are given. This cross-task transfer strongly indicates that the model has learned to establish generalizable semantic connections, linking perceptual evidence to abstract connotations

Table 4. **Training data composition comparison.** $+\Delta$ denotes gain over L3-only.

Data	#QA	Acc_{perc}	Acc_{bridge}	Acc_{conn}	Acc_{full}
L3-only	10k	89.25	85.75	59.50	47.00
Full-hierarchy	10k	89.75 $+0.50$	86.25 $+0.50$	60.75 $+1.25$	48.25 $+1.25$
Full-data	30k	93.50 $+4.25$	89.25 $+3.50$	63.75 $+4.25$	50.00 $+3.00$

through intermediate factual reasoning, rather than memorizing task-specific templates. The overall HVCU-Bench score demonstrates an improvement of $+6.17\%$, further confirming that bottom-up data generation with validation effectively teaches structured visual reasoning. Critically, our data generation pipeline operates independently of HVCU-Bench’s specific task formulations and answer patterns, relying solely on generic hierarchical reasoning prompts. This ensures that the observed gains reflect genuine improvements in hierarchical inference rather than dataset-specific overfitting (detailed results in Appendix D.1).

To assess whether these hierarchical reasoning improvements generalize beyond HVCU-Bench, we evaluate Qwen3-VL-4B-Bridge on four established general benchmarks: MMBench, HallusionBench, MMStar, and MMMU. As shown in Figure 4 (right), Qwen3-VL-4B-Bridge demonstrates strong generalization, achieving substantial improvements on MMStar ($+7.26\%$) and MMMU ($+3.22\%$) while maintaining competitive performance on MMBench and HallusionBench, even though it was trained exclusively on HVCU-Bench’s connotation-focused data (see Appendix D.1 for complete results). This cross-benchmark transfer demonstrates that grounding abstract interpretations in perceptual evidence through a structured semantic bridge enhances reasoning skills, benefiting tasks beyond connotative understanding.

Impact of Hierarchical Training Data. We compare three training setups in Table 4 to examine how hierarchical supervision affects model learning. **Full-data** uses complete three-level supervision across L_{perc} , L_{bridge} , and L_{conn} , yielding approximately 30k QA pairs as described in appendix C. **L3-only** trains exclusively on L_{conn} pairs, producing 10k examples that span all data entries in Full-data but only at the top reasoning level. **Full-hierarchy** samples a balanced 10k subset from Full-data by selecting one-third of its complete hierarchical entries, where each retained entry includes all three levels for that specific reasoning chain. This ensures Full-hierarchy matches L3-only’s count while preserving multi-level structure, but exposes the model to fewer reasoning chains than L3-only.

Results demonstrate the effectiveness of hierarchical supervision. Full-hierarchy outperforms L3-only by $+1.25\%$ despite training on fewer complete entries. This confirms that L_{perc} and L_{bridge} supervision provide critical learning signals that explicitly teach the model how to ground abstract connotations in perceptual evidence through struc-

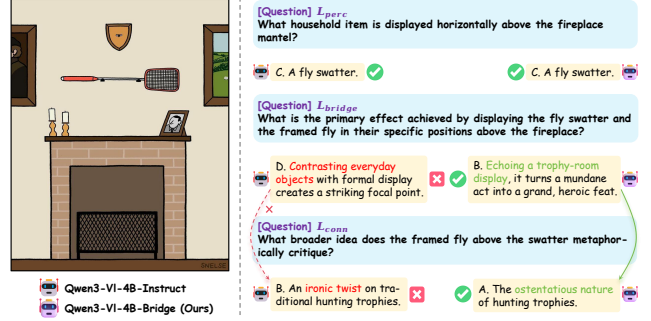


Figure 5. **Representative case study.** Training on hierarchical data corrects semantic bridging failures.

tured intermediate steps. Full-data further improves performance by $+1.75\%$, showing that scaling hierarchical supervision amplifies these benefits.

4.4. Case Study

Figure 5 demonstrates how training on hierarchical data corrects semantic bridging failures. In this fly swatter display example, the base model Qwen3-VL-4B-Instruct correctly identifies the object at L_{perc} but fails at L_{bridge} , misinterpreting the display as visual contrast rather than recognizing it as a trophy-room parody, which propagates to incorrect interpretation at L_{conn} . In contrast, Qwen3-VL-4B-Bridge successfully constructs the complete reasoning chain by establishing the correct semantic bridge, demonstrating that our hierarchical data explicitly teaches models to ground abstract connotations through structured intermediate reasoning. This directly addresses the L_{bridge} bottleneck, where the improvement reflects genuine learning of systematic connections between perceptual evidence and abstract meanings rather than task-specific memorization.

5. Conclusion

This paper addressed the challenge of bridging visual perception and abstract interpretation in MLLMs. We introduced VCU-Bridge, a hierarchical framework progressing from perception through semantic bridging to connotation, explicitly modeling the inferential process that connects concrete visual evidence to abstract meanings. Through HVCU-Bench, we revealed that all evaluated models exhibit universal performance degradation across levels, with semantic bridging emerging as the critical bottleneck, while hierarchical context experiments demonstrated that lower-level perceptual foundations provide crucial grounding for higher-level interpretation. To address this gap, our MCTS-driven data generation pipeline produces hierarchical data for instruction tuning, yielding substantial improvements on both HVCU-Bench and general benchmarks, demonstrating that explicit hierarchical supervision is essential for teaching models human-like visual understanding.

Acknowledgments

This work was supported in part by the New Generation Artificial Intelligence-National Science and Technology Major Project (2025ZD0123003), the National Key R&D Program of China (2024YFA1014003), the National Natural Science Foundation of China Enterprise Innovation and Development Joint Fund (Artificial Intelligence Field) Key Support Projects (U25B2072), the National Natural Science Foundation of China (92470121, 62402016), the Major Key Project of PCL (Grant No. PCL2025A17), Zhongguancun Academy (C20250204, C20250602), the Beijing Major Science and Technology Project (Z251100008125043, Z251100008425023), and the High-performance Computing Platform of Peking University.

References

- [1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 6
- [2] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024. 6
- [3] David Acuna, Ximing Lu, Jaehun Jung, Hyunwoo Kim, Amal Kar, Sanja Fidler, and Yejin Choi. Socratic-MCTS: Test-time visual reasoning by asking the right questions. In *EMNLP*, pages 24147–24160, 2025. 1
- [4] Ruichuan An, Sihan Yang, Ming Lu, Kai Zeng, Yulin Luo, Ying Chen, Jiajun Cao, Hao Liang, Qi She, Shanghang Zhang, and Wentao Zhang. MC-LLaVA: Multi-concept personalized vision-language model. *arXiv preprint arXiv:2411.11706*, 2024. 3
- [5] Ruichuan An, Kai Zeng, Ming Lu, Sihan Yang, Renrui Zhang, Huitong Ji, Qizhe Zhang, Yulin Luo, Hao Liang, and Wentao Zhang. Concept-as-tree: Synthetic data is all you need for VLM personalization. In *Synthetic Data for Computer Vision Workshop @ CVPR 2025*, 2025. 3
- [6] Ruichuan An, Sihan Yang, Ziyu Guo, Wei Dai, Zijun Shen, Haodong Li, Renrui Zhang, Xinyu Wei, Guopeng Li, Wenshan Wu, and Wentao Zhang. GENIUS: Generative fluid intelligence evaluation suite. *arXiv preprint arXiv:2602.11144*, 2026. 3
- [7] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *ICCV*, pages 2425–2433, 2015. 3
- [8] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 6
- [9] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhao-hai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6
- [10] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. ShareGPT4V: Improving large multi-modal models with better captions. In *ECCV*, pages 370–387, 2024. 1
- [11] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *NeurIPS*, 2024. 3, 6
- [12] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023. 1
- [13] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xianwu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A comprehensive evaluation benchmark for multimodal large language models. In *NeurIPS*, 2025. 3
- [14] Lancheng Gao, Ziheng Jia, Yunhao Zeng, Wei Sun, Yiming Zhang, Wei Zhou, Guangtao Zhai, and Xiongkuo Min. EEmo-Bench: A benchmark for multi-modal large language models on image evoked emotion assessment. In *ACM MM*, pages 7064–7073, 2025. 3, 4
- [15] Wentao Ge, Shunian Chen, Hardy Chen, Nuo Chen, Junying Chen, Zhihong Chen, Wenya Xie, Shuo Yan, Chenghao Zhu, Ziyue Lin, Dingjie Song, Xidong Wang, Anningzhe Gao, Zhiyi Zhang, Jianquan Li, Xiang Wan, and Benyou Wang. MLLM-Bench: Evaluating multimodal LLMs with per-sample criteria. In *NAACL*, pages 4951–4974, 2025. 1
- [16] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hal-lusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*, pages 14375–14385, 2024. 3, 6
- [17] Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 6
- [18] Zhuohang Jiang, Pangjing Wu, Ziran Liang, Peter Q. Chen, Xu Yuan, Ye Jia, Jiancheng Tu, Chen Li, Peter H. F. Ng, and Qing Li. HiBench: Benchmarking LLMs capability on

- hierarchical structure reasoning. In *KDD*, pages 5505–5515, 2025. 3
- [19] Guanzhen Li, Yuxi Xie, and Min-Yen Kan. MVP-Bench: Can large vision-language models conduct multi-level visual perception like humans? In *Findings of EMNLP*, pages 13505–13527, 2024. 3
- [20] Shuangtao Li, Shuaihao Dong, Kexin Luan, Xinhan Di, and Chaofan Ding. Enhancing reasoning through process supervision with monte carlo tree search. *arXiv preprint arXiv:2501.01478*, 2025. 3
- [21] Weifeng Lin, Xinyu Wei, Ruichuan An, Peng Gao, Bocheng Zou, Yulin Luo, Siyuan Huang, Shanghang Zhang, and Hongsheng Li. Draw-and-understand: Leveraging visual prompts to enable MLLMs to comprehend what you want. In *ICLR*, 2025. 3
- [22] Weifeng Lin, Xinyu Wei, Ruichuan An, Tianhe Ren, Tingwei Chen, Renrui Zhang, Ziyu Guo, Wentao Zhang, Lei Zhang, and Hongsheng Li. Perceive anything: Recognize, explain, caption, and segment anything in images and videos. In *NeurIPS*, 2025. 3
- [23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, pages 34892–34916, 2023. 6, 1
- [24] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pages 26286–26296, 2024.
- [25] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, 2024. 6
- [26] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *ECCV*, pages 216–233, 2024. 3, 6
- [27] Ziqiang Liu, Feiteng Fang, Xi Feng, Xinrun Du, Chenhao Zhang, Zekun Wang, Yuelin Bai, Qixuan Zhao, Liyang Fan, Chengguang Gan, Hongquan Lin, Jiaming Li, Yuansheng Ni, Haihong Wu, Yaswanth Narsupalli, Zhigang Zheng, Chengming Li, Xiping Hu, Ruifeng Xu, Xiaojun Chen, Min Yang, Jiaheng Liu, Ruibo Liu, Wenhao Huang, Ge Zhang, and Shiwen Ni. II-Bench: An image implication understanding benchmark for multimodal large language models. In *NeurIPS*, 2024. 3, 4
- [28] Yulin Luo, Ruichuan An, Bocheng Zou, Yiming Tang, Jiaming Liu, and Shanghang Zhang. LLM as dataset analyst: Subpopulation structure discovery with large language model. In *ECCV*, pages 235–252, 2024. 3
- [29] Chenhui Qiang, Zhaoyang Wei, Xumeng Han, Zipeng Wang, Siyao Li, Xiangyuan Lan, Jianbin Jiao, and Zhenjun Han. VER-Bench: Evaluating MLLMs on reasoning with fine-grained visual evidence. In *ACM MM*, pages 12698–12705, 2025. 1
- [30] Frank Serafini. Reading multimodal texts: Perceptual, structural and ideological perspectives. *Children’s Literature in Education*, 41(2):85–104, 2010. 1, 2
- [31] Tomoyuki Suzuki, Kang-Jun Liu, Naoto Inoue, and Kota Yamaguchi. Layerd: Decomposing raster graphic designs into layers. In *ICCV*, pages 17783–17792, 2025. 4
- [32] Yuwen Tan, Yuan Qing, and Boqing Gong. Vision llms are bad at hierarchical visual understanding, and llms are the bottleneck. *arXiv preprint arXiv:2505.24840*, 2025. 3
- [33] Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, An-Lan Wang, Chunhui Lin, Hao Feng, Zhen Zhao, Yanjie Wang, Yuliang Liu, Hao Liu, Xiang Bai, and Can Huang. MTVQA: Benchmarking multilingual text-centric visual question answering. In *Findings of ACL*, pages 7748–7763, 2025. 1
- [34] Zinan Tang and Qiyao Sun. Big escape benchmark: Evaluating human-like reasoning in language models via real-world escape room challenges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM)*, pages 488–503, 2025. 1
- [35] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. 6
- [36] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 6
- [37] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingting Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhao Wang, and Gen Luo. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 6
- [38] Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Xuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Fuxiao Liu, Gedas Bertasius, Mohit Bansal, Huaxiu Yao, and Furong Huang. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. In *ACL*, pages 416–442, 2024. 1
- [39] Wolfgang Wildgen. The cognitive roots of multimodal symbolic forms with an analysis of multimodality in movies. *Frontiers in Communication*, 9:1352252, 2024. 2

- [40] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. DeepSeek-VL2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. 6
- [41] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 6
- [42] Qi Yang, Chenghao Zhang, Lubin Fan, Kun Ding, Jieping Ye, and Shiming Xiang. Re-ranking reasoning context with tree search makes large vision-language models stronger. In *ICML*, 2025. 3
- [43] Ruilin Yao, Bo Zhang, Jirui Huang, Xinwei Long, Yifang Zhang, Tianyu Zou, Shili Xiong, Yi Rong, Yufei Wu, Shichao Su, Yifan Xu, Wenxi Zeng, Zhaoyu Yang, Guoyou Li, Shilan Zhang, Zichan Li, Yaxiong Chen, Shengwu Xiong, Peng Xu, Jiajun Zhang, Bowen Zhou, David A. Clifton, and Luc Van Gool. LENS: Multi-level evaluation of multimodal reasoning with large language models. In *ICLR*, 2026. 3
- [44] Xiaofei Yin, Yijie Hong, Ya Guo, Yi Tu, Weiqiang Wang, Gongshen Liu, and Huijia Zhu. InsightVision: A comprehensive, multi-level chinese-based benchmark for evaluating implicit visual semantics in large vision language models. *arXiv preprint arXiv:2502.15812*, 2025. 3
- [45] Jiakang Yuan, Tianshuo Peng, Yilei Jiang, Yiting Lu, Renrui Zhang, Kaituo Feng, Chaoyou Fu, Tao Chen, Lei Bai, Bo Zhang, and Xiangyu Yue. MME-Reasoning: A comprehensive benchmark for logical reasoning in MLLMs. *arXiv preprint arXiv:2505.21327*, 2025. 1, 3
- [46] Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhua Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *CVPR*, pages 9556–9567, 2024. 3, 6
- [47] Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. ReST-MCTS*: LLM self-training via process reward guided tree search. In *NeurIPS*, pages 64735–64772, 2024. 1
- [48] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of NAACL*, pages 881–916, 2025. 3
- [49] Zhenhao Zhang, Ye Shi, Lingxiao Yang, Suting Ni, Qi Ye, and Jingya Wang. OpenHOI: Open-world hand-object interaction synthesis with multimodal large language model. In *NeurIPS*, 2025. 3
- [50] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyang Luo. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *ACL Demos*, pages 400–410, 2024. 2
- [51] Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Hao-han Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning, acting, and planning in language models. In *ICML*, pages 62138–62160, 2024. 3
- [52] King Zhu, Qianbo Zang, Shian Jia, Siwei Wu, Feiteng Fang, Yizhi Li, Shuyue Guo, Tianyu Zheng, Jiawei Guo, Bo Li, Haoning Wu, Xingwei Qu, Jian Yang, Ruibo Liu, Xiang Yue, Jiaheng Liu, Chenghua Lin, Hamid Alinejad-Rokny, Min Yang, Shiwen Ni, Wenhao Huang, and Ge Zhang. LIME: Less is more for MLLM evaluation. In *Findings of ACL*, pages 9086–9121, 2025. 1