Beyond Counting: Evaluating Abstract and Emotional Reasoning in Vision-Language Models

Yuan Zhou[1,2], Yan Zhang[3], Jianlong Chang[4], Xin Gu[5], Ying Wang[1], Kun Ding[1*], Guangwen Yang[3†], Shiming Xiang[1, 2]

[1] MAIS, Institute of Automation, Chinese Academy of Sciences
[2] School of Artificial Intelligence, UCAS
[3] Department of Computer Science and Technology, Tsinghua University
[4] Huawei
[5] Research and Development Department of China Academy of Launch Vehicle Technology

Abstract

Despite the rapid progress of Vision-Language Models (VLMs), existing benchmarks still concentrate on coarse-grained object recognition or simple relational reasoning, leaving the fine-grained and higher-order reasoning abilities of these systems largely unexamined. To bridge this critical evaluation gap, we introduce EmojiGrid, a novel diagnostic benchmark specifically designed to probe these fine-grained and higher-order skills. Leveraging the universal and semantically rich nature of emojis, we synthesize a grid-based visual dataset paired with 29,000+ QA pairs. Each pair is explicitly anchored in a three-level cognitive taxonomy comprising (i) *Perception and Information Extraction*, (ii) *Relational and Structural Reasoning*, and (iii) *Abstraction and Advanced Cognition*. These dimensions further decompose into nine categories covering a broad range of cognitive skills, including counting, spatial relations, compositional logic, semantic sentiment, and related higher-order reasoning tasks. Our extensive evaluation of 25 leading and proprietary VLMs reveals a significant performance gap between foundational perceptual tasks and higher-level cognitive abilities, particularly in abstraction and advanced emotional reasoning. Notably, all models struggle with compositional logic, spatial consistency, and especially emotional and semantic understanding. EmojiGrid provides a quantifiable, fine-grained benchmark to diagnose VLM limitations and guides future progress toward models that can truly perceive, reason about, and interpret complex, symbol-rich visual scenes.

Code — https://github.com/yz413/EmojiGrid

1 Introduction

Vision-Language Models (VLMs) such as GPT-4V (OpenAI et al. 2024), Gemini (Reid et al. 2024; Team et al. 2025a), and LLaVA (Liu et al. 2023) represent a major milestone in artificial intelligence, exhibiting impressive proficiency in describing images, answering questions about visual content, and engaging in multi-modal conversations. Their success is largely built upon and measured by benchmarks like VQA v2 (Goyal et al. 2019) and GQA (Hudson and Manning 2019), which have driven progress in general-purpose visual-linguistic alignment. While these models excel at identifying common objects and their basic attributes in natural images, this very success often masks deeper, more fundamental limitations. We argue that the current evaluation paradigm falls short in two critical areas: fine-grained perception and higher-order cognitive reasoning.

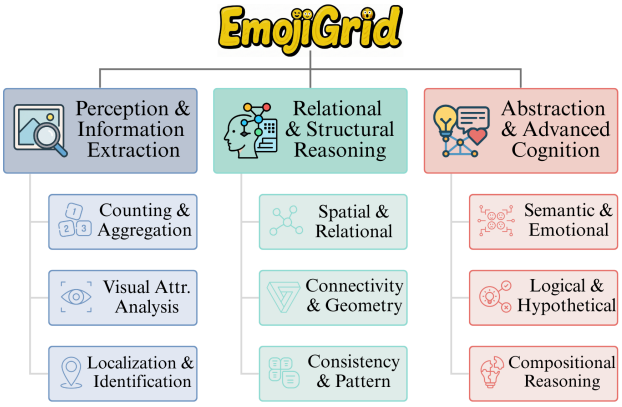

Figure 1: Hierarchical cognitive taxonomy underlying EmojiGrid, with three top-level dimensions further decomposed into nine sub-tasks that guide all question–answer generation and annotation.

First, current VLMs often struggle when faced with scenes containing small, dense, and numerous objects. Standard benchmarks, with their focus on salient, well-separated objects, do not adequately stress-test this aspect of visual acuity. This "can't see the trees for the forest" problem is a significant bottleneck for applications requiring meticulous detail, such as document analysis, medical imaging, or industrial inspection(Liu et al. 2024; Wu et al. 2025; Yu et al. 2025). Second, the reasoning required by most VQA datasets is often shallow, boiling down to object presence, counting, or simple spatial relations (Fu et al. 2024; Yue et al. 2024; Li et al. 2024). They seldom probe a model's capacity for multi-step logical inference, pattern recognition,

*Corresponding author. *kun.ding@ia.ac.cn*
†Corresponding author. *ygw@tsinghua.edu.cn*

or understanding compositional rules—skills that are hallmarks of human intelligence (Li et al. 2024; Fu et al. 2025). Beyond perception and logic, a crucial dimension of intelligence remains almost entirely absent from VLM evaluation: the understanding of abstract semantics and emotion (Ma et al. 2023; Hsieh et al. 2023). Human cognition is not just about identifying what is in an image, but interpreting what it means. We effortlessly infer intent, mood, and narrative from symbolic representations, a skill essential for nuanced human-computer interaction. *Can a VLM that recognizes a "crying face" emoji also infer that the overall sentiment of a scene is "sadness"? Can it distinguish a scene depicting "celebration" from one depicting "commiseration"?* This leap from recognition to interpretation is a grand challenge that current evaluation are ill-equipped to address.

To systematically investigate these shortcomings, we propose EmojiGrid, a new diagnostic benchmark designed to perform a "cognitive stress test" on modern VLMs. We choose emojis as our visual primitive for three key reasons: (1) they are inherently small objects with rich semantic and emotional content; (2) their discrete and symbolic nature allows for the procedural generation of complex, unambiguous scenes; and (3) their grid-based arrangement provides a precise coordinate system for evaluating spatial and structural reasoning without the ambiguity of natural scenes.

The EmojiGrid benchmark, illustrated in Fig. 1, is structured around a hierarchical taxonomy of nine distinct capabilities, from foundational *Perception and Information Extraction* (e.g., localization, counting) to intermediate *Relational and Structural Reasoning* (e.g., spatial relations, pathfinding) and culminating in advanced *Abstraction and Advanced Cognition* (e.g., semantic association, overall emotional assessment). Rather than being limited to a single accuracy score, this decomposable structure provides insights into the model's cognitive strengths and weaknesses through a detailed diagnostic profile. Our main contributions are as follows:

- We introduce EmojiGrid, a novel benchmark for VLMs that unifies fine-grained perception, complex spatial reasoning, and abstract semantic and emotional understanding within a structured evaluation framework spanning nine distinct cognitive dimensions, enabling comprehensive diagnosis of model capabilities beyond traditional accuracy-based metrics.
- Through extensive experiments, we empirically demonstrate significant performance deficits of leading VLMs in higher-order reasoning tasks, highlighting a previously overlooked but critical gap in semantic and emotional comprehension.
- Over 29k QA pairs and 500 synthesized images will be made publicly available, hoping to support ongoing advancements in the VLM research community.

2 Related Work

2.1 Vision-Language Models

Vision-language models are designed to jointly interpret and reason over visual and textual modalities, constituting a major advance toward general multimodal intelligence. Early VLMs such as Flamingo (Alayrac et al. 2022) and BLIP-2 (Li et al. 2023) pioneered the integration of large visual encoders with language decoders, demonstrating the feasibility of vision-language pre-training. A significant milestone was set by LLaVA (Liu et al. 2023), which introduced instruction tuning using LLM-synthesized, instruction-following chat data in the visual question answering (VQA) format, further enhancing the interactive and generalization capabilities of VLMs. Following these foundational efforts, a new wave of VLMs—including Qwen-VL (Bai et al. 2023), InternVL (Chen et al. 2024), and others—has expanded the functional spectrum of multimodal LLMs. These models not only support traditional VQA tasks, but also enable multimodal dialogue systems (Xiao et al. 2023), visually-grounded document understanding (Li et al. 2025; Feng et al. 2025), and cross-modal generation tasks (Wu, Zheng et al. 2025; Jin Xu et al. 2025). However, despite these remarkable advancements, current VLMs still struggle to move beyond surface-level associations toward deeper compositional reasoning and emotional cognition.

2.2 VLM Evaluation Benchmarks

The rapid advancement of VLMs has spurred the creation of a diverse array of benchmarks to systematically assess their capabilities across multiple dimensions. Traditional evaluation tasks such as image captioning (Onoe et al. 2024; Masry et al. 2022) and VQA (Antol et al. 2015; Mathew, Karatzas, and Jawahar 2021) remain central, gauging a model's ability to generate natural language descriptions and answer image-based queries by integrating visual perception, linguistic understanding, and external knowledge. In recent years, a suite of comprehensive benchmarks—including MME (Fu et al. 2024), SEED-Bench (Li et al. 2024), MMBench (Liu et al. 2024), and MMMU (Yue et al. 2024)—has emerged to evaluate a broader spectrum of VLM abilities, ranging from instruction following and basic perception to multilingual comprehension and expert-level reasoning. While these benchmarks have revealed issues such as object hallucination, limited spatial understanding, and challenges in text recognition, their primary focus still remains on general visual or semantic understanding.

Although some benchmarks, such as BLINK (Fu et al. 2025) and MMVP (Tong et al. 2024), include question–answer pairs related to compositional reasoning, these are typically mixed with other task types, making it challenging to precisely evaluate a model's compositional capabilities. Moreover, several specialized benchmarks—including Capture (Pothiraj et al. 2025), VGRP (Ren et al. 2025), and LEGO-Puzzles (Tang et al. 2025)—have been developed to probe core aspects of human cognition such as spatial relations, occluded object counting, and multi-step reasoning. However, results consistently show that state-of-the-art VLMs struggle with perceptual complexity, compositional logic, and nuanced spatial reasoning. Furthermore, many of these benchmarks rely on simplified or artificial scenarios, limiting their ability to reflect the complexity of real-world environments and hindering comprehensive assessment of human-level spatial cognition. In contrast, EmojiGrid offers a cognitively-

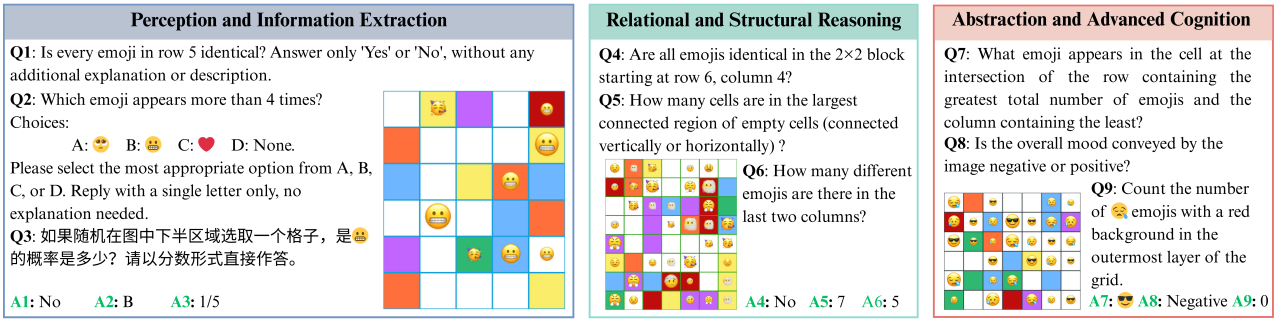

Figure 2: Task examples of EmojiGrid.

informed benchmark that systematically assesses VLMs' abilities in perceiving, reasoning about, and interpreting complex, symbol-rich scenes, with a focus on spatial reasoning, small-object recognition, and semantic, emotional, and abstract understanding.

3 The EmojiGrid Benchmark

3.1 Task Definition

Formally, the primary task of EmojiGrid is visual question answering. Given an emoji-based image I and an associated textual question Q, a model M is tasked with generating an answer A that semantically matches the ground-truth answer. However, unlike traditional VQA benchmarks, the fundamental goal of EmojiGrid is not merely to measure a single accuracy score, but to diagnose a model's performance across a spectrum of pre-defined cognitive skills. Each question-answer pair is meticulously designed to probe a specific cognitive dimension. At the core of our benchmark lies a hierarchical cognitive taxonomy (Fig.1), which decomposes visual understanding into three major dimensions, each encompassing three specific subskills. This structure enables a fine-grained diagnostic evaluation of vision-language models.

Perception and Information Extraction. This dimension assesses a model's ability to perceive, localize, and extract basic visual elements from a structured grid. It focuses on low-level visual parsing, symbol recognition, and quantitative understanding. ***Task 1****: Localization & Identification.* This task evaluates whether the model can accurately locate and identify individual emoji symbols in a spatially organized layout, including disambiguation of visually similar symbols. ***Task 2****: Visual Attribute Analysis.* Recognition of symbolic attributes—such as facial expressions, color hues, and accessories—is examined, reflecting sensitivity to subtle visual cues in abstract emoji designs. ***Task 3****: Counting & Aggregation.* Enumeration and aggregation of objects under specified conditions are assessed, requiring the integration of visual perception with basic numerical reasoning.

Relational and Structural Reasoning. This dimension examines a model's ability to reason about spatial relationships and structural organization among visual elements, extending beyond recognition to the inference of implicit topologies and patterns. ***Task 4****: Spatial & Relational Reasoning.* Reasoning over spatial proximity, directional relations (e.g., above or beside), and containment is required, involving implicit simulation of relative layouts and spatial grammar. ***Task 5****: Connectivity & Geometry.* Identification of connected components, aligned formations, and geometric configurations—such as rows, clusters, or loops—is assessed, reflecting internal representations of spatial regularity and relational structure. ***Task 6****: Consistency & Pattern.* Detection of repetitive motifs, symmetries, and analogical patterns is emphasized, requiring inference of governing rules (e.g., alternation or mirroring) and their application to novel instances. By isolating relational and structural reasoning, this dimension evaluates whether models can progress from local observations to a coherent global understanding of layout and structure.

Abstraction and Advanced Cognition. This dimension targets high-level cognitive capabilities, requiring models to interpret semantics, reason hypothetically, and perform multi-step inference—abilities approaching advanced visual-language understanding. ***Task 7****: Semantic & Emotional Understanding.* Comprehension of emotional states and symbolic meanings conveyed by emoji compositions is evaluated, including inference of mood, intent, and social context. ***Task 8****: Logical & Hypothetical Reasoning.* Conditional, counterfactual, and logical inferences grounded in visual input are assessed, such as reasoning about outcomes under hypothetical modifications. ***Task 9****: Compositional Reasoning.* Integration of multiple visual cues to derive a unified conclusion is required, testing symbolic abstraction, multi-hop reasoning, and cross-dimensional attention.

To operationalize these cognitive abilities, a set of question templates is designed for each task. Fig. 2 presents representative examples spanning from basic identification to holistic semantic and emotional assessment. Each example consists of a clear visual input, a precise question, and a structured answer format—either multiple-choice or binary (Yes/No)—to ensure objectivity and reproducibility.

3.2 Dataset Curation

Our EmojiGrid benchmark is underpinned by a meticulously designed three-step pipeline, as delineated in Fig. 3. This pipeline pioneers a "scene-first, question-second" methodology. This approach prioritizes the programmatic generation

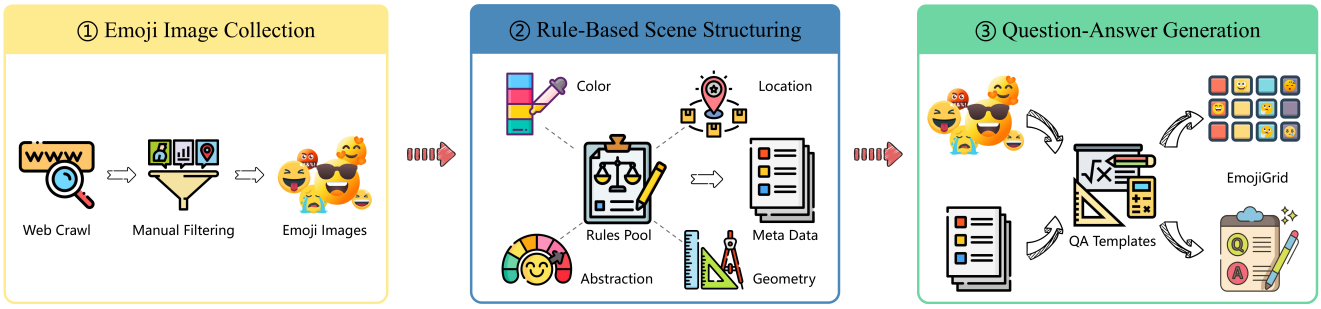

Figure 3: EmojiGrid benchmark construction pipeline.

of visually rich scenes with verifiable ground-truth, which subsequently serve as a foundation for the automated construction of questions. These questions are formulated based on both the intrinsic logic of the scene and our bespoke cognitive taxonomy. In the first step, we curate and filter publicly available emoji images with a primary focus on facial representations. The second step involves the creation of a comprehensive rule set designed to extract metadata information. In the final step, questions are generated by randomly combining rules from the pool, yielding unique answers while simultaneously synthesizing emoji grid images.

Step 1: Emoji Collection. Our benchmark begins with large-scale web crawling and filtering to identify emojis suitable for visual reasoning. We selectively retained 150 distinct human-face emojis, each with clearly recognizable emotional expressions, enabling definitive categorization into positive, negative, or neutral sentiment groups. To further enhance visual complexity and diversity, a small number of non-face emojis were also incorporated.

Step 2: Scene Structuring. Following the curation of the Emoji Lexicon, we developed a parameter-driven pipeline to systematically generate visually diverse scenes. Each scene comprises an $N \times N$ grid image (typically 4×4 to 8×8) populated with a randomly selected subset of emojis. Scene composition is rigorously governed by randomized yet controlled parameters drawn from predefined rule pools, including emoji density, variety, quantity, size, background coloration, and spatial arrangement strategies such as random distribution, clustering, and continuous path formation, enabling targeted exploration of distinct cognitive capabilities.

Concurrently with visual synthesis, we generate comprehensive structured metadata that serves as definitive ground-truth data. These metadata records comprehensively capture foundational attributes such as precise grid coordinates, emoji quantities, scene density, and so forth, while critically encoding higher-order relational information, including spatial adjacency patterns. This rich metadata architecture provides the logical foundation for subsequent programmatic question generation, enabling the construction of complex queries with unique, verifiable answers while facilitating systematic evaluation of multimodal reasoning capabilities across diverse spatial and semantic dimensions.

Step 3: Question-Answer Generation. QA pairs are generated through a template-based system, grounded in the comprehensive metadata constructed in the previous step. We developed an extensive rule pool containing over 60 distinct question templates, each explicitly designed to instantiate one of nine cognitive sub-categories. Each template functions as a logical transformation, taking the scene's rich ground-truth data (e.g., precise emoji types, quantities, positions, and intricate relational properties derived from the metadata) as input. This unique integration of structured metadata with the rule pool enables the batch generation of diverse questions for every scene image, ensuring comprehensive coverage across all cognitive categories. For instance, a counting template might randomly select an emoji type and query its total count, while a 'spatial relation' template probes the relative positioning of two distinct emojis. Critically, each generated question is syntactically correct, logically sound, and inherently produces a unique, verifiable answer, eliminating the need for manual review. This approach enables the automatic batch generation of high-quality, diverse question–answer pairs for each scene image, ensuring both randomness and diversity to support robust, fine-grained evaluation of cognitive capabilities.

3.3 Dataset Analysis

The pipeline described above yields **500** grid images and **29,339** QA pairs. Tab. 1 reports the core statistics.

Distribution and Balance. A key design principle of the EmojiGrid benchmark is its balanced question distribution, crafted to assess a wide array of reasoning skills. As shown in Fig. 4, questions are strategically allocated among three high-level dimensions: *Perception and Information Extraction* (40.43%), *Relational and Structural Reasoning* (32.12%), and *Abstraction and Advanced Cognition* (27.44%). This principle of balance extends to the nine subtasks within these dimensions. We deliberately avoided the dominance of any single category—the share ranges from 23.62% for *Counting and Aggregation* down to 3.41% for *Localization and Identification*. This hierarchical balance is crucial: it prevents models from inflating their scores by overfitting to a few dominant task types and instead compels a holistic evaluation of their reasoning abilities.

Token Length. Token length serves as a practical proxy for the linguistic and cognitive load a question imposes on a model. As illustrated in Fig. 5, the distribution of question lengths in EmojiGrid, for both English and Chinese, ex-

Type	Perception and Information Extraction			Relational and Structural Reasoning			Abstraction and Advanced Cognition		
Task	Loc. & Id.	V. Attr. Anal.	Count. & Agg.	Spa. & Rel.	Conn. & Geom.	Cons. & Patt.	Sem. & Emo.	Log. & Hypo.	Comp. Reas.
Count	1000	3933	6930	2802	2622	4000	3000	2870	2182

Table 1: Distribution of EmojiGrid questions by cognitive dimension and subcategory.

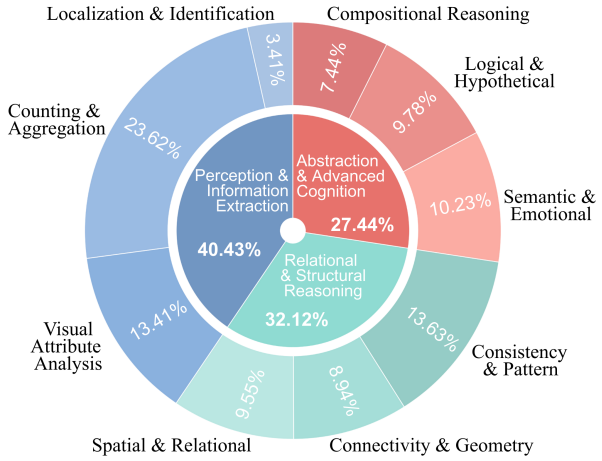

Figure 4: Proportional distribution of EmojiGrid questions across three cognitive dimensions and nine sub-categories.

hibits a pronounced long-tail pattern. While a majority of questions are concentrated in the 25–60 token range, ideal for clear and unambiguous queries, a significant fraction extends towards 125 tokens. This long tail is a deliberate design choice, creating a graduated difficulty that compels models to process not only concise commands but also more verbose, compositional, and multi-part queries. This structural diversity, combined with an average of 58.6 questions per image spanning all nine sub-tasks, ensures a comprehensive and robust evaluation of a VLM's ability to comprehend a wide spectrum of instructions.

Emoji Density. In grid-based scenes, emoji density directly affects visual clutter and thus task difficulty. We systematically control in EmojiGrid using emoji density—the percentage of grid cells occupied by an emoji—as a proxy for visual complexity. As illustrated in Fig. 5, the benchmark is deliberately designed to cover a wide and balanced spectrum of visual conditions. A small set of 17 images, constituting just 3.4% of the benchmark, exhibits low density (<30%), providing a baseline for simple scenes. In stark contrast, the remaining images are spread almost uniformly across densities from 30% to 100%, each successive 10% interval contains roughly 60–80 images. This design purposefully includes highly saturated grids, with some reaching 100% occupancy, to create scenarios of extreme visual congestion that stress a model's perceptual and reasoning capabilities. By spanning from sparse layouts to densely packed scenes in a balanced manner, this design ensures EmojiGrid serves as a rigorous stress test for evaluating a VLM's robustness against varying levels of visual complexity.

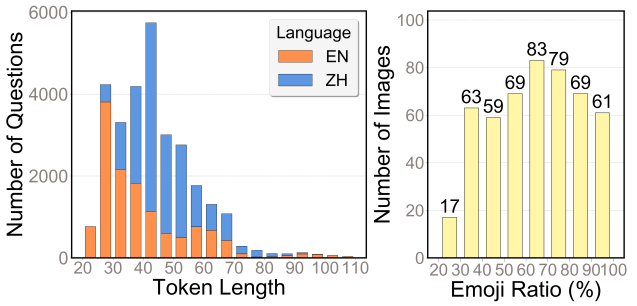

Figure 5: Linguistic and visual complexity in EmojiGrid. (Left) The distribution of question token lengths for English (EN) and Chinese (ZH). The deliberate long-tail design ensures models are evaluated on a spectrum of query complexities, from concise commands to verbose, multi-part questions. (Right) The distribution of images by emoji density (ratio). The near-uniform coverage of densities from 30% to 100% creates a rigorous testbed for VLM performance across varying levels of visual clutter.

4 Evaluation on EmojiGrid

4.1 Experimental Setting

Benchmark Models. We evaluate 25 VLMs on EmojiGrid, covering a diverse range of model scales and training paradigms. For open-source models, we evaluate DeepSeek-VL2 (Lu et al. 2024), Gemma3-[4B/12B/27B] (Team et al. 2025b), GLM-4.1V-9B-Thinking (Team et al. 2025d), Intern-VL3-[8B/14B/38B/78B] (Zhu et al. 2025), Kimi-VL-A3B-[Instruct/Thinking] (Team et al. 2025c), LLaVA-V1.6-Mistral-7B-HF (Liu et al. 2023), Llama-3.2-11B-Vision-Instruct (Meta AI 2024), Mistral-Small-3.1-24B-Instruct (Mistral AI 2025), Phi-4-Multimodal-Instruct (Microsoft et al. 2025), Qwen2.5-VL-[7B/32B/72B] (Bai et al. 2025), and QVQ-72B-Preview (Team 2024). For proprietary models, we evaluate Claude-4-Sonnet, Gemini-2.5-[Flash/Pro], GPT-4o, GPT-o4-mini, and Qwen-VL-Max. Notably, GLM-4.1V-9B-Thinking, Kimi-VL-A3B-Thinking, QVQ72B-Preview, Claude-4-Sonnet, Gemini-2.5-Flash, Gemini-2.5-Pro, and GPT-o4-mini are categorized as reasoning-enhanced models.

Evaluation Metric. Since each QA pair in EmojiGrid has only one correct answer, we adopt exact-match accuracy (%) as the primary evaluation metric for VLMs' performance on each task. Specifically, for numeric answers, multiple-choice questions (e.g., options A, B, C, D), and fact-checking tasks (yes/no), we apply a strict exact-match criterion for scoring.

Human Evaluation. To establish a robust benchmark for human-level performance, we constructed EmojiGrid-lite,

Model	Perception and Information Extraction				Relational and Structural Reasoning				Abstraction and Advanced Cognition				Overall
	Loc. & Id.	V. Attr. Anal.	Count. & Agg.	Overall	Spa. & Rel.	Conn. & Geom.	Cons. & Patt.	Overall	Sem. & Emo.	Log. & Hypo.	Comp. Reas.	Overall	
Human Baseline	98.30	94.00	95.75	96.10	90.45	88.72	95.60	92.84	94.90	90.02	91.70	93.20	94.05
Closed-Source Models													
Claude-Sonnet-4*	72.45	71.18	41.31	53.84	67.32	49.68	59.89	59.26	57.84	53.21	45.89	52.95	55.34
Gemini-2.5-Falsh*	**94.58**	**77.43**	**52.96**	**64.60**	**78.24**	**62.58**	69.84	**70.32**	**66.16**	**67.90**	**63.68**	**66.11**	**66.85**
Gemini-2.5-Pro*	**94.43**	**84.79**	**60.04**	**71.23**	**81.34**	**70.35**	**76.99**	**76.44**	**69.01**	**75.46**	**75.03**	**72.93**	**73.37**
GPT-4o	85.00	57.08	33.74	45.80	54.21	41.42	69.90	57.31	57.33	46.90	34.97	47.55	49.98
GPT-o4-mini*	87.18	68.19	42.82	54.97	66.91	52.63	59.72	59.89	62.54	58.06	44.50	56.06	56.85
Qwen-VL-Max	69.20	58.07	36.25	46.26	60.21	38.94	**71.60**	59.13	59.80	50.38	42.35	51.71	51.89
Open-Source Models													
DeepSeek-VL2	42.70	35.90	19.77	27.05	33.98	20.94	54.07	38.88	49.13	34.08	23.37	36.79	33.52
Gemma3-4B	45.20	36.44	19.49	27.28	38.54	24.52	65.25	45.98	40.03	33.48	10.31	29.64	33.93
Gemma3-12B	49.70	41.34	25.35	32.71	42.15	24.94	58.45	44.28	50.87	40.35	20.62	38.92	38.13
Gemma3-27B	57.10	52.91	26.93	38.08	49.50	31.46	52.98	45.96	49.60	42.47	19.29	38.85	40.82
GLM-4.1V-9B-Thinking*	68.81	**66.04**	**36.13**	**48.79**	**64.62**	**43.48**	53.89	**54.19**	53.45	**47.10**	**37.84**	**46.96**	**50.02**
Intern-VL3-8B	64.60	53.45	27.92	39.47	55.64	32.34	59.95	50.99	55.53	41.11	28.23	43.00	44.14
Intern-VL3-14B	67.50	59.80	29.03	42.47	57.57	30.32	63.38	52.45	56.43	44.84	33.09	45.98	46.64
Intern-VL3-38B	**74.60**	53.39	30.65	42.89	51.89	31.12	64.03	52.26	**57.30**	43.55	29.84	45.96	46.74
Intern-VL3-78B	71.70	55.45	31.13	42.61	55.25	32.38	**64.60**	52.85	56.90	45.12	35.66	46.94	47.09
Kimi-VL-A3B-Instruct	48.50	37.40	24.30	30.68	36.62	24.83	53.30	40.42	50.27	33.21	20.99	36.25	35.34
Kimi-VL-A3B-Thinking*	55.21	57.23	31.05	41.75	54.12	**40.41**	54.39	50.42	50.84	42.72	28.22	41.82	44.55
LLaVA-V1.6-Mistral-7B-HF	27.80	33.87	18.51	24.39	35.26	21.05	54.07	39.29	42.10	30.28	18.19	31.41	31.10
Llama-3.2-11B-Vision-Instruct	43.20	31.12	20.45	25.90	34.12	19.95	52.62	38.03	40.93	28.33	13.29	28.95	30.63
Mistral-Small-3.1-24B-Instruct	55.80	58.23	27.89	40.30	51.46	28.83	55.88	47.04	46.33	41.95	37.40	42.35	43.03
Phi-4-Multimodal-Instruct	44.00	42.26	25.31	32.50	47.25	23.23	48.98	41.30	45.50	35.40	22.59	35.69	36.20
Qwen-VL2.5-7B	44.50	51.56	27.90	37.15	53.10	22.27	44.73	40.97	46.47	36.90	28.00	38.05	38.62
Qwen-VL2.5-32B	48.60	50.65	30.04	38.44	42.54	27.08	62.92	46.89	53.13	39.09	25.76	40.71	41.78
Qwen-VL2.5-72B	**73.70**	**61.00**	**36.16**	**47.56**	**57.67**	36.61	**68.05**	**56.22**	**57.47**	**50.87**	**38.73**	**50.04**	**51.02**
QVQ-72B-Preview*	39.59	35.69	20.60	27.19	39.09	26.61	60.74	44.82	49.10	33.29	15.56	34.43	34.83

Table 2: Performance comparison of VLMs on EmojiGrid (%). Models marked with an asterisk (*) possess reasoning capabilities. Dark green and Light green indicate the top-ranked and second-ranked performance within each group, respectively.

a curated subset of our main dataset. This subset was formulated through stratified sampling of 50 images, yielding over 3,000 question-answer pairs that maintain comprehensive coverage all three dimensions and nine sub-tasks. Three human participants answered every question under the same constraints imposed on VLMs—no external tools or internet access. Their accuracy provides an empirical ceiling on task performance, especially for spatial and relational reasoning. Furthermore, EmojiGrid-lite serves as a specialized benchmark for evaluating the reasoning proficiency of advanced VLMs, particularly by measuring the average length of their generated thought-process tokens.

4.2 Main Results

Tab. 2 presents the performance of 25 VLMs across nine tasks spanning three cognitive dimensions on the EmojiGrid benchmark. Overall, larger models demonstrate superior performance, with closed-source models generally outperforming their open-source counterparts. Among closed-source models, reasoning-enhanced variants such as Gemini-2.5 series, GPT-o4-mini, and Claude-Sonnet-4 significantly outperform standard models like Qwen-VL-Max and GPT-4o. This performance gap clearly demonstrates the effectiveness of reasoning augmentation in enhancing model accuracy. Notably, GPT-4o ranks lowest among closed-source models with only 49.98%, further highlighting the importance of advanced reasoning capabilities.

Within the open-source model category, Qwen-VL2.5-72B exhibits remarkable reasoning capabilities, achieving the highest scores across five tasks. Meanwhile, GLM-4.1V-9B-Thinking demonstrates competitive overall performance despite having only 9B parameters, securing top rankings in three tasks. This suggests that architectural innovations and reasoning enhancements can compensate for parameter limitations. Across model families including Gemma3, Intern-VL3, and Qwen-VL2.5, larger parameter versions consistently deliver significant performance improvements over their smaller counterparts, underscoring the benefits of scale in vision-language modeling.

Domain-aspect Performance. A closer examination of Tab. 2 reveals a clear progression in task difficulty across the three core cognitive domains of the EmojiGrid benchmark. Both open-source and closed-source models attain substantially higher scores on the *Perception and Information Extraction* dimension, with accuracy declining consistently as tasks advance to the *Relational and Structural Reasoning* domain and, most notably, the *Abstraction and Advanced Cognition* domain. This marked performance drop-off pinpoints a critical and underexplored deficiency in contemporary VLMs: a limited capacity for handling abstract semantic, emotional, and compositional reasoning.

Empirical results show a systematic degradation in performance as task complexity increases, underscoring EmojiGrid's diagnostic strength in revealing critical weaknesses in current vision–language models. Unlike benchmarks that emphasize near-saturated perceptual tasks, EmojiGrid adopts a tiered cognitive design to probe advanced reasoning beyond simple recognition. This structure enables a more nuanced assessment of model understanding, particularly for abstract and complex visual cognition.

Language-aspect Performance. Our analysis isolates the impact of language on VLM performance, leveraging the

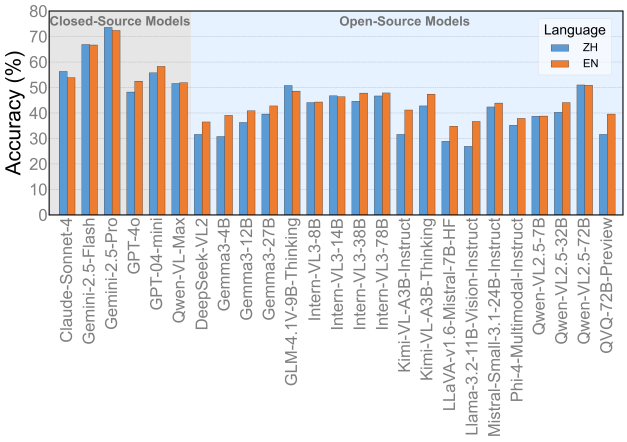

Figure 6: Cross-lingual performance on EmojiGrid.

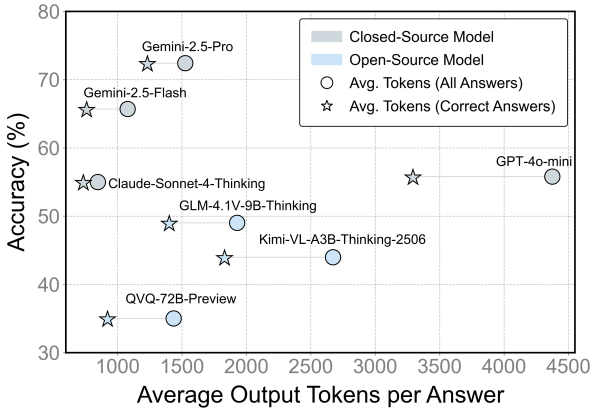

Figure 7: Cost–benefit trade-off between average reasoning length and accuracy.

EmojiGrid benchmark, which consists of questions in both English (54.9%) and Chinese (45.1%). As illustrated in Fig. 6, distinct performance patterns emerge between the two languages. Top-tier closed-source models, particularly the Gemini-2.5 series and Claude-Sonnet-4, demonstrate robust cross-lingual capabilities, achieving high accuracy in both languages. Interestingly, these leading models show exceptional proficiency in Chinese. However, the predominant trend observed across both closed-source and open-source models is a clear performance bias in favor of English. This suggests an underlying English-centric optimization in their development. A striking exception is GLM-4.1V-9B-Thinking, which clearly demonstrates superior performance in Chinese, likely reflecting its bilingual training foundation. In contrast, many open-source models, often trained on English-dominant corpora, exhibit a more significant degradation in accuracy on Chinese tasks, underscoring the critical role of training data in multilingual reasoning.

Reasoning Efficiency Analysis. Beyond final-answer accuracy, the computational cost of a model's reasoning process—measured by the average token length of its full output—represents a critical yet underappreciated dimension of performance, particularly for VLMs with reasoning capabilities. Fig. 7 presents the trade-off between accuracy and average output token count across a representative set of VLMs with explicit reasoning capabilities, distinguishing between the global average (◯ , all responses) and the token cost restricted to correct responses (☆ , correct answers only).

In Fig. 7, excessive verbosity does not indicate deeper reasoning but rather reflects model uncertainty or failure, since incorrect answers are often accompanied by longer, less productive outputs. For instance, models like GPT-4o mini expend substantial computational resources on unproductive reasoning traces without achieving higher accuracy, challenging the assumption that longer reasoning chains lead to better outcomes. In contrast, the Gemini-2.5 series combines high accuracy with efficient token usage, exemplifying the benefits of concise reasoning. Conversely, models such as QVQ-72B-preview, despite their brevity, exhibit unsatisfactory accuracy, highlighting the need to balance correctness and conciseness for optimal cost–performance trade-offs. These findings collectively underscore that reasoning efficiency is a vital, yet underexplored, axis in VLM evaluation. They challenge the implicit assumption that "more reasoning is better," revealing instead that verbosity often signals inefficiency or internal uncertainty.

Key Insights and Observations. We summarize several key findings as below.

- **Abstract cognition remains a frontier challenge.** Abstract semantic and affective reasoning is still difficult for current VLMs. Performance is often brittle and sensitive to prompt phrasing, suggesting limited robustness to nuanced concepts.
- **A persistent open–closed performance gap.** Leading proprietary VLMs consistently outperform open-source models, especially on tasks requiring complex, multi-step reasoning, underscoring advantages in scale and training sophistication.
- **Spatial reasoning scales with model capacity.** Spatial reasoning improves with model size. Reasoning-oriented architectures further strengthen performance, indicating the importance of both capacity and specialized design.
- **The inefficiency of error in reasoning.** Longer reasoning chains do not reliably improve accuracy. Incorrect predictions are often accompanied by more verbose outputs, implying that verbosity can reflect uncertainty or failure rather than deeper cognition.

5 Conclusion

Current VLM evaluations favor coarse recognition while masking deficits in fine-grained perception and higher-order reasoning. We introduce EmojiGrid, a benchmark with synthesized images and hierarchical QA spanning nine cognitive domains. Evaluations of 25 VLMs reveal substantial failures in compositional, spatial, semantic, and emotional reasoning despite strong recognition performance. EmojiGrid enables targeted diagnosis and progress toward cognitively robust vision–language models.

6 Acknowledgments

This document is the result of the research project funded by the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No. XDB0500103) and the National Natural Science Foundation of China (Grant No. 62306310). The authors gratefully acknowledge this support.

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual Question Answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv:2308.12966.

Bai, S.; Chen, K.; Liu, X.; Wang; et al. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24185–24198.

Feng, H.; Wei, S.; Fei, X.; Shi, W.; Han, Y.; Liao, L.; Lu, J.; Wu, B.; Liu, Q.; Lin, C.; Tang, J.; Liu, H.; and Huang, C. 2025. Dolphin: Document Image Parsing via Heterogeneous Anchor Prompting. In *Proceedings of the 65rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Yang, J.; Zheng, X.; Li, K.; Sun, X.; Wu, Y.; and Ji, R. 2024. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. arXiv:2306.13394.

Fu, X.; Hu, Y.; Li, B.; et al. 2025. BLINK: Multimodal Large Language Models Can See but Not Perceive. In *Computer Vision – ECCV 2024*, 148–166. Cham: Springer Nature Switzerland. ISBN 978-3-031-73337-6.

Goyal, Y.; Khot, T.; Agrawal, A.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2019. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. *Int. J. Comput. Vision*, 127(4): 398–414.

Hsieh, C.-Y.; Zhang, J.; Ma, Z.; Kembhavi, A.; and Krishna, R. 2023. SugarCrepe: Fixing Hackable Benchmarks for Vision-Language Compositionality. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Hudson, D. A.; and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6700–6709.

Jin Xu, J. H. H. H. T. H. S. B., Zhifang Guo; et al. 2025. Qwen2.5-Omni Technical Report. *arXiv preprint arXiv:2503.20215*.

Li, B.; Ge, Y.; Chen, Y.; Ge, Y.; Zhang, R.; and Shan, Y. 2024. SEED-Bench-2-Plus: Benchmarking Multimodal Large Language Models with Text-Rich Visual Comprehension. *arXiv preprint arXiv:2404.16790*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. 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, Z.; Liu, Y.; Liu, Q.; Ma, Z.; Zhang, Z.; Zhang, S.; Guo, Z.; Zhang, J.; Wang, X.; and Bai, X. 2025. MonkeyOCR: Document Parsing with a Structure-Recognition-Relation Triplet Paradigm. arXiv:2506.05218.

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

Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. 2024. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, 216–233. Springer.

Lu, H.; Liu, W.; Zhang, B.; Wang, B.; Dong, K.; Liu, B.; Sun, J.; Ren, T.; Li, Z.; Yang, H.; et al. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Ma, Z.; Hong, J.; Gul, M. O.; Gandhi, M.; Gao, I.; and Krishna, R. 2023. @ CREPE: Can Vision-Language Foundation Models Reason Compositionally? In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10910–10921.

Masry, A.; Long, D.; Tan, J. Q.; Joty, S.; and Hoque, E. 2022. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2263–2279. Dublin, Ireland: Association for Computational Linguistics.

Mathew, M.; Karatzas, D.; and Jawahar, C. V. 2021. DocVQA: A Dataset for VQA on Document Images. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Meta AI. 2024. Llama-3.2-11B-Vision-Instruct. https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct. Accessed: 2025-12-20.

Microsoft; :; Abouelenin, A.; Ashfaq, A.; Atkinson, A.; Awadalla, H.; et al. 2025. Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs. arXiv:2503.01743.

Mistral AI. 2025. Mistral-Small-3.1-24B-Instruct. https://mistral.ai/news/mistral-small-3-1. Accessed: 2025-12-20.

Onoe, Y.; Rane, S.; Berger, Z.; Bitton, Y.; Cho, J.; Garg, R.; Ku, A.; Parekh, Z.; Pont-Tuset, J.; Tanzer, G.; Wang, S.; and Baldridge, J. 2024. DOCCI: Descriptions of Connected and Contrasting Images. arXiv:2404.19753.

OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; et al. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Pothiraj, A.; Stengel-Eskin, E.; Cho, J.; and Bansal, M. 2025. CAPTURe: Evaluating Spatial Reasoning in Vision Language Models via Occluded Object Counting. arXiv:2504.15485.

Reid, M.; Savinov, N.; Teplyashin, D.; Lepikhin, D.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv*, abs/2403.05530.

Ren, Y.; Tertikas, K.; Maiti, S.; Han, J.; Zhang, T.; Süsstrunk, S.; and Kokkinos, F. 2025. VGRP-Bench: Visual Grid Reasoning Puzzle Benchmark for Large Vision-Language Models. arXiv:2503.23064.

Tang, K.; Gao, J.; Zeng, Y.; Duan, H.; Sun, Y.; Xing, Z.; Liu, W.; Lyu, K.; and Chen, K. 2025. LEGO-Puzzles: How Good Are MLLMs at Multi-Step Spatial Reasoning? *arXiv preprint arXiv:2503.19990*.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; et al. 2025a. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805.

Team, G.; Kamath, A.; Ferret, J.; et al. 2025b. Gemma 3 Technical Report. arXiv:2503.19786.

Team, K.; Du, A.; Yin, B.; Xing, B.; Qu, B.; et al. 2025c. Kimi-VL Technical Report. arXiv:2504.07491.

Team, Q. 2024. QVQ: To See the World with Wisdom.

Team, V.; Hong, W.; Yu, W.; et al. 2025d. GLM-4.1V-Thinking: Towards Versatile Multimodal Reasoning with Scalable Reinforcement Learning. arXiv:2507.01006.

Tong, S.; Liu, Z.; Zhai, Y.; Ma, Y.; LeCun, Y.; and Xie, S. 2024. Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs. arXiv:2401.06209.

Wu, C.; Zheng, P.; et al. 2025. OmniGen2: Exploration to Advanced Multimodal Generation. *arXiv preprint arXiv:2506.18871*.

Wu, X.; Ding, Y.; Li, B.; Lu, P.; Yin, D.; Chang, K.-W.; and Peng, N. 2025. Visco: Benchmarking fine-grained critique and correction towards self-improvement in visual reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 9527–9537.

Xiao, B.; Wu, H.; Xu, W.; Dai, X.; Hu, H.; Lu, Y.; Zeng, M.; Liu, C.; and Yuan, L. 2023. Florence-2: Advancing a unified representation for a variety of vision tasks. *arXiv preprint arXiv:2311.06242*.

Yu, H.-T.; Wei, X.-S.; Peng, Y.; and Belongie, S. 2025. Benchmarking Large Vision-Language Models on Fine-Grained Image Tasks: A Comprehensive Evaluation. arXiv:2504.14988.

Yue, X.; Ni, Y.; Zheng, T.; et al. 2024. MMMU: A Massive Multi-Discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9556–9567.

Zhu, J.; Wang, W.; Chen, Z.; et al. 2025. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. arXiv:2504.10479.