DiCoBench: Benchmarking Multi-Image Fine-Grained Perception via Differential and Commonality Visual Cues

Geng Li[1] and Yuxin Peng[1⋆]

Wangxuan Institute of Computer Technology, Peking University, Beijing, China
ligeng@stu.pku.edu.cn, pengyuxin@pku.edu.cn

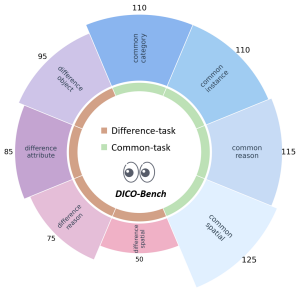
Abstract. Recent advancements in Multimodal Large Language Models (MLLMs) have demonstrated impressive fine-grained perception capabilities. However, existing benchmarks predominantly rely on explicit textual cues or low-resolution inputs, failing to evaluate a model's ability to autonomously perceive implicit visual cues in high-resolution. To bridge this gap, we introduce **DiCoBench**, a comprehensive, multi-image high-resolution benchmark designed for cross-image fine-grained perception. DiCoBench consists of 765 meticulously curated samples categorized into two progressive tracks: Differential Visual Cues and Commonality Visual Cues, covering 8 distinct perception tasks. By formulating the benchmark as a multiple-choice question task and utilizing high-resolution imagery (approaching 2K), we eliminate evaluation metric bias and pose a substantial challenge to current state-of-the-art MLLMs. Our extensive evaluation of 18 diverse MLLMs reveals a striking performance gap compared to human accuracy (98.3%), with top-performing models struggling significantly with micro-scale detail capture. We believe DiCoBench will serve as a challenging testbed to drive future research in autonomous, high-resolution multi-image perception. Dataset: https://huggingface.co/datasets/oking0197/DiCoBench

Keywords: Multimodal Large Language Models · Fine-Grained Perception · Implicit Visual Cues

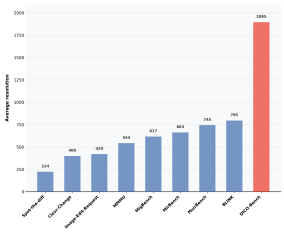
1 Introduction

Driven by the evolution of high-resolution visual encoders and massive vision-language alignment data, recent Multimodal Large Language Models (MLLMs) have achieved remarkable breakthroughs in single-image fine-grained perception [1,3,4,11,15,26,34,35]. As demonstrated by recent benchmarks like V* [31], HR-Bench [29], TreeBench [26] and Visual Probe [15], advanced MLLMs exhibit remarkable proficiency in localizing and recognizing minute details in single high-resolution image. However, these existing fine-grained evaluations fundamentally rely on explicit textual cues contained by question (*e.g.*, "Where is the *umbrella*?"

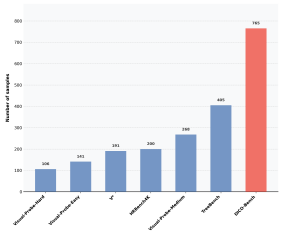
⋆ Corresponding author.



(a) Composition of our proposed DiCoBench.



(b) Comparison of image resolutions across multi-image multimodal benchmarks.



(c) Comparison of sample sizes across fine-grained perception benchmarks.

Fig. 1: Overview of our proposed DiCoBench. (a) DiCoBench covers 2 major perception categories and 8 specific perception tasks. (b) We observe that the average resolution of existing multi-modal benchmarks remains primarily in low-resolution scenarios. In contrast, our proposed DiCoBench reaches approaching 2K. (c) Due to high-resolution constraints, the largest existing single-image fine-grained perception benchmarks contain only around 400 samples. DiCoBench with a sample size and diversity nearly double that of the current largest benchmark.

or "Who the *person in red* is?"). Essentially, they mainly assess a model's passive text-to-image grounding capability. In contrast, real-world visual cognition is rarely instruction-driven; when navigating complex, dynamic environments, human perception proactively captures spontaneous *implicit visual cues*, such as subtle visual differences or commonalities across multiple observations. Current single-image benchmarks fail to measure this "autonomous visual-cue guided perception." Consequently, we find when deployed without explicit textual guidance and forced to discover extremely minute visual cues purely through cross-image comparison, even state-of-the-art MLLMs suffer from severe performance degradation.

To address the limitations of single-image evaluation, the community has increasingly focused on multi-image multimodal benchmarks. Comprehensive benchmarks like MuirBench [25], MIR Benchmark [7], BLINK [9], and MLLM-COMPBENCH [14] evaluate broad capabilities such as STEM knowledge, scene understanding and temporal ordering. Concurrently, domains like Image Difference Captioning (IDC) [6, 13, 19, 20, 22] require models to describe local changes between image pairs. Despite these advancements, existing multi-image multimodal benchmarks face three bottlenecks.

1. *Evaluation Metric Bias*: IDC tasks predominantly frame perception as text generation, evaluated by strict n-gram matching metrics like ROUGE-L or CIDEr. As demonstrated by G-VEval [24], these metrics severely penalize MLLMs for formatting differences rather than genuine perception errors, thereby suppressing and misrepresenting the true perceptual limits of advanced models like Qwen2-VL [27].

Fig. 2: Comparison between DiCoBench and previous benchmarks. Previous fine-grained perception benchmarks primarily rely on explicit textual prompts (highlighted in red) related to objects to guide perception. In contrast, the key distinction of DiCoBench is its emphasis on evaluating a model's ability to drive multi-image perception through implicit visual cues of commonalities and differences (highlighted in green boxes), which more closely resembles the human perceptual process in dynamic, real-world environments without text cues available.

2. *Lack of Systematic Taxonomy*: Existing multi-image benchmarks fail to consider implicit visual cues as a systematic evaluation dimension, resulting in fragmented and unsystematic assessments. For instance, tasks like IDC exclusively focus on capturing visual differences, completely neglecting the equally crucial cognitive dimension of visual commonalities.

3. *Insufficient Image Resolution*: Current multi-image multimodal datasets predominantly rely on low-resolution scenes like Figure 1. They severely lack the capacity to evaluate a model's perceptual limits under high-resolution conditions, making it extremely difficult to ascertain whether models can genuinely capture minute visual cues to achieve accurate fine-grained perception.

To bridge the gap mentioned above, we propose a novel and highly challenging task: *Vision-cue guided Cross-Image Fine-grained Perception*, accompanied by a new multi-image multimodal benchmark, DiCoBench (Difference-Common Bench). Innovatively, DiCoBench systematically categorizes the visual cues captured by humans without text guidance into two parallel tracks: *Differential Visual Cues* and *Commonality Visual Cues*. To comprehensively evaluate a model's ability to acquire these cues, we specifically design 4 perception tasks for each track. This yields a total of 8 distinct task categories, ranging from attributes and instances to spatial relationships and logical reasoning. To fundamentally eliminate the evaluation metric bias caused by n-gram text generation penalties,

we formulate DiCoBench as a Multiple-Choice Question (MCQ) task, universally appending a robust "No visible difference" or "No visible commons" option to complete the logical space.

To ensure the images are sufficiently high-definition to pose a substantial challenge to existing models, DiCoBench collects base images with average resolution of 1895, approaching 2K.

We systematically evaluate 18 MLLMs of diverse architectures and scales on the DiCoBench. Our results reveal a striking paradox: **while these tasks are intuitive and trivial for humans (98.3% average accuracy), they remain exceptionally challenging for current SOTA models**. Even the most capable closed-source model, Gemini-3-Pro, achieves only 58.1% average accuracy, trailing behind human performance by 40.2%, and representing only a modest advancement over the capabilities of current open-source alternatives. Furthermore, we observe significant performance volatility across different task types; for instance, while models demonstrate emerging proficiency in categorical tasks, their ability to conduct high-level logical reasoning during perception remains severely constrained, with scores often plummeting toward random-guessing levels. Our findings indicate that the cross-image fine-grained perceptual capabilities of current MLLMs have been significantly overestimated. We believe DiCoBench highlights a critical frontier in MLLM perception, serving as an effective testbed to bridge the profound gap in high-resolution, visual-cue guided, and multi-image perception.

2 Related Works

2.1 Text-Guided Fine-Grained Perception

Driven by the evolution of high-resolution visual encoders and dynamic spatial pooling strategies (*e.g.*, Qwen2-VL [27], InternVL-1.5 [5]), recent MLLMs have demonstrated remarkable proficiency in perception. Consequently, benchmarks such as V* [31], HR-Bench [29], TreeBench [26], and Visual Probe [15] have been proposed to further evaluate the localization and recognition of minute details in high-resolution images, also known as the fine-grained perception task. However, these benchmarks fundamentally rely on *explicit textual cues* (*e.g.*, "Where is the red cup in the image?"). They evaluate a model's passive text-to-image grounding capability rather than its proactive ability to mine and interpret spontaneous visual cues in the wild. We find that although existing MLLMs have already performed excellently in the aforementioned tasks (e.g., the best model on V* has reached 90% accuracy), when deployed in complex, dynamic environments without explicit textual guidance, existing MLLMs often suffer from severe performance degradation, highlighting a critical gap in autonomous high-resolution fine-grained perception.

2.2 Multi-Image Perception

To push MLLMs beyond single-image constraints, researchers have increasingly focused on multi-image perception and reasoning tasks. Comprehensive bench-

marks such as MMMU [32], BLINK [9], MuirBench [25], MLLM-CompBench [14], MileBench [21] and MIR Benchmark [7] have been introduced to evaluate broad capabilities, including STEM knowledge, scene understanding, multi-view reasoning and *etc*. Concurrently, specialized datasets have emerged to probe specific multi-image perception dimensions: MIG-Bench [18] explores free-form multi-image grounding, while MIHBench [16] specifically evaluates multi-image hallucinations regarding object existence and identity consistency. Despite these advancements, existing multi-image perception benchmarks face structural limitations. They rely on low-resolution images, which inevitably lose fine-grained visual details. More importantly, their task designs predominantly focus on macro-level logical relationships or the association of highly salient objects, inherently failing to assess the fine-grained ability to discover micro-level visual cues under high-resolution, complex background interference.

2.3 Image Difference Captioning

The most closely related domain to our work is "spot-the-difference" or Image Difference Captioning (IDC). Early foundational works introduced datasets like CLEVR-Change [20], Spot-the-Diff [13], and Image Editing Request (IER) [22] benchmarks to explore pixel-level modifications or synthetic visual changes. Building upon this, recent efforts have constructed various datasets to train and evaluate MLLMs on capturing realistic visual changes, including DiffTell [6] for image manipulations, M3-Verse [30] for 3D environments, OmniDiff [19], and OneDiff [12]. Other works [10, 17, 33] have focused on enhancing MLLMs' architectures or pre-training strategies for difference captioning. VisMin [2] further introduces visual minimal-change understanding. However, this paradigm suffers from two fatal flaws. First, evaluation metric bias. Existing works predominantly frame this as a text generation task evaluated by n-gram matching metrics like ROUGE-L or CIDEr. As demonstrated by G-VEval [24], these metrics are highly unsuitable for MLLMs, severely penalizing them for formatting differences rather than genuine perception errors. Second, limited scope of task types. These datasets strictly focus on visual differences, entirely ignoring the equally critical cognitive dimension of implicit commonalities (*e.g.*, entities that consistently appear across two different scenes). Furthermore, their "minimal changes" are mostly restricted to salient object attributes within relatively low-resolution contexts, falling entirely short of the extreme "micro" scale required for fine-grained perception under complex high-resolution backgrounds.

3 The Difference Commonality Bench

DiCoBench is designed to address the critical gap in evaluating "visual-cue guided cross-image fine-grained perception" by establishing the first corresponding comprehensive benchmark. Specifically, DiCoBench systematically evaluates MLLMs across two progressive tracks: *Differential Visual Cues* and *Commonality Visual Cues*. The benchmark comprises 765 meticulously constructed high-resolution

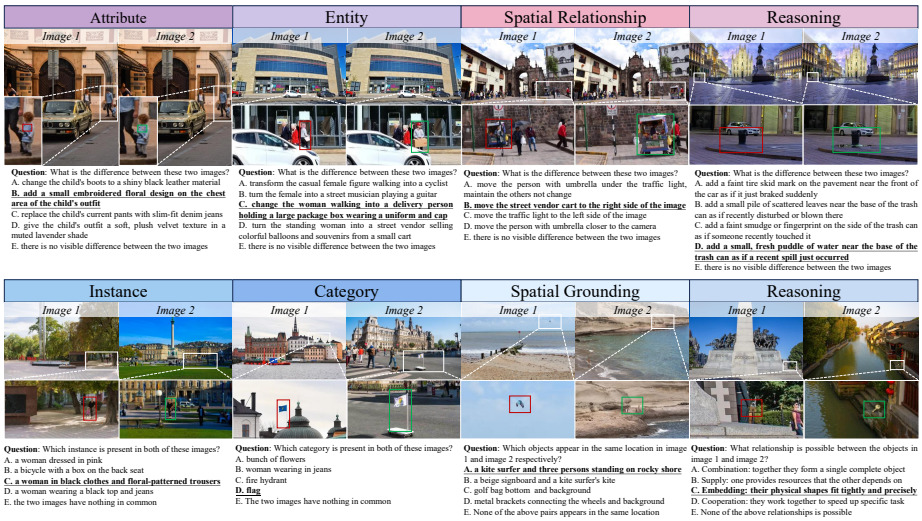


Fig. 3: Qualitative results on DiCoBench. The first and second rows illustrate examples of the four task types within the Differential Visual Cues Tasks and Commonality Visual Cues Tasks. Notably, the question for each task contains no explicit text cues. To answer correctly, models must actively perceive the visual cues representing differences or commonalities directly from the image pair.

samples across 8 distinct task categories, challenging models to discover visual cues from image pairs without explicit textual prompts.

3.1 Task Definition

Differential Visual Cues evaluates the model's ability to perceive minute changes within highly aligned or complex high-resolution backgrounds. It includes four progressive categories:

1. **Attribute** evaluates the capacity to discern microscopic alterations in fine-grained visual properties (*e.g.*, spectral reflectance, surface material, or morphological shape) of a specific target, strictly preserving its categorical identity and spatial coordinates. This demands precise semantic decoupling and rigorous analysis of localized visual indicators against complex background interference. **Example:** In an extremely cluttered electronic workbench, a millimeter-scale DIP switch changes from red to green (or flips from ON to OFF). The implicit question forces the model to focus on the attribute state rather than merely detecting the object.
2. **Entity** measures the proficiency in detecting categorical substitution, anomalous disappearance, or spontaneous emergence of microscopic entities within dense local regions. Success requires meticulous parsing of localized semantic shifts and robust feature discrimination to identify subtle structural or categorical deviations rather than mere pixel-level noise. **Example:** In a

high-definition street view, a distant "Speed Limit 60" sign is replaced by a visually similar "Speed Limit 80"; or a tiny white circular pill in a medicine box is replaced by a white circular button.

3. **Spatial Relationship** probes the advanced comprehension of physical-world dynamics by tracking the micro-displacement or topological reorganization of an entity, maintaining strict invariance in its identity and intrinsic attributes. This necessitates robust spatial grounding and the ability to map relative positional shifts within a stable contextual framework, distinguishing genuine spatial semantics from trivial viewpoint variations. **Example:** A small bunch of keys moves from inside a drawer to hanging on the wall. This strictly differentiates from traditional pixel-level differences by emphasizing explicit spatial semantics.
4. **Reasoning** assesses the pinnacle of visual discrepancy interpretation, where micro-inconsistencies serve not as static pixel changes, but as "visual traces" indicative of latent physical interactions or temporal events. This capability demands second-order cognitive operations, integrating causal inference with precise extraction of visual cues (*e.g.*, footprint deposition) to reconstruct unobserved physical occurrences. **Example:** Across two high-res beach images, a tiny human footprint appears in the corner of Image B; or a microscopic stress crack emerges on the edge of a perfectly intact glass.

Commonality Visual Cues introduces a novel challenge: finding the unique intersection of entities, locations, or logic across two high-resolution images with completely different global semantics, lighting, and perspectives. In this scenario, the MLLMs are faced with the fundamental challenge of isolating associative cues between objects while navigating through vast amounts of visual interference and disparity.

1. **Instance** evaluates the capability to achieve physical-instance re-identification under extreme scene and viewpoint variance. **Example:** Finding a specific Swiss Army knife with unique wear marks hidden in both a highly messy student dormitory (Image A) and an outdoor picnic mat (Image B).
2. **Category** measures the generalization capability to align micro-targets sharing highly specific taxonomic semantics, despite undergoing domain shifts in scale, chromaticity, material composition, or physical state. **Example:** A short, red, knotted Ethernet cable dropped on a server room floor versus a long, blue, straight Ethernet cable in a recycling station.
3. **Spatial Grounding** benchmarks the model's aptitude for identifying spatially equivalent correspondences between disparate objects. Instead of matching appearances, it requires the model to determine whether objects located at identical relative positions in Image A and Image B. **Example:** Identifying spatially corresponding objects across disparate scenes, such as a kite surfer and three persons on a rocky shore versus a beige signboard and a kite, where the model must pinpoint the pair occupying equivalent relative locations.
4. **Reasoning** represents the apex of Commonality Visual Cues tasks, characterized by the absence of any explicit visual or structural overlap. It requires

the model to first develop a fine-grained understanding of nearly every detail within both images. Subsequently, based on this comprehensive perception, the model must deduce whether objects across the two images form a specific functional relationship. Currently, this category encompasses four relationship types:

- *Combination:* The objects together form a single complete entity (*e.g.*, a bottle body and its cap).
- *Supply:* One object provides essential energy or resources that the other depends on (*e.g.*, a charger and a smartphone).
- *Embedding:* The physical shapes of the objects are designed to fit tightly and precisely into each other (*e.g.*, a memory card and a card slot).
- *Cooperation:* The objects must work together to accomplish a specific task (*e.g.*, a hammer and a nail).

Examples: A solitary, weathered door lock in an old alleyway versus a rusted, vintage key discarded on a park bench; despite the complete lack of visual similarity, the model must recognize the functional "combination" relationship between the two. Similarly, a high-voltage power outlet on a modern office wall versus a drained laptop battery in a dark drawer exemplifies a "supply" relationship, necessitating deep logical reasoning beyond mere visual perception.

3.2 Dataset Construction Pipeline

The DiCoBench dataset is constructed through rigorous human supervision and a multi-stage pipeline that leverages state-of-the-art (SOTA) image editing models and MLLMs for semi-automated data synthesis. To ensure high-quality baseline scenes, we source high-resolution base images from V* [31], which is widely recognized in the field of fine-grained perception.

General Image Editing & MCQ Formulation To establish the foundation of our benchmark, we employ an automated local-editing paradigm. First, **Automated Instruction & Mask Generation:** Based on grounding annotations, we extract micro-target masks and enlarge them by a $2\times$ context margin to ensure smooth blending. We utilize GPT-5.1 to generate 6 diverse, task-aware modification instructions per mask. To ensure diversity and prevent mode collapse, previously generated instructions are iteratively fed back into the MLLM's context window. Second, **Micro-scale Editing:** The instructions and expanded masks are fed into the FLUX.2 Klein model. This ensures that only the intrinsic properties of the micro-target are altered, while the complex high-resolution background remains perfectly preserved. Finally, **Multiple-Choice Construction:** To eliminate penalties arising from text formatting during evaluation, we structure the benchmark as a Multiple-Choice Question (MCQ). For each valid sample, we randomly select four successful image-instruction pairs as options A, B, C, and D. To prevent the model from exploiting linguistic biases to guess the answer, we ensure that the correct answer is balanced across all options for any

given text prompt. Furthermore, to ensure logical completeness and mitigate spurious successes through random guessing, we universally append Option E: *"There is no visible difference between the two images"* or *"There is no visible commonality between the two images"*.

Task-Specific Construction Protocols To fulfill the distinct requirements of the 8 sub-tasks, we design customized generation and filtering workflows, deliberately injecting manual verification and fallback mechanisms at vulnerable nodes.

Track 1: Differential Visual Cues.

- *Attribute:* The MLLMs generates attributes-altering instructions (e.g., color, material). Following FLUX.2 editing, an automated MLLMs check filters out failed edits, followed by human annotators who verify if the categorical identity and spatial coordinates remain strictly unchanged.
- *Entity:* The MLLMs drafts instructions for microscopic object substitution or removal. After generation and editing, human experts inspect the local regions to ensure the substituted entity does not introduce semantic contradictions with the surrounding context.
- *Spatial Relationship:* The MLLMs generates instructions to displace the target. After editing and MLLMs-based verification, human annotators conduct a strict review. If the MLLMs-generated spatial instructions violate physical constraints (e.g., objects floating in mid-air) or fall out of logical bounds, human experts directly intervene to supplement and rewrite valid spatial instructions.
- *Reasoning:* The MLLMs generates instructions to create "visual traces" (e.g., stress cracks, footprints). The generated images are rigorously reviewed by humans to ensure the traces are visually realistic and logically imply a latent physical event.

Track 2: Commonality Visual Cues.

- *Instance:* We crop a specific local region from Image A and utilize FLUX.2 Klein's multi-image editing capability to seamlessly blend the exact entity into a structurally disparate high-res Image B. We prepare 6 candidate blends per source image. After manual inspection, if fewer than 4 valid candidate pairs remain, human annotators take over to supervise the generation and supplement the dataset.
- *Category:* Building upon the Instance pipeline, we apply local editing to the entity in Image A to explicitly alter its instance-level attributes (e.g., color, material) while preserving its taxonomic category, ensuring the model matches based on conceptual category rather than identical pixels.
- *Spatial Grounding:* Images are divided into 6×6 grids, and each grid is captioned by MLLMs, followed by human correction. We compute text-embedding similarities using OpenAI's `text-embedding-3` and select pairs of grids with the *lowest* semantic similarity (ensuring maximum background disparity). Finally, the model is queried to determine whether a target pair of objects occupies the corresponding spatial location across the two images.

– *Reasoning:* We pre-define four abstract physical/logical relations: *Combination*, *Supply*, *Embedding*, and *Cooperation*. These relations are translated into pairs of visually recognizable objects. After filtering suitable masks in two entirely different images, the respective objects are synthesized. Finally, human experts verify the quality of the synthesized objects and confirm that the intended answer can be accurately deduced from the images.

Strict Human Verification & Quality Control To ensure the dataset acts as a flawless gold standard, all synthesized pairs undergo a final, exhaustive manual review based on four strict criteria: (1) *Textual Fidelity:* The modification must be strictly faithful to the instruction; (2) *Visual Naturalness:* The edited image must be free of blurriness, artifacts, or boundary degradation; (3) *Mutual Distinguishability:* Modifications across different options (A to D) for the same base image must be mutually exclusive and distinguishable; (4) *Scale Constraint:* The altered region must remain rigorously "micro" (accounting for $< 5\%$ of the total image area). Any sample failing to meet all four criteria is either routed back to the human fallback mechanism for manual regeneration or permanently discarded.

4 Experiments

In this section, we first describe the experimental setup and the baselines (§4.1). Then, we present a comprehensive evaluation of 18 latest SOTA MLLMs (§4.2). We demonstrate that while humans can answer the questions with high accuracy, DiCoBench poses significant challenges to existing models. Finally, we provide detailed analyses on multiple experimental settings, examining how high-resolution inputs impact the difficulty of perception tasks. We further reveal the performance patterns of humans during fine-grained perception to inspire future research directions, and perform an error analysis of existing models on DiCoBench (§4.3).

4.1 Experimental Setup

Multimodal LLMs: We evaluate DiCoBench on 18 recent MLLMs, including state-of-the-art closed-source models such as Gemini-3 (Flash, Pro), and the GPT family (4o, o4-mini, 4.1-mini, 4.1, 5); state-of-the-art open-source models such as Qwen3-VL (8B, 30B-A28B) [3], Qwen2.5-VL (7B, 32B) [4], Gemma3 (12B, 27B) [23], InternVL3.5 (241B-A28B) [28], and Qwen3.5; as well as models specifically optimized for fine-grained perception, including DeepEyes [35], Thyme [34], TreeVGR [26], and Mini-O3 [15].

Evaluation setup: We follow standard setups as in the VLMEvalKit [8], where the temperature is set to 0. Specifically, we require the models to directly output the corresponding answer letters for the MCQs (A, B, C, D and E) and evaluate them using exact letter matching. We find that in most cases, existing models demonstrate strong instruction-following capabilities.

Table 1: Performance comparison of SOTA MLLMs on DiCoBench. The highest and lowest scores for each model type across task types are highlighted in blue and red, respectively. The highest performance achieved by the model in each column is indicated in **bold**.

	Difference Tasks				Commonality Tasks				
Model	**Attr.**	**Ent.**	**Spa.**	**Rea.**	**Ins.**	**Cat.**	**Spa.**	**Rea.**	**Avg**
Human	97.6	98.9	98.0	97.3	99.1	99.1	98.4	97.4	98.3
Closed-source Models									
Gemini-3-Pro	**49.4**	**89.5**	**82.0**	**28.0**	**72.7**	**63.6**	**58.4**	**36.4**	**58.1**
Gemini-3-Flash	25.0	55.6	40.0	14.3	45.5	**63.6**	41.7	**36.4**	41.9
GPT-4o	32.9	46.3	26.0	25.33	28.2	24.6	26.0	20.0	29.0
GPT-o4-mini	38.8	53.7	54.0	22.7	71.8	60.9	48.0	17.4	46.3
GPT-4.1-mini	42.4	55.8	54.0	**28.0**	70.9	60.9	25.6	20.0	44.1
GPT-4.1	28.3	30.5	24.0	21.3	22.7	26.4	27.2	19.1	25.0
GPT-5	35.3	54.7	48.0	22.7	66.4	53.6	36.0	22.6	42.6
Open-source Models									
Qwen2.5-VL-7B [4]	33.0	53.7	34.0	25.3	74.6	59.1	23.2	20.0	41.0
Qwen2.5-VL-32B [4]	22.4	35.8	38.0	20.0	45.5	48.2	21.6	20.0	31.4
Gemma3-12B [23]	22.4	34.7	30.0	20.0	50.0	39.1	29.6	20.0	31.4
Gemma3-27B [23]	20.0	30.5	24.0	21.3	47.3	44.6	36.8	20.0	31.9
InternVL3.5-241B-A28B [28]	22.4	28.4	36.0	18.7	46.4	43.6	21.6	20.0	29.7
Qwen3-VL-8B [3]	29.4	51.6	40.0	21.3	71.8	**63.6**	20.8	20.0	40.3
Qwen3-VL-30B-A3B [3]	31.8	46.3	36.0	22.7	21.8	51.8	20.8	20.0	30.9
Qwen3.5-35B-A3B	28.3	42.1	**64.0**	26.7	64.6	59.1	49.6	20.0	44.1
DeepEyes-7B [35]	31.8	55.8	36.0	26.7	75.5	60.9	29.6	20.0	42.9
Thyme-7B [34]	23.5	42.1	22.0	20.0	66.4	53.6	24.8	20.0	35.6
TreeVGR-7B [26]	**35.3**	**57.9**	34.0	**29.3**	**76.4**	62.7	**58.4**	20.0	**48.8**

4.2 Main Results

Overall Performance: DiCoBench poses a formidable challenge to existing MLLMs. As shown in Table 1, even the most advanced closed-source models (e.g., Gemini-3-Pro) and open-source models (e.g., TreeVGR-7B) achieve average accuracies of only 58.1% and 48.8%, respectively, indicating a massive performance gap compared to the 98.3% average accuracy of humans. In particular, most models perform poorly in the reasoning (Rea.) sub-tasks within the Difference Tasks category, with scores for many models falling below 20-30. Notably, the reasoning tasks within Commonality Tasks present an exceptionally high level of difficulty. Although these tasks are easily solvable for human participants, we observe that nearly all existing open-source models mistakenly perceive no commonalities between the two images, thus predominantly selecting Option E, which leads to trivialized, identical results. This highlights a significant limitation in existing models when handling fine-grained cross-image comparison.

Closed-source vs. Open-source Models: Within the closed-source camp, Gemini-3-Pro leads with an average score of 58.1%, demonstrating strong capabilities in entity (Ent.) and spatial (Spa.) understanding for Difference Tasks. In contrast, the performance of the GPT series is highly inconsistent; for instance, GPT-4.1-mini excels in attribute (Attr.) recognition within Difference Tasks (42.4%) but scores only 25.6% in spatial (Spa.) understanding within Commonality Tasks. This reveals an uneven distribution of capabilities across different visual reasoning tasks among closed-source models. Regarding open-source models, TreeVGR-7B demonstrates remarkable competitiveness, with its 48.8% average score outperforming several closed-source models (e.g., GPT-4o at 29.0% and GPT-4.1 at 25.0%). This underscores the potential for lightweight, task-specific optimized models to compete effectively with large-scale closed-source counterparts.

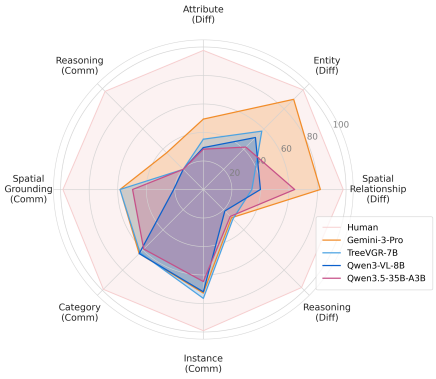


Fig. 4: Accuracies of MLLMs on DiCoBench.

Performance Disparities across Tasks: A comparison across task types reveals that the categorization (Cat.) task within Commonality Tasks is a relative strength for current models (with most scoring above 50%), whereas the reasoning (Rea.) task remains a universal "Achilles' heel," with most models hovering around 20% in Figure 4. This phenomenon suggests that while current MLLMs have made progress in fundamental object recognition and classification, they still lack the ability to capture complex visual cues required for deep logical analysis and multi-image fine-grained perception.

Is human perception instantaneous? Although existing benchmarks commonly include human performance result, the relationship between human perceptual processes and accuracy has been absent, particularly in fine-grained perception tasks. We attempt to analyze and demonstrate this relationship on DiCoBench by tracking human perceptual duration against accuracy. We invited eight Ph.D. students, who had no prior exposure to the test, to serve as participants. Each two were required to complete the tasks under one of the four time constraints: 30s, 60s, 120s, and unlimited time; the final results represent the average. As shown in Figure 5, we find that human precision in perception is not instantaneous; when perceptual time is limited, the performance gap between humans and state-of-the-art models begins to narrow. However, as the human time investment increases, perceptual accuracy improves consistently, eventually reaching a near-perfect state. This suggests that the perceptual capabilities of existing MLLMs, like their reasoning abilities, may benefit from increased computational investment.

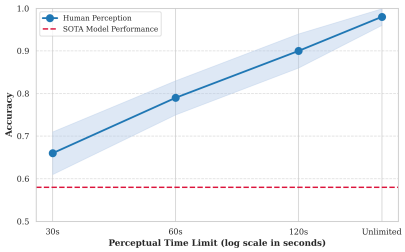


Fig. 5: Human accuracy on DiCoBench as a function of perceptual duration. The SOTA model Gemini-3-Pro is shown for comparison.

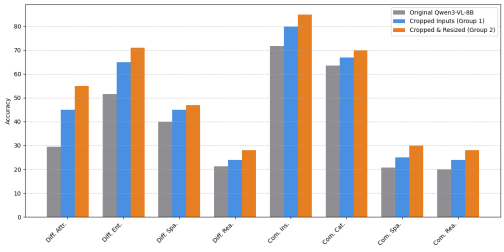


Fig. 6: Performance comparison of the Qwen-3-VL model on DiCoBench under different settings: vanilla inputs, cropped inputs, and resized inputs.

4.3 Analysis

How does high resolution challenge fine-grained perception? Compared to low-resolution perception tasks, high-resolution images, on one hand, increase the volume of visual information that models must process; on the other hand, they lower the relative scale at which visual cues must be identified to be effectively observed. To explore how high resolution challenges existing MLLMs, we conducted comparative experiments using the Qwen-3-VL-8B model. We sampled 1/10 of the instances from each task in DiCoBench, annotated the positions of the visual cues manually, and performed two sets of controlled experiments. In the first set, we used the cropped images based on the annotated positions as inputs. In the second set, we cropped the original images and then resized them back to the original image dimensions as inputs. As shown in Figure 6, the first set demonstrates a significant improvement compared to the original evaluation results. Notably, the second set achieves an even greater performance gain compared to the first. These results indicate that, on one hand, filtering out irrelevant visual regions significantly enhances model performance, suggesting that high-resolution images increase evaluation difficulty by introducing an excessive volume of extraneous visual information. On the other hand, high resolution poses the challenge of excessively low visual cue ratios. Therefore, increasing the relative proportion of these cues contributes substantially to improved perceptual performance.

Error analysis: To investigate the reasons behind the potential failures of existing models, we employed Gemini-3-pro and Qwen-3-VL as representatives of closed-source and open-source models respectively. Since our evaluation protocol requires models to output the final answer directly, which hinders the assessment of their reasoning processes, we allowed the models to generate brief descriptions before providing the answers during the error analysis. We categorized the errors into four types:

1. *Loss of visual cues:* Ignoring visual cues regarding the differences or commonalities between two images.

2. *Factual descriptive hallucination:* Misjudgments such as the orientation of a dog or the position of a person in the images.
3. *Hallucination from nothing (ex nihilo):* Perceiving differences between identical images, or identifying commonalities where none exist.
4. *Miscellaneous:* Refusals to answer or errors with indeterminate causes.

As shown in Figure 7, we found that although there is a significant performance gap between Gemini-3-pro and Qwen-3-VL, their error distributions are quite similar. The loss of visual cues is the primary issue for existing models, suggesting a fundamental problem in these models regarding visual-cue-guided perception tasks that have not been evaluated in the past, likely due to a systemic lack of training data. Beyond the loss of visual cues, we observed distinct hallucination preferences: Gemini-3-pro is more prone to "hallucination from nothing," while Qwen-3-VL tends to produce "factual descriptive hallucinations." These errors indicate that future MLLMs should prioritize optimization for recognizing visual cues in perception tasks to prevent the neglect of subtle visual information. Concurrently, it is also essential to enhance general perceptual accuracy to mitigate both factual descriptive hallucinations and hallucinations from nothing.

5 Conclusion

In this paper, we have presented DiCoBench, the first comprehensive benchmark tailored to evaluate high-resolution, cross-image fine-grained perception in MLLMs. By synthesizing datasets that necessitate the detection of implicit visual cues without explicit textual guidance, we have demonstrated that current SOTA models, both closed-source and open-source still struggle particularly in tasks requiring the integration of micro-scale visual information. Our analysis highlights that while these models have achieved proficiency in fundamental object recognition, they still struggle significantly with the proactive interpretation of spontaneous visual discrepancies and commonalities. We believe that DiCoBench not only provides a rigorous framework to expose these limitations but also offers a critical frontier for developing more robust, perceptually aware multimodal systems capable of bridging the gap between current machine performance and human-level visual cognition.

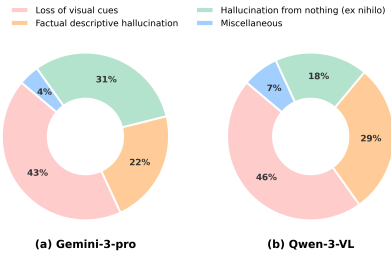


Fig. 7: Proportional distribution of error types for Gemini-3-Pro and Qwen-3-VL on DiCoBench.

Acknowledgements

This work was supported by the grants from Beijing Natural Science Foundation (L2470060) and the National Natural Science Foundation of China (62525201, 62132001, 62432001).

References

1. An, X., Xie, Y., Yang, K., Zhang, W., Zhao, X., Cheng, Z., Wang, Y., Xu, S., Chen, C., Zhu, D., Wu, C., Tan, H., Li, C., Yang, J., Yu, J., Wang, X., Qin, B., Wang, Y., Yan, Z., Feng, Z., Liu, Z., Li, B., Deng, J.: Llava-onevision-1.5: Fully open framework for democratized multimodal training (2025), https://arxiv.org/abs/2509.23661
2. Awal, R., Ahmadi, S., Zhang, L., Agrawal, A.: Vismin: Visual minimal-change understanding. Advances in Neural Information Processing Systems **37**, 107795–107829 (2024)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report (2025), https://arxiv.org/abs/2511.21631
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923
5. Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., Cui, E., Tong, W., Hu, K., Luo, J., Ma, Z., et al.: How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. Science China Information Sciences **67**(12), 220101 (2024)
6. Di, Z., Shi, J., Fan, Y., Tan, H., Black, A., Collomosse, J., Liu, Y.: Difftell: A high-quality dataset for describing image manipulation changes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 24580–24590 (2025)
7. Du, H., Zhang, J., Nan, G., Deng, W., Chen, Z., Zhang, C., Xiao, W., Huang, S., Pan, Y., Qi, T., Leng, S.: From easy to hard: The mir benchmark for progressive interleaved multi-image reasoning. arXiv preprint arXiv:2509.17040 (2025)
8. Duan, H., Yang, J., Qiao, Y., Fang, X., Chen, L., Liu, Y., Dong, X., Zang, Y., Zhang, P., Wang, J., et al.: Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 11198–11201 (2024)
9. Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: Blink: Multimodal large language models can see but not perceive. arXiv preprint arXiv:2404.12390 (2024)
10. Guo, Z., Sun, J., Wang, T.J.J., Radman, A., Pehlivan, S., Cao, M., Laaksonen, J.: Learning to describe implicit changes: Noise-robust pre-training for image difference captioning. In: Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng,

V. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2025. pp. 10125–10145. Association for Computational Linguistics, Suzhou, China (Nov 2025). https://doi.org/10.18653/v1/2025.findings-emnlp.537, https://aclanthology.org/2025.findings-emnlp.537/
11. Hong, J., Zhao, C., Zhu, C., Lu, W., Xu, G., Yu, X.: Deepeyesv2: Toward agentic multimodal model (2026), https://arxiv.org/abs/2511.05271
12. Hu, E., Guo, L., Yue, T., Zhao, Z., Xue, S., Liu, J.: Onediff: A generalist model for image difference captioning. In: Proceedings of the Asian Conference on Computer Vision. pp. 2439–2455 (2024)
13. Jhamtani, H., Berg-Kirkpatrick, T.: Learning to describe differences between pairs of similar images. arXiv preprint arXiv:1808.10584 (2018)
14. Kil, J., Mai, Z., Lee, J., Chowdhury, A., Wang, Z., Cheng, K., Wang, L., Liu, Y., Chao, W.L.H.: Mllm-compbench: A comparative reasoning benchmark for multimodal llms. Advances in Neural Information Processing Systems **37**, 28798–28827 (2024)
15. Lai, X., Li, J., Li, W., Liu, T., Li, T., Zhao, H.: Mini-o3: Scaling up reasoning patterns and interaction turns for visual search. arXiv preprint arXiv:2509.07969 (2025)
16. Li, J., Wu, M., Jin, Z., Chen, H., Ji, J., Sun, X., Cao, L., Ji, R.: Mihbench: Benchmarking and mitigating multi-image hallucinations in multimodal large language models. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 3143–3152 (2025)
17. Li, Y., Tu, Y., Li, L., Su, L., Huang, Q.: Change entity-guided heterogeneous representation disentangling for change captioning. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) Findings of the Association for Computational Linguistics: ACL 2025. pp. 17050–17060. Association for Computational Linguistics, Vienna, Austria (Jul 2025). https://doi.org/10.18653/v1/2025.findings-acl.876, https://aclanthology.org/2025.findings-acl.876/
18. Li, Y., Huang, H., Chen, C., Huang, K., Huang, C., Guo, Z., Liu, Z., Xu, J., Li, Y., Li, R., et al.: Migician: Revealing the magic of free-form multi-image grounding in multimodal large language models. arXiv preprint arXiv:2501.05767 (2025)
19. Liu, Y., Hou, S., Hou, S., Du, J., Meng, S., Huang, Y.: Omnidiff: A comprehensive benchmark for fine-grained image difference captioning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21440–21449 (2025)
20. Park, D.H., Darrell, T., Rohrbach, A.: Robust change captioning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4624–4633 (2019)
21. Song, D., Chen, S., Chen, G.H., Yu, F., Wan, X., Wang, B.: Milebench: Benchmarking mllms in long context. arXiv preprint arXiv:2404.18532 (2024)
22. Tan, H., Dernoncourt, F., Lin, Z., Bui, T., Bansal, M.: Expressing visual relationships via language. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 1873–1883 (2019)
23. Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., Mesnard, T., Cideron, G., bastien Grill, J., Ramos, S., Yvinec, E., Casbon, M., Pot, E., Penchev, I., Liu, G., Visin, F., Kenealy, K., Beyer, L., Zhai, X., Tsitsulin, A., Busa-Fekete, R., Feng, A., Sachdeva, N., Coleman, B., Gao, Y., Mustafa, B., Barr, I., Parisotto, E., Tian, D., Eyal, M., Cherry, C., Peter, J.T., Sinopalnikov, D., Bhupatiraju, S., Agarwal, R., Kazemi, M., Malkin, D., Kumar, R., Vilar, D., Brusilovsky, I., Luo, J., Steiner, A., Friesen, A., Sharma, A., Sharma, A., Gilady, A.M., Goedeckemeyer,

A., Saade, A., Feng, A., Kolesnikov, A., Bendebury, A., Abdagic, A., Vadi, A., György, A., Pinto, A.S., Das, A., Bapna, A., Miech, A., Yang, A., Paterson, A., Shenoy, A., Chakrabarti, A., Piot, B., Wu, B., Shahriari, B., Petrini, B., Chen, C., Lan, C.L., Choquette-Choo, C.A., Carey, C., Brick, C., Deutsch, D., Eisenbud, D., Cattle, D., Cheng, D., Paparas, D., Sreepathihalli, D.S., Reid, D., Tran, D., Zelle, D., Noland, E., Huizenga, E., Kharitonov, E., Liu, F., Amirkhanyan, G., Cameron, G., Hashemi, H., Klimczak-Plucińska, H., Singh, H., Mehta, H., Lehri, H.T., Hazimeh, H., Ballantyne, I., Szpektor, I., Nardini, I., Pouget-Abadie, J., Chan, J., Stanton, J., Wieting, J., Lai, J., Orbay, J., Fernandez, J., Newlan, J., yeong Ji, J., Singh, J., Black, K., Yu, K., Hui, K., Vodrahalli, K., Greff, K., Qiu, L., Valentine, M., Coelho, M., Ritter, M., Hoffman, M., Watson, M., Chaturvedi, M., Moynihan, M., Ma, M., Babar, N., Noy, N., Byrd, N., Roy, N., Momchev, N., Chauhan, N., Sachdeva, N., Bunyan, O., Botarda, P., Caron, P., Rubenstein, P.K., Culliton, P., Schmid, P., Sessa, P.G., Xu, P., Stanczyk, P., Tafti, P., Shivanna, R., Wu, R., Pan, R., Rokni, R., Willoughby, R., Vallu, R., Mullins, R., Jerome, S., Smoot, S., Girgin, S., Iqbal, S., Reddy, S., Sheth, S., Põder, S., Bhatnagar, S., Panyam, S.R., Eiger, S., Zhang, S., Liu, T., Yacovone, T., Liechty, T., Kalra, U., Evci, U., Misra, V., Roseberry, V., Feinberg, V., Kolesnikov, V., Han, W., Kwon, W., Chen, X., Chow, Y., Zhu, Y., Wei, Z., Egyed, Z., Cotruta, V., Giang, M., Kirk, P., Rao, A., Black, K., Babar, N., Lo, J., Moreira, E., Martins, L.G., Sanseviero, O., Gonzalez, L., Gleicher, Z., Warkentin, T., Mirrokni, V., Senter, E., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Matias, Y., Sculley, D., Petrov, S., Fiedel, N., Shazeer, N., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Alayrac, J.B., Anil, R., Dmitry, Lepikhin, Borgeaud, S., Bachem, O., Joulin, A., Andreev, A., Hardin, C., Dadashi, R., Hussenot, L.: Gemma 3 technical report (2025), https://arxiv.org/abs/2503.19786

24. Tong, T.C., He, S., Shao, Z., Yeung, D.Y.: G-veval: A versatile metric for evaluating image and video captions using gpt-4o. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 7419–7427 (2025)
25. Wang, F., Fu, X., Huang, J.Y., Li, Z., Liu, Q., Liu, X., Ma, M.D., Xu, N., Zhou, W., Zhang, K., et al.: Muirbench: A comprehensive benchmark for robust multi-image understanding. arXiv preprint arXiv:2406.09411 (2024)
26. Wang, H., Li, X., Huang, Z., Wang, A., Wang, J., Zhang, T., Zheng, J., Bai, S., Kang, Z., Feng, J., et al.: Traceable evidence enhanced visual grounded reasoning: Evaluation and methodology. arXiv preprint arXiv:2507.07999 (2025)
27. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
28. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., Wang, Z., Chen, Z., Zhang, H., Yang, G., Wang, H., Wei, Q., Yin, J., Li, W., Cui, E., Chen, G., Ding, Z., Tian, C., Wu, Z., Xie, J., Li, Z., Yang, B., Duan, Y., Wang, X., Hou, Z., Hao, H., Zhang, T., Li, S., Zhao, X., Duan, H., Deng, N., Fu, B., He, Y., Wang, Y., He, C., Shi, B., He, J., Xiong, Y., Lv, H., Wu, L., Shao, W., Zhang, K., Deng, H., Qi, B., Ge, J., Guo, Q., Zhang, W., Zhang, S., Cao, M., Lin, J., Tang, K., Gao, J., Huang, H., Gu, Y., Lyu, C., Tang, H., Wang, R., Lv, H., Ouyang, W., Wang, L., Dou, M., Zhu, X., Lu, T., Lin, D., Dai, J., Su, W., Zhou, B., Chen, K., Qiao, Y., Wang, W., Luo, G.: Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency (2025), https://arxiv.org/abs/2508.18265

29. Wang, W., Ding, L., Zeng, M., Zhou, X., Shen, L., Luo, Y., Yu, W., Tao, D.: Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In: AAAI. vol. 39, pp. 7907–7915 (2025)
30. Wei, K., Hu, B., Cao, J., Chen, X., Lu, Z., Xia, W., Xu, W., Wu, J., He, J., Jia, M., et al.: $m3-verse$: A" spot the difference" challenge for large multimodal models. arXiv preprint arXiv:2512.18735 (2025)
31. Wu, P., Xie, S.: V*: Guided visual search as a core mechanism in multimodal llms. In: CVPR. pp. 13084–13094 (2024)
32. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: CVPR (2024)
33. Zhang, X., Wen, H., Wu, J., Qin, P., Xue', H., Nie, L.: Differential-perceptive and retrieval-augmented MLLM for change captioning. In: ACM Multimedia 2024 (2024), https://openreview.net/forum?id=eiGs5VCsYM
34. Zhang, Y.F., Lu, X., Yin, S., Fu, C., Chen, W., Hu, X., Wen, B., Jiang, K., Liu, C., Zhang, T., Fan, H., Chen, K., Chen, J., Ding, H., Tang, K., Zhang, Z., Wang, L., Yang, F., Gao, T., Zhou, G.: Thyme: Think beyond images (2025), https://arxiv.org/abs/2508.11630
35. Zheng, Z., Yang, M., Hong, J., Zhao, C., Xu, G., Yang, L., Shen, C., Yu, X.: Deepeyes: Incentivizing" thinking with images" via reinforcement learning. arXiv preprint arXiv:2505.14362 (2025)