

CompareBench: A Benchmark for Visual Comparison Reasoning in Vision-Language Models

Jie Cai, Kangning Yang, Lan Fu, Jiaming Ding, Jinlong Li, Huiming Sun,
Daitao Xing, Jinglin Shen, Zibo Meng

OPPO AI Center

caijie0620@gmail.com

Abstract

We introduce CompareBench, a benchmark for evaluating visual comparison reasoning in vision-language models (VLMs), a fundamental yet understudied skill. CompareBench consists of 1,000 QA pairs across four tasks: quantity (600), temporal (100), geometric (200), and spatial (100). It is derived from two auxiliary datasets that we constructed: TallyBench (2,000 counting images with QA) and HistCaps (515 historical images with bilingual captions). We evaluate both closed-source APIs (OpenAI, Gemini, Claude) and open-source models (Qwen2.5-VL and Qwen3-VL series). Results show clear scaling trends but also reveal critical limitations: even the strongest models consistently fail at temporal ordering and spatial relations, and they often make mistakes in basic counting and geometric comparisons that are trivial for humans. These findings demonstrate that visual comparison remains a systematic blind spot for current VLMs. By providing controlled, diverse, and diagnostic evaluation, CompareBench establishes a foundation for advancing more reliable multimodal reasoning. All code, data, and instruction prompts will be released at <https://github.com/caijie0620/CompareBench>.

1. Introduction

Vision-language models (VLMs) have achieved remarkable progress in captioning, visual question answering (VQA), and multimodal reasoning. However, their ability to perform basic visual comparison, a fundamental human skill, remains underexplored. Humans can effortlessly compare objects with respect to quantity, temporal sequence, geometric properties, and spatial relations, yet these tasks remain highly challenging for current VLMs.

Visual comparison is not only essential for everyday perception (e.g., counting objects, judging depth, comparing

sizes), but also underpins higher-level reasoning tasks in education, science, and decision making. Despite its importance, existing benchmarks mainly emphasize recognition, description, or commonsense reasoning, while few provide systematic evaluation of comparative ability. For example, VQA-style datasets focus on open-domain knowledge, CLEVR [9] emphasizes synthetic logical reasoning, and recent holistic benchmarks (e.g., MMBench [13], MM-Vet [21], BLINK [5]) test diverse capabilities but lack targeted assessment of comparison reasoning.

To address this gap, we introduce CompareBench, a new benchmark specifically designed to evaluate visual comparison reasoning in VLMs. CompareBench is constructed from 1,000 image-based question-answer pairs that systematically cover four fundamental tasks: quantity comparison, temporal ordering, geometric property comparison, and spatial relation reasoning. These tasks are derived from two self-built auxiliary resources that provide the foundation for benchmark construction: TallyBench, a large-scale dataset of 2,000 images annotated with object counting questions, and HistCaps, a curated set of 515 historical images with bilingual captions covering a wide temporal range of events. By integrating these resources, CompareBench enables controlled yet diverse evaluation across complementary dimensions of comparison reasoning. As illustrated in Fig. 1, even state-of-the-art models such as GPT-5 fail on tasks that are trivial for humans, including simple object counting, identifying geometric differences, reasoning spatial relations, and ordering historical events, underscoring the necessity of a dedicated benchmark to diagnose and advance this capability. Table 1 summarizes the four sub-benchmarks of CompareBench, including their task types, scales, formats, and representative example questions.

Our contributions are summarized as follows:

- We construct two auxiliary datasets: TallyBench (2,000 counting images with QA) and HistCaps (515 historical images with bilingual captions).
- From these resources, we derive CompareBench, a bench-



Figure 1. Overview of TallyBench, HistCaps, and CompareBench with representative GPT-5 failure cases. CompareBench (bottom) encompasses four fundamental comparison tasks: geometric, spatial, quantity, and temporal sequence reasoning. TallyBench (top-left) is designed for object counting and also forms the basis of the quantity comparison task in CompareBench. HistCaps (top-right), annotated with temporal tags and bilingual captions, serves as the foundation for the temporal sequence comparison task. Although trivial for humans, GPT-5 consistently fails on these tasks. For example, it underestimates stacked cups, overestimates the number of chickens, misjudges book thickness, misestimates relative object height, miscounts in comparative settings, and incorrectly orders historical events, highlighting systematic limitations in visual comparison reasoning.

mark of 1,000 QA pairs spanning four tasks: quantity, temporal, geometric, and spatial comparison.

- We evaluate closed-source APIs (OpenAI, Gemini, Claude) and open-source models (Qwen2.5-VL and Qwen3-VL), revealing consistent scaling trends but persistent failures in visual comparison reasoning.

2. Related Work

2.1. Vision-Language Models

In recent years, vision-language models (VLMs) have achieved rapid progress. Early works such as CLIP [17] demonstrated the effectiveness of contrastive learning for joint image–text representations. Instruction-tuned models including BLIP-2 [10] and LLaVA [12] extended these approaches to interactive multimodal tasks, enabling VLMs to follow natural language queries about images. More recent large-scale systems such as Qwen2.5-VL [3], Gemini 2.5 [6], and GPT-4o [15] have further advanced visual reasoning by leveraging larger training corpora, stronger model

backbones, and multi-stage alignment strategies. Nevertheless, despite these advances in captioning and open-ended visual question answering, current VLMs still struggle with basic comparative reasoning, tasks that are trivial for humans. This overlooked capability is the focus of our work.

2.2. Benchmarks for Comparative Reasoning

Existing multimodal benchmarks emphasize recognition, description, or commonsense reasoning, but do not explicitly target comparative abilities. Captioning datasets such as MSCOCO [11] and Flickr30k [20] are designed for descriptive generation, focusing on aligning visual content with natural language sentences. VQA benchmarks (e.g., VQA, GQA [8]) extend beyond description to factual and compositional question answering, testing attributes, objects, or relationships within an image. Reasoning datasets such as RealWorldQA [19] and ScienceQA [14] incorporate external knowledge and multimodal reasoning, but rarely test direct comparison across multiple visual entities. Counting has also been studied extensively, often as a standalone

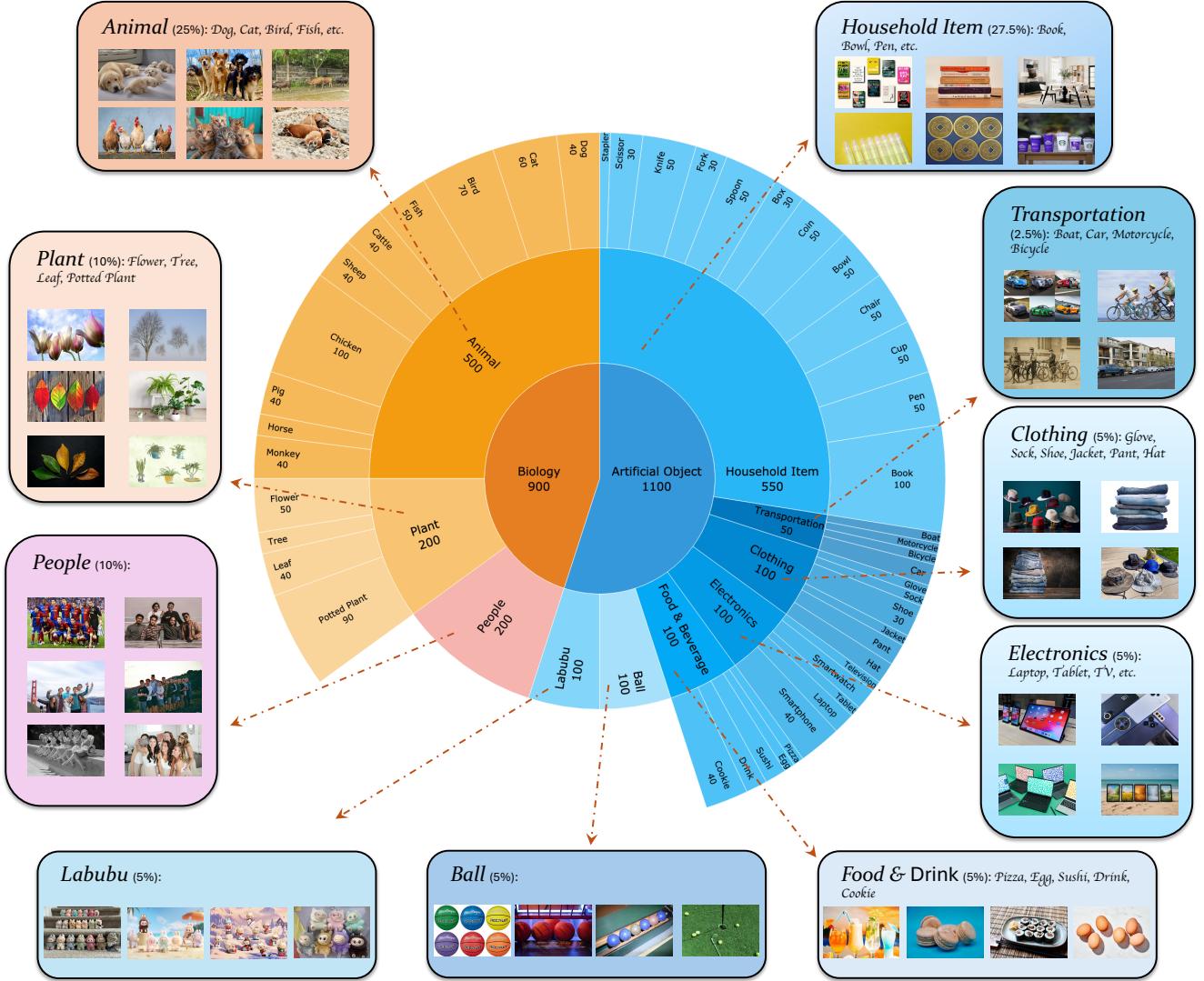


Figure 2. Distribution of TallyBench categories. The top level splits into Biology (900) and Artificial Objects (1100). Subcategories include animal, plant, people, food & beverage, electronics, clothing, transportation, ball, household item, etc. The outer ring further specifies around 50 fine-grained classes, such as Dog (40), Cat (60), Chicken (100), Book (100), Spoon (50), and Knife (50).

task. Early works addressed counting in natural scenes or specific domains (e.g., drone-based object counting [7]), while TallyQA [1] posed complex counting questions in a VQA setting. More recent benchmarks explore open-world counting [2], dense counting [18], or teaching CLIP to count [16]. However, these benchmarks remain limited to numerosity and do not generalize to broader comparative reasoning tasks such as size, distance, or temporal order.

Synthetic datasets such as CLEVR [9] provide carefully designed tests for compositional logic and spatial reasoning, but their artificial nature and limited diversity hinder transfer to real-world applications. More holistic benchmarks such as MMBench [13], MM-Vet [21], BLINK [5], and SimpleVQA [4] have recently emerged to test a wide

range of multimodal skills, from recognition and perception to knowledge grounding and factuality. Yet, even in these comprehensive evaluations, comparative reasoning is not systematically represented, and questions about “which object is larger/closer/earlier” are often absent. In contrast, CompareBench is purposefully designed to fill this gap: it provides a controlled and diverse evaluation of four fundamental comparison tasks (quantity, temporal, geometric, and spatial reasoning), derived from two self-built real-world datasets (TallyBench, HistCaps) and annotated into QA format. This makes it a complementary and focused diagnostic tool for evaluating a core but underexplored dimension of VLMs.

3. Methodology

3.1. TallyBench (2,000 Samples)

To construct the TallyBench dataset, we collected 2,000 images paired with object counting questions. Each image is annotated with a JSON entry containing its metadata, including: `image_name`, `vlm_question`, `gt_answer`, `categories`, and `image_type` (e.g., real, synthetic, artificial). As shown in Fig. 2, the `categories` field spans approximately 50 fine-grained object classes across diverse domains, such as animals, plants, people, food, electronics, clothing, transportation, household items, etc. For example, an entry might specify a question like “*How many dogs are in the image?*” with a ground-truth answer of 7.

To ensure consistent evaluation, each image-question pair is paired with a unified instruction template. The instruction enforces strict counting rules, requiring models to:

- Count all clearly identifiable and distinct instances of the target category.
- Include both physical objects and visual representations (e.g., drawings, prints).
- Consider partially occluded, cropped, or blurred objects if they are recognizable to a human observer.
- Exclude reflections in mirrors, water, glass, or other reflective surfaces.
- Answer with a single, precise integer only (e.g., 1, 7, 10, or 120), with no additional text or explanation.

These guidelines are embedded in a single standardized prompt, instructing the model to output only a precise integer without any additional text or explanation. By combining diverse categories and strict instructions, TallyBench provides a robust resource for evaluating counting ability in VLMs, serving as a foundation for the more complex comparative reasoning tasks in CompareBench.

3.2. HistCaps (515 Samples)

As shown in Fig. 1, HistCaps is a curated dataset of historical visual content, providing bilingual (*English–Chinese*) annotations for each image. Unlike TallyBench, no QA pairs are included. Instead, each entry is enriched with:

- A historical tag summarizing the key event (e.g., *Union Act, Naval War, Revolution, Inauguration Ceremony*).
- A long caption in both English and Chinese that provides detailed visual and historical context.
- A short caption in both languages for concise reference.

HistCaps covers a wide temporal span of historical events, including political treaties, wars, revolutions, inaugurations, disasters, and cultural milestones. Representative examples range from the 1707 Articles of Union and 18th-century naval battles to the 1776 Declaration of Independence, the 1789 Washington inauguration, and iconic 20th-century moments such as the 1945 WWII surrender and the 1969 Moon Landing. Each image is enriched with tempo-

ral information and bilingual captions, making HistCaps a valuable resource for studying chronological understanding and temporal reasoning in VLMs.

3.3. CompareBench (1,000 Samples)

CompareBench is built upon TallyBench, HistCaps, and additional human annotations, and is organized into four complementary sub-benchmarks that target quantity, temporal, geometric, and spatial comparison reasoning (Table 1 and Fig. 3). Each sub-benchmark is paired with a standardized instruction template to ensure consistent task formulation across settings. The final input to the VLM is constructed by concatenating the instruction with the task-specific question, and models are required to produce only a single choice (A–D) without additional text or explanation.

CompareGeometryBench (200) targets geometric comparison. Each sample consists of a single image with four objects labeled A–D, and the questions focus on dimensional properties such as length, width, height, thickness, or diameter. The task follows strict comparison rules, requiring models to:

- Consider only the four target objects explicitly marked with colored dots. Ignore all other objects in the image.
- Compare the specified dimensional property (length, width, diameter, thickness, or height) precisely between the marked objects.
- Base the judgment solely on the visible geometry of the objects, regardless of texture, shading, or semantic category.
- Do not infer from unmarked context or surrounding objects.
- Answer with a single, precise letter only (e.g., A, B, C, or D), with no additional text or explanation.

CompareSpatialBench (100) evaluates spatial relation reasoning. Each sample is a single annotated image containing four labeled points (A–D), with questions such as “Which object is closer to the camera?” or “Which object is higher above ground?”. The task enforces strict spatial comparison rules, requiring models to:

- Consider only the points or objects explicitly marked with colored dots, ignoring all others in the image.
- Treat a marked location as either a point in space (e.g., in the sky, on the ground, or on the ocean) or, if placed on an object, as representing that object as a whole.
- Compare the specified spatial property (vertical height above ground or distance to the camera) only among the marked points or objects.
- Provide a single, precise letter as the answer (e.g., A, B, C, or D), with no additional text or explanation.

CompareTallyBench (600) targets quantity comparison. Each sample is a 1600×1600 four-image grid derived from TallyBench images, with questions such as “Which image shows the most dogs?”. The task enforces strict

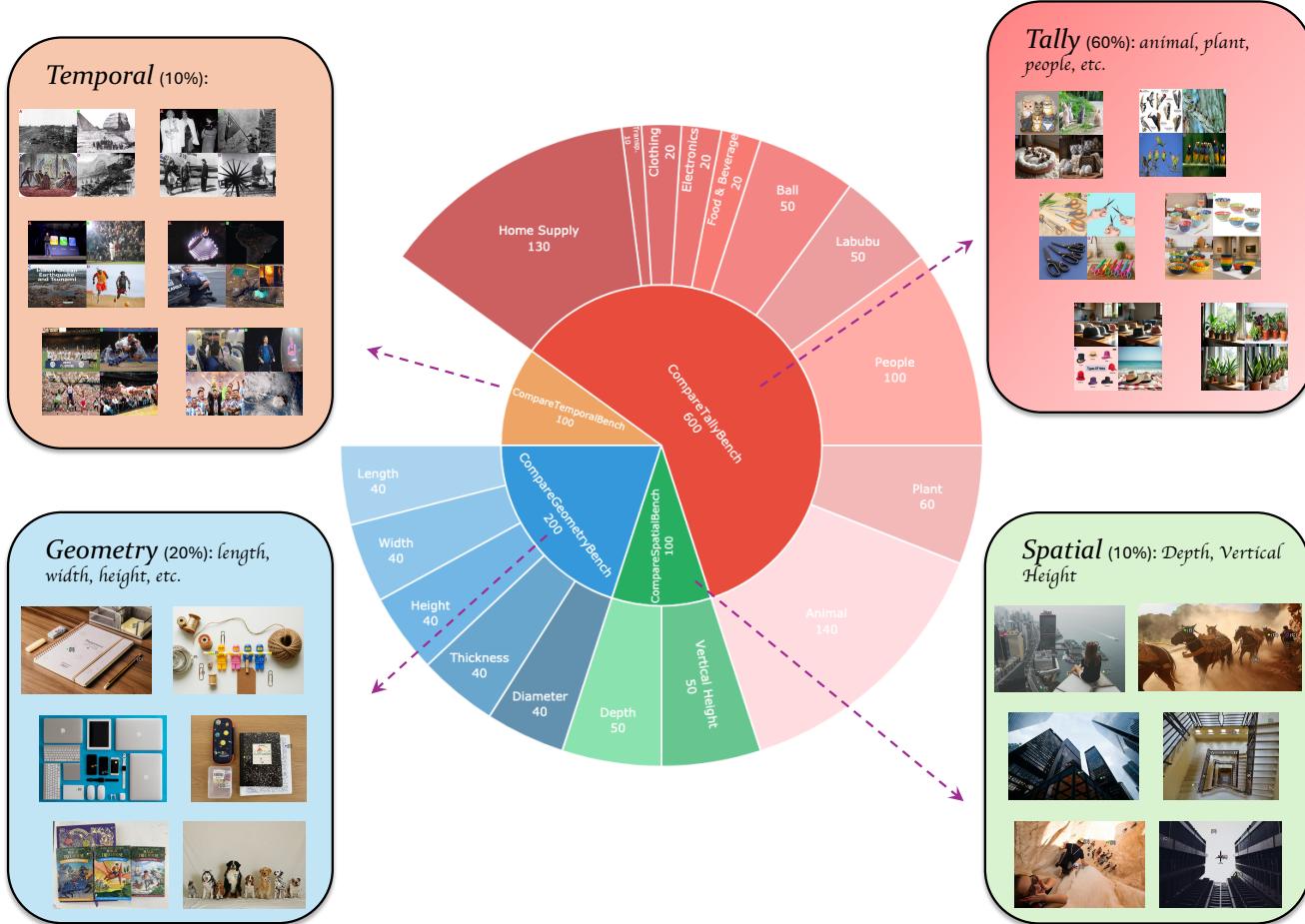


Figure 3. Category distribution of **CompareBench**. The inner ring represents the four sub-benchmarks: CompareTallyBench (600), CompareTemporalBench (100), CompareGeometryBench (200), and CompareSpatialBench (100). CompareTallyBench inherits diverse categories from TallyBench, including animals, plants, people, food & beverages, electronics, clothing, transportation, household items, etc. The outer ring further decomposes the geometric tasks into five fine-grained types that capture intrinsic object properties, including length, width, height, thickness, and diameter (40 samples each). The spatial tasks are divided into depth (object/point distance to the camera) and vertical height (object/point distance above the ground), with 50 samples each.

Table 1. Overview of CompareBench Subsets

| Subset | Task | Scale | Format | Example Question |
|----------------------|----------|-------|------------------------------|---|
| CompareTallyBench | Quantity | 600 | 4-image grid, 1600×1600 | Which image shows the most dogs? |
| CompareTemporalBench | Temporal | 100 | 4-image grid, 1920×1440 | Which scene appears earlier in history? |
| CompareGeometryBench | Geometry | 200 | Single image with A-D labels | Which object is longer in length? |
| CompareSpatialBench | Spatial | 100 | Single image with A-D labels | Which object is closer to the camera? |

counting comparison rules, requiring models to:

- Count all clearly identifiable and distinct instances of the specified target category in each sub-image.
- Include both physical objects and visual representations such as drawings or prints.
- Consider objects that are partially occluded, cropped,

or blurred, as long as they remain recognizable to a human observer.

- Exclude reflections in mirrors, water, glass, or other reflective surfaces.
- Take into account potential resizing distortions across sub-images that may affect object proportions when

Table 2. Accuracy (%) on TallyBench and CompareBench. CTally = CompareTallyBench, CTemp = CompareTemporalBench, CGeom = CompareGeometryBench, CSpat = CompareSpatialBench. Numbers in parentheses indicate the number of samples in each dataset. The best results are in **bold**, and the second-best results are underlined. All Qwen-VL models are of the Instruct version.

| Model | Tally (2,000) | CGeom (200) | CSpat (100) | CTally (600) | CTemp (100) | Compare (1,000) |
|---------------------------|---------------|--------------|--------------|--------------|--------------|-----------------|
| Closed-source APIs | | | | | | |
| Claude Sonnect 4 | 72.40 | 38.50 | 35.00 | 59.17 | 31.00 | 49.80 |
| OpenAI GPT-4o mini | 64.95 | 43.50 | 49.00 | 48.00 | 26.00 | 45.00 |
| OpenAI GPT-4o | 65.95 | 59.00 | 71.00 | 66.83 | 38.00 | 62.80 |
| OpenAI GPT-4.1 nano | 52.10 | 30.50 | 41.00 | 30.67 | 33.00 | 31.90 |
| OpenAI GPT-4.1 mini | 74.00 | 60.50 | 71.00 | 70.00 | 27.00 | 63.90 |
| OpenAI GPT-4.1 | 75.20 | 70.50 | 76.00 | 75.83 | 36.00 | 70.80 |
| OpenAI GPT-5 nano | 64.10 | 67.00 | 65.00 | 75.33 | 34.00 | 68.50 |
| OpenAI GPT-5 mini | <u>80.25</u> | <u>74.50</u> | 78.00 | 84.17 | 49.00 | 78.10 |
| OpenAI GPT-5 | 74.85 | 72.50 | 86.00 | 81.17 | 74.00 | <u>79.20</u> |
| OpenAI o4-mini | 79.30 | 73.50 | <u>81.00</u> | 85.83 | 47.00 | 79.00 |
| OpenAI o3 | 78.05 | 71.50 | 78.00 | 80.67 | 65.00 | 77.00 |
| OpenAI o3-pro | 78.90 | 69.50 | 77.00 | 83.83 | <u>72.00</u> | 79.10 |
| Gemini 2.5 Flash-Lite | 69.35 | 47.00 | 58.00 | 66.50 | 30.00 | 58.10 |
| Gemini 2.5 Flash | 78.40 | 71.50 | 70.00 | <u>86.33</u> | 58.00 | 78.90 |
| Gemini 2.5 Pro | 87.35 | 82.00 | 81.00 | 90.83 | 64.00 | 85.40 |
| Open-source Models | | | | | | |
| Qwen2.5-VL-3B | 56.05 | 34.50 | 36.00 | 37.50 | 27.00 | 35.70 |
| Qwen2.5-VL-7B | 67.35 | 36.00 | 54.00 | 51.50 | 29.00 | 46.40 |
| Qwen2.5-VL-32B | 69.45 | 46.50 | 64.00 | 56.83 | 24.00 | 52.20 |
| Qwen2.5-VL-72B | 75.60 | 50.50 | 68.00 | 62.83 | 29.00 | 57.50 |
| Qwen3-VL-235B-A22B | 82.25 | 71.00 | 81.00 | 66.50 | 32.00 | 65.40 |
| Human | 98.00 | 99.00 | 98.00 | 99.00 | 30.00 | 92.00 |

comparing counts.

- Provide a single, precise letter as the answer (e.g., A, B, C, or D), with no additional text or explanation.

CompareTemporalBench (100) focuses on temporal ordering. Each sample is a 1920×1440 four-image grid composed of HistCaps entries. The task requires models to identify which scene corresponds to the earliest historical event by leveraging both visual evidence and prior knowledge. To ensure consistent evaluation, the task enforces strict temporal reasoning rules, requiring models to:

- Use visible visual cues such as clothing styles, architectural features, technology, or environmental context to estimate the historical period of each scene.
- Combine these observations with relevant world knowledge of historical events or eras to determine which occurred earliest.
- Account for possible resizing distortions across sub-images when making comparisons.
- Provide a single, precise letter as the answer (e.g., A, B, C, or D), with no additional text or explanation.

Together, the four sub-benchmarks provide a systematic framework for evaluating visual comparison in VLMs.

By encompassing quantity, temporal, geometric, and spatial reasoning, CompareBench targets a core but underexplored capability and offers a complementary perspective to recognition- and description-oriented benchmarks.

4. Experiments

We evaluate both closed-source and open-source VLMs. The closed-source systems include OpenAI, Gemini, and Claude, while the open-source models are drawn from the Qwen2.5-VL family at four scales (3B, 7B, 32B, and 72B). Together, these models span a broad range of training scales and architectural designs.

Metrics. Accuracy (%) is used as the unified evaluation metric across all benchmarks. For TallyBench, accuracy is defined as the percentage of exact integer matches between predicted and ground-truth counts. For CompareBench, each question is formulated as a single-choice (A–D) with exactly one correct answer, and accuracy is measured as the proportion of correct predictions.

Results. Table 2 reports the performance of all evaluated models on TallyBench and the four sub-benchmarks of CompareBench: CompareTallyBench, CompareTempo-

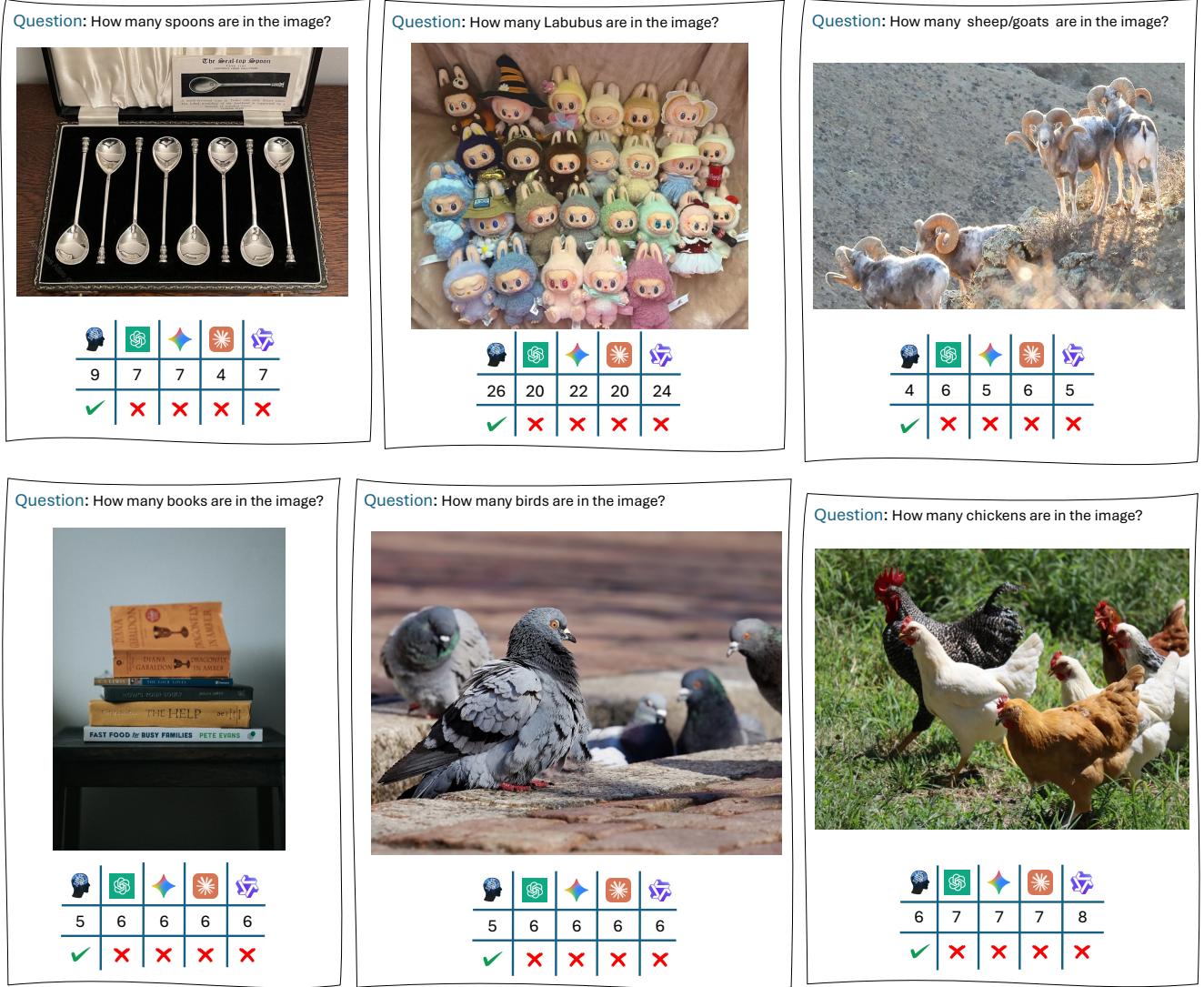


Figure 4. TallyBench hard cases where all four models fail. Each panel shows the image, the counting question (e.g., “How many spoons are in the image?”), and the predictions from four models (Claude Sonnet 4, Gemini 2.5 Pro, GPT-5, Qwen2.5-VL-72B-Instruct), all of which are incorrect. The six examples (top-left to bottom-right) cover *spoons*, *Labubu* instances, *sheep/goats*, *stacked books*, *birds*, and *chickens*. These cases illustrate typical counting failure modes, including confusing visually similar instances, missing partially occluded objects, and misreading fine-scale duplicates, despite such tasks being trivial for humans.

ralBench, CompareGeometryBench, and CompareSpatialBench. Closed-source APIs consistently outperform open-source counterparts, with Gemini 2.5 Pro achieving the highest overall accuracy and OpenAI GPT-5 following closely. The Qwen2.5-VL models show clear scaling behavior, with larger variants achieving higher scores, but still lag behind the best closed-source systems. Human performance is near 100% on most tasks but drops to 30% on CompareTemporalBench, where several large-scale models (e.g., GPT-5 and o3-pro) surpass human accuracy, reflecting the importance of prior knowledge in temporal reasoning.

Analysis. As shown in Fig. 4, even the strongest VLMs

(Claude Sonnet 4, Gemini 2.5 Pro, GPT-5, and Qwen2.5-VL-72B-Instruct) simultaneously fail on simple TallyBench cases such as spoons, books, birds, and chickens, which are trivial for humans. Similarly, Fig. 5 illustrates typical failure cases from CompareBench across geometric, spatial, quantity, and temporal reasoning. Despite progress in scaling, all four strong VLMs misinterpret tasks that require distinguishing object length versus height, or fail to classify foreground and background correctly in spatial reasoning. Counting-related tasks (TallyBench and CompareTallyBench) are relatively easier, while temporal and spatial reasoning remain the most challenging. These results indi-

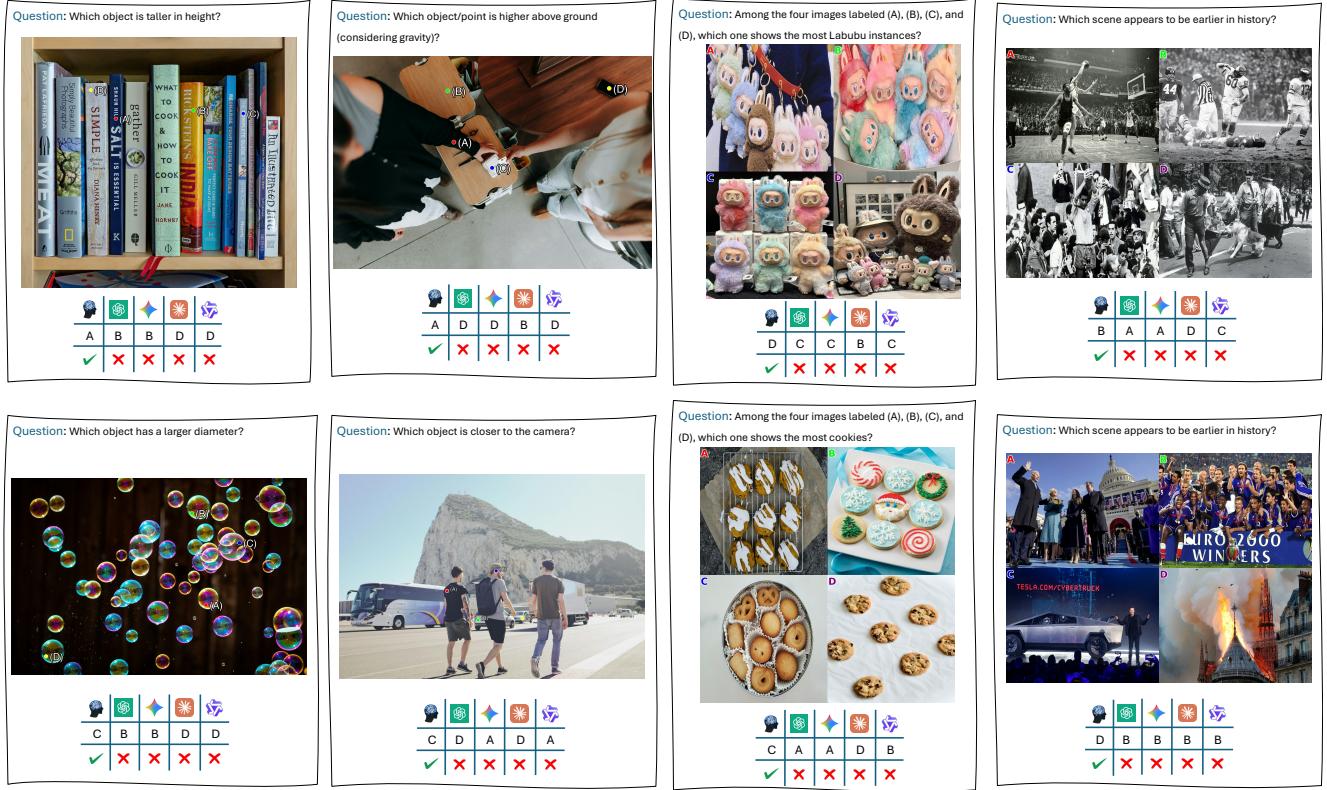


Figure 5. **Failure cases from the four CompareBench sub-benchmarks.** Each panel shows a sample question and predictions from four state-of-the-art VLMs, all of which are incorrect. From left to right on the first line: (*Geometry*) identifying the taller book; (*Spatial*) deciding which marked point is higher above ground; (*Tally*) comparing the quantity of Labubus; (*Temporary*) selecting the earliest historical scene. These cases highlight persistent weaknesses in comparative reasoning across dimensions of size, space, quantity, and time.

cate that current VLMs still lack robust comparative reasoning capabilities: they struggle not only on tasks trivial for humans but also, in certain cases such as temporal reasoning, succeed only by leveraging background knowledge that humans may not apply without external context.

A striking observation from CompareBench is the discrepancy between human and model performance on temporal reasoning. While humans achieve near-perfect accuracy on counting, geometric, and spatial comparisons, their performance drops sharply to 30% on CompareTemporalBench. This is because many temporal samples involve visually similar events where chronological ordering requires historical knowledge beyond the image itself (e.g., subtle differences in clothing, architecture, or technology). In contrast, large-scale VLMs such as Gemini 2.5 Pro and GPT-5 achieve substantially higher scores, likely by leveraging background knowledge learned from massive pretraining corpora. This divergence highlights that CompareTemporalBench does not merely test perceptual comparison but also evaluates the integration of visual evidence with external world knowledge. Consequently, it provides a unique diagnostic setting where VLMs can, in some cases, surpass

human annotators constrained to vision-only reasoning.

5. Conclusion

We introduced CompareBench, a benchmark consisting of 1,000 QA pairs spanning four fundamental tasks: quantity, temporal, geometric, and spatial comparison. These tasks are derived from TallyBench and HistCaps to evaluate visual comparison reasoning in VLMs. Our experiments on both closed- and open-source VLMs reveal clear scaling trends, but also show that all systems struggle with temporal and spatial reasoning, and even with basic counting and geometric comparison that are trivial for humans. These findings highlight visual comparison as a fundamental yet underexplored capability, exposing systematic blind spots in current multimodal systems.

Beyond evaluation, CompareBench provides a principled testbed for diagnosing both perception-driven limitations (e.g., confusion between object length and height or foreground and background) and knowledge-driven challenges (e.g., ordering visually similar historical events). While its scope is focused on core comparative reasoning rather than broader multimodal tasks, this targeted design

makes it a valuable complement to existing benchmarks. We hope CompareBench will serve as a diagnostic tool to guide the development of more robust, transparent, and trustworthy VLMs.

References

- [1] Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: Answering complex counting questions. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8076–8084, 2019. [3](#)
- [2] Niki Amini-Naieni, Kiana Amini-Naieni, Tengda Han, and Andrew Zisserman. Open-world text-specified object counting. *BMVC*, 2023. [3](#)
- [3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. [2](#)
- [4] Xianfu Cheng, Wei Zhang, Shiwei Zhang, Jian Yang, Xiangyuan Guan, Xianjie Wu, Xiang Li, Ge Zhang, Jiaheng Liu, Yuying Mai, et al. Simplevqa: Multimodal factuality evaluation for multimodal large language models. *arXiv preprint arXiv:2502.13059*, 2025. [3](#)
- [5] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer, 2024. [1, 3](#)
- [6] Google Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. [2](#)
- [7] Meng-Ru Hsieh, Yen-Liang Lin, and Winston H Hsu. Drone-based object counting by spatially regularized regional proposal network. In *Proceedings of the IEEE international conference on computer vision*, pages 4145–4153, 2017. [3](#)
- [8] Drew A Hudson and Christopher D Manning. 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*, pages 6700–6709, 2019. [2](#)
- [9] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. [1, 3](#)
- [10] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. [2](#)
- [11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. [2](#)
- [12] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. [2](#)
- [13] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. [1, 3](#)
- [14] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. [2](#)
- [15] OpenAI. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. [2](#)
- [16] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3170–3180, 2023. [3](#)
- [17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. [2](#)
- [18] Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3394–3403, 2021. [3](#)
- [19] xAI. Realworldqa: A benchmark for real-world spatial understanding in vision-language models. <https://huggingface.co/datasets/visheratin/realworldqa>, 2024. Accessed: 2025-08-25. [2](#)
- [20] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2:67–78, 2014. [2](#)
- [21] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. [1, 3](#)