VLM-SUBTLEBENCH: HOW FAR ARE VLMS FROM HUMAN-LEVEL SUBTLE COMPARATIVE REASONING?

Anonymous authors
Paper under double-blind review

ABSTRACT

The ability to distinguish subtle differences between visually similar images is essential for diverse domains such as industrial anomaly detection, medical imaging, and aerial surveillance. While comparative reasoning benchmarks for vision-language models (VLMs) have recently emerged, they primarily focus on images with large, salient differences and fail to capture the nuanced reasoning required for real-world applications. In this work, we introduce **VLM-SubtleBench**, a benchmark designed to evaluate VLMs on *subtle comparative reasoning*. Our benchmark covers ten difference types—Attribute, State, Emotion, Temporal, Spatial, Existence, Quantity, Quality, Viewpoint, and Action—and curate paired question–image sets reflecting these fine-grained variations. Unlike prior benchmarks restricted to natural image datasets, our benchmark spans diverse domains, including industrial, aerial, and medical imagery. Through extensive evaluation of both proprietary and open-source VLMs, we reveal systematic gaps between model and human performance across difference types and domains, and provide controlled analyses highlighting where VLMs' reasoning sharply deteriorates. Together, our benchmark and findings establish a foundation for advancing VLMs toward human-level comparative reasoning.

1 INTRODUCTION

The ability to discern *subtle visual differences*, i.e., minor discrepancies between otherwise similar objects, scenes, or situations, is central to human cognition. It enables us to perform a wide range of fine-grained comparative tasks, such as *recognizing micro-expression changes* in daily life (Li et al., 2022), *detecting anomalies* in manufacturing (Bergmann et al., 2019), *assessing subtle variations* in satellite imagery (Asokan & Anitha, 2019), and *distinguishing disease stages* in medical imaging (Litjens et al., 2017). Beyond everyday life, the capacity to detect and reason about such subtle differences has expanded human abilities toward higher forms of intelligence, underpinning advances from scientific discovery to complex social organization.

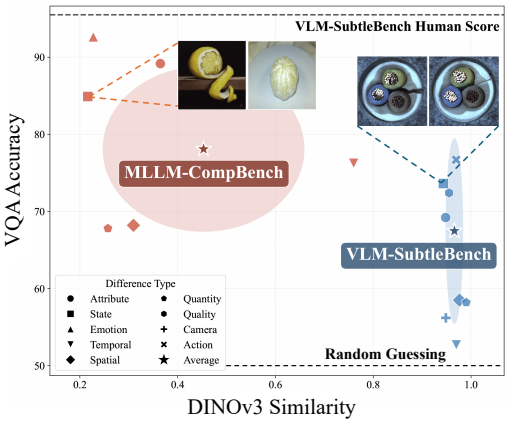


Figure 1: Comparison of VLM-SubtleBench and MLLM-CompBench with GPT-4o.

Recently, vision-language models (VLMs) have shown remarkable progress toward artificial general intelligence (AGI), showing promising results in various tasks, such as visual question answering (VQA) and scene description (Zhang et al., 2024). Yet, most progress has primarily centered on single visual inputs, e.g., an image or a video, while comparative tasks that require comparison over multiple inputs, e.g., two images, have received relatively little attention. However, narrowing this gap is increasingly important, as recent applications often deploy VLMs as agents to perform complex tasks involving comparative reasoning across multiple observations, e.g., self-reflection over previously observed scenes (Hong et al., 2024). To truly serve as human-level surrogates, advancing VLMs' capacity for advanced comparative reasoning is becoming indispensable.

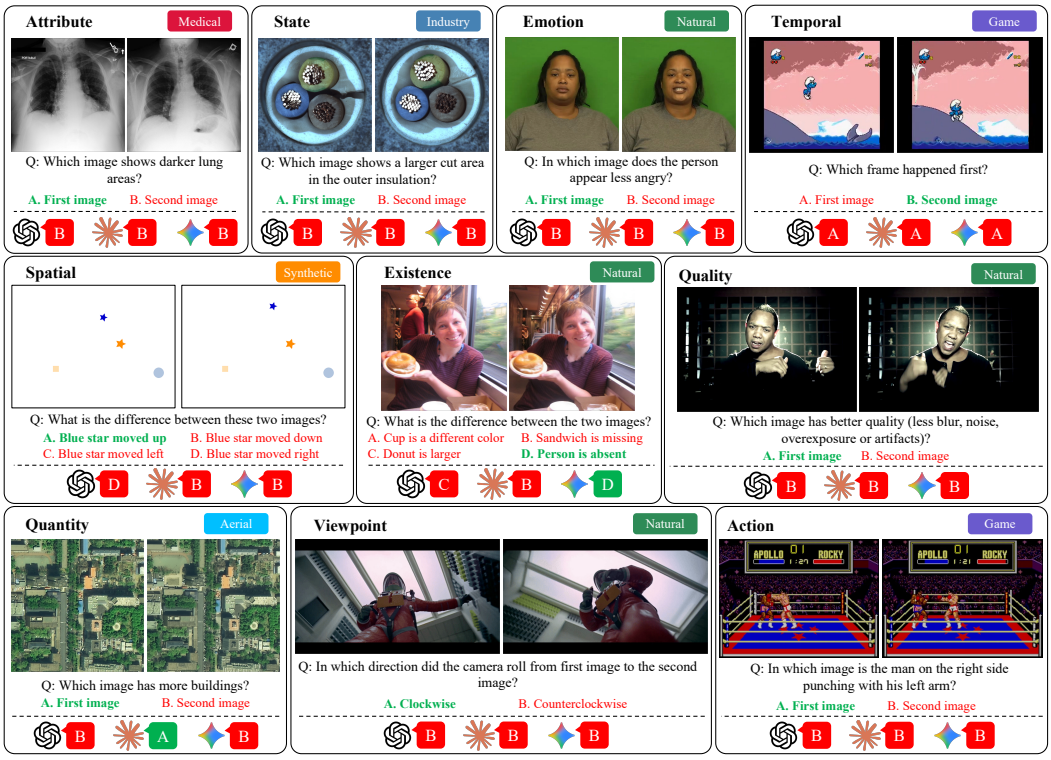


Figure 2: Example tasks from the VLM-SubtleBench, covering ten difference categories (Attribute, State, Emotion, Temporal, Spatial, Existence, Quality, Quantity, Viewpoint, Action) and six domains (natural, game, medical, industry, aerial, synthetic). For each example, the correct answer is highlighted in **bold green**. Model responses from GPT-5, Claude-Sonnet-4, and Gemini-2.5-Pro are shown beneath each question in order. Some VQA instances are simplified due to space constraints; full versions and additional examples are provided in the appendix.

A few benchmarks have been presented to evaluate the comparative reasoning capabilities of VLMs. However, as shown in Figure 1, prior benchmarks focus on relatively simple comparisons between fairly dissimilar scenes, e.g., identifying differences in salient features (states) of distinct objects (two lemons), as indicated by the low average subtleness scores of MLLM-CompBench measured by embedding similarity with DINOv3 (Siméoni et al., 2025). As a result, they are easily solved by state-of-the-art VLMs, such as GPT-4o. In addition, most prior benchmarks are composed of natural images, and thus fail to assess performance in specialized domains such as industry or medical (See Table 1). Therefore, this calls for a new benchmark that can evaluate human-level subtle comparative reasoning across diverse domains and high-difficulty tasks.

To this end, we present VLM-SubtleBench, a benchmark designed to evaluate human-level comparative reasoning capabilities of VLMs. As illustrated in Figure 2, VLM-SubtleBenchcomprises 11.7k triplets of image pairs, questions, and answers, covering 10 representative difference types, i.e., *Attribute*, *State*, *Emotion*, *Temporal*, *Spatial*, *Existence*, *Quality*, *Quantity*, *Viewpoint*, and *Action*, collected from diverse image domains, i.e., *Natural*, *Game*, *Industry*, *Aerial*, *Synthetic*, and *Medical*. Each instance presents a non-trivial challenge even for advanced proprietary VLMs such as GPT-5 and Gemini-2.5-Pro, while remaining straightforward for humans to solve.

We conduct systematic studies on the performance of open-source and proprietary VLMs, the effectiveness of test-time prompting strategies on comparative tasks, and controlled experiments with synthetic image pairs, which reveal the extent to which VLMs align with human performance and where their reasoning sharply deteriorates. Our findings show that (1) proprietary VLMs still struggle with subtle visual comparison, leaving large gaps from human performance; (2) simple prompting strategies, such as chain-of-thought prompting, grid layouts, and overlapping images, yield only limited improvements; and (3) VLMs are highly sensitive to difficulty factors such as object size and count. Together, VLM-SubtleBench and our experimental studies provide critical insights toward narrowing the gap between VLMs and humans in subtle comparative reasoning.

Table 1: **Summary of Comparative Reasoning Benchmarks.** 'Is Subtle?' indicates whether the benchmark contains image pairs whose DINOv3 similarity averages at least 0.8, meaning the differences are subtle. VLM-SubtleBench is the *only* benchmark that focuses on subtle comparison, spans diverse domains, and includes both multiple-choice questions and captioning tasks.

Benchmarks	Is Subtle?	# Domains	# Diff. Types	MCQ	Captioning
Birds-to-Words (Forbes et al., 2019)	✗	1	1	✗	✓
Spot-the-Diff (Jhamtani et al., 2018)	✓	1	1	✗	✓
MLLM-CompBench (Kil et al., 2024)	✗	1	8	✓	✗
VLM-SubtleBench (Ours)	✓	6	10	✓	✓

2 Related Work

Vision-Language Models. Vision-Language Models (VLMs) have been introduced to address the limitation of Large Language Models (LLMs), which are restricted to processing text-only data (Bai et al., 2025; Zhu et al., 2025; Li et al., 2024a). VLMs integrate pre-trained vision encoders (e.g., CLIP (Radford et al., 2021)) with LLMs via vision-to-language adaptors (Liu et al., 2023; Alayrac et al., 2022), enabling the joint processing and reasoning over visual and textual modalities. VLMs extend the applicability of LLMs to multimodal tasks, including visual question answering (Goyal et al., 2017), image captioning (Agrawal et al., 2019), and visual grounding (Chen et al., 2023).

Benchmarks for Vision-Language Models. A variety of benchmarks have been proposed to assess VLMs, primarily focusing on single-image understanding tasks such as VQA, captioning, and visual grounding (Liu et al., 2024b; Ying et al., 2024). More recently, multi-image benchmarks have emerged to examine models' abilities in cross-image comparison, relational reasoning, and integrating information across visual contexts (Kazemi et al., 2024; Zhang et al., 2025; Tong et al., 2024). Specifically, MLLM-CompBench (Kil et al., 2024) investigates comparative reasoning by asking models to judge relative properties. However, many of its comparisons rely on salient differences between images that involve different subjects or settings, and a considerable portion of its VQAs can be solved by inspecting individual images rather than comparing the two together. In contrast, our benchmark emphasizes subtle differences within nearly identical contexts, with VQAs that can only be answered by simultaneously examining both images, thus providing a more fine-grained and challenging evaluation of the comparative capabilities of VLMs. Table 1 shows the summary of comparative reasoning benchmarks for VLMs.

Difference Understanding in Classical Vision Tasks. Prior work has examined difference understanding across diverse domains. Spot-the-Diff (Jhamtani & Berg-Kirkpatrick, 2018) studied textual descriptions of fine-grained differences between surveillance frames. In the medical domain, MIMIC-Diff-VQA (Johnson et al., 2019) introduced a large-scale chest X-ray dataset for disease and difference-focused visual question answering. In remote sensing, GeoBench (Lacoste et al., 2023) provided a benchmark of classification and segmentation tasks to assess foundation models for Earth monitoring.

Beyond these domain-specific efforts, more general approaches to Image Difference Captioning (IDC) have also emerged. Img-Diff (Jiao et al., 2025) proposed an automated data synthesis pipeline that generates object replacement samples to improve MLLMs on image difference and VQA tasks. OneDiff (Hu et al., 2025) introduced a generalist model with a Visual Delta Module and a new DiffCap dataset, achieving strong performance across multiple IDC benchmarks. DiffTell (Di et al., 2025) provided a large-scale dataset covering global, object-level, and text-based differences, demonstrating significant gains in IDC performance. While these works substantially improve IDC modeling and resources, their coverage of difference types remains relatively limited compared to the wide range of variations that can occur between pairs of images.

3 VLM-SubtleBench

In this section, we introduce **VLM-SubtleBench**, a benchmark designed to evaluate subtle comparative reasoning capabilities of VLMs. VLM-SubtleBench focuses on whether models can reliably

identify **subtle differences** between two highly similar images, a key aspect of comparative visual reasoning. In the following, Section 3.1 describes the scope of the benchmark, covering the visual domains and difference categories included. Section 3.2 details the dataset curation process, and Section 3.3 explains how caption annotations were performed to ensure high-quality textual descriptions. Section 3.4 presents dataset statistics that highlight the diversity of VLM-SubtleBench.

3.1 SCOPE OF VLM-SUBTLEBENCH

Covered Image Domains. To evaluate whether a model possesses human-level subtle comparative reasoning across diverse, cognitively demanding tasks, it is essential to cover images from a wide range of domains. Thus, we design VLM-SubtleBench to span *six* representative image domains: **Natural Scenes**, capturing everyday real-world photographs (Abu-El-Haija et al., 2016; Lin et al., 2014; Souček et al., 2022; Cao et al., 2014; Livingstone & Russo, 2018; Kossaifi et al., 2017; Gupta et al., 2016; Zhou et al., 2025; Wang et al., 2024; Idrees et al., 2018; Lin et al., 2025); **Game Environments**, simulated yet realistic scenes that test generalization beyond natural images (Abu-El-Haija et al., 2016; Lin et al., 2025); **Aerial Imagery**, covering remote sensing and overhead views where subtle spatial differences are critical (Liu et al., 2024a; Huang et al., 2022); **Industrial Inspection**, representing structured settings where fine-grained defects or anomalies need to be detected (Bergmann et al., 2019; 2022); **Medical Imaging**, where diagnostic reasoning often requires distinguishing subtle changes across visits (Johnson et al., 2019; Hu et al., 2023); and **Synthetic Primitives**, consisting of abstract 2D shapes with varying colors and arrangements on plain backgrounds, which further allows controlled analysis.

Covered Difference Types. We also design VLM-SubtleBench to cover diverse type of differences. Specifically, we follow the categorization proposed in Kil et al. (2024), while extending it by adding two new types of differences, Viewpoint and Action. In total, VLM-SubtleBench encompasses *ten* difference types: **Attribute** captures variations in object properties such as color, size, or shape; **State** reflects the condition of an object, such as whether an apple is peeled; **Emotion** addresses comparative judgments of facial expressions; **Temporal** involves identifying which image corresponds to a later stage of an event; **Spatial** describes changes in arrangement or relative position; **Existence** refers to whether an object is missing; **Quantity** handles whether the number of objects differs across images; **Quality** captures degradations such as blur, noise, or overexposure; **Viewpoint** reflects changes in camera perspective; and **Action** denotes differences in human or animal poses or activities. Together, these ten categories establish a comprehensive taxonomy of subtle differences, spanning from low-level visual variations to high-level semantic changes.

3.2 DATASET CURATION

For each difference category, we curate paired images from diverse sources and generate comparative question-answer pairs through a mix of rule-based, annotation-driven, and model-assisted methods. Existing datasets with rich ground-truth labels and metadata enable systematic pairing and QA construction, while synthetic edits and primitives provide controlled settings for specific attributes. Below we briefly summarize the curation strategy for each category. Refer to appendix A for details.

Attribute. We use four sources: MVTEC-AD (Bergmann et al., 2019) for industrial inspection, COCO (Lin et al., 2014) for natural images, MIMIC-Diff-VQA (Hu et al., 2023) for medical domain, and synthetic primitives. In MVTEC-AD, pairs are formed by selecting anomalies of the same type but with different severity, using pixel-level defect annotations to order defect size. In COCO, we apply image editing model, namely *gemini-2.5-flash-image-preview* (also known as "nano-banana") to change object colors, producing minimally different pairs. For primitives, we render shapes on a blank canvas with controlled variations in size or color. Question-answers (QAs) then ask which image has a larger defect, which object's color is stronger, or which shape is bigger. For the medical domain, we leverage MIMIC-Diff-VQA dataset, which provides chest X-ray pairs from different visits of the same patient and comparative questions about clinical changes. Since our benchmark focuses on models' ability to detect visually perceptible differences independent of medical knowledge, we reformulated the original clinical questions into layperson-friendly forms (e.g., converting "lung opacity progression" into "Which image shows whiter lungs?"). The reformulation was performed using GPT-4o, and annotators verified the consistency between questions and images. For synthetic data, we render primitive scenes containing objects with distinctive shapes and colors.

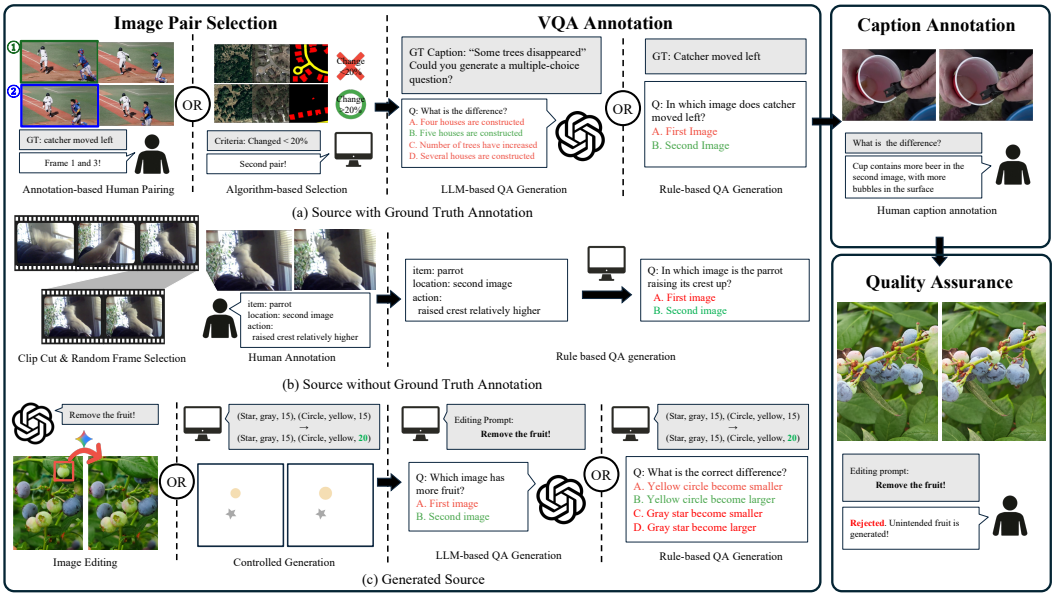


Figure 3: Data Curation Pipeline of VLM-SubtleBench.

We then select one object and apply a transformation, either brightness adjustment (brightening or darkening) or size modification (increasing or decreasing). The resulting QA pairs ask what change occurred and to which object.

State. From MVTEC-AD, we focus on object breakage or cracks. Akin to attribute, image pairs are constructed by sampling images with different levels of damage, with annotations guiding the relative severity. The resulting QAs ask which crack is larger or which state shows more breakage. We also include natural domain pairs from ChangeIt (Souček et al., 2022), where human annotators manually annotated the state-modifying action together with the object states in a set of internet videos. From each video, we sample multiple frames based on the provided state information, and human annotators then select frame pairs that capture the object before and after the state-modifying action. The QA pairs then ask which image reflects a greater degree of state modification. For example, if the object is an apple and the action is peeling, the QA would ask "In which image is the apple peeled more?".

Emotion. We draw images from emotional video datasets—CREMA-D, RAVDESS, AFEW-VA, and DAiSEE (Cao et al., 2014; Livingstone & Russo, 2018; Kossaifi et al., 2017; Gupta et al., 2016)—all of which provide clip-level emotion annotations. From these clips, we randomly sample frames and construct paired examples based on the relative intensity of expressed emotion. The QA pairs ask which image conveys a stronger or weaker emotion, based on the annotations.

Temporal. From YT8M (Abu-El-Haija et al., 2016) and VLM4D (Zhou et al., 2025) video datasets, we randomly select two frames from the same clip. Their temporal order is determined by timestamps, and QA pairs ask which image depicts the earlier event. This task requires models to capture temporal progression rather than static differences, sometimes relying on commonsense knowledge (e.g., a boat cutting through water can only move forward, not backward).

Spatial. We use VLM4D, which provides 4D annotations of the object's translational and rotational motions in video. From each video, we uniformly sample multiple frames. Human annotators then select frame pairs that visually align with the ground-truth motion annotations. Using these pairs and their associated motion data, we generate QA that ask the spatial changes resulting from the motion. The model is required to identify which of these six transformations the object has undergone.

Existence. For aerial imagery, LEVIR-MCI (Liu et al., 2024a) provides image pairs of the same location across years, along with segmentation maps capturing object-level appearance or disap-

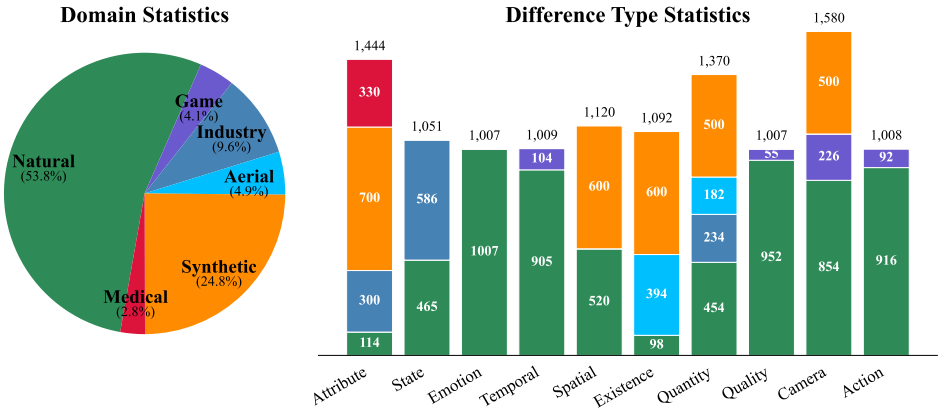


Figure 4: Statistics of the VLM-SubtleBench Test Split.

pearance and human-labeled difference captions. We also construct synthetic settings by rendering primitives and removing one object to create a pair. Additionally, COCO is used to remove objects from natural images, similar to attribute. For synthetic data, we render primitive scenes containing multiple types of shapes and colors, and then add or remove one instance. In all cases, QAs ask what has disappeared or appeared between the two images, grounded in annotations or edit prompts.

Quantity. We combine multiple domains to capture differences in object counts. MVTEC-LOCO (Bergmann et al., 2022) provides annotated anomaly and normal images, enabling comparisons of object multiplicity. UCF-QNRF (Idrees et al., 2018) offers crowded street scenes with dense human annotations, from which we sample pairs with count differences. Aerial imagery datasets (LEVIR-MCI, UBC (Huang et al., 2022)) allow comparisons of building counts. Synthetic datasets and Gemini-edited images (MegaFruit (Wang et al., 2024)) further introduce controlled object additions. QAs consistently ask which image contains more objects, people, or buildings. For synthetic data, we render primitive scenes containing a single type of shape and color, and then create quantity differences by adding or removing one instance of the shape. The QA pairs then ask which image contains more (or fewer) shapes.

Quality. From YT8M, we randomly sample two frames that are temporally proximate within the same video. When the two frames differ in visual quality, e.g., blur, noise, overexposure, or artifacts, human annotators label which image is of higher or lower quality. Based on these annotations, we construct multiple-choice QA pairs that ask which image has better or worse quality.

Viewpoint. CameraBench (Lin et al., 2025) provides camera-centric and object-centric annotations describing translations and rotations. From each video, we uniformly sample frames, and human annotators then select pairs of frames that are visually aligned with ground-truth camera annotations (e.g., sufficient visual cues for viewpoint change). Based on these pairs, we construct binary QA tasks: for example, if the camera rotated to the right, the question asks "In which direction did the camera move?" with answer choices right or left. Similarly, for object-centric annotations, QAs ask whether the camera orbited an object clockwise or counterclockwise.

In case of synthetic data, we render primitive scenes and simulate camera motion by translating the objects or by rotating them clockwise or counterclockwise around the center. From the six possible transformations, we randomly sample four options to form multiple-choice QAs, where the model must identify the ground-truth motion.

Action. Finally, from YT8M, we sample pairs of frames that capture changes in human or object actions and interactions. Human annotators identify objects that exhibit different actions in the two frames and provide the corresponding action labels. Based on these annotations, we generate QA pairs that compare the type or intensity of the actions, thereby extending comparative reasoning beyond static attributes.

3.3 Difference Captions

In practical scenarios across diverse VLM application domains, the ability to directly describe differences between two images is crucial. To enable such forms of evaluation, we additionally collected human-written captions explicitly highlighting the differences between paired images, thereby complementing the comparative QA tasks. We sampled 1,200 random image pairs (10% of the test split) for human caption annotation. Annotators were instructed to identify at least one difference between the two images and to write a comparative caption using the annotation interface.

3.4 Dataset Statistics

Each difference category contains at least **1K** question–answer pairs, resulting in a total of **13K pairs**. Every difference type includes data from the natural domain, where foundation models are most commonly applied in practice. We split the dataset into a test set (11.7K) and a validation set (1.3K). The test set is used for evaluation, while the validation set is used for fine-tuning models in our experiments. Figure 4 presents the statistics of VLM-SubtleBench test set.

4 Experiment

4.1 Experiment Setup

Models. We evaluated both open-source and proprietary vision–language models. For the open-source side, We used the Qwen2.5-VL (Bai et al., 2025) family at three scales (7B, 32B, and 72B), as well as LLaVA-NeXT and LLaVA-OneVision (both 7B). For proprietary models, we included GPT-4o (Achiam et al., 2023), o3, GPT-5-main, GPT-5-thinking, Claude Sonnet 4 (Anthropic, 2025), Gemini-2.5-Flash (Team et al., 2023), and Gemini-2.5-Pro. This set spans both non-reasoning models (e.g., GPT-4o, GPT-5-main) and reasoning-oriented models (e.g., o3, GPT-5-thinking).

Prompting Strategies. To better understand the role of prompting, we experimented with several strategies. We adopted the standard Chain-of-Thought (CoT) approach, which encourages models to generate intermediate reasoning before producing final answers (Wei et al., 2022). We further introduce a two-step reasoning setup in which the VLM generates responses in two stages. In the first step, the VLM is prompted to describe the differences between the two images that are relevant to the question; in the second stage, the two images, the question, and the output from the first step are provided together to answer the question. We also augmented images with a grid and instructed the models to parse them sequentially along the horizontal axis (Izadi et al., 2025). To investigate how models handle multiple images, we tested different fusion techniques: (i) horizontally concatenating the two images into a single composite input, (ii) creating an overlap image by averaging the pixel values of the two input images and using it together with the original images, (iii) generating a grayscale subtraction image by computing the absolute pixel-wise difference, normalizing it by the maximum value to highlight regions of change, and providing it along with the original images, and (iv) highlighting regions of interest by retaining pixels with large differences, clustering adjacent pixels, and drawing bounding boxes around at most three of the largest clusters to emphasize the main regions of change. Further details of these prompting techniques and their comparative results are provided in Appendix B.2.

Evaluation Metric. We used task-appropriate metrics to evaluate model performance. For multiple-choice questions, performance was measured by accuracy, capturing the proportion of correct answers. For the captioning task, we applied CSS (Reimers & Gurevych, 2019) and LLM-as-a-judge (Zheng et al., 2023), to assess the quality and relevance of generated captions.

4.2 Benchmark Results

Multiple-Choice Questions. Table 2 summarizes the performance of proprietary and open-source VLMs on VLM-SubtleBench. Among proprietary models, GPT-5-thinking achieves the strongest results overall, ranking first in 7 out of the 10 difference types and yielding the highest average accuracy. Other reasoning-oriented models such as o3 and Gemini also show strong performance, highlighting the advantage of models explicitly designed for reasoning. Within the open-source models, Qwen2.5-VL-72B stands out with competitive accuracy, in some cases approaching that of

Table 2: Performance of open-source and proprietary vision-language models in VLM-SubtleBench. Human evaluation was conducted on a randomly selected 10% of the samples.

Model	AT	ST	EM	TM	SP	EX	QN	QL	VP	AC	AVG
Random Guess	35.9	50.0	50.0	50.0	36.6	23.2	48.9	50.0	42.1	50.0	43.3
Human Eval	92.0	93.0	93.0	93.0	95.0	97.0	97.0	99.0	98.0	98.0	95.5
Open-source											
LLaVA-NeXT-7B	37.0	51.3	51.8	47.4	37.3	25.6	49.5	48.0	43.7	46.9	43.6
LLaVA-OneVision-7B	41.6	56.8	73.9	48.7	35.5	44.2	54.9	62.7	49.1	60.5	52.0
Qwen2.5-VL-7B	46.5	63.7	87.8	50.2	39.5	73.8	58.0	70.9	47.5	69.3	59.4
Qwen2.5-VL-32B	48.3	64.0	85.3	50.4	43.6	84.2	67.5	72.5	47.4	72.0	62.2
Qwen2.5-VL-72B	53.9	68.9	85.9	49.9	47.8	81.7	67.7	78.4	56.2	74.1	65.4
Proprietary											
GPT-4o	51.5	73.6	89.5	52.7	42.4	60.6	58.2	72.4	51.4	76.7	61.6
o3	78.0	79.5	92.9	**60.4**	55.1	82.2	78.2	**87.6**	64.6	84.8	75.7
GPT-5-main	72.9	78.4	92.7	53.6	50.1	75.4	72.6	84.5	57.5	83.6	71.3
GPT-5-thinking	**83.6**	**80.7**	**93.1**	60.2	**59.9**	85.4	**79.9**	84.8	**68.5**	**84.9**	**77.8**
Claude-sonnet-4	48.9	64.7	83.3	49.3	48.7	**87.5**	63.1	70.8	53.5	66.3	62.6
gemini-2.5-flash	49.3	72.5	88.4	53.9	40.7	73.2	60.0	77.1	51.8	72.3	62.5
gemini-2.5-pro	55.3	76.4	89.8	57.6	44.8	79.9	68.0	84.8	60.3	76.8	68.2

proprietary systems. Among the 7B-scale models, Qwen2.5-VL-7B achieved the highest accuracy, followed by LLaVA-OneVision-7B, while LLaVA-NeXT-7B showed lowest performance.

Across different difference types, VLMs show strong performance on emotion, with GPT-5-thinking achieving 93.1% accuracy. In contrast, all models perform weakly on temporal, spatial, and viewpoint differences, which require common-sense reasoning (e.g., predicting the future position of a person or distinguishing between object and camera motion) and spatial understanding. These findings underscore the need for VLMs to incorporate richer spatial–temporal representations.

Captioning. Table 3 presents the performance of VLMs on VLM-SubtleBench's captioning task. Similar to VQA tasks, GPT-5-thinking achieved the strongest overall performance across all metrics. However, when captions were evaluated with the LLM-as-a-judge metric, its accuracy reached only 43.0%, leaving a noticeable gap compared to ground-truth captions. Large open-source VLMs, such as Qwen2.5-VL-32B/72B, performed on par with proprietary models in terms of CSS score, but exhibited substantially lower performance under the LLM-as-a-judge evaluation. In particular, Qwen2.5-VL-72B scored 24.3, which is significantly behind GPT-5-thinking's 43.0.

4.3 Effect of Prompting

Table 4 reports the effect of different prompting strategies. Adding reasoning steps before the answer improved performance in 9 out of 10 domains. While such gains are intuitive in tasks like *temporal* that require world knowledge, it is particularly interesting that reasoning also boosts performance in tasks such as *attribute* and *quality*, where success hinges on capturing fine-grained visual differences. This suggests that explicit textual reasoning supports not only abstract inference but also subtle perceptual discrimination, consistent with our main finding that models with stronger inherent reasoning achieve higher accuracy. In contrast, the two-step reasoning approach leads to a slight decrease in performance. We observe that the model frequently produces intermediate descriptions indicating "no difference" in the first stage, which results in incorrect final predictions. The highlighting method yields a modest improvement in performance. It is particularly effective on datasets with limited variations (e.g., synthetic data); however, its performance declines on datasets exhibiting substantial variations in brightness or image quality (e.g., YT8M), where bounding boxes often fail to accurately localize regions of change.

Other strategies generally led to performance drops. In particular, concatenating two images into a single input, a common setup in prior work (Kil et al., 2024; Jiao et al., 2025), degraded accuracy in 9 out of 10 domains. Overlap and subtract showed mixed effects: they yielded clear gains in spatial and existence tasks where only objects change under fixed views, and in viewpoint tasks where the scene is mostly static with camera movements. However, in other domains these strategies provided little or no benefit, reflecting their dependence on highlighting layout differences between images.

Table 3: Performance of open-source and proprietary vision-language models in VLM-SubtleBench captioning.

Model	CSS	LLM-Judge
Open-source		
Qwen2.5-VL-7B	0.42	12.7
Qwen2.5-VL-32B	0.52	22.7
Qwen2.5-VL-72B	0.54	24.3
Proprietary		
GPT-4o	0.54	25.2
o3	0.56	38.1
GPT-5-main	**0.57**	37.2
GPT-5-thinking	**0.57**	**43.0**
Claude-sonnet-4	0.54	24.8
gemini-2.5-flash	0.50	25.9
gemini-2.5-pro	0.52	29.4

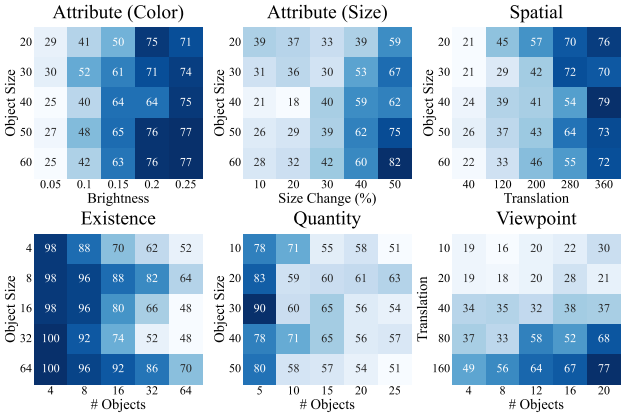


Figure 5: Performance heatmaps of GPT-4o on synthetic data under controlled difficulty factors.

Table 4: Effect of prompting strategies and fine-tuning in VLM-SubtleBench.

Model	AT	ST	EM	TM	SP	EX	QN	QL	VP	AC	AVG
Random Guess	35.9	50.0	50.0	50.0	36.6	23.2	48.9	50.0	42.1	50.0	43.3
Prompting Strategies											
GPT-5-main	72.9	78.4	92.7	53.6	50.1	75.4	72.6	84.5	57.5	83.6	71.3
+ Reasoning	**76.5**	**79.1**	91.2	56.1	51.6	80.2	**75.8**	**86.1**	57.0	**85.4**	**73.1**
+ Two-Step Reasoning	70.8	**79.1**	**93.4**	**56.9**	47.4	81.3	66.4	83.5	58.0	83.6	71.0
+ Grid	71.6	77.5	89.1	52.8	51.0	75.8	72.5	82.6	57.2	84.2	70.6
+ Concat	70.3	77.9	92.2	51.6	44.6	75.5	64.8	81.2	52.3	82.4	68.2
+ Overlap	69.5	76.7	91.8	52.4	53.6	76.1	69.1	79.0	58.8	83.0	70.2
+ Subtract	73.8	76.4	91.8	50.9	**55.4**	78.0	69.8	80.1	**60.0**	82.0	71.2
+ Highlight	71.1	75.2	92.0	51.1	54.9	**86.5**	74.3	77.8	57.3	82.9	71.5
Fine-Tuning											
Qwen-2.5-VL-7B	46.5	63.7	87.8	50.2	39.5	73.8	58.0	70.9	47.5	69.3	59.4
+ fine-tuned	**62.0**	**69.1**	**92.2**	**52.5**	**47.0**	**85.3**	**77.0**	**85.9**	**57.5**	**75.4**	**69.5**

4.4 Controlled Evaluation with Synthetic Data

Setup. We leverage synthetic data generation to systematically manipulate task difficulty, which allows precise control over the factors that may cause VLMs to fail. For each difference type, we selected two primary factors that strongly influence difficulty and varied them along a controlled axis. Specifically, for *attribute*, we considered the size of changed objects and the magnitude of variation (brightness shifts in [0, 1] for color, or size-change ratios for scale). For *spatial*, we manipulated object size and the degree of translation. For *existence* and *quantity*, the two axes were object size and scene complexity, defined as the total number of objects. For *viewpoint*, we varied camera translation and scene complexity. For each configuration, we generated 100 paired images to probe VLM performance. Evaluation was performed using GPT-4o. Notably, for *quantity* differences, random guessing achieves a 50%, while for the other categories the baseline is 25%.

Results. Our experiments reveal distinct failure modes across difference types. For *attribute (color)*, the decisive factor is the magnitude of brightness change: the model requires shifts of roughly 25% to show strong performance above 70%, while smaller changes, e.g., 5%, lead to random-like performance. In the *attribute (size)* condition, the model depends more on the absolute size of the changed object than on relative scaling, achieving reliable accuracy only when large objects undergo substantial transformations. For *spatial* differences, accuracy is largely influenced by both translation and object size, with the model responding more strongly to relative displacement than absolute pixel shifts; notably, smaller objects moving larger relative distances are easier to detect. In the *existence* setting, scene complexity emerges as the dominant factor: accuracy is nearly perfect with four or fewer objects but rapidly degrades to below 60% once scenes exceed 32 objects, though larger disappearing objects remain easier to track. Similarly, in *quantity*, performance remains high nearly 80% in simple scenes with 5 objects, but drops to close to 60% in complex scene

Table 5: Rank correlation of our benchmark and MLLM-CompBench with MMAD and QAG.

	MMAD	QAG
VLM-SubtleBench	**0.8424**	**0.7212**
MLLM-CompBench	0.8110	0.7195

Table 6: Downstream performance after fine-tuning on each benchmark.

	MMAD	QAG
Qwen2.5-VL-7B	65.0	34.4
+ VLM-SubtleBench	**69.6**	**35.5**
+ MLLM-CompBench	66.3	32.2

containing ten or more objects, approaching the 50% random baseline. Finally, for *viewpoint*, the model shows an interesting opposite trend: performance improves as scene complexity increases, benefiting from richer visual cues, and stable accuracy requires camera translations of around 160 pixels (which is 27% of the image height).

4.5 Effect of Fine-tuning

To evaluate whether additional supervision can mitigate the comparative reasoning challenge, we fine-tune Qwen2.5-VL-7B using the validation set. Table 4 presents the results. Fine-tuning yields consistent performance improvements across all difference types, with particularly notable gains in existence, quantity, and quality categories. In contrast, the spatial and temporal dimensions showed more modest gains, suggesting that richer spatial–temporal reasoning rather than in-distribution adaptation may be required for further progress. Despite these improvements, a substantial gap remains compared to GPT-5-thinking and human performance, indicating that broader data diversity and advanced training method could offer promising future directions beyond the present scope. Details on fine-tuning are provided in Appendix B.4.

4.6 Real-world Relevance Analysis

We assess the real-world relevance of VLM-SubtleBench through correlation and transfer studies on industrial anomaly detection (MMAD (Jiang et al., 2025)) and aerial surveillance (QAG-360k (Li et al., 2024b)), both requiring fine-grained visual discrimination. Table 5 reports the rank correlations (Spearman, 1904) between each benchmark and the downstream tasks. Across models, our benchmark shows higher rank correlations with MMAD and QAG-360K than MLLM-CompBench, suggesting that it better captures the comparative cues underlying downstream performance. Model-wise results are summarized in Appendix C.1.

To evaluate practical transfer, we fine-tune Qwen2.5-VL-7B on a validation split of VLM-SubtleBench and compare it with an equally sized subset of MLLM-CompBench. Table 6 reports the resulting downstream accuracies on MMAD and QAG-360K. Fine-tuning on VLM-SubtleBench consistently yields larger gains on both application benchmarks, whereas fine-tuning on MLLM-CompBench provides limited or even negative transfer. These results indicate that the subtle, fine-grained difference types in VLM-SubtleBench more effectively encode cues required for real-world perceptual reasoning.

5 Conclusion

In this paper, we introduce VLM-SubtleBench, a benchmark for evaluating *subtle comparative reasoning* in vision–language models. VLM-SubtleBench deliberately focuses on *pairs with subtle changes*—those that humans can spot but are highly challenging for current VLMs. The dataset comprises both VQA and captioning instances spanning ten difference types (attribute, state, emotion, temporal, spatial, existence, quantity, quality, viewpoint, and action) and six domains (natural, game, industrial, aerial, medical, synthetic), thus covering both everyday and specialized settings. Our evaluation of proprietary and open-source models shows that even state-of-the-art proprietary systems struggle on VLM-SubtleBench. We further find that explicit reasoning can enhance comparative performance. Controlled studies with synthetic pairs reveal consistent failure modes and sensitivities, highlighting the challenges posed by subtle differences. Together, these findings position VLM-SubtleBench as both a rigorous benchmark for measuring subtle comparative reasoning and a diagnostic tool that reveals where current VLMs fall short, offering valuable insights for guiding future model development and dataset design.

REFERENCES

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark, 2016. URL https://arxiv.org/abs/1609.08675.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/960a172bc7fbf0177ccccbb411a7d800-Paper-Conference.pdf.

Anthropic. Claude 3.7 sonnet: Our most capable model yet. https://www.anthropic.com/news/claude-3-7-sonnet, 2025. Accessed: 2025-09-08.

Anju Asokan and JJESI Anitha. Change detection techniques for remote sensing applications: A survey. *Earth Science Informatics*, 12(2):143–160, 2019.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9592–9600, 2019.

Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. *International Journal of Computer Vision*, 130(4):947–969, 2022.

Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic, 2023. URL https://arxiv.org/abs/2306.15195.

Zonglin Di, Jing Shi, Yifei Fan, Hao Tan, Alexander Black, John Collomosse, and Yang Liu. Difftell: A comprehensive dataset for image difference captioning, 2025. URL https://openreview.net/forum?id=86uYj8DcfK.

Maxwell Forbes, Christine Kaeser-Chen, Piyush Sharma, and Serge Belongie. Neural naturalist: Generating fine-grained image comparisons. *arXiv preprint arXiv:1909.04101*, 2019.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Abhay Gupta, Arjun D'Cunha, Kamal Awasthi, and Vineeth Balasubramanian. Daisee: Towards user engagement recognition in the wild. *arXiv preprint arXiv:1609.01885*, 2016.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. Cogagent: A visual language model for gui agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14281–14290, 2024.

Erdong Hu, Longteng Guo, Tongtian Yue, Zijia Zhao, Shuning Xue, and Jing Liu. Onediff: A generalist model for image difference captioning, 2025. URL https://arxiv.org/abs/2407.05645.

Xinyue Hu, Lin Gu, Qiyuan An, Mengliang Zhang, Liangchen Liu, Kazuma Kobayashi, Tatsuya Harada, Ronald M Summers, and Yingying Zhu. Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4156–4165, 2023.

Xingliang Huang, Libo Ren, Chenglong Liu, Yixuan Wang, Hongfeng Yu, Michael Schmitt, Ronny Hänsch, Xian Sun, Hai Huang, and Helmut Mayer. Urban building classification (ubc)-a dataset for individual building detection and classification from satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1413–1421, 2022.

Nam Hyeon-Woo, Moon Ye-Bin, Wonseok Choi, Lee Hyun, and Tae-Hyun Oh. Vlm's eye examination: Instruct and inspect visual competency of vision language models, 2024. URL https://arxiv.org/abs/2409.14759.

Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 532–546, 2018.

Amirmohammad Izadi, Mohammad Ali Banayeeanzade, Fatemeh Askari, Ali Rahimiakbar, Mohammad Mahdi Vahedi, Hosein Hasani, and Mahdieh Soleymani Baghshah. Visual structures helps visual reasoning: Addressing the binding problem in vlms. *arXiv preprint arXiv:2506.22146*, 2025.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. *arXiv preprint arXiv:1808.10584*, 2018.

Xi Jiang, Jian Li, Hanqiu Deng, Yong Liu, Bin-Bin Gao, Yifeng Zhou, Jialin Li, Chengjie Wang, and Feng Zheng. Mmad: A comprehensive benchmark for multimodal large language models in industrial anomaly detection, 2025. URL https://arxiv.org/abs/2410.09453.

Qirui Jiao, Daoyuan Chen, Yilun Huang, Bolin Ding, Yaliang Li, and Ying Shen. Img-diff: Contrastive data synthesis for multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9296–9307, 2025.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

Mehran Kazemi, Nishanth Dikkala, Ankit Anand, Petar Devic, Ishita Dasgupta, Fangyu Liu, Bahare Fatemi, Pranjal Awasthi, Dee Guo, Sreenivas Gollapudi, and Ahmed Qureshi. Remi: A dataset for reasoning with multiple images, 2024. URL https://arxiv.org/abs/2406.09175.

Jihyung Kil, Zheda Mai, Justin Lee, Arpita Chowdhury, Zihe Wang, Kerrie Cheng, Lemeng Wang, Ye Liu, and Wei-Lun Harry Chao. Mllm-compbench: A comparative reasoning benchmark for multimodal llms. *Advances in Neural Information Processing Systems*, 37:28798–28827, 2024.

Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and Maja Pantic. Afew-va database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65:23–36, 2017.

Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan David Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Andrew Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, Mehmet Gunturkun, Gabriel Huang, David Vazquez, Dava Newman, Yoshua Bengio, Stefano Ermon, and Xiao Xiang Zhu. Geo-bench: Toward foundation models for earth monitoring, 2023. URL https://arxiv.org/abs/2306.03831.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024a. URL https://arxiv.org/abs/2408.03326.

Ke Li, Fuyu Dong, Di Wang, Shaofeng Li, Quan Wang, Xinbo Gao, and Tat-Seng Chua. Show me what and where has changed? question answering and grounding for remote sensing change detection, 2024b. URL https://arxiv.org/abs/2410.23828.

Yante Li, Jinsheng Wei, Yang Liu, Janne Kauttonen, and Guoying Zhao. Deep learning for micro-expression recognition: A survey. *IEEE Transactions on Affective Computing*, 13(4):2028–2046, 2022.

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*, pp. 740–755. Springer, 2014.

Zhiqiu Lin, Siyuan Cen, Daniel Jiang, Jay Karhade, Hewei Wang, Chancharik Mitra, Tiffany Ling, Yuhan Huang, Sifan Liu, Mingyu Chen, et al. Towards understanding camera motions in any video. *arXiv preprint arXiv:2504.15376*, 2025.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42: 60–88, 2017.

Chenyang Liu, Keyan Chen, Haotian Zhang, Zipeng Qi, Zhengxia Zou, and Zhenwei Shi. Change-agent: Towards interactive comprehensive remote sensing change interpretation and analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

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*, pp. 216–233. Springer, 2024b.

Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Nils Reimers and Iryna Gurevych. Sentence-bert: sentence embeddings using siamese bert-networks. pp. 3982–3992, 2019.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. URL https://arxiv.org/abs/2508.10104.

Tomáš Souček, Jean-Baptiste Alayrac, Antoine Miech, Ivan Laptev, and Josef Sivic. Look for the change: Learning object states and state-modifying actions from untrimmed web videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13956–13966, 2022.

C. Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15(1):72–101, 1904. doi: 10.2307/1412159.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024.

Yanan Wang, Zhenghao Fei, Ruichen Li, and Yibin Ying. Learn from foundation model: Fruit detection model without manual annotation. *arXiv preprint arXiv:2411.16196*, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, Jiayi Lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi, 2024. URL https://arxiv.org/abs/2404.16006.

Jianshu Zhang, Dongyu Yao, Renjie Pi, Paul Pu Liang, and Yi R. Fung. Vlm2-bench: A closer look at how well vlms implicitly link explicit matching visual cues, 2025. URL https://arxiv.org/abs/2502.12084.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Shijie Zhou, Alexander Vilesov, Xuehai He, Ziyu Wan, Shuwang Zhang, Aditya Nagachandra, Di Chang, Dongdong Chen, Xin Eric Wang, and Achuta Kadambi. Vlm4d: Towards spatiotemporal awareness in vision language models. *arXiv preprint arXiv:2508.02095*, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479.

A Benchmark Curation Detail

A.1 Attribute

MVTEC-AD (Bergmann et al., 2019) data was used to construct the attribute dataset. We compared images containing anomalies and selected pairs with small cosine similarity differences. Question–answer pairs were then generated by comparing the relative size of the anomalies in the images.

COCO (Lin et al., 2014) data was also used. We modified attributes related to color or size using nano-banana, and generated question–answer pairs by paraphrasing the changes applied to the images.

MIMIC (Johnson et al., 2019), a large-scale medical dataset, was also utilized. Since the original captions contain advanced medical terminology that is difficult for non-experts to understand, we simplified them into accessible language (e.g., whether the chest region appears darker).

A.2 State

MVTEC-AD pairs constructed in Section A.1 were also used for state differences, since some anomalies pertain to state rather than attribute.

ChangeIt (Souček et al., 2022) contains annotations of state-modifying actions and resulting object states in internet videos. A subset of the videos is manually annotated, and we use this portion of the data. From each video, we sample multiple frames corresponding to either state1, action, or state2. Based on object and action, we automatically generate questions. For example, if the object is an apple and the action is peeling, the question becomes: In which image is the apple peeled more? Annotators then select frame pairs in which the state-modifying action is more evident in the second frame (Figure 6).

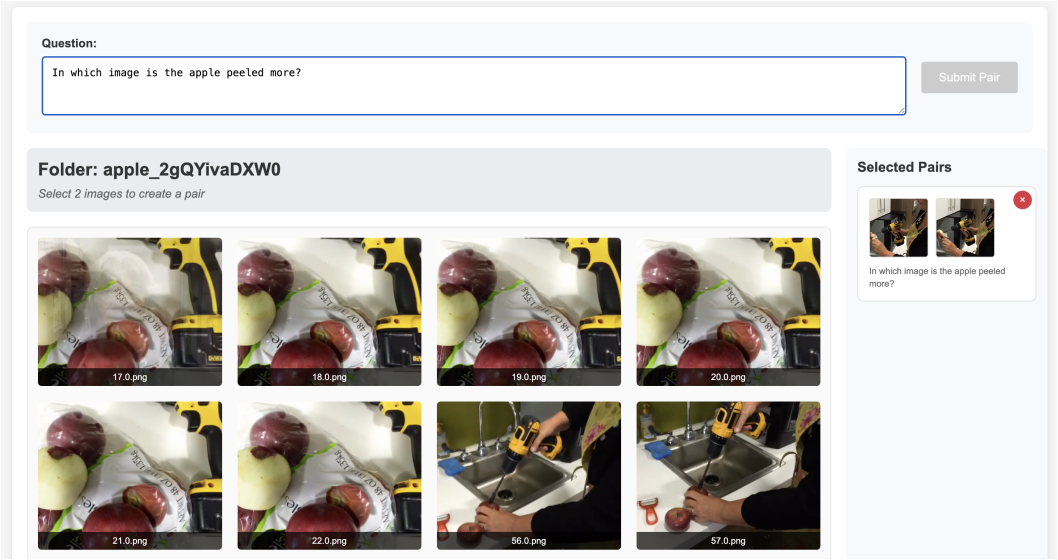


Figure 6: Annotation interface for ChangeIt.

A.3 Emotion

CREMA-D, RAVDESS, AFEW-VA, and DAiSEE (Cao et al., 2014; Livingstone & Russo, 2018; Kossaifi et al., 2017; Gupta et al., 2016) are clip-level video datasets annotated for emotion. Specifically, CREMA-D and RAVDESS consist of actors speaking sentences with specified emotions and intensity levels. AFEW-VA contains movie clips annotated with valence and arousal, while DAiSEE provides short video snippets capturing users' emotions. From these clips, we randomly sample frames and construct paired examples based on the relative intensity of expressed emotions. For

question generation, we use the labeled emotion categories; in the case of AFEW-VA, which lacks explicit emotion labels, annotators are asked to choose the most appropriate emotion depicted in the scene.

A.4 TEMPORAL

VLM4D (Zhou et al., 2025) is a benchmark designed to evaluate the spatiotemporal reasoning capabilities of video LLMs. It provides video clips depicting exocentric and egocentric movements of people, animals, or objects, along with VQA tasks about their motion. We adapt these VQA questions into comparative form using GPT-4o. For example, if the original question is "Which direction is the horse moving towards?" with the answer "left", we reformulate it as "In which image is the horse relatively in a left position?" Annotators then select the frame pairs that reflect the corresponding spatial change. In some cases, such as a person riding a bicycle or running, the movement cannot be reversed; in these instances, annotators mark the sample as temporal, and it is included as part of the temporal pairs.

YT8M (Abu-El-Haija et al., 2016) contains a diverse collection of YouTube videos spanning numerous domains. For each video, we segment clips using histogram matching and construct image pairs by randomly selecting frames with cosine similarity below 0.99, thereby avoiding frame extraction from frozen screens. Annotators then select frame pairs whose order cannot be reversed, marking them as temporal pairs (Figure 7).

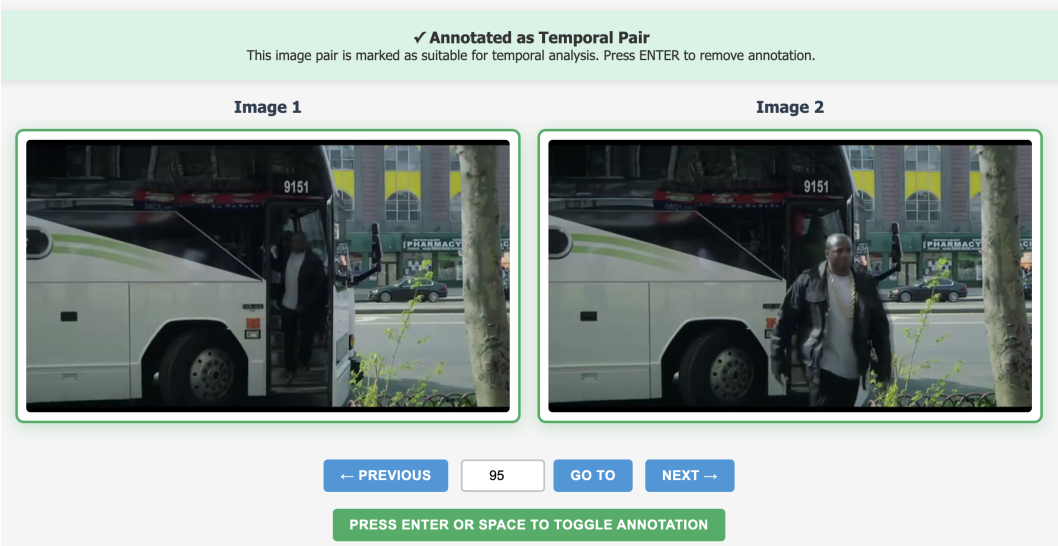


Figure 7: Annotation interface for YT8M temporal pairs.

A.5 SPATIAL

VLM4D is used to construct spatial data. We collect pairs of frames and corresponding questions in the same manner as in the Section A.4. However, in this case, reversible pairs are also used.

A.6 EXISTENCE

LEVIR-MCI (Liu et al., 2024a) is an aerial dataset for image change detection. A subset of the data includes captions describing changes within the same region. We construct image pairs in which less than 20% of the total area has changed, further filtering for pairs with cosine similarity greater than 0.8. GPT-4o is then employed to paraphrase the captions in order to generate answers and distractors for the before and after images, and to determine whether each question pertains to existence or quantity.

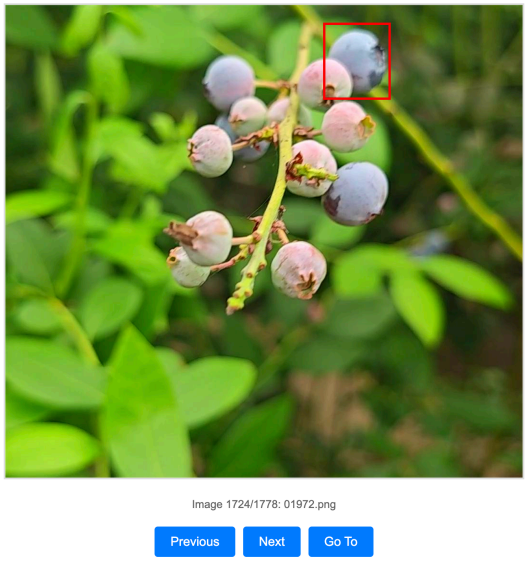


Figure 8: Annotation interface for quantity pairs.

A.7 QUANTITY

MVTEC-LOCO data was used to construct the quantity dataset. Using the annotated anomalies related to quantity, we formed pairs by selecting those with small cosine similarity.

ChangeIt data with ”pouring” as action can be used as quantity. We used the same annotation method as we used in the Section A.2

LEVIR-MCI data related to quantity is being used. We used the same annotation method as we used in the Section A.6

MegaFruits, UCF-QNRF-ECC, and UBC (Wang et al., 2024; Idrees et al., 2018; Huang et al., 2022) are datasets containing multiple instances of fruits, people, or buildings. Annotators first draw a bounding box around a selected object in the image (Figure 8). A black box of the same size is then overlaid on the image, after which nano-banana is tasked with inpainting the region while ensuring that no new object is introduced. In this way, we obtain modified images in which one or more objects have been removed.

A.8 QUALITY

YT8M was also used to construct the quality dataset. Many videos in YT8M contain frames of varying quality due to object motion, camera motion, lighting, or smoke. We constructed pairs in the same manner as described in Section A.4. Human annotators then assessed whether a quality difference was present and, if so, identified which frame had better quality.

A.9 VIEWPOINT

CameraBench (Lin et al., 2025) provides extensive annotations of ground-truth camera movements in video clips. Human annotators selected appropriate frame pairs that capture the ground-truth camera movement, and ground-truth annotations were used to generate corresponding questions and answers.

A.10 ACTION

YT8M was also used to construct the action dataset. We formed pairs in the same manner as described in Section A.4. Human annotators then specified the item, the location (first image or second image), and the action (i.e., the relative change in location).

A.11 Synthetic Data

Synthetic data were generated using a simple Python script that draws circles, squares, and triangles on a white background. During generation, we stored the metadata to enable controlled experiments.

B Evaluation Detail

B.1 Prompt Design for Evaluation Tasks

Figures 9–17 provide the exact prompts used in our experiments. Unless otherwise noted, temperature was fixed at 0.5 and the repetition penalty at 1.0.

System Prompt

You are a helpful assistant that answers multiple-choice questions about differences between two images.
Your task is to carefully analyze both images and identify the main difference between them.

Guidelines:
- Unless specified in the options, the difference is described in terms of the second image relative to the first.
- Respond **only** with the answer letter (A, B, C, D, etc.). Do not provide any reasoning or explanation.

User Prompt

Question: {question_text}

Carefully examine the images and choose the best description of the key visual difference.

Options:
{options_text}

Figure 9: *Standard prompt* used in our benchmark.

System Prompt

You are a helpful assistant that answers multiple-choice questions about differences between two images.
Your task is to carefully analyze both images and identify the main difference between them.

Guidelines:
- Unless specified in the options, the difference is described in terms of the second image relative to the first.
- Respond **only** in the following format. The answer should be a single letter.

Reasoning
[explanation of the key visual difference between the two images]

Answer
[answer (single letter)]

User Prompt

Question: {question_text}

Carefully examine the images and choose the best description of the key visual difference.

Options:
{options_text}

Figure 10: *Reasoning prompt* used in our benchmark.

System Prompt (1st step)
You are an expert image analyst. Your task is to carefully describe the differences between two images. Do not answer any questions about the images yet - only analyze and describe what is different between them.

User Prompt (1st step)
Please provide a careful and detailed description of the differences between these two images. Focus on what changed, what's different, or what distinguishes the first image from the second image. Using the description of the differences, you will be asked the following question and you will need to choose one correct answer. However, do not answer the question yet - just analyze the differences:

{question_text}

{options_text}

System Prompt (2nd step)
You are a helpful assistant that answers multiple-choice questions about differences between two images.
Your task is to carefully analyze both images and identify the main difference between them. You will be provided with a description of the differences to help guide your analysis.

Guidelines:
- Use the provided difference description to understand what has changed between the images.
- Verify the described difference by examining the images.
- Unless specified in the options, the difference is described in terms of the second image relative to the first.
- Respond **only** with the answer letter (A, B, C, D, etc.). Do not provide any reasoning or explanation.

User Prompt (2nd step)
I am showing you two images (first and second).

Description of the differences: {diff_description}

Question: {question_text}

Based on the images and the description of differences, choose the best answer.

Options:
{options_text}

Figure 11: *Two-step reasoning prompt* used in our benchmark.

System Prompt
You are a helpful assistant that answers multiple-choice questions about differences between two images. The grid lines are added to both images to help you compare the objects better.
Your task is to carefully analyze both images and identify the main difference between them.

Guidelines:
- Unless specified in the options, the difference is described in terms of the second image relative to the first.
- Respond **only** with the answer letter (A, B, C, D, etc.). Do not provide any reasoning or explanation.

User Prompt
Question: {question_text}

Carefully examine the images and choose the best description of the key visual difference.

Options:
{options_text}

Figure 12: *Grid prompt* used in our benchmark.

System Prompt

You are a helpful assistant that answers multiple-choice questions about differences between two images that are concatenated horizontally (first image on the left and second image on the right, separated by a black line). Your task is to carefully analyze both images and identify the main difference between them.

Guidelines:
- Unless specified in the options, the difference is described in terms of the second image relative to the first.
- Respond **only** with the answer letter (A, B, C, D, etc.). Do not provide any reasoning or explanation.

User Prompt

Question: {question_text}

Carefully examine the images and choose the best description of the key visual difference.

Options:
{options_text}

Figure 13: *Concatenating prompt* used in our benchmark.

System Prompt

You are a helpful assistant that answers multiple-choice questions about differences between two images. Your task is to carefully analyze the images and identify the main difference between them. I am showing you four images:
1. Original first image
2. Original second image
3. Highlighted first image (with areas of significant change marked with green boxes, and other areas dimmed)
4. Highlighted second image (with the same areas marked)

The highlighted images help you focus on the most significant differences between the two images. Use them to quickly identify where the changes occur, then examine those areas carefully in the original images.

Guidelines:
- Unless specified in the options, the difference is described in terms of the second image relative to the first.
- Focus on the green-boxed regions in the highlighted images to identify where changes occur.
- Respond **only** with the answer letter (A, B, C, D, etc.). Do not provide any reasoning or explanation.

User Prompt

I am showing you four images:
1. Original first image
2. Original second image
3. Highlighted first image (green boxes mark significant change areas, other areas dimmed)
4. Highlighted second image (same areas marked)

The highlighted images (3 and 4) show you WHERE the main differences are located. The green boxes indicate the top 2-3 most significant change regions. Use these to guide your attention, then carefully examine those specific areas in the original images (1 and 2) to determine WHAT the difference is.

Question: {question_text}

Carefully examine the images and choose the best description of the key visual difference.

Options:
{options_text}

Figure 14: *Highlighting prompt* used in our benchmark.

System Prompt

You are a helpful assistant that answers multiple-choice questions about differences between two images. Your task is to carefully analyze first and second images and identify the main difference between them. The third image is the overlay of the first and second images. You may use the third image to help you analyze the difference between the first and second images.

Guidelines:
- Unless specified in the options, the difference is described in terms of the second image relative to the first.
- Respond **only** with the answer letter (A, B, C, D, etc.). Do not provide any reasoning or explanation.

User Prompt

I am showing you three images:
1. First image
2. Second image
3. Overlapped image (50/50 blend of first and second images)

Question: {question_text}

Carefully examine the images and choose the best description of the key visual difference of first and second images.

Options:
{options_text}

Figure 15: *Overlapping prompt* used in our benchmark.

System Prompt

You are a helpful assistant that answers multiple-choice questions about differences between two images. Your task is to carefully analyze first and second images and identify the main difference between them. The third image is a black-and-white difference map between the first and second images, where brighter areas indicate larger differences. You may use the third image to help you analyze the difference between the first and second images.

Guidelines:
- Unless specified in the options, the difference is described in terms of the second image relative to the first.
- Respond **only** with the answer letter (A, B, C, D, etc.). Do not provide any reasoning or explanation.

User Prompt

I am showing you three images:
1. First image
2. Second image
3. Black-and-white difference map between the first and second images

Question: {question_text}

Carefully examine the images and choose the best description of the key visual difference of first and second images.

Options:
{options_text}

Figure 16: *Subtracting prompt* used in our benchmark.

System Prompt
You are an expert at visual analysis and image comparison. Compare the two images and briefly describe what's different between them.

Keep your response concise and direct. Use simple phrases like "In the first image, X, however in the second image, Y" or "X appeared in the second image" or "In the first image, X is relatively Y". Avoid detailed explanations or structured lists.

User Prompt
Describe the differences between the two images

Figure 17: *Captioning prompt* used in our benchmark.

B.2 Visual Input Construction and Examples

We describe how visual inputs are constructed for prompting, including grid overlays, image concatenation, image overlap, and image subtraction.

Grid Overlays. A 4×4 grid overlay is generated by drawing black lines with 30% opacity and a line width of 3 pixels on both images.

Concatenated Images. Image pairs are concatenated horizontally with a 1-pixel-wide black separator, forming a single composite image that is then used as the VQA input.

Overlap Images. Two aligned inputs are blended with equal weights (50% contribution each) in pixel space, producing a composite image that visually merges both inputs. The generated overlap image is provided to the VLM together with the original input pair. Figure 18 shows an example of the overlap image construction, illustrating the two aligned inputs and the resulting blended composite.

(a) First image (b) Second image (c) Overlap result

Figure 18: Example of overlap-image construction. Two aligned inputs are blended with equal weights (50% each) to produce a composite image that visually merges both inputs.

Subtraction Images. Subtraction images are generated by computing the absolute pixel-wise difference between the aligned inputs, converting the result to grayscale, and normalizing it to highlight regions of maximal change. Specifically, for two images $I_1, I_2 \in [0, 255]^{H \times W \times 3}$,

$$G(x, y) = \frac{1}{3} \sum_{c=1}^{3} \left| I_1(x, y, c) - I_2(x, y, c) \right|,$$

$$S(x, y) = 255 \cdot \frac{G(x, y)}{\max_{u,v} G(u, v)},$$

where S denotes the grayscale subtraction image. For both overlap and subtraction variants, the generated images are provided to the VLM alongside the original inputs. Figure 19 shows an example of the subtraction images, illustrating the two input images and the resulting difference map.

Highlight Images. Highlight images are generated by drawing bounding boxes on the largest regions of change. Given the pixel-wise difference map G, we obtain a mask of considerable change by computing a percentile threshold: letting τ_p be the p-th percentile of values in G (we use $p = 90$),

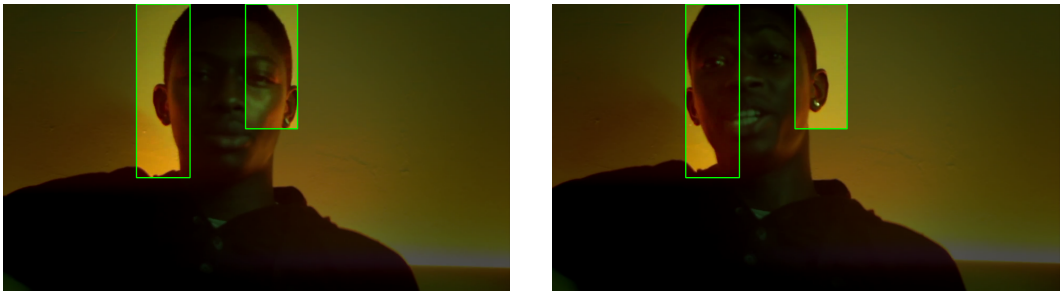
(a) First image (b) Second image (c) Subtraction result

Figure 19: Example of subtraction-image construction. Given two aligned images (I_1, I_2), the subtraction map S highlights regions with maximal pixel-wise change.

we define the binary mask

$$M(x,y) = \begin{cases} 1, & G(x,y) > \tau_p, \\ 0, & \text{otherwise.} \end{cases}$$

Morphological closing and opening are applied to M to connect nearby changes and remove noise, producing a cleaned mask with connected components $\{C_k\}$. The clusters $\{C_k\}$ are sorted by area, and the three largest clusters are selected. To avoid highlighting insignificant regions, we retain the second and third clusters only if their areas are at least 50% of the area of the largest cluster; otherwise, they are discarded. Highlight images are then constructed by dimming the background (using $\alpha = 0.5$) while preserving the original appearance only inside the selected bounding boxes. Finally, we draw a green-colored boundary around each box to emphasize the regions of change. Figure 20 shows example highlight images produced on a pair, illustrating how bounding boxes are drawn over the largest regions of change.



(a) Highlighted first image (b) Highlighted second image

Figure 20: Example highlight images. Significant regions of change, extracted via percentile thresholding and morphological filtering, are emphasized by dimming the background and drawing green bounding-box boundaries.

B.3 Human Evaluation

Human evaluation was conducted by sampling 1,000 examples from each dataset. For each difference type, 100 examples were selected, and the sampling ratio across sources was adjusted to align with the random-guess baseline. During evaluation, annotators were asked to answer questions using the interface shown in Figure 21. The interface displays two images side by side for comparison, and evaluators can also view the images sequentially in the bottom left corner by pressing the left and right arrow keys.

B.4 Fine-tuning Setup

Training is performed on 4 NVIDIA A100 GPUs, each equipped with 80GB memory. We use a learning rate of 1e-5, a per-device batch size of 8, resulting in an effective batch size of 32. The

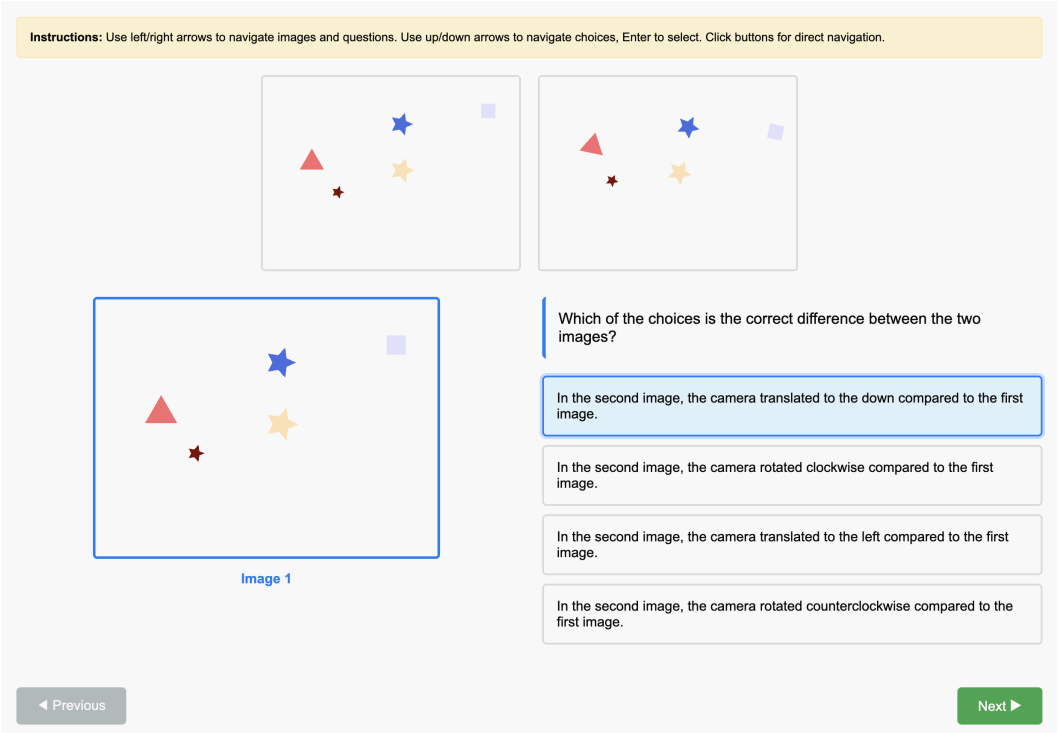


Figure 21: human evaluation interface.

models were trained for 3 epoch, with all parameters including vision encoder, projector, and language model jointly optimized.

For transferability study (table 6), we fine-tune Qwen-2.5-VL-7B on our 1,277-sample training split of SubtleBench and a size-matched subset of MLLM-CompBench for comparison. For MMAD evaluation, all training instances are converted to the standardized MMAD task format, and for QAG-360K, all training samples are adapted to our standard prompt template. Aside from this dataset-specific formatting, the fine-tuning procedure is identical across all settings, and we use the same hyperparameters described above.

C Additional Results

C.1 Correlation Analysis with Downstream Benchmarks

Table 7 shows the accuracy of all evaluated models across VLM-SubtleBench, MLLM-CompBench, MMAD, and QAG-360K. For MMAD evaluation, we randomly sample 10% of the test set and evaluate models using this subset. For QAG-360K evaluation, we exclude the change-ratio category due to its continuous answer format and converted the remaining question types into multiple-choice queries following our base prompt template. We then randomly sampled 723 validation examples for evaluation.

C.2 Additional Prompting Strategy

We explore an additional prompting strategy based on pure language-only comparison.

For the *emotion*, *existence*, and *quality* categories—similar to MLLM-CompBench (Kil et al., 2024)—we ask the following questions for each image:

- **Emotion:** "Describe the emotion expressed in the image in detail and rate its intensity on a scale of 1–10."

Table 7: Model-wise accuracy on VLM-SubtleBench, MLLM-CompBench, MMAD, and QAG-360K.

Models	VLM-SubtleBench	MLLM-CompBench	MMAD	QAG-360K
Qwen-2.5-VL-7B	59.4	73.6	65.0	34.4
Qwen-2.5-VL-32B	62.2	74.6	67.6	35.5
Qwen-2.5-VL-72B	65.4	76.9	68.9	41.2
GPT-4o	61.6	75.7	67.7	35.2
o3	75.7	86.3	72.9	39.7
GPT-5-main	71.3	83.9	70.6	36.3
GPT-5-thinking	77.8	86.3	73.5	42.1
Claude-sonnet-4	62.6	73.6	70.9	30.7
gemini-2.5-flash	62.5	85.2	71.4	36.8
gemini-2.5-pro	68.2	87.2	72.2	36.3

Table 8: Effect of pure language-based comparison for emotion, existence, and quality categories.

Category	GPT-5-main	Two-stage Reasoning
Emotion	92.7	87.8
Existence	75.4	63.2
Quality	84.5	84.6

- **Existence:** "Carefully list all objects visible in the image, including their approximate locations."
- **Quality:** "Analyze the quality of the image and rate it on a scale of 1–10, considering blur, noise, overexposure, compression artifacts, and other quality issues."

We then perform VQA using only the generated descriptions, without providing the original images, to evaluate the model's ability to answer purely from language-based reasoning.

Results. The results for the pure language-based comparison are shown in Table 8. For the *emotion* and *existence* categories, this method exhibits lower performance, consistent with the findings reported in MLLM-CompBench. Interestingly, performance on the *quality* category remains nearly identical, suggesting that explicit 1–10 rating prompts may serve as a reliable intermediate representation for comparative judgment in this aspect.

C.3 Domain-wise Performance Analysis

Table 9 shows domain-wise accuracy for proprietary VLMs. Among proprietary models, o3 and GPT-5-thinking achieved the highest scores in most categories. In particular, they exhibited a substantial performance gap over other models in the synthetic and medical domains, while both showed comparatively weaker results on the aerial domain.

C.4 Extended Color-Sensitivity Analysis

Inspired by (Hyeon-Woo et al., 2024), we extended our synthetic-control analysis to include a color-sensitivity axis, probing whether hue-level perceptual biases compound subtle comparative reasoning failures. We incorporated five representative colors (two green tones and three non-green: blue, red, and magenta) and systematically varied ΔE (hue shift in OKLAB space), brightness (L channel), size, count, and translation. OKLAB was chosen for its perceptual uniformity, where L encodes lightness and (a, b) correspond to green–magenta and blue–yellow opponent axes. To isolate hue effects, ΔE adjustments were applied by modifying (a, b) while keeping L constant, and brightness variation fixed hue while sampling $L \in [0.4, 0.8]$. All other setup conditions followed those in Figure 5.

Table 9: Domain-wise accuracy (%) of open-source and proprietary VLMs.

Model	Natural	Game	Industry	Aerial	Synthetic	Medical
Random Guess	49.2	50.0	50.0	26.9	29.3	50.0
Open-source						
LLaVA-NeXT-7B	48.6	47.9	53.8	35.4	30.0	41.5
LLaVA-OneVision-7B	59.3	56.0	54.7	62.2	32.4	50.9
Qwen2.5-VL-7B	65.2	61.8	66.3	62.8	43.8	50.3
Qwen2.5-VL-32B	66.1	63.4	64.4	64.6	53.0	54.5
Qwen2.5-VL-72B	69.6	65.1	69.6	64.1	54.8	65.2
Proprietary						
GPT-4o	68.4	65.8	71.5	46.9	45.0	62.4
o3	**77.3**	**76.4**	79.3	71.4	72.5	68.5
GPT-5-main	74.1	72.1	77.9	60.8	63.6	78.8
GPT-5-thinking	77.2	75.5	**81.2**	70.7	**78.8**	**82.4**
Claude-sonnet-4	64.2	60.3	66.3	74.1	56.3	54.8
Gemini-2.5-flash	68.1	66.5	73.7	73.4	43.3	62.4
Gemini-2.5-pro	73.2	71.8	79.7	**75.7**	50.6	68.8

Results are presented in Figure 22. Consistent with prior work (Hyeon-Woo et al., 2024), models exhibited pronounced difficulty in distinguishing green hues, showing significantly lower accuracy than for red or blue tones. Magenta elicited the most severe degradation (approaching 0%), revealing a systematic color-specific weakness in GPT-4o. In contrast, brightness variation produced no notable color-dependent gap, suggesting that VLMs are relatively invariant to luminance when performing comparative reasoning. Cross-factor analyses (color × size/count/viewpoint) showed minimal interaction, implying that these tasks are largely unaffected by color sensitivity, as hue discrimination itself contributes little to the reasoning objective.

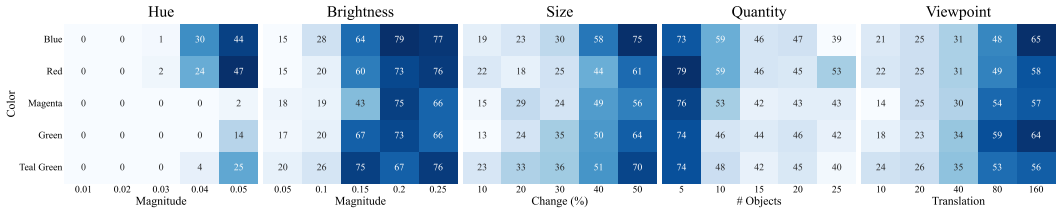


Figure 22: Color-sensitivity analysis of GPT-4o under controlled synthetic settings.

C.5 Source Validation and Distributional Consistency

To examine whether the use of nano-banana may introduce stylistic artifacts or distribution shifts that could confound evaluation, we conduct an analysis on our caption dataset. Specifically, we compare approximately 1K original (non-edited) caption pairs with subsets in which one image of each pair was reconstructed using nano-banana. Evaluation is performed on both VQA accuracy and captioning metrics (CSS and LLM-Judge). Table 10 reports the comparison between edited and non-edited pairs. Across all metrics, the differences are negligible, indicating that the use of nano-banana editing does not introduce measurable stylistic or distributional bias that could affect evaluation.

C.6 Effect of Prompting in Open-source Vision-Language Models

Table 11 shows effect of prompting in open-source Vision-Language models.

Table 10: Comparison between real and *nano-banana*-reconstructed image pairs on VQA and captioning metrics.

Type	**Accuracy**	**CSS**	**LLM-Judge**
Real	60.6	0.51	26.3
Reconstructed	60.6	0.51	27.3

Table 11: Effect of different prompting strategies in VLM-SubtleBench.

Model	AT	ST	EM	TM	SP	EX	QN	QL	VP	AC
Random Guess	35.9	50.0	50.0	50.0	36.6	23.2	48.9	50.0	42.1	50.0
Qwen2.5-VL-72B	53.9	**68.9**	85.9	49.9	47.8	81.7	67.7	**78.4**	**56.2**	74.1
+ Reasoning	50.6	68.0	**87.5**	**51.0**	44.4	81.7	65.0	77.1	54.7	**75.0**
+ Grid	54.9	67.6	85.6	49.8	46.3	**83.4**	**70.1**	75.7	54.4	71.1
+ Concat	55.7	68.0	86.4	50.0	39.6	67.5	61.0	65.4	47.0	73.7
+ Overlap	46.3	63.2	85.3	49.7	48.8	78.7	63.1	72.5	49.7	71.6
+ Subtract	**55.7**	64.4	85.6	49.5	**51.1**	79.8	63.4	63.0	49.9	71.2

C.7 Performance of Proprietary models by Data Source

Table 12 shows performance of proprietary models by data source.

Table 12: Performance of proprietary models by data source

Dataset	Random	GPT-4o	GPT-5-main	GPT-5-thinking	GPT-o3	Claude-sonnet-4	Gemini-2.5-flash	Gemini-2.5-pro
Attribute								
MVTEC-AD	50.0	76.7	79.0	82.0	79.0	66.3	76.3	83.3
COCO	25.0	78.1	82.5	86.0	84.2	71.1	69.3	77.2
MIMIC-DIFF-VQA	50.0	62.4	78.8	82.4	68.5	54.8	62.4	68.8
Synthetic	25.0	31.3	66.0	84.4	81.1	35.0	28.3	33.4
State								
MVTEC-AD	50.0	67.2	73.5	75.9	74.9	64.0	71.5	75.1
ChangeIt	50.0	81.7	84.5	86.7	85.4	65.6	73.8	78.1
Emotion								
CREMA-D, RAVDESS, AFEW-VA, DAiSE	50.0	89.5	92.7	93.1	92.9	83.3	88.4	89.8
Temporal								
YT8M	50.0	52.7	55.4	62.3	63.0	49.4	55.8	60.4
VLM4D	50.0	52.7	49.3	54.8	53.8	49.0	49.3	50.7
Spatial								
VLM4D	50.0	58.5	57.3	62.5	60.4	52.1	56.5	58.5
Synthetic	25.0	28.5	43.8	57.7	50.5	45.7	27.0	33.0
Existence								
LEVIR-MCI	20.0	43.7	59.1	74.6	73.6	79.2	80.5	79.7
COCO	25.0	77.6	86.7	80.6	88.8	83.7	75.5	92.9
Synthetic	25.0	69.0	84.2	93.3	86.8	93.5	68.0	77.8
Quantity								
MVTEC-LOCO	50.0	75.6	87.6	93.2	90.6	71.8	75.6	86.8
MegaFruits	50.0	50.9	59.7	66.1	67.0	57.9	51.5	62.7
UCF-QNRF-ECC	50.0	51.5	71.2	78.8	83.3	50.0	56.1	71.2
UBC	50.0	51.9	60.9	59.4	66.9	61.7	52.6	59.4
LEVIR-MCI	20.0	59.2	73.5	69.4	65.3	67.3	73.5	87.8
ChangeIt	50.0	41.4	58.6	55.2	56.9	50.0	60.3	62.1
Synthetic	50.0	59.4	79.0	92.4	86.0	65.8	59.4	63.2
Quality								
YT8M	50.0	72.4	84.5	84.8	87.6	70.8	77.1	84.8
Camera								
CameraBench	50.0	56.2	63.8	69.9	68.5	57.6	58.1	64.8
Synthetic	25.0	41.0	44.0	65.4	56.2	44.8	38.2	50.4
Action								
YT8M	50.0	76.7	83.6	84.9	84.8	66.3	72.3	76.8