# MVP-Bench: Can Large Vision–Language Models Conduct Multi-level Visual Perception Like Humans?

**Guanzhen Li** **Yuxi Xie** * **Min-Yen Kan**
National University of Singapore
{guanzhen,xieyuxi}@u.nus.edu kanmy@comp.nus.edu.sg

## Abstract

Humans perform visual perception at multiple levels, including low-level object recognition and high-level semantic interpretation such as behavior understanding. Subtle differences in low-level details can lead to substantial changes in high-level perception. For example, substituting the shopping bag held by a person with a gun suggests violent behavior, implying criminal or violent activity. Despite significant advancements in various multimodal tasks, Large Visual Language Models (LVLMs) remain unexplored in their capabilities to conduct such multi-level visual perceptions.

To investigate the perception gap between LVLMs and humans, we introduce MVP-Bench, the first visual–language benchmark systematically evaluating both low- and high-level visual perception of LVLMs. We construct MVP-Bench across natural and synthetic images to investigate how manipulated content influences model perception. Using MVP-Bench, we diagnose the visual perception of 10 open-source and 2 closed-source LVLMs, showing that high-level perception tasks significantly challenge existing LVLMs. The state-of-the-art GPT-4o only achieves an accuracy of 56% on Yes/No questions, compared with 74% in low-level scenarios. Furthermore, the performance gap between natural and manipulated images indicates that current LVLMs do not generalize in understanding the visual semantics of synthetic images as humans do. Our data and code are publicly available at https://github.com/GuanzhenLi/MVP-Bench.

## 1 Introduction

Visual perception (VP) refers to the ability to transform visual signals into meaningful perceptions (de Wit and Wagemans, 2012; Gordon et al., 2019). When humans parse visual signals, they initially engage in high-level perception to grasp the overarching concept using commonsense knowledge. This serves as context guidance for exploring further low-level details aligned with their intentions (Wang et al., 2024; Garner, 1987). For example, given an image of a man in a bar, humans first grasp the high-level concept, such as the behaviour of drinking, and focus on low-level details, such as the type of alcohol, to obtain specific information. Existing Large Vision–Language Models (LVLMs) demonstrate an exceptional understanding of such low-level visual clues. However, it remains unexplored whether they have similar hierarchical visual perceptions at both levels, like humans.

Recently, several benchmarking works have considered evaluating visual perceptions (Liu et al., 2023c; Fu et al., 2024; Chow et al., 2021). However, such holistic evaluation benchmarks lack the critical specialization needed to assess visual perceptions. Specifically, most of their tasks focus on low-level perception such as *Counting* and *Existence Detection* questions on single images. Besides, existing benchmarks are mostly designed based on individual question–image samples, failing to evaluate the consistency and accuracy of understanding an image with different forms of perceptions. Furthermore, most of the current benchmarks are built on real-world natural image data, making it hard to disentangle reliance on prior knowledge from the visual perception of specific contexts, such as synthetic images (Bitton-Guetta et al., 2023). Motivated by the challenges of interpreting LVLMs' visual perception capabilities, we propose MVP-Bench, the first benchmark systematically evaluating multi-level visual perceptions of LVLMs. As shown in Figure 1, each sample is accompanied by questions at both levels. We thoroughly design five high-level and thirteen low-level perception categories, detailed in Section 3. Furthermore, we construct {natural, manipulated} image pairs which convey contrasting perceptions as a more challenging task for visual perception.

In this work, with our constructed MVP-Bench,

*Corresponding author



*Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13505–13527
November 12-16, 2024 ©2024 Association for Computational Linguistics

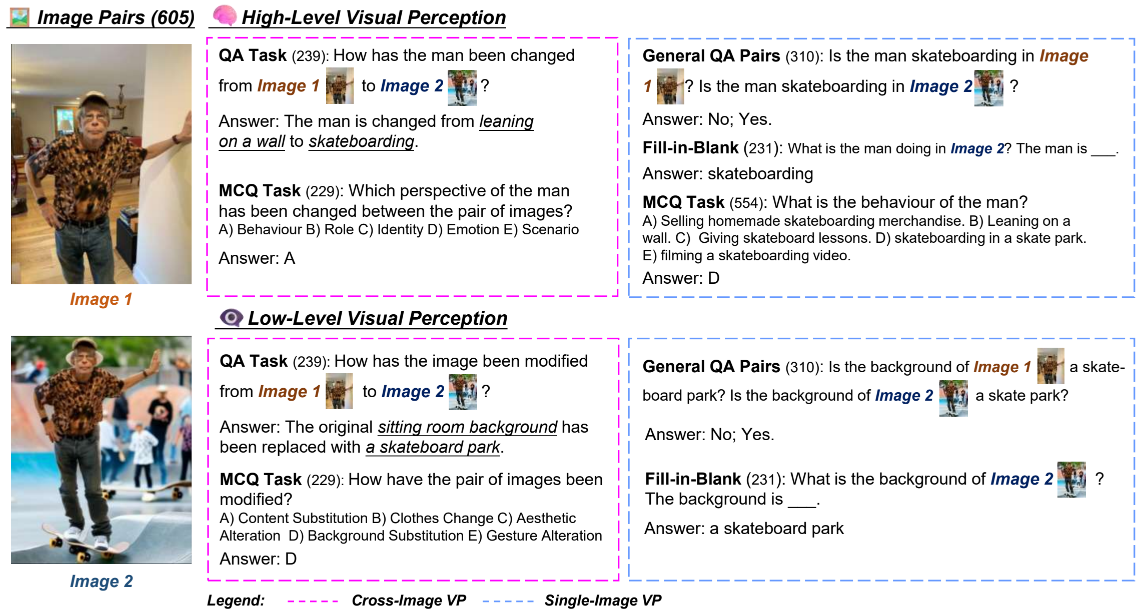



Figure 1: A sample of MVP-Bench manifesting both high- and low-level visual perception. *Image 1* and *Image 2* form an image pair. Their different backgrounds indicate that the man is engaged in different behaviours.

we evaluate twelve LVLMs and find that there is a significant performance gap between high- and low-level visual perception in LVLMs. Furthermore, we observe that manipulated visual contents are more challenging than natural images for LVLMs to understand and interpret. Our further qualitative analysis reveals the deficiency of current LVLMs and the gap between open- and closed-source models.

## 2 Related Work

**Visual Perception.** Visual Perception represents how the human brain transforms the pattern of information on the retina into a meaningful perception of the world (de Wit and Wagemans, 2012; Cornsweet, 2012). This process involves interactions among sensory and cognitive processes across hierarchical levels in the brain (Gordon et al., 2019; Rouw et al., 1997). Low-level visual features refer to the properties like colors and spatial attributes, while high-level visual processing integrates with human cognitive functions (*e.g.* commonsense knowledge, personal experiences) related to recognized objects (Akcelik et al., 2022; Wu et al., 2023b; Kandel et al., 2021; Schindler et al., 2021). Both perception competences are crucial, as human visual perception begins with grasping the image's main idea at a high level, and then delving into low-level features motivated by particular intentions (Garner, 1987). In MVP-Bench, we define five high-level categories and thirteen low-level categories. The mapping relationships between levels indicate that certain low-level features can support the high-level perception (illustrated in Section 3).

**Vision–Language Benchmarks.** Some recent benchmarks contain visual perception as a section, but their aim to offer a comprehensive evaluation of LVLMs' various capabilities leads to an inadequate exploration of visual perception. MMBench (Liu et al., 2023c) and MME (Fu et al., 2024) categorize visual perception based on question granularity. Although coarse perception questions are general, their questions like *Counting* or *Existence Detection* cannot reflect an image's main idea as high-level visual perception. Additionally, they evaluate different categories of visual perception individually, making it unavailable to compare an LVLM's different perceptions. II-Bench(Liu et al., 2024) and DEEPEVAL(Yang et al., 2024) focus on understanding deep image semantics, requiring LVLMs to perform complex commonsense reasoning based on low-level details. While they also reveal LVLMs' performance gaps across levels, these gaps are primarily due to limited reasoning abilities rather than intuitive visual perception. The definition of perception in PCA-Bench (Chen et al., 2024) resembles our benchmark, emphasizing how perception offers a guiding context in decision-making domains. However, their images depicting environments normally do not require significant high-level perception. MVP-Bench systematically evaluates LVLMs' multi-level visual perception, with each image accompanied by high- and low-level questions simultaneously. As perceptions related to humans normally require significant perception at both levels (such as misinformation understanding or emotion recognition) (Peng et al., 2023; Thomson et al., 2022), we construct image pairs containing humans to ensure that the cases can assess LVLMs' multi-level perception.

**Synthetic Images.** Recent advancements in image generation tools (Ramesh et al., 2021; Rombach et al., 2021) and image editing models (Brooks et al., 2023; Zhang et al., 2023) have led to synthetic datasets for different tasks, such as Whoops (Bitton-Guetta et al., 2023) and StableRep (Tian et al., 2024). In the process of utilizing text-to-image tools for generating synthetic images, a prompt aligned with the expected image content is essential. In previous works, the source of such prompts can be manually-crafted prompts (Bitton-Guetta et al., 2023), text annotations in existing datasets (Tian et al., 2024) or prompts generated by LLMs (Aboutalebi et al., 2024; Li et al., 2023; Wu et al., 2023a). In MVP-Bench, we generate manipulated images for constructing image pairs. To obtain a prompt tailored to each case while minimizing human effort, we employ ChatGPT to generate the prompts (*cf.* Section 4.1).

## 3 MVP-Bench Evaluation Suite

MVP-Bench comprises 530 {natural, manipulated} image pairs accompanied by questions at multiple perception levels. Using MVP-Bench, we diagnose LVLMs by investigating (1) the performance gap between high- and low-level visual perceptions and (2) the difference in visual understanding abilities on natural and manipulated images.

### 3.1 Evaluation across Perception Levels

We prioritize the perception of humans as high-level perception, *e.g.*, misinformation understanding (Da et al., 2021) and emotion recognition (Hari and Kujala, 2009), where high-level perception is commonly engaged.

We categorize high-level ($L_h$) perceptions of humans into five dimensions, including *Behaviour*, *Role*, *Identity*, *Emotion*, *Scenario*. Each dimension corresponds to several low-level ($L_l$) perception types. As shown in Figure 3 (a), certain low-level perceptions (*e.g.*, *attire* such as a police uniform or *group association* with firefighters) can support the high-level perception (*e.g.*, *Role*).

We design Yes/No questions and Cross-Image questions at both levels. Constructed on the same set of images, the multi-level perception tasks enable us to diagnose the perception gap in LVLMs across different levels. Specifically, we calculate the accuracy on Yes/No questions based on the correctness of each individual question–image pair (represented as $aAcc$), while all multiple-choice questions within MVP-Bench are evaluated with Circular Strategy (Liu et al., 2023c) to alleviate the model prediction bias from the option order.

### 3.2 Evaluation with Image Pairs

Each {natural, manipulated} image pair in MVP-Bench conveys significantly different multi-level perceptions. Specifically, the two images differ only in one of the $L_l$ perception categories (in Figure 3 (a)), leading to distinct $L_h$ perceptions. To mitigate the effect of the LVLMs' biased tendency to answer Yes/No questions (Liu et al., 2023a), we examine if LVLMs can elicit different perceptions given an image pair with the same question. We further explore the performance gap in LVLMs on natural and manipulated images in Section 5.

For Yes/No questions, we ask the same question on pairwise image data. As the two images are manipulated to convey different perceptions, they have opposite corresponding ground truth answers. We calculate $qAcc$ and $iAcc$ based on question- and image-level accuracy, respectively, following (Liu et al., 2023a). We design a holistic metric $mAcc$, requiring answering all questions corresponding to an image pair correctly.

For single-image multiple-choice questions, we focus on model understanding of manipulated images as a more challenging task. We include the answer to the natural image as a distractor to assess the discriminability of LVLMs in discerning the differences between the image pair. Additionally, we leverage ChatGPT[1] to generate three other options aligned with the low-level clues in the manipulated image to heighten our task difficulty.

## 4 MVP-Bench Construction

We now present our construction process of image manipulation and the designs of corresponding multi-level questions for MVP-Bench.

[1]We used `gpt-3.5-turbo-1106`.

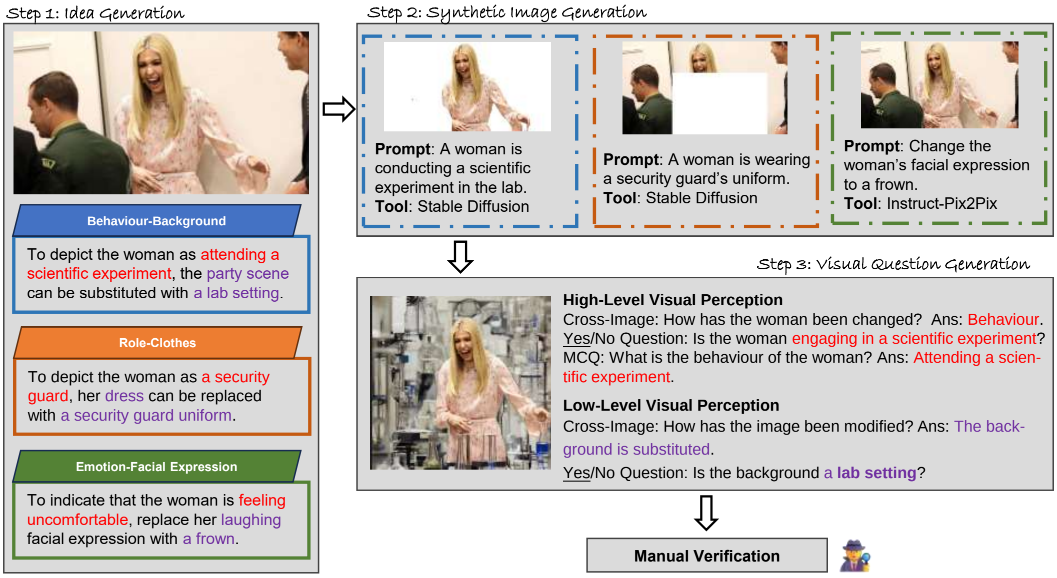



Figure 2: MVP-Bench three-step construction pipeline (best viewed in color). Step 1 uses three categories ('Behaviour-Background', 'Role-Clothes', 'Emotion-Facial Expression') as examples to illustrate how high-level perception guides the identification of low-level perception. Step 2 demonstrates three categories of manipulated image generation: *Overall Background Substitution*, *Partial Component Substitution*, and *Direct Alteration* (from left to right). Step 3 explains how to generate questions based on the ideas obtained in Step 1, with the same colour indicating that the generated question is based on the corresponding part from the expected perception.

### 4.1 Construction Pipeline

We select images from the EMU dataset (Da et al., 2021) as natural images for constructing image pairs. EMU focuses on visual misinformation, portraying cases involving humans and complex social scenes that require perceptions at both levels. Based on the natural image, we generate synthetic manipulations following one of the $L_l$ categories.

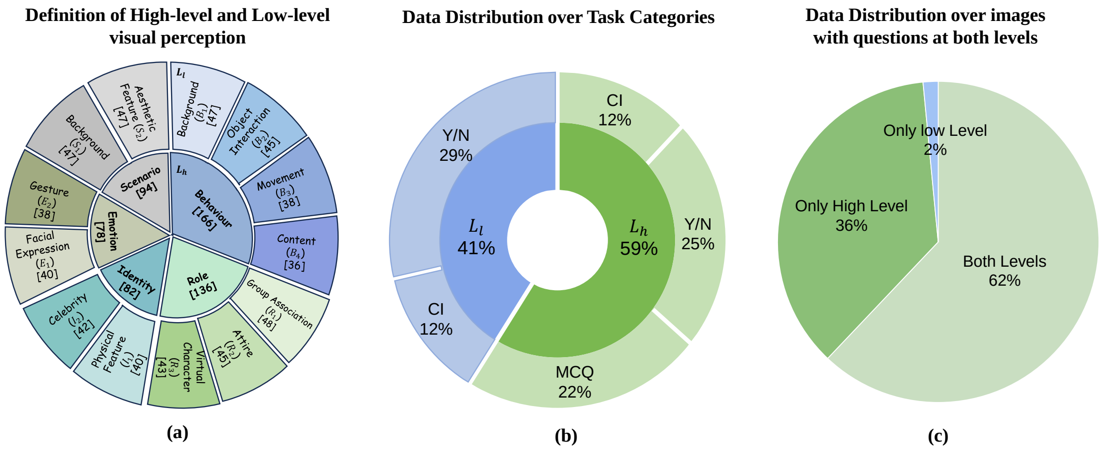
However, to alter manipulated images' $L_h$ perceptions in certain categories, it is challenging to constrain the manipulation applied exactly to a specific $L_l$ category without significant modification on other details. Besides, it is also hard to ensure consistency between the image pairs and the questions. We propose a three-step benchmark construction pipeline to meet the two requirements.

**Step one: Idea Generation.** We utilize ChatGPT to generate ideas on how to manipulate natural images via Chain of Thoughts (CoT). Given an initially determined $L_h$ category, we prompt ChatGPT to identify a corresponding low-level perception to support it. For instance, in Figure 2, considering the "Behaviour-Background Substitution" category, ChatGPT first generates an idea to change the woman's behaviour from attending a party to engaging in an experiment. Under this guidance, the background of the manipulated image should be a laboratory environment. Specifically, we provide auxiliary information such as the description of the manipulated image, which is incorporated into the textual prompt for image generation in Step 2.

To ensure coherence between the generated idea and the subsequent visual editing, we fixate on a specific subject at this initial step utilizing the visual grounding ability of Shikra (Chen et al., 2023). Specifically, we employ Shikra to retrieve the coordinates of a selected subject ($C_{sub}$) and utilize it to query low-level features (*e.g.*, "What is the man holding?") from the image in the subsequent steps.

**Step two: Manipulated Image Generation.** We define three categories of manipulated image generation based on the image-editing type: Partial Component Substitution, Overall Background Substitution, and Direct Manipulation.

*2.1 Partial Component Substitution.* This refers to manipulating an image by substituting an object or a part of the main subject. The pipeline utilizes Shikra to extract the target object's coordinates ($C_{obj}$), with $C_{sub}$ serving as a constraint. After masking $C_{obj}$ as a blank, we apply the Stable-Diffusion-Inpaint (Stacchio, 2023) as a tool, using the edited image's caption obtained from step one as the prompt to generate a manipulated image. A set of defined $L_l$ categories, $\{B_2, B_3, B_4, R_2, I_1, I_2, I_3, E_1\}$, can be executed in this process.




Top 5 Most Frequent Adjectives in captions of Source images and Manipulated images

| Source Image | Manipulated Image |
|---|---|
| happy | angry |
| excited | frustrated |
| innocent | aggressive |
| peaceful | violent |
| good | suspicious |

(d)

Top 5 Most Frequent Nouns in captions of Source images and Manipulated images

| Source Image | Manipulated Image |
|---|---|
| police | business |
| president | firefighter |
| people | criminal |
| street | protest |
| car | event |

(e)

Figure 3: MVP-Bench statistics. (a) shows 5 high-level ($L_h$) categories and 13 low-level ($L_l$) categories, where the mapping relationship indicates that the low-level features can support certain high-level perceptions. (b) shows the distribution of questions. Y/N, CI, MCQ denote Yes/No questions, cross-image questions, and single-image multiple-choice questions respectively. (c) demonstrates the distribution of images with questions at different levels. (d) and (e) demonstrate that our pipeline successfully generates pairs of images with significantly distinct content.

*2.2 Overall Background Substitution.* This represents generating a manipulated image by retaining solely the main subject while replacing the entire background. In these cases, a standard rectangle cannot exactly mask the subject, potentially remaining unexpected elements and distorting the background generation. To address this limitation, we employ the Segment Anything Model (Kirillov et al., 2023) to produce a set of detected object masks ($\mathbb{M} = \{M_1, M_2, ..., M_n\}$) in irregular shapes for a given image. We identify a mask with the greatest overlap with $C_{sub}$.

$$mask = \arg\max_{M_i \in \mathbb{M}} Overlap(M_i, C_{sub}) \quad (1)$$

Here, $Overlap$ refers to a function that calculates the overlapping square between two regions. To enhance flexibility and increase the case difficulty, we randomly translate the location of $C_{sub}$, rescale the $C_{sub}$, and resize the entire mask. Finally, with the new mask and the manipulated image's caption obtained from Step 1, we utilize Stable-Diffusion-Inpaint to generate a new image with a different background from the original natural image. This process can handle $\{B_1, R_1, S_1\}$.

*2.3 Direct Alteration.* This addresses situations where nothing can be substituted, yet some alteration is necessary, such as changing facial expressions. With the original natural image and the manipulation instruction obtained from Step 1, we directly utilize the image-editing model InstructPix2Pix (Brooks et al., 2023) to generate a manipulated image for $\{E_2, S_2\}$. However, since this process cannot focus on specific subjects, we mainly apply it to images containing a single person or cases requiring overall manipulations.

**Step three: Visual Question Generation.** We generate Yes/No questions, Single- and Cross-Image multiple-choice questions using ChatGPT based on the ideas generated in Step 1. Single-Image questions focus on the discrepancy between image pairs, while Cross-Image tasks focus on the differences across each pair of images. To ensure the quality of generated questions, two of this paper's authors manually verified all 3205 questions.

| Yes/No | MCQ | Cross-Image | Overall |
|---|---|---|---|
| 89.64 | 85.14 | 95.38 | 89.67 |

Table 1: The high agreement between annotators across different tasks ensures the quality of the retained cases.

A question was retained only when both annotators accepted it, and the annotators demonstrated a high level of agreement throughout the process (shown in Table 1). Finally, 1872 questions are retained within the MVP-Bench. While verifying Yes/No questions, we focused on: (1) the quality of manipulation and (2) the consistency between images and ground truths. For multiple-choice questions, we paid additional attention to cases where distractors were not discrepant with the ground truth. We manually adjusted these distractors and double-checked the cases to ensure both annotators accepted them.

### 4.2 MVP-Bench Statistics

Of the final 605 retained images, 62% are accompanied by questions at both levels, supporting our MVP-Bench's novel contribution to assess LVLMs' performance gaps across levels. Figure 3 shows the balanced distribution of different question types at both low and high levels. Additionally, to create image pairs for evaluating LVLMs' heterogeneous performance on natural and manipulated images, we designed an automated pipeline using ChatGPT to generate conflicting captions for each image pair. For verifying whether the automatic process can lead to conflicting captions as expected, we identified and listed the top 5 most frequent adjectives and nouns in the captions of natural and manipulated images (in Figure 3). The significant polarity differences (e.g., {innocent, aggressive}, {police, criminal}) between two sets of tokens indicate that our pipeline successfully generated image pairs with contrasting contents.

Furthermore, to ensure the generated content aligns with human perception, we compared human performance with state-of-the-art LVLMs on a randomly sampled subset.[2] Human annotators achieved over 95% accuracy, significantly outperforming LVLMs, indicating that our MVP-Bench offers a convincing evaluation.

## 5 Experiments

We use MVP-Bench to diagnose and compare the visual perception capabilities of LVLMs belonging to two categories: (1) *Open-Source LVLMs* including MiniCPM-V-2 (OpenBMB, 2024), DeepSeek-VL (Lu et al., 2024), MiniGPT4 (Zhu et al., 2023), mPLUG-Owl2 (Ye et al., 2023), InstructBLIP (Dai et al., 2023), and LLaVA-1.5 (Liu et al., 2023b); (2) *Proprietary LVLMs* including GPT-4V and GPT-4o. All the experiments are conducted with VLMEvalKit (Contributors, 2023) under the zero-shot setting for a fair comparison.

### 5.1 Result Analysis

As outlined in Section 3, we compare the performance of LVLMs at multiple perception levels (Table 2). We also investigate the performance variation when given manipulated images in Table 3.

**Performance at Different Perception Levels.** As shown in Table 2, both open- and closed-source models perform worse on high-level perception tasks than low-level ones, *e.g.*, 55%, 52%, and 56% compared to 69%, 67%, and 74% of *qAcc* on MiniCPM-V-2, LLaVA-1.5-13B, and GPT-4o, respectively. Specifically, we observe that closed-source models present a larger relative performance gap between high-level and low-level perception. For example, GPT-4o achieves an accuracy of 34% (relatively reduced by 53% from 74%) on cross-image MCQ, compared to 18% ( relatively reduced by 30% from 26%) of LLaVA-1.5-13B. This indicates that the performance gains from closed models mainly come from their superior low-level perceptions, yet they still encounter challenges in high-level tasks. We further discuss the potential cause of this observation in Section 5.2.

**Impact of Model Sizes.** Small models can outperform the larger ones in Table 2. Among open-source models, MiniCPM-V-2-3B and DeepSeek-VL-7B achieve the best performance on high-level and low-level tasks respectively. As MiniCPM-V-2 is aligned with fine-grained correctional human feedback, it shows excellent trustworthiness and reduced hallucination. This implies that LVLMs' trustworthiness may benefit their high-level visual perception. DeepSeek-VL demonstrates a strong capability of perceiving specific details with additional visual encoders for processing low-level features, indicating these features are crucial to low-level visual perception. Besides, comparing LLaVA and InstructBLIP with different sizes reveals that increasing parameters from 7B to 13B does not notably enhance their visual perception at either level. Therefore, to enhance LVLMs' single-

[2]Appendix 6 compares the performance of human annotators and LVLMs on MVP-Bench.

| Models | Single-Image | | | | | | | Cross-Image | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | *qAcc* | | | *aAcc* | | | *mAcc* | *CircularEval* | | *VanillaEval* | |
| | $L_l$ | $L_h$ | $L_m$ | $L_l$ | $L_h$ | $L_m$ | $L_m$ | $L_l$ | $L_h$ | $L_l$ | $L_h$ |
| DeepSeek (1.3B) | 63.33 | 53.04 | 58.60 | 81.48 | 75.87 | 78.90 | 28.40 | 19.38 | 18.94 | 40.97 | 29.07 |
| MiniCPM-2 (3B) | 68.52 | 55.22 | 62.40 | 84.07 | **76.30** | 80.50 | 34.91 | 29.51 | 11.45 | 43.61 | 31.72 |
| DeepSeek (7B) | 70.00 | 54.35 | 62.80 | 84.82 | 76.09 | 80.00 | 33.73 | 36.12 | 25.99 | 47.58 | 36.56 |
| InstructBLIP (7B) | 49.63 | 40.00 | 45.20 | 74.82 | 69.13 | 72.20 | 17.75 | 0.00 | 1.32 | 27.31 | 23.79 |
| LLaVA-1.5 (7B) | 68.89 | 51.74 | 61.00 | 84.45 | 75.44 | 80.30 | 31.36 | 20.26 | 14.10 | 39.21 | 26.87 |
| MiniGPT4 (8.2B) | 14.44 | 8.26 | 11.60 | 39.26 | 33.70 | 36.70 | 0.59 | 0.00 | 0.00 | 2.64 | 5.73 |
| MiniGPT4-v2 (8.2B) | 52.59 | 40.87 | 47.20 | 73.70 | 67.40 | 70.80 | 14.20 | 0.00 | 0.00 | 21.59 | 24.67 |
| mPLUG-Owl2 (8.2B) | 69.26 | 54.78 | 62.60 | 84.63 | **76.30** | 80.80 | 36.09 | 21.14 | 13.22 | 34.80 | 25.99 |
| InstructBLIP (13B) | 50.37 | 36.09 | 43.80 | 75.19 | 67.61 | 71.70 | 15.98 | 1.76 | 0.44 | 25.99 | 18.50 |
| LLaVA-1.5 (13B) | 66.67 | 52.17 | 60.00 | 83.34 | 76.09 | 80.00 | 28.40 | 25.99 | 18.06 | 41.85 | 32.60 |
| GPT-4V | 66.30 | 39.57 | 54.00 | 82.23 | 69.13 | 76.20 | 23.08 | 44.50 | 14.10 | 63.00 | 37.44 |
| GPT-4o | **74.44** | **56.09** | **66.00** | **86.85** | 76.09 | **81.90** | **39.05** | **74.01** | **34.80** | **87.22** | **51.54** |

Table 2: Results comparison across low-level ($L_l$), high-level ($L_h$), and multi-level ($L_m$) tasks. *CircularEval* and *VanillaEval* refer to Circular and Direct evaluation for multiple-choice questions. We highlight the problematic results ($< 5\%$) and best performance across **all models** and on open-source models only. *qAcc*, *aAcc*, and *mAcc* represent question-level, individual, and holistic accuracies, repectively.

| Method | Yes/No | | | | | | | MCQ | |
|---|---|---|---|---|---|---|---|---|---|
| | *iAcc* | | | *aAcc* | | | *mAcc* | *CircularEval* | *VanillaEval* |
| | N | M | N+M | N | M | N+M | N+M | N+M | N+M |
| DeepSeek (1.3B) | 60.95 | 44.38 | 52.66 | 83.20 | 74.60 | 78.90 | 28.40 | 43.78 | 62.44 |
| MiniCPM-2 (3B) | 68.64 | 53.85 | 61.24 | 85.20 | 75.80 | 80.50 | 34.91 | 44.74 | 62.20 |
| DeepSeek (7B) | 68.05 | 52.07 | 60.06 | 85.00 | 76.60 | 80.80 | 33.73 | 59.33 | 74.40 |
| InstructBLIP (7B) | 44.38 | 44.97 | 44.68 | 72.40 | 72.00 | 72.20 | 17.75 | 4.07 | 19.14 |
| LLaVA-1.5 (7B) | 64.50 | 52.66 | 58.58 | 83.20 | 77.40 | 80.30 | 31.36 | 57.18 | 71.29 |
| MiniGPT4 (8.2B) | 10.06 | 4.73 | 7.40 | 41.80 | 31.60 | 36.70 | 0.59 | 0.00 | 2.63 |
| MiniGPT-v2 (8.2B) | 53.85 | 31.95 | 42.90 | 79.60 | 62.00 | 70.80 | 14.20 | 1.91 | 29.43 |
| mPLUG-Owl2 (8.2B) | 66.27 | 54.44 | 60.36 | 84.20 | 77.40 | 80.80 | 36.09 | 50.72 | 67.70 |
| InstructBLIP (13B) | 41.42 | 46.15 | 43.79 | 70.60 | 72.80 | 71.70 | 15.98 | 3.83 | 11.96 |
| LLaVA-1.5 (13B) | 58.58 | **55.62** | 57.10 | 81.20 | **78.80** | 80.00 | 28.40 | 55.02 | 72.25 |
| GPT-4V | 71.07 | 30.77 | 50.92 | 87.80 | 65.98 | 76.20 | 23.08 | 59.81 | 72.25 |
| GPT-4o | **76.92** | 48.52 | **62.72** | **90.00** | 73.80 | **81.90** | **39.05** | **64.83** | **77.27** |

Table 3: Result comparison across natural (N) and manipulated (M) images. *iAcc* refers to the image-level accuracy.

image visual perception, focusing on their ability to provide trustworthy answers and capture low-level features is more effective than simply scaling up.

**Analysis on the Cross-Image Task.** Table 2 shows that closed-source models significantly surpass open-source models on cross-image tasks, especially at low perception level. For instance, GPT-4V and GPT-4o achieve accuracies of 45% and 74% respectively at the low level, significantly surpassing the accuracy of LLaVA-1.5-13B (26%). Furthermore, this performance gap is larger than that observed in single-image tasks. In the cross-image task, GPT-4o outperforms LLaVA-1.5-13B relatively by 93% and 185% on each of the two levels separately, compared to just 8% and 12% in single-image tasks. The significant gap indicates open-source LVLMs' insufficient contextual attention, due to a lack of cross-image training data.

**Comparison between {natural, manipulated} Images.** As shown in Table 3, both open- and closed-source models show inferior performance on manipulated images compared to natural images. For example, MiniCPM-V-2, LLaVA-1.5-13B, and GPT-4o achieve an $iAcc$ of 69%, 59%, and 77% on natural images, while exhibiting lower $iAcc$ of 54%, 56%, and 49% on manipulated images. We attribute this observation to the discrepancy between the visual perception of manipulated images and LVLMs' training data. Besides, closed-source models demonstrate a larger performance gap across image pairs than open-source models. The $iAcc$ gap of GPT-4V and GPT-4o is 40.3% and 28.4% separately, while LLaVA-1.5-13B and MiniCPM-V-2 have gaps of only 2.96% and 14.79%. One reason for this is the rigorous manner of GPT-4V and GPT-4o in interpreting the high-level semantics of visual content, which we will discuss in Section 5.2. Besides, these models equally scrutinize all the details with their prior knowledge. This tendency to provide critical and reasonable answers impedes better visual perception on manipulated images.

**Yes/No v.s. MCQ** GPT-4V and GPT-4o present conflicting results on different tasks. Although both tasks are based on the manipulated images, two models perform poor on Yes/No task with an $iAcc$ of 31% and 49%, while outperforming all open-sourced models on the MCQ task. From Table 3, we can witness that the results of MCQ and $iAcc$ on natural images share the same trend, which suggests that closed-source models' inferior performance on manipulated images is owing to the nature of Yes/No questions. As an open-ended generative task, these models tend to perform rigorously and safely, while the MCQ task is less influenced by their rigorous manner. This is also a motivation for us to design both tasks for single-image perception.

## 5.2 Discussion

In this section, we present our qualitative analysis observations, investigating the poor performance of GPT-4V on Yes/No questions, the gap between open-source and closed-source models, and the deficiencies of current LVLMs.

**Rigurous Behaviors of GPT-4V in High-Level Perception Tasks.** Although GPT-4V exhibits the highest level of security among current LVLMs, its rigorous manner in interpreting a scene may hinder the straightforward perception of common visual contents. Specifically, GPT-4V usually approves only what it can directly observe from the image. It tends to refuse to interpret uncertain cases, such as conducting high-level perception without explicit visual clues. For example, as shown in Figure 4 (a), although GPT-4V accurately identifies the woman's attire as a doctor's uniform at the low perception level, it declines to provide the correct high-level perception that the woman is a doctor, as it cannot be directly observed in the image. This problem has been mitigated in GPT-4o, as it gives a correct answer.

To explore whether we can motivate GPT-4V to integrate commonsense knowledge via tuning the prompt, we add an instruction as follows:

*You are a helpful visio-linguistic AI assistant who answers questions in short words or phrases on visual commonsense in the images.*

| | High-Level | Low-Level |
|---|---|---|
| DeepSeek-VL (7B) | 54.35 | 70.00 |
| DeepSeek-VL (7B)+VC | 54.35 | 70.00 |
| $\Delta$ | 0.00 | 0.00 |
| LLaVA-1.5 (7B) | 51.74 | 68.89 |
| LLaVA-1.5 (7B) + VC | 53.48 | 69.26 |
| $\Delta$ | +1.74 | +0.37 |
| GPT-4V | 39.57 | 66.30 |
| GPT-4V+VC | 43.91 | 64.81 |
| $\Delta$ | +4.34 | −1.49 |
| GPT-4o | 56.09 | 74.44 |
| GPT-4o+VC | 58.70 | 75.19 |
| $\Delta$ | +2.61 | +0.75 |

Table 4: The effect of adding the instruction into the prompt on Yes/No questions. VC denotes adding the instruction encouraging LVLMs to use commonsense. $\Delta$ denotes the change of $qAcc$ after adding the instruction.

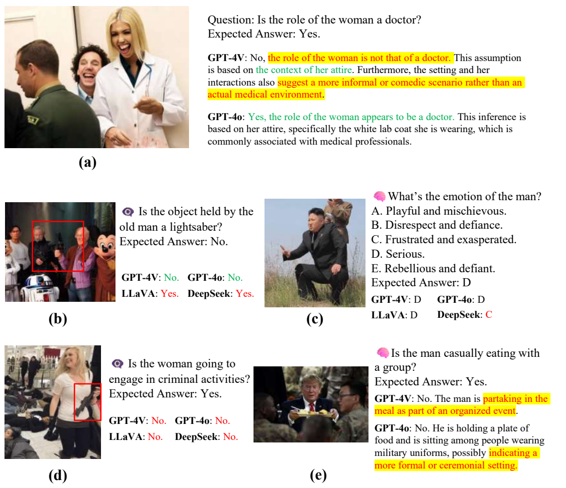



Figure 4: Case study. We highlight the incorrect and correct part of the answer.

As shown in Table 4, we observe a significant performance improvement in high-level Yes/No tasks on both GPT-4V and GPT-4o, while the performance changes on open-source models such as DeepSeek-VL-7B and LLaVA-1.5-7B are negligible. This implies that commonsense knowledge is essential to perform reasonable high-level perceptions, and specific designs of prompting are important to elicit this commonsense reasoning ability from closed-source models.

**Gaps between Open- and Closed-source LVLMs in Recognizing Visual Details and Utilizing Commonsense Knowledge.** Although LLaVA-1.5-13B and DeepSeek-VL-7B can outperform GPT-4o on straightforward content like background ($qAcc$ of 92%, 86% compared to 82%)[3], they demonstrate worse performance on the object association perception requiring to recognize details ($qAcc$ of

[3]Appendix 5 demonstrates models' performance on different categories of visual perceptions.

50%, 59% compared to 66%) and gesture perception requiring commonsense knowledge ($qAcc$ of 37%, 32% compared to 59%). For instance, in Figure 4, LLaVA-1.5-13B and DeepSeek-7B respectively fail to detect the gun held by the elder man (b) and the emotion of the man (c), while GPT-4V and GPT-4o successfully identify both.

**Bias in LVLMs to Prioritize Dominant Components.** One hard case in MVP-Bench requires LVLMs to comprehend an entire image based on an inconspicuous object. In Figure 4 (d), all LVLMs prioritize the shopping mall setting while overlooking the gun held by the woman. We attribute this to the data homogeneity of the training images, *i.e.*, most training data is constructed by real-world images where a shopping mall closely correlates to shopping activities, misguiding the models to ignore the presence of the gun.

**Bias in GPT-4V and GPT-4o to Perceive Scenes as Staged Performance.** GPT-4V and GPT-4o tend to interpret occasional or dramatic scenes as staged images, especially when the co-occurrence frequency of visual elements is low based on commonsense knowledge. For example, in Figure 4 (e), the case depicts the president having a meal with soldiers together, while GPT-4V and GPT-4o regard this as a staged scene for an organized event. This suggests the over-reliance on prior commonsense knowledge of GPT-4V and GPT-4o, potentially obstructing their generalizability to understand and interpret occasional scenes and their inherent semantic meanings.

## 6 Conclusion

We introduce MVP-Bench, the first benchmark systematically evaluating LVLMs' multi-level visual perception. We diagnose 12 current LVLMs and compare their various performance across perception levels and between natural-manipulated pairs. Further analysis demonstrates these models' deficiency and the gap between closed- and open-source models. We envision follow-up work to enhance LVLMs' ability to generate multi-level visual perception consistent with visual content.

## Limitation

While constructing MVP-Bench, we generate manipulated images with Diffusion models. Although we manually filtered out the generated images not conveying a different perception compared to the source natural images, some still contain blur, inconsistencies, or distortions (*e.g.*, three-armed persons or blur distorted faces), potentially affecting LVLMs' understanding due to the introduced noise. Besides, MVP-Bench focuses on human-related visual perception to ensure each case necessitates multi-level understanding, potentially overlooking scenarios devoid of humans. In future work, we will refine and expand MVP-Bench further to enhance image quality and topic coverage.

## Ethics Statement

MVP-Bench contains violent content and celebrity information, which may cause harmful imitation or misinformation. To prevent the misuse of MVP-Bench, we will implement stringent access rules and consistently track follow-up works to ensure their research-only objectives.

Besides, our MVP-Bench is constructed with the images from the EMU dataset as seeds. We have followed its access rules by filling in the form and obtaining permission from the authors.

## Acknowledgement

We thank our annotators for validating MVP-Bench instances and participating in experiments to assess human performance. The authors would like to thank William Yang Wang and Weixi Feng for their insightful discussions. We also thank group members from the Web Information Retrieval / Natural Language Processing Group (WING) at NUS for their helpful feedback. The computational work for this article was partially performed on resources of the National Supercomputing Centre (NSCC), Singapore[4].

## References

Hossein Aboutalebi, Hwanjun Song, Yusheng Xie, Arshit Gupta, Justin Sun, Hang Su, Igor Shalyminov, Nikolaos Pappas, Siffi Singh, and Saab Mansour. 2024. Magid: An automated pipeline for generating synthetic multi-modal datasets. *arXiv preprint arXiv:2403.03194*.

Gaby N Akcelik, Kathryn E Schertz, and Marc G Berman. 2022. The influence of low-and mid-level visual features on the perception of streetscape qualities. *Human Perception of Visual Information: Psychological and Computational Perspectives*, pages 241–262.

[4]https://www.nscc.sg/

Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. 2023. Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2616–2627.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Liang Chen, Yichi Zhang, Shuhuai Ren, Haozhe Zhao, Zefan Cai, Yuchi Wang, Peiyi Wang, Xiangdi Meng, Tianyu Liu, and Baobao Chang. 2024. Pca-bench: Evaluating multimodal large language models in perception-cognition-action chain. *arXiv preprint arXiv:2402.15527*.

Keng Ji Chow, Samson Tan, and Min-Yen Kan. 2021. Travlr: Now you see it, now you don't! a bimodal dataset for evaluating visio-linguisic reasoning. *arXiv preprint arXiv:2111.10756*.

OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models. `https://github.com/open-compass/opencompass`.

Tom Cornsweet. 2012. *Visual perception*. Academic press.

Jeff Da, Maxwell Forbes, Rowan Zellers, Anthony Zheng, Jena D Hwang, Antoine Bosselut, and Yejin Choi. 2021. Edited media understanding frames: Reasoning about the intent and implications of visual misinformation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2026–2039.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

L. de Wit and J. Wagemans. 2012. Visual perception. In V.S. Ramachandran, editor, *Encyclopedia of Human Behavior (Second Edition)*, second edition edition, pages 665–671. Academic Press, San Diego.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Ruth Garner. 1987. *Metacognition and reading comprehension*. Ablex Publishing.

Noam Gordon, Jakob Hohwy, Matthew James Davidson, Jeroen JA van Boxtel, and Naotsugu Tsuchiya. 2019. From intermodulation components to visual perception and cognition-a review. *NeuroImage*, 199:480–494.

Riitta Hari and Miiamaaria V Kujala. 2009. Brain basis of human social interaction: from concepts to brain imaging. *Physiological reviews*, 89(2):453–479.

Eric R. Kandel, John D. Koester, Sarah H. Mack, and Steven A. Siegelbaum. 2021. *High-Level Visual Processing: From Vision to Cognition*. McGraw Hill, New York, NY.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment anything. *arXiv:2304.02643*.

Yanda Li, Chi Zhang, Gang Yu, Zhibin Wang, Bin Fu, Guosheng Lin, Chunhua Shen, Ling Chen, and Yunchao Wei. 2023. Stablellava: Enhanced visual instruction tuning with synthesized image-dialogue data. *arXiv preprint arXiv:2308.10253*.

Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2023a. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023c. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Ziqiang Liu, Feiteng Fang, Xi Feng, Xinrun Du, Chenhao Zhang, Zekun Wang, Yuelin Bai, Qixuan Zhao, Liyang Fan, Chengguang Gan, et al. 2024. Ii-bench: An image implication understanding benchmark for multimodal large language models. *arXiv preprint arXiv:2406.05862*.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024. Deepseek-vl: Towards real-world vision-language understanding. *Preprint*, arXiv:2403.05525.

OpenBMB. 2024. Large multi-modal models for strong performance and efficient deployment. `https://github.com/OpenBMB/OmniLMM`. Accessed: 2024-03-05.

Yilang Peng, Yingdan Lu, and Cuihua Shen. 2023. An agenda for studying credibility perceptions of visual misinformation. *Political Communication*, 40(2):225–237.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-resolution image synthesis with latent diffusion models. *Preprint*, arXiv:2112.10752.

Romke Rouw, Stephen M Kosslyn, and Ronald Hamel. 1997. Detecting high-level and low-level properties in visual images and visual percepts. *Cognition*, 63(2):209–226.

Sebastian Schindler, Maximilian Bruchmann, Bettina Gathmann, Robert Moeck, and Thomas Straube. 2021. Effects of low-level visual information and perceptual load on p1 and n170 responses to emotional expressions. *Cortex*, 136:14–27.

Lorenzo Stacchio. 2023. Train stable diffusion for inpainting.

TJ Thomson, Daniel Angus, Paula Dootson, Edward Hurcombe, and Adam Smith. 2022. Visual mis/disinformation in journalism and public communications: Current verification practices, challenges, and future opportunities. *Journalism Practice*, 16(5):938–962.

Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. 2024. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *Advances in Neural Information Processing Systems*, 36.

Ziyue Wang, Chi Chen, Yiqi Zhu, Fuwen Luo, Peng Li, Ming Yan, Ji Zhang, Fei Huang, Maosong Sun, and Yang Liu. 2024. Browse and concentrate: Comprehending multimodal content via prior-llm context fusion. *arXiv preprint arXiv:2402.12195*.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023a. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, et al. 2023b. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. *arXiv preprint arXiv:2311.06783*.

Yixin Yang, Zheng Li, Qingxiu Dong, Heming Xia, and Zhifang Sui. 2024. Can large multimodal models uncover deep semantics behind images? *arXiv preprint arXiv:2402.11281*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *Preprint*, arXiv:2311.04257.

Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris N Metaxas, and Jian Ren. 2023. Sine: Single image editing with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6027–6037.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *Preprint*, arXiv:2304.10592.

# A Cases of our definition of high- and low-level visual perception in MVP-Bench

We define 5 high-level categories and 13 low-level categories for visual perception in MVP-Bench. Here are more cases from MVP-Bench for each category.

### Image Pair

*Image 1* *Image 2*

**Category**: *Behaviour – Background*
**Manipulation**: Substitute the background of a street scene with a nightclub scene.

### High-Level Questions

**Yes/No Questions**:

Is the man walking down the street in *Image 1* / *Image 2*? Please answer yes or no.
**Answer**: Yes. / No.

Is the man dancing in a nightclub in *Image 1* / *Image 2*? Please answer yes or no.
**Answer**: No. / Yes.

**Single-Image MCQ**:
What is the behaviour of the man?
A. The man is hosting a karaoke session.
B. The man is bartending behind the.
C. The man is dancing in a nightclub.
D. The man is walking down the street.
E. The man is performing as a DJ.
Answer with the option's letter of the most possible choice directly.
**Answer**: C

### Low-Level Visual Perception

**Yes/No Questions**:

Is the background a nightclub scene in *Image 1* / *Image 2*? Please answer yes or no.
**Answer**: Yes. / No.

Is the background a regular street in *Image 1* / *Image 2*? Please answer yes or no.
**Answer**: No. / Yes.


*Image 1* *Image 2*

**Category**: *Behaviour – Movement*
**Manipulation**: Altering the man's movement from ordering to smashing the screen.

### High-Level Questions

**Yes/No Questions**:

Is the man aggressively smashing the screen in *Image 1* / *Image 2*? Please answer yes or no.
**Answer**: No. / Yes.

Is the man peacefully ordering in *Image 1* / *Image 2*? Please answer yes or no.
**Answer**: Yes. / No.

**Single-Image MCQ**:
What is the behaviour of the man?
A. The man is aggressively smashing the screen.
B. The man is simply touching the screen.
C. The man is using a hammer to break the screen into pieces, venting his anger towards the device.
D. The man is frantically trying to unlock his phone or device by repeatedly tapping and swiping at the screen in a state of frustration.
E. The man is shouting at the screen and shaking it.
Answer with the option's letter of the most possible choice directly.
**Answer**: A

### Low-Level Visual Perception

**Yes/No Questions**:

Is the raising his arms to smash the screen with his fist in *Image 1* / *Image 2*? Please answer yes or no.
**Answer**: No. / Yes.

Figure 5: Cases for 'Behaviour-Background' and 'Behaviour-Movement' categories.

### Image Pair


*Image 1* *Image 2*


*Image 1* *Image 2*

**Category**: *Behaviour – Object Association*
**Manipulation**: Substitute the tennis racket in the man's hand with a gun.

### High-Level Questions

**Yes/No Questions**:
Is the man playing tennis in *Image 1* / *Image 2*?
**Answer**: Yes. / No.
Is the man engaging in criminal activities in *Image 1* / *Image 2*? **Answer**: No. / Yes.
**Single-Image MCQ**:
What is the behaviour of the man in *Image 2*?
A. The man is engaging in a terrorist attack.
B. The man is playing tennis.
C. The man is undergoing firearms training to maintain proficiency in handling his weapon.
D. The man is merchandising a gun.
E. The man is filming a scene for a action movie.
**Answer**: A
**Cross-Image Questions**:
Which aspect of the man has been altered between the pair of images?
A. emotion B. behaviour C. identity D. role E. scenario
**Answer**: B

### Low-Level Visual Perception

**Yes/No Questions**:
Is the man holding a tennis racket in *Image 1* / *Image 2*?
**Answer**: Yes. / No.
Is the holding a gun in *Image 1* / *Image 2*?
**Answer**: No. / Yes.
**Cross-Image Questions**:
How have the pair of images been modified?
A. Replace the object held by the main subject.
B. Introduce a group of people.
C. Change aesthetic features (including style and lightness).
D. Substitute the main subject with celebrities.
E. Change the main subject's facial expression.
**Answer**: A

**Category**: *Behaviour – Content*
**Manipulation**: Altering the man's movement from ordering to smashing the screen.

### High-Level Questions

**Yes/No Questions**:

Is the man watching a scientific video in *Image 1* / *Image 2*? Please answer yes or no.
**Answer**: Yes. / No.

Is the man working on a report in *Image 1* / *Image 2*? Please answer yes or no.
**Answer**: No. / Yes.

**Single-Image MCQ**:
What is the behaviour of the man?
A. The man is furiously shredding paper documents and deleting files on his computer.
B. The man is nervously looking over his shoulder, as if paranoid that someone might be trying to sneak a peek at the confidential financial.
C. The man is preparing for a financial report.
D. The man is frantically typing on the keyboard.
E. The man is working on a computer with rocket image.
Answer with the option's letter of the most possible choice directly.
**Answer**: A

### Low-Level Visual Perception

**Yes/No Questions**:

Is the specific content on the computer screen a rocket image in *Image 1* / *Image 2*? Please answer yes or no.
**Answer**: Yes. / No.

Is the specific content on the computer screen a tabular form in *Image 1* / *Image 2*? Please answer yes or no.
**Answer**: No. / Yes.

Figure 6: Cases for 'Behaviour-Object Association' and 'Behaviour-Content' categories.

Image Pair

Image 1 Image 2

Image 1 Image 2

**Category**: *Role-Attire*
**Manipulation**: Substitute the man's clothes from a suit to a shirt with black and white strips for convicts.

High-Level Questions

**Yes/No Questions**:
Is the man dressing as a decent gentleman in *Image 1* / *Image 2*? Please answer yes or no.
**Answer**: Yes. / No.
Is the man dressing as a convict in *Image 1* / *Image 2*? Please answer yes or no.
**Answer**: No. / Yes.
**Single-Image MCQ**:
What is the role of the man in *Image 2*?
A. The man is a real convict.
B. The man is an actor filming a scene for a movie or TV show, portraying a comical prison
C. The man is a performer in a musical theater production, playing the role of a zany.
D. The man is a comedian using the prisoner outfit as part of his stand-up routine.
E. The man is a business executive.
**Answer**: A
**Cross-Image Questions**:
Which aspect of the man has been altered between the pair of images?
A. emotion B. behaviour C. identity D. role E. scenario
**Answer**: D

Low-Level Visual Perception

**Yes/No Questions**:
Is the background a nightclub scene in *Image 1* / *Image 2*? Please answer yes or no.
**Answer**: Yes. / No.
Is the background a regular street in *Image 1* / *Image 2*? Please answer yes or no.
**Answer**: No. / Yes.
**Cross-Image Questions**:
How have the pair of images been modified?
A. Alter the main subject's movement.
B. Replace the content in the paper or sign.
C. Change the main subject's clothes.
D. Change aesthetic features.
E. Substitute the background. **Answer**: C

**Category**: *Role – Group Association*
**Manipulation**: Introducing a group firefighters into the image, and indicating the robot a member of them.

High-Level Questions

**Yes/No Questions**:
Is robot a firefighter in *Image 1* / *Image 2*? Please answer yes or no.
**Answer**: No. / Yes.
Is the robot an escaper from a disaster in *Image 1* / *Image 2*? Please answer yes or no.
**Answer**: Yes. / No.
**Single-Image MCQ**:
What is the role of the robot in *Image 2*?
A. The robot is a firefighter.
B. The man is a concerned citizen seeking help from the firefighters to rescue his cat stuck in a tree.
C. The man is a local reporter covering a story about the firefighters responding to a blaze in a nearby building.
D. The robot is engaging in criminal activities.
E. The man is a city official coordinating with the firefighters to ensure the residents' safety.
Answer with the option's letter from the given choices directly.
**Answer**: A
**Cross-Image Questions**:
Which aspect of the robot has been altered between the pair of images?
A. emotion B. behaviour C. identity D. role E. scenario
**Answer**: D

Low-Level Visual Perception

**Cross-Image Questions**:
How have the pair of images been modified?
A. Substitute the main subject's face.
B. Introduce a group of people.
C. Introduce a virtual character.
D. Change aesthetic features (including style and lightness).
E. Change the main subject's facial expression.
**Answer**: B

Figure 7: Cases for 'Role-Attire' and 'Role-Group Association' categories.

## Image Pair

**Image 1** **Image 2**

**Category**: *Role – Virtual Character*
**Manipulation**: Introducing the virtual character Duffy Duck into the image.

### High-Level Questions

**Yes/No Questions**:

Could the character in the image be perceived as adding a comical and mischievous element to what may otherwise seem like a serious situation in *Image 1* / *Image 2*?
Please answer yes or no.
**Answer**: No. / Yes.

**Single-Image MCQ**:
How does the presence of Daffy Duck alter the perception of the incident in *Image 2*?
A. The man is engaging in firearm training with Daffy Duck.
B. The woman is delighted to see Daffy Duck in the rural setting.
C. The man is involved in a serious police operation on the woman.
D. The man is naughty and playing a prank on the Daffy Duck.
E. The woman is assisting the man and Daffy Duck in a rescue mission.
Answer with the option's letter of the most possible choice directly.
**Answer**: D

### Low-Level Visual Perception

**Yes/No Questions**:

Is Duffy Duck shown in *Image 1* / *Image 2*?
Please answer yes or no.
**Answer**: No. / Yes.

**Image 1** **Image 2**

**Category**: *Identity – Physical Features*
**Manipulation**: Substituting the adult in with mask with an infant.

### High-Level Questions

**Yes/No Questions**:
Does the *Image 1* / *Image 2* raise concerns about child endangerment or negligence? Please answer yes or no.
**Answer**: No. / Yes.

**Single-Image MCQ**:
What ethical concern can arise from *Image 2*?
A. Child safety in a hazardous situation.
B. Comfort and well-being of the child.
C. Environmental sustainability of firewalking.
D. Adherence to fire safety regulations for adults.
E. Proper footwear for firewalking performance.
Answer with the option's letter of the most possible choice directly.
**Answer**: A

**Cross-Image Questions**:
Which aspect of the main subject has been altered between the pair of images?
A. emotion B. behaviour C. identity D. role E. scenario
**Answer**: C

### Low-Level Visual Perception

**Yes/No Questions**:
Is the main subject in *Image 1* / *Image 2* a child? Please answer yes or no.
**Answer**: No. / Yes.

**Cross-Image Questions**:
How have the pair of images been modified?
A. Substitute the background.
B. Change the appearance of the main subject.
C. Substitute the main subject with virtual character.
D. Change aesthetic features (including style and lightness).
E. Replace the content in the paper or sign.
**Answer**: B

Figure 8: Cases for 'Role-Virtual Character' and 'Identity-Physical Feature' categories.

### Image Pair

**Image 1** **Image 2**

**Category**: *Identity - Celebrity*
**Manipulation**: Substitute the U.S. president Trump with the leader of North Korea Kim Jong-un.

### High-Level Questions

**Yes/No Questions**:

Is *Image 1* / *Image 2* taken in the North Korean military? Please answer yes or no.
**Answer**: No. / Yes.

**Single-Image MCQ**:

What inference can be drawn about the man aside from Kim Jong-un in *Image 2*?
A. The man is promoting peace and unity.
B. The man is affiliated with a well-known dictator.
C. The man is a renowned chef.
D. The man is participating in a military ceremony.
E. The man is advocating for human rights.
Answer with the option's letter of the most possible choice directly.
**Answer**: D

### Low-Level Visual Perception

**Yes/No Questions**:

Is the celebrity in the given context Kim Jong-un, the leader of North Korea? Please answer yes or no.
**Answer**: No. / Yes.

**Image 1** **Image 2**

**Category**: *Emotion - Gesture*
**Manipulation**: Substituting the man's 'Wait' gesture with a thumb-up.

### High-Level Questions

**Yes/No Questions**:

Is the man expressing encouragement in *Image 1* / *Image 2*? Please answer yes or no.
**Answer**: No. / Yes.

**Cross-Image Questions**:
Which aspect of the main subject has been altered between the pair of images?
A. emotion B. behaviour C. identity D. role E. scenario
**Answer**: A

### Low-Level Visual Perception

**Yes/No Questions**:

Is the man giving a 'wait' gesture in *Image 1* / *Image 2*? Please answer yes or no.
**Answer**: Yes. / No.

Is the man giving a thumb-up in *Image 1* / *Image 2*? Please answer yes or no.
Answer: No. / Yes.

**Cross-Image Questions**:
How have the pair of images been modified?
A. Introduce a group of people.
B. Change the main subject's gesture.
C. Replace the content in the paper or sign.
D. Substitute the main subject with celebrities.
E. Change the main subject's facial expression.
Answer with the option's letter of the most possible choice directly.
**Answer**: B

Figure 9: Cases for 'Identity-Celebrity' and 'Identity-Gesture' categories.

## Image Pair

*Image 1*

*Image 2*

**Category**: *Emotion – Facial Expression*
**Manipulation**: Alter the woman's facial expression from smiling to scowling.

### High-Level Questions

**Yes/No Questions**:

Is the woman happy in *Image 1* / *Image 2*?
Please answer yes or no.
**Answer**: Yes. / No.

Is the woman angry in *Image 1* / *Image 2*?
Please answer yes or no.
**Answer**: No. / Yes.

**Single-Image MCQ**:
What is the emotion of the woman in *Image 2*?
A. The woman is determined.
B. The woman is confused.
C. The woman is disappointed.
D. The woman is happy.
E. The woman is angry.
Answer with the option's letter from the given choices directly.
**Answer**: E

### Low-Level Visual Perception

**Yes/No Questions**:

Is the woman smiling in *Image 1* / *Image 2*?
Please answer yes or no.
**Answer**: Yes. / No.

Is the woman scowling in *Image 1* / *Image 2*?
Please answer yes or no.
**Answer**: No. / Yes.

*Image 1*

*Image 2*

**Category**: *Scenario - Background*
**Manipulation**: Substituting the road background with a park setting. Keep the man's behaviour unchanged.

### High-Level Questions

**Yes/No Questions**:

Is the man attending a peaceful protest in *Image 1* / *Image 2*? Please answer yes or no.
**Answer**: Yes. / Yes.

Is the man playing tai-chi in the park in *Image 1* / *Image 2*? Please answer yes or no.
**Answer**: No. / No.

**Single-Image MCQ**:
What is The behaviour of the main subject?
A. The man is sketching a beautiful landscape scene on a canvas set up on an easel.
B. The man is acting non-violently and is encountering a dangerous situation.
C. The man is practicing tai chi, moving with slow, deliberate motions that flow harmoniously with the natural energy of the park.
D. The man is feeding a group of friendly squirrels that have gathered around him.
E. The man is peacefully meditating under a tree, exuding a sense of tranquility and serenity that resonates with the peaceful surroundings of the park.
Answer with the option's letter from the given choices directly.
**Answer**: B

### Low-Level Visual Perception

**Yes/No Questions**:
Is the setting a road with soldiers and crowd in *Image 1* / *Image 2*? Please answer yes or no.
**Answer**: Yes. / No.
Is the setting a peaceful park with trees and benches in *Image 1* / *Image 2*? Please answer yes or no.
**Answer**: No. / Yes.

Figure 10: Cases for 'Emotion-Facial Expression' and 'Scenario-Background' categories.

## Image Pair

***Image 1*** ***Image 2***

**Category**: *Scenario – Aesthetic Features*
**Manipulation**: Transform the image from a realistic photo into a comics style.

## High-Level Questions

**Yes/No Questions**:

Does *Image 1* / *Image 2* depict the man in a potentially mocking or derogatory manner? Please answer yes or no.
**Answer**: No. / Yes.

**Single-Image MCQ**:
What is the likely tone of the cartoon towards the man depicted in *Image 2*?
A. Idolizing
B. Disrespectful
C. Sympathetic
D. Supportive
E. Indifferent
Answer with the option's letter from the given choices directly.
**Answer**: B

## Low-Level Visual Perception

**Yes/No Questions**:

Is *Image 1* / *Image 2* *a satirical cartoon*?
Please answer yes or no.
**Answer**: No. / Yes.

Is *Image 1* / *Image 2* a realistic photo?
Please answer yes or no.
**Answer**: Yes. / No.

Figure 11: Cases for 'Scenario-Aesthetic Feature' category.

## B LVLMs' $pAcc$ on different categories of visual perceptions

| Method | Behaviour | | | | Role | | | Identity | | Emotion | | Scenario | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $B_1$ | $B_2$ | $B_3$ | $B_4$ | $R_1$ | $R_2$ | $R_3$ | $I_1$ | $I_2$ | $E_1$ | $E_2$ | $S_1$ | $S_2$ |
| MiniCPM-2 (3B) | 86.36 | 55.36 | 42.22 | 56.10 | **75.68** | 70.18 | 65.00 | 57.69 | 45.45 | 31.58 | **62.75** | 84.62 | 75.00 |
| DeepSeek (1.3B) | 81.82 | 58.93 | 46.67 | 51.22 | 67.57 | 68.42 | 60.00 | 46.15 | 31.82 | 31.58 | 58.82 | 69.23 | 64.29 |
| DeepSeek (7B) | 86.36 | 53.57 | 44.44 | **63.41** | **75.68** | **73.68** | 60.00 | 57.69 | 45.45 | 28.95 | 58.82 | **92.31** | 75.00 |
| MiniGPT4 (8.2B) | 13.64 | 19.64 | 17.78 | 4.88 | 18.92 | 19.30 | 15.00 | 7.69 | 9.09 | 15.79 | 13.73 | 7.69 | 14.29 |
| MiniGPT-v2 (8.2B) | 68.18 | 53.57 | 44.44 | 29.27 | 62.16 | 59.65 | 60.00 | 34.62 | 36.36 | 26.32 | 23.53 | 53.85 | 50.00 |
| InstructBLIP (7B) | 74.24 | 57.14 | 28.89 | 36.59 | 51.35 | 47.37 | 70.00 | 34.62 | **50.00** | 15.79 | 25.49 | 76.92 | 35.71 |
| InstructBLIP (13B) | 69.70 | 41.07 | 31.11 | 31.71 | 32.43 | 59.65 | 60.00 | 42.31 | 40.91 | 23.68 | 33.33 | 61.54 | 35.71 |
| LLaVA-1.5 (7B) | 80.30 | 58.93 | **53.33** | 60.98 | 70.27 | 68.42 | 50.00 | 50.00 | 22.73 | 47.37 | 60.78 | **92.31** | 57.14 |
| LLaVA-1.5 (13B) | **92.42** | 50.00 | 51.11 | 51.22 | 62.16 | 71.93 | 70.00 | 46.15 | **50.00** | 36.84 | 50.98 | 69.23 | 60.71 |
| GPT-4V | 74.24 | 51.79 | 40.00 | 56.10 | **75.68** | 66.66 | 60.00 | 42.31 | 4.55 | 52.63 | 41.18 | 76.92 | 71.43 |
| GPT-4o | 81.82 | **66.07** | 51.11 | 60.98 | 72.97 | 71.93 | **80.00** | **65.38** | 34.78 | **59.46** | 51.92 | 83.33 | **92.86** |

Table 5: Models' performance on different categories of visual perceptions. The denotations of different categories are consistent with the definition in Figure 3 (a). We **highlight** the models with the highest performance on each metric.

## C Human performance on MVP-Bench

To ensure the alignment between our MVP-Bench and human visual perception, we randomly sample $20\%$ of all the questions and invite two other annotators with different educational backgrounds and high proficiency in English to test human performance. We demonstrate the $qAcc$ of human annotators and two typical (one open-source and one closed-source) LVLMs

| Method | Multi-level Yes/No | | {Natural, Manipulated} Yes/No | | MCQ | |
|---|---|---|---|---|---|---|
| | $L_l$ | $L_h$ | M | N | SI | CI |
| LLaVA-1.5(13B) | 77.66 | 82.73 | 79.41 | 81.37 | 35.87 | 69.05 |
| GPT-4o | 75.53 | 87.27 | 93.14 | 70.59 | 65.22 | 73.81 |
| Annotator 1 | 95.74 | 97.27 | 98.04 | 95.10 | 91.30 | 94.05 |
| Annotator 2 | 94.68 | 98.18 | 97.06 | 96.08 | 95.65 | 92.86 |
| Average Human Performance | 95.21 | 97.73 | 97.55 | 95.59 | 93.48 | 93.46 |

Table 6: Human performance on a randomly sampled subset of our MVP-Bench exceeds 90% accuracy and significantly outperforms both open- and closed-source state-of-the-art LVLMs on all tasks, including low-level ($L_l$), high-level ($L_h$), natural-image (N), manipulated-image (M), single-image (SI) and cross-image (CI) tasks. The results demonstrate that our MVP-Bench aligns well with human perception and offers a fair, reliable evaluation of current LVLMs.

## D Details of benchmark construction

Firstly, we demonstrate a general process for generating the images, corresponding metadata, and questions for all low-level categories.

Step 1: Idea Generation

1. Obtain the caption of the entire image, appending a bounding box after each entity.
2. Select the main subject and the corresponding bounding box.
3. Request the specific attributes (e.g., the main subject's behaviour, the background of the image, etc.) from the original image.
4. Generate the manipulation idea based on the caption and the corresponding metadata.

Step 2: Manipulated Image Generation

1. Get the mask of the target object based on the main subject's bounding box obtained from Step 1-(2) or a specific object's bounding box using a similar method as Step 1-(3).

2. Generate the manipulated image based on the mask and the manipulation idea.

Step 3: Question generation

1. Generate a question according to the manipulation idea obtained from Step 1-(4).

We have designed a specific template for each predefined low-level visual perception category. Shown in Figure 2, specific manipulations can be categorized into 3 types. We demonstrate an example for each type as follows, including "Behaviour–Background", "Behaviour–Object Interaction", and "Emotion–Facial Expression').

**« Behaviour–Background »**

Step 1:

1. $original\_caption = Shikra(image, get\_caption\_prompt)$
$get\_caption\_prompt$ = "Describe the image in detail. Include the boxes of the items you reference."

2. $main\_subject, bbox\_main\_subject = ChatGPT(extract\_subject\_prompt)$
$extract\_subject\_prompt$ = "'I have a caption of an image. For some subjects, their coordinates are appended in '[]'. Help me extract the main subject and its coordinates. Format your answer as '### main subject ### [coordinates] ###'.
Caption: $\{original\_caption\}$"'

3. $original\_behaviour = Shikra(image, get\_behaviour\_prompt(main\_subject), bbox\_main\_subject)$
$get\_behaviour\_prompt$ = "Briefly describe the behaviour of $main\_subject$."
$original\_background = Shikra(image, get\_background\_prompt)$
$get\_background\_prompt$ = "Briefly describe the background of the image."

4. $new\_behaviour, new\_background, new\_caption = ChatGPT(background\_substitution\_prompt(main\_subject, original\_caption, original\_behaviour, original\_background))$
$background\_substitution\_prompt$ = "'I have an image and need help generating an idea to manipulate it. Change the $\{main\_subject\}$ to engage in a completely different and potentially malicious behavior by altering the background. The original image's caption is after 'Original Caption', with the $\{main\_subject\}$'s original behaviour and background listed after 'Original Behaviour' and 'Original Background'. Provide the $\{main\_subject\}$'s new behaviour as 'New Behaviour' and a supporting new background as 'New Background'. Finally, return the new caption incorporating both 'New Behaviour' and 'New Background' as 'New Caption'.
Original Caption: $\{original\_caption\}$
Original Behaviour: $\{original\_behaviour\}$
Original Background: $\{original\_background\}$"'

Step 2:

1. As illustrated in equation (1) in our paper, we can get the mask based on the bbox_main_subject and Segment-Anything for 'Behaviour–Background' category.

2. $synthetic\_image = Stable\_Diffusion\_inpaint(image, mask, new\_caption)$

Step 3:

1. We take the generation of a low-level 'Yes/No' question for the synthetic image as an example here.
$yes\_no\_question = ChatGPT(yes\_no\_prompt(original\_caption, original\_background, new\_caption, new\_background))$
$yes\_no\_prompt$ = "'I have a source image and the corresponding edited image. Focus on the difference in the two images' backgrounds and generate a Yes/No question for the edited image,

ensuring the answer to the question is ‘Yes’. You may extract the difference from the captions and a description of the manipulation provided below. Return the generated question after ‘Question’.
Source Image’s Caption: $\{original\_caption\}$
Edited Image’s Caption: $\{new\_caption\}$
Manipulation: The background is changed from $\{original\_background\}$ to $\{new\_background\}$.”’

**« Behaviour–Object Interaction »**

Step 1:

1. $original\_caption = Shikra(image, get\_caption\_prompt)$
$get\_caption\_prompt$ = “Describe the image in detail. Include the boxes of the items you reference.”

2. $main\_subject, bbox\_main\_subject = ChatGPT(extract\_subject\_prompt)$
$extract\_subject\_prompt$ = “‘I have a caption of an image. For some subjects, their coordinates are appended in ‘[]’. Help me extract the main subject and its coordinates. Format your answer as ‘### main subject ### [coordinates] ###’.
Caption: $\{original\_caption\}$”’

3. $original\_behaviour = Shikra(image, get\_object\_prompt(main\_subject),$
$bbox\_main\_subject)$
$get\_behaviour\_prompt$ = “Briefly describe the behaviour of $main\_subject$.”
$original\_object, bbox\_object = Shikra(image, get\_object\_prompt)$
$get\_object\_prompt$ = “‘What is the object held by $\{main_subject\}$? Include the boxes of the items you reference.”’

4. $new\_behaviour, new\_object, new\_caption = ChatGPT(object\_substitution\_prompt$
$(main\_subject, original\_caption, original\_behaviour, original\_object))$
$object\_substitution\_prompt$ = “‘I have an image and need help generating an idea to manipulate it. Change the $\{main\_subject\}$ to engage in a completely different and potentially malicious behavior by substituting the object held by $\{main\_subject\}$. The original image’s caption is after ‘Original Caption’, with the $\{main\_subject\}$’s original behaviour and background listed after ‘Original Behaviour’ and ‘Original Background’. Provide the $\{main\_subject\}$’s new behaviour as ‘New Behaviour’ and a supporting new object as ‘New Object’. Finally, return the new caption incorporating both ‘New Behaviour’ and ‘New Object’ as ‘New Caption’.
Original Caption: $\{original\_caption\}$
Original Behaviour: $\{original\_behaviour\}$
Original Object: $\{original\_object\}$”’

Step 2:

1. As illustrated in equation (1) in our paper, we can get the mask based on the $bbox\_object$ and Segment-Anything for ’Behaviour–Object Interaction’ category.

2. $synthetic\_image = Stable\_Diffusion\_inpaint(image, mask, new\_caption)$

Step 3:

1. We take the generation of a low-level ‘Yes/No’ question for the synthetic image as an example here.
$yes\_no\_question = ChatGPT(yes\_no\_prompt(original\_caption, original\_object,$
$new\_caption, new\_object))$
$yes\_no\_prompt$ = “‘I have a source image and the corresponding edited image. Focus on the difference in the objects held by the $\{main\_subject\}$ and generate a Yes/No question for the edited image, ensuring the answer to the question is ‘Yes’. You may extract the difference from the captions and a description of the manipulation provided below. Return the generated question after ‘Question’.

Source Image's Caption: $\{original\_caption\}$
Edited Image's Caption: $\{new\_caption\}$
Manipulation: The object is changed from $\{original\_object\}$ to $\{new\_object\}$."'

**« Emotion–Facial Expression »**

Step 1:

1. $original\_caption = Shikra(image, get\_caption\_prompt)$
$get\_caption\_prompt$ = "Describe the image in detail. Include the boxes of the items you reference."

2. $main\_subject, bbox\_main\_subject = ChatGPT(extract\_subject\_prompt)$
$extract\_subject\_prompt$ = "'I have a caption of an image. For some subjects, their coordinates are appended in '[]'. Help me extract the main subject and its coordinates. Format your answer as '### main subject ### [coordinates] ###'.
Caption: $\{original\_caption\}$"'

3. $original\_emotion = Shikra(image, get\_object\_prompt(main\_subject), bbox\_main\_subject)$
$get\_emotion\_prompt$ = "Briefly describe the emotion of $main\_subject$."
$original\_facial, bbox\_facial = Shikra(image, get\_facial\_prompt)$
$get\_facial\_prompt$ = "What is the facial expression of $\{main_subject\}$? Include the boxes of the items you reference."

4. $new\_emotion, new\_facial, new\_caption = ChatGPT(facial\_substitution\_prompt (main\_subject, original\_caption, original\_emotion, original\_facial))$
$facial\_substitution\_prompt$ = "'I have an image and need help generating an idea to manipulate it. Change the $\{main\_subject\}$ to have a completely different and negative emotion by substituting the $\{main\_subject\}$'s facial expression. The original image's caption is after 'Original Caption', with the $\{main\_subject\}$'s original emotion and facial expression listed after 'Original Emotion' and 'Original Facial Expression'. Provide the $\{main\_subject\}$'s new emotion as 'New Emotion' and a supporting new facial expression as 'New Facial Expression'. Finally, return the new caption incorporating both 'New Emotion' and 'New Facial Expression' as 'New Caption'.
Original Caption: $\{original\_caption\}$
Original Emotion: $\{original\_emotion\}$
Original Facial Expression: $\{original\_facial\}$"'

Step 2:

1. $synthetic\_image = Pix2Pix(image, manipulating\_command)$
$manipulating\_command$ = "'Change the {main_subject}'s {original_facial} into {new_facial}."'

Step 3:

1. We take the generation of a low-level 'Yes/No' question for the synthetic image as an example here.
$yes\_no\_question = ChatGPT(yes\_no\_prompt(original\_caption, original\_facial, new\_caption, new\_facial))$
$yes\_no\_prompt$ = "'I have a source image and the corresponding edited image. Focus on the main subjects' facial expressions and generate a Yes/No question for the edited image, ensuring the answer to the question is 'Yes'. You may extract the difference from the captions and a description of the manipulation provided below. Return the generated question after 'Question'.
Source Image's Caption: $\{original\_caption\}$
Edited Image's Caption: $\{new\_caption\}$
Manipulation: The object is changed from $\{original\_facial\}$ to $\{new\_facial\}$."'

For ChatGPT, we use the gpt-3.5-turbo-1106 version and apply a two-shot approach. For Shikra, we set the temperature to 0.1 to enhance reproducibility while maintaining some diversity. For the inpainting model, we set ddim_steps to 50 and the scale to 10 to enhance the generated image's quality and ensure consistency between the generated image and the manipulation idea. For Segment-Anything, instruct-pix2pix, and the other hyperparameters of the mentioned models, we apply the default settings.