AutoBench-V: Can Large Vision-Language Models Benchmark Themselves?

Anonymous authors
Paper under double-blind review

Abstract

Large Vision-Language Models (LVLMs) have become essential for advancing the integration of visual and linguistic information, facilitating a wide range of complex applications and tasks. However, the evaluation of LVLMs presents significant challenges as the evaluation benchmark always demands lots of human cost for its construction, and remains static, lacking flexibility once constructed. Even though automatic evaluation has been explored in textual modality, the visual modality remains under-explored. As a result, in this work, we address a question: "Can LVLMs serve as a path to automatic benchmarking?". We introduce AutoBench-V, an automated framework for serving evaluation on demand, *i.e.*, benchmarking LVLMs based on specific aspects of model capability. Upon receiving an evaluation capability, AutoBench-V leverages text-to-image models to generate relevant image samples and then utilizes LVLMs to orchestrate visual question-answering (VQA) tasks, completing the evaluation process efficiently and flexibly. Through an extensive evaluation of seven popular LVLMs across five demanded user inputs (*i.e.*, evaluation capabilities), the framework shows effectiveness and reliability. We observe the following: (1) Our constructed benchmark accurately reflects varying task difficulties; (2) As task difficulty rises, the performance gap between models widens; (3) While models exhibit strong performance in abstract level understanding, they underperform in detailed reasoning tasks; and (4) Constructing a dataset with varying levels of difficulties is critical for a comprehensive and exhaustive evaluation. Overall, AutoBench-V not only successfully utilizes LVLMs for automated benchmarking but also reveals that LVLMs as judges have significant potential in various domains.

1 Introduction

The flourishing of Large Language Models (LLMs) (Touvron et al., 2023; Achiam et al., 2023; Liu et al., 2024a; Anthropic, 2024) has paved the way for significant advancements in the field of natural language processing (NLP) (Brown et al., 2020; Vaswani et al., 2023). As the capabilities of LLMs grew, researchers began to explore the integration of visual information understanding capabilities into LLMs, giving rise to the development of Large Vision-Language models (LVLMs) (Achiam et al., 2023). These models are trained on extensive paired image-text datasets, enabling them to perform sophisticated multimodal reasoning by effectively integrating visual and textual information (Zou et al., 2023; Ghandi et al., 2023; Karras et al., 2019; Agrawal et al., 2016).

With the widespread adoption of LVLMs, evaluating these models has become increasingly important, for understanding their limitations and reliability better. Recent research (Xu et al., 2023; Liu et al., 2023; Ying et al., 2024; Li et al., 2023b;a; Yin et al., 2023) emphasize the urgent need for comprehensive and sophisticated evaluation standards that accurately assess LVLMs' abilities across various modalities. Various benchmarks are aiming to evaluate a range of capabilities of LVLMs including 3D understanding (Yin et al., 2023), perception and cognition capacity (Liu et al., 2023; Fu et al., 2024), multi-discipline understanding and reasoning (Yue et al., 2024). Even though these works have solidly evaluated certain aspects of LVLMs' capabilities, they lack the flexibility to support on-demand evaluation across various capability aspects. Recent studies have explored

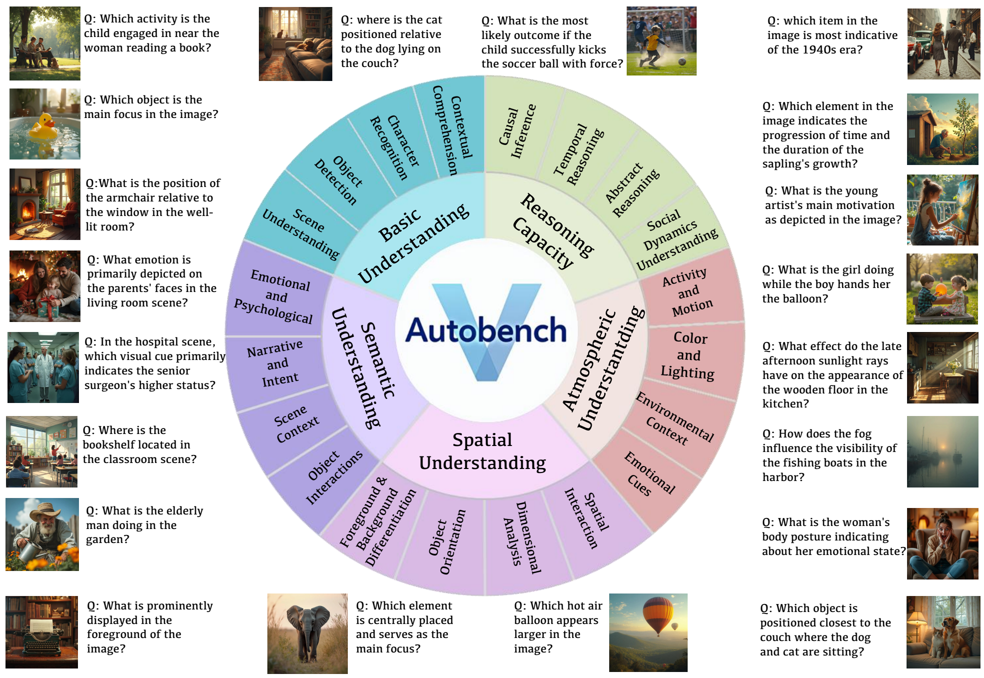


Figure 1: Five key evaluation dimensions supported by AUTOBENCH-V, along with their fine-grained sub-aspects, accompanied by questions and images to assist in understanding.

the usage of generative AI in automating evaluation, which offers flexibility in varying evaluation dimensions and reduces the human cost of benchmark dataset construction (Wu et al., 2024; Zhu et al., 2024a; Li et al., 2024). While these studies focus on the automatic evaluation of LLMs, we aim to extend this to visual modality by answering addressing this question: ***"Can LVLMs serve as a path to automatic benchmarking?"***

Automating the evaluation of LVLMs presents several key challenges. First, the targeted capabilities to be evaluated must be clearly identified based on the input demand. This is the foundation that relevant images and appropriate visual question-answering (VQA) tasks can be generated to accurately assess the LVLMs' performance in those specific aspects. Second, the generated images and VQA tasks should be relevant and accurately reflect the evaluation target. Third, the risk of answer leakage from *Examiner LVLM* during question generation should be mitigated. This issue arises when the model responsible for generating questions exhibits self-enhancement bias (Ye et al., 2024; Zheng et al., 2023), wherein the model being evaluated is also employed to generate the evaluation cases.

To address the above challenges, we propose AUTOBENCH-V, which supports automated evaluation of LVLMs based on a user demand regarding specific aspects of model capability (*e.g.*, *Spatial Understanding*). Initially, the input demand is processed by an *examiner LVLM*, which categorizes it into several overarching aspects. Each aspect is further divided into several fine-grained components, for which image descriptions of varying difficulty levels are generated. To ensure that the descriptions align with their corresponding images, a self-validation mechanism is applied using VQA (Agrawal et al., 2016). Furthermore, an error control mechanism is implemented to prevent a negative impact on the generation of questions and reference answers. The generated questions and images are then presented to the evaluated LVLM to generate responses, which are assessed against reference answers (Liu et al., 2024b). The pipeline of AUTOBENCH-V is shown in Figure 2.

By leveraging AUTOBENCH-V, we conduct extensive evaluation of seven popular LVLMs across five demanded evaluation capabilities (see Figure 1). The results show that LVLMs exhibit declining performance as task difficulty rises, with varied performances over distinctive LVLMs. While excelling in high-level understanding, they struggle with detailed reasoning, revealing a key area for improvement in future research. We also carried out several human evaluation experiments on the generated cases, which yielded positive results, demonstrating the reliability of our approach. To summarize, our key contributions are three-fold:

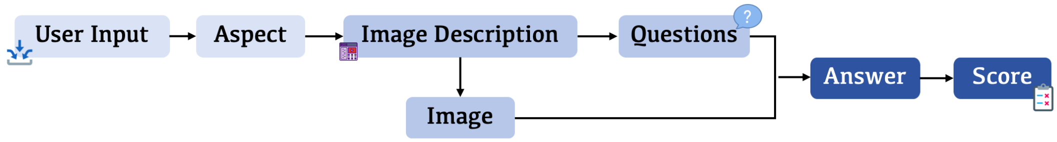


Figure 2: An overview of the AUTOBENCH-V pipeline, illustrating the automated evaluation process. It starts from user input intake, then aspect generation, followed by the generation of corresponding images and questions, and finally outputs the evaluation score of LVLMs.

▷ **An automated LVLM evaluation framework.** This proposed AUTOBENCH-V is the first automated framework for benchmarking LVLMs' capability. The framework leverages text-to-image models to generate images for evaluation and employs `GPT-4o` as an examiner to conduct VQA evaluations. This automation significantly reduces human involvement, enhancing the efficiency and objectivity of the evaluation process.

▷ **Extensive experiments validating the framework's effectiveness.** We conducted comprehensive experiments, including main evaluations on multiple models, examiner superiority tests, option position bias analysis, and human assessments. The results confirm the framework's robustness and effectiveness in evaluating LVLMs.

▷ **In-depth analysis of LVLMs' performance across diverse visual tasks.** Through systematic evaluation with varied user inputs, we find that LVLMs demonstrate strong proficiency in abstract conceptual understanding while exhibiting comparatively lower performance in concrete visual reasoning tasks. These insights offer a perspective on the current state of LVLM technology, highlighting areas with potential for future development and exploration.

2 RELATED WORKS

Benchmark for LVLMs. The emergence of the LVLMs greatly promoted the development of the multimodal model, demonstrating exceptional progress in their multimodal perception and reasoning capabilities. This makes the past, focused on isolated task performance benchmarks (Karpathy & Fei-Fei, 2015; Agrawal et al., 2016) insufficient to provide a comprehensive evaluation. Subsequent studies have introduced benchmarks for assessing LVLMs across a range of multimodal tasks (Goyal et al., 2017; Lin et al., 2015; Russakovsky et al., 2015). However, these benchmarks often fall short in providing fine-grained assessments of abilities and robust evaluation metrics.

Hence, recent works (Xu et al., 2023; Liu et al., 2023; Ying et al., 2024; Fu et al., 2024; Yin et al., 2023; Chen et al., 2024; Yu et al., 2023; 2024) highlight the critical need for developing advanced, comprehensive benchmarks to more accurately assess LVLMs' multimodal understanding and reasoning capabilities. However, these benchmarks still have different kinds of limitations. For example, LVLM-eHub (Xu et al., 2023) and LAMM (Yin et al., 2023) have utilized several classical datasets that are widely recognized but not sufficiently novel for current advancements, overlooking the possibility of data leakage during LVLM training. Hence, MMStar (Chen et al., 2024) aims to solve the unnecessity of visual content and unintentional data leakage that exists in LVLM training via constructing an elite vision-indispensable dataset.

Compared to previous work, AUTOBENCH-V not only automates the entire benchmarking process for LVLMs—significantly reducing human workload and minimizing subjective biases—but also scales up and customizes the evaluation process to address fine-grained user needs.

Automatic benchmarks. The significant early advancements in LLMs have driven the development of various benchmarks designed to automate evaluation processes. For example, LMExamQA (Bai et al., 2023b) employs the concept of a Language-Model-as-an-Examiner to create a comprehensive and scalable evaluation framework. In addition, DYVAL (Zhu et al., 2024a) and DYVAL2 (Zhu et al., 2024b) both highlight the importance of dynamic assessment, with DYVAL focusing on reasoning tasks and DYVAL2 adopting a broader psychometric approach. AutoBencher (Li et al., 2024) automates the generation of novel, challenging, and salient datasets for evaluating LLMs, further expanding the scope of automated benchmarking. Other efforts, such as UNIGEN (Wu et al., 2024) and Task Me Anything (Zhang et al., 2024a), focus on developing more tailored and relevant benchmarks for assessing LLM/LVLMs performance across diverse tasks.

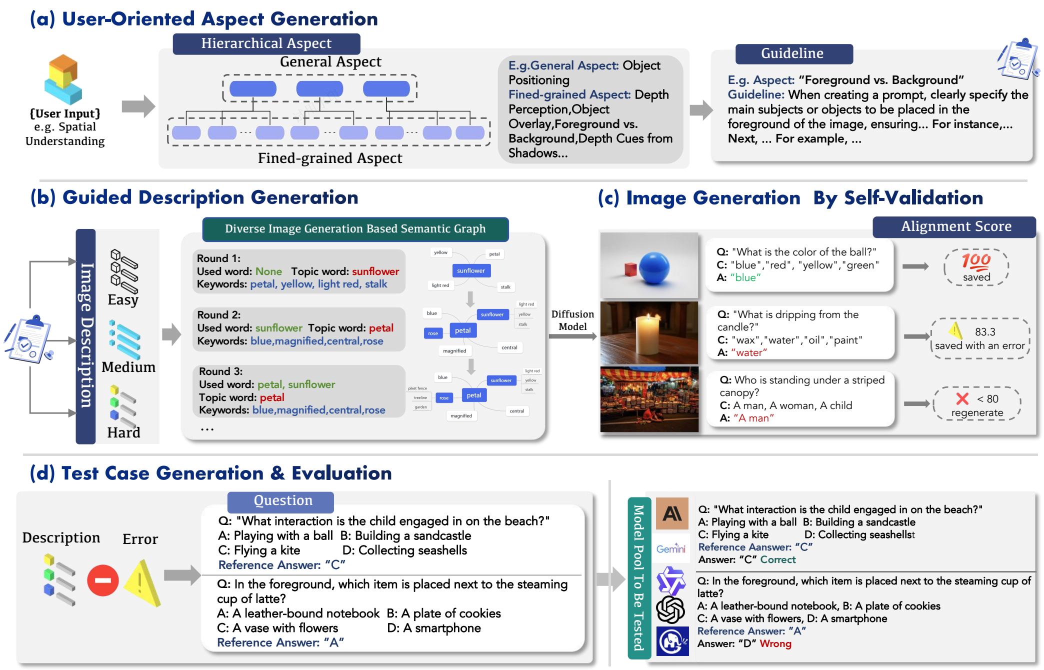


Figure 3: A comprehensive overview of the AUTOBENCH-V framework.

3 AUTOBENCH-V

In this section, we introduce AUTOBENCH-V, a framework designed for automating the process of benchmarking LVLMs, empowered by a LVLM $\mathcal{M}_v$ and a text-to-image model $\mathcal{M}_d$. As shown in Figure 3, AUTOBENCH-V consists of four modules: user-oriented aspect generation, guided description generation, image generation by self-validation, and test case generation & evaluation.

3.1 USER-ORIENTED ASPECT GENERATION

User input. The user input can specify an evaluation target focused on certain aspects of LVLMs' capability. AUTOBENCH-V covers the following key evaluation aspects, which are the most crucial for assessing the capabilities of LVLMs: ***Basic Understanding***, ***Spatial Understanding*** (Li et al., 2023b), ***Semantic Understanding*** (Meng et al., 2024), ***Reasoning Capacity*** (Liu et al., 2023), and ***Atmospheric Understanding*** (Geetha et al., 2024). Notably, the user input is not limited to the above kinds, and can be customized as needed.

Hierarchical aspect generation. For each user input, we derive a set of aspects representing specific capability items. For example, as shown in Figure 1, *contextual comprehension* is an aspect under *Basic Understanding*. However, directly generating aspects from user input can lead to excessive repetition, reducing both diversity and reliability by overlapping in semantics and repeatedly evaluating the same capability. To mitigate this, we propose hierarchical aspect generation inspired by the previous study (Qin et al., 2023) to constrain the aspect generation process. Formally, given the user input q, we first generate n general aspects $\{A_1^{(g)}, A_2^{(g)}, \ldots, A_n^{(g)}\}$ by $\mathcal{M}_v$, which can be formulated as: $\{A_1^{(g)}, A_2^{(g)}, \ldots, A_n^{(g)}\} = \mathcal{M}_v(q)$. These general aspects represent high-level evaluation dimensions based on q. Next, for each general aspect $A_i^{(g)}$, we further generate m fine-grained aspects $\{A_{i1}^{(f)}, A_{i2}^{(f)}, \ldots, A_{im}^{(f)}\}$, where each fine-grained aspect provides more specific criteria related to the general aspect. The fine-grained aspects are also generated by $\mathcal{M}_v$ and depend on both the user input q and the corresponding general aspect $A_i^{(g)}$. The fine-grained aspect of generation can be represented as $\{A_{i1}^{(f)}, A_{i2}^{(f)}, \ldots, A_{im}^{(f)}\} = \mathcal{M}_v(q, A_i^{(g)})$. Thus, the hierarchical aspect generation yields a structured set of evaluation aspects (*i.e.*, fine-frained aspect) $\mathcal{A} = \bigcup_{i=1}^{n} \left(\{A_i^{(g)}\} \cup \bigcup_{j=1}^{m} \{A_{ij}^{(f)}\} \right)$, where $|\mathcal{A}| = mn$.

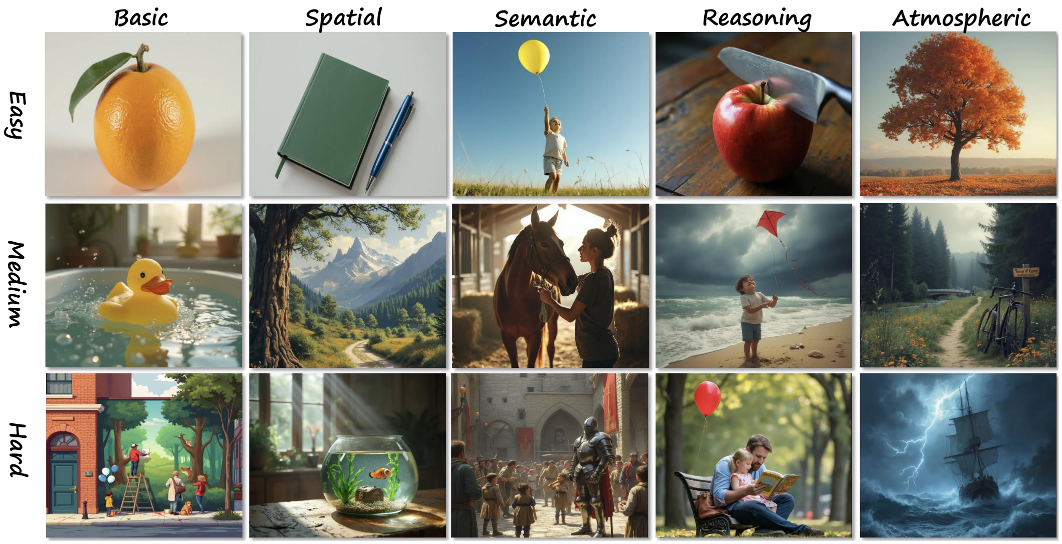


Figure 4: Images examples corresponding to different user inputs at varying difficulty levels.

3.2 Guided Description Generation

Guidelines formulation. To avoid the generation of *irrelevant*, *abstract*, or *vague details* that could lead to discrepancies in image descriptions, we introduce a guideline generation step. Before generating image descriptions, the LVLM model $\mathcal{M}_v$ formulates a guideline $\mathcal{D}_{ij}$ (*e.g.*, in *Background vs Foreground* aspect, it is essential to distinguish between the elements present in the background and those in the foreground) for each fine-grained aspect $A_{ij}^{(f)}$. This guideline acts as a guideline for $\mathcal{M}_v$, ensuring that the generated descriptions are coherent, clear, and specific to the fine-grained aspects under evaluation. The process can be expressed as $\mathcal{D}_{ij} = \mathcal{M}_v(A_{ij}^{(f)})$. The generated guideline $\mathcal{D}_{ij}$ is then utilized to guide the subsequent image description $\mathcal{T}_{ij}$.

Image description with difficulty grading. To enable a more comprehensive evaluation, we introduce a difficulty-grading mechanism for the image descriptions, which includes the evaluation cases from different difficulties. This is achieved by classifying the generated image descriptions into three difficulty levels: easy, medium, and hard. We show the examples across different difficulties in Figure 4. The difficulty level d is determined by key factors such as background complexity, element relationships, and the intricacy of textures. The generation of ω image descriptions $\{\mathcal{T}_{ij1}^d, \mathcal{T}_{ij2}^d, \ldots, \mathcal{T}_{ij\omega}^d\}$ for $A_{ij}^{(f)}$ at a specific difficulty level d can be defined as: $\bigcup_{k=1}^{\omega}\{\mathcal{T}_{ijk}^d\} = \mathcal{M}_v(q, A_{ij}^{(f)}, \mathcal{D}_{ij}, d)$, where $d \in \{\text{easy}, \text{medium}, \text{hard}\}$. This grading strategy allows for a nuanced understanding of the model's capabilities across a range of challenges with details provided in Appendix C.

Diverse description generation strategy. A key challenge when generating image descriptions at the same difficulty level is minimizing repetitive elements and backgrounds, which can reduce the diversity and generalization of the evaluation. For example, given a user input q related to spatial understanding, the model $\mathcal{M}_v$ might tend to produce descriptions centered around urban landscapes, potentially compromising the variety of test cases. To address this, we introduce a description optimization strategy using a semantic graph (Quillian, 1966) to enhance the diversity of image prompts generated by $\mathcal{M}_v$, with significant results referred to in Appendix Figure 10. For a visualization of specific words, see Appendix Figure 11 and Figure 12. The process is iterative, and during the e'th iteration of prompt generation, a topic word t_e and a set of $|c|$ related keywords $K_e = \{k_{e1}, k_{e2}, \ldots, k_{ec}\}$ are selected. These keywords are added as nodes to the semantic graph G, where nodes are connected by edges representing semantic relationships between them.

Formally, let $G_e = (V_e, E_e)$ be the semantic graph generated at iteration e, and let $S_e = (V_{e-1} \cup \{t_e\} \cup K_e)$. Then $V_e = S_e \setminus f(S_e)$ represents the node set of topic words and keywords, and E_e is the set of edges capturing the relationships between them. After each round of prompt generation, we apply a degree-based exclusion mechanism, where the number of excluded nodes is determined by a function $f(S_e)$. This function defines the number of top-degree nodes to be excluded, allowing flexibility in adjusting how many frequently used words are removed as the iterations progress. The function $f(S_e)$ could be a simple function such as $f(S_e) = \underset{V' \subseteq S_e, |V'|=e}{\arg\max} \sum_{v \in V'} \deg(v)$, where $\deg(v)$ represents the degree of node $v \in V_i$, or it could take a more complex form based on specific conditions. We mitigate redundancy and promote diversity in the generated prompts by excluding these high-degree nodes, which correspond to the most commonly used words. The function $f(S_e)$ offers the flexibility to control how aggressively the exclusion process operates based on the round number e.

Overall, the generation of an image description $\mathcal{T}_{ij}^{e}$ can be formalized as follows:

$$\bigcup_{k=1}^{\omega} \{\mathcal{T}_{ijk}^{de}\} = M_v(q, A_{ij}^{(f)}, \mathcal{D}_{ij}, V_e, d),$$

where V_e represents the refined and diverse set of topic words and keywords after the exclusion mechanism has been applied. We show the detailed procedure in Algorithm 1.

Algorithm 1 Diverse Description Generation Strategy

Input: User input q, model $\mathcal{M}_v$, initial set of topic words and keywords V_0, exclusion function $f(S_e)$, number of iterations N

Output: Set of diverse image descriptions $\{\mathcal{T}(1), \mathcal{T}(2), \ldots, \mathcal{T}(N)\}$

```
Initialize iteration counter e := 1
Initialize the set of topic words and keywords V_1 := V_0
while e ≤ N do
    Select a topic word t_e and a set of related keywords K_e = {k_e1, k_e2, ..., k_ec}
    Form the node set S_e = V_{e-1} ∪ {t_e} ∪ K_e
    Formulate E_e, where edge represent semantic relationship.
    Identify exclusion set f(S_e) = argmax_{V' ⊆ S_e, |V'|=e} Σ_{v∈V'} deg(v)
    Update the node set as V_e = S_e \ f(S_e)
    Set superparameter d and ω
    Set T(e) = ⋃_{k=1}^{ω} {T_ijk^de} = M_v(q, A_ij^(f), D_ij, V_e, d)
    Increment the iteration counter: i := i + 1
end while
return Set of diverse image descriptions {T(1), T(2), ..., T(N)}
```

3.3 Image Generation By Self-Validation

Self-validation. The image descriptions $\mathcal{T}_{ij}^{d}$ and their corresponding aspects $A_{ij}^{(f)}$ are subsequently provided to the text-to-image model for image generation. At this stage, a potential issue is the possibility of generated images $\mathcal{I}_{ij}^{d}$ not aligning with the descriptions, due to hallucinations inherent in the text-to-image model (Lee et al., 2023). To tackle this issue, drawing inspiration from TIFA (Hu et al., 2023), we employed a self-validation process to evaluate the consistency of images with their descriptions via VQA.

In the self-validation process $\mathcal{F}$, for each image $\mathcal{I}_{ij}^{d}$, based on its image description, $\mathcal{M}_v$ is prompted to generate a set of simple questions $\Phi_{ij}^{d} = \{\phi_{ij1}^{d}, \phi_{ij2}^{d}, \ldots, \phi_{ijp}^{d}\}$ (*e.g.*,*"Is there a wooden chair in the image?"*), where p denotes the question number to evaluate the alignment. The function $\mathcal{F}$ takes the image $\mathcal{I}_{ij}^{d}$, its description $\mathcal{T}_{ij}^{d}$, and the set of questions Φ_{ij}^{d} as inputs and outputs an alignment score S_{ij}^{d}, which is calculated as the ratio of correctly answered questions to the total number of questions:

$$S_{ij}^{d} = \mathcal{F}(\mathcal{I}_{ij}^{d}, \mathcal{T}_{ij}^{d}, \Phi_{ij}^{d})$$

We set a threshold ζ, where: (i) If $S_{ij}^{d} < \zeta$, the image $\mathcal{I}_{ij}$ will be reworked in line with the description until it meets the required standard; (ii) If $\zeta \leq S_{ij}^{d} < 1$, the image meets the basic criteria but contains an error $\mathcal{E}_{ij}^{d}$, which will be documented; and (iii) If $S_{ij}^{d} = 1$, the image is considered to fully align with the description and is deemed acceptable.

3.4 Test Case Generation & Evaluation

Q&A generation with error control. To enhance the accuracy of question generation, particularly when addressing potential flaws in images, we propose error control. Despite thorough self-

Table 1: Effectiveness of hierarchical aspect generation under various hyperparameter settings.

m=3, n=5		m=3, n=6		m=3, n=7		m=4, n=5		m=4, n=6		m=4, n=7	
Raw	+Hierarchy	Raw	+Hierarchy	Raw	+Hierarchy	Raw	+Hierarchy	Raw	+Hierarchy	Raw	+Hierarchy
0.767	0.778 (1.4% ↑)	0.773	0.780 (1.0% ↑)	0.779	0.825 (5.9% ↑)	0.780	0.790 (1.3% ↑)	**0.786**	**0.849** (**10.2%** ↑)	0.798	0.842 (5.5% ↑)

validation, it's not guaranteed that every image will be flawless. Furthermore, when generating problems, we aim to avoid introducing biases stemming from the visual capabilities of examiner LVLM (Zhang et al., 2024b). Therefore, when generating questions, we will only include the image description $\mathcal{T}_{ij}^d$ and any identified defects $\mathcal{E}_{ij}^d$ in the input to the examiner $\mathcal{M}_v$. The function $\mathcal{M}_v$ generates the question Q_{ij}^d based on the image description and errors:

$$Q_{ij}^d = \mathcal{M}_v(\mathcal{T}_{ij}^d, \mathcal{E}_{ij}^d).$$

This will enable the creation of a diverse set of questions, along with reference answers, that specifically target the defective elements. For each image, we will provide a related question Q_{ij}^d (*e.g., multiple-choice or true/false*). These questions, along with the accompanying images, will be presented to the LVLMs under evaluation for their response.

Evaluation. The response $\mathcal{P}_{ij}^d$ from the tested LVLMs was compared to the reference answer A_{ij}^d to determine accuracy. If $\mathcal{P}_{ij}^d$ matched A_{ij}^d, it was marked correct ($\text{Acc}_{ij}^d = 1$); otherwise, it was marked incorrect ($\text{Acc}_{ij}^d = 0$). The overall accuracy $\text{Acc}_{\text{total}}^d$ was calculated as the average accuracy over all N questions.

4 Experiment

In this section, we evaluate seven of the latest models using AutoBench-V and perform human evaluations to validate our experimental findings. First, we demonstrate how AutoBench-V significantly reduces potential answer leakage and self-enhancement bias, as evidenced by the experimental results in Figure 5. Next, based on Figure 6, Table 1, Table 2, and Table 3, we analyze the impact of various evaluation factors (*e.g., user input*) and question difficulty on model performance, which reveals several insightful findings. We then present the model rankings across five user input categories with varying difficulty levels in Figure 8, followed by a discussion on human evaluations regarding alignment during AutoBench-V's generation process. Lastly, we investigate position bias in the evaluation process, as illustrated in Figure 9.

4.1 Experimental Setup

Selected models. In evaluating LVLMs, we selected seven representative models: `GPT-4o`, `GPT-4o mini` (Achiam et al., 2023), `Claude-3.5-Sonnet`, `Claude-3-Haiku` (Anthropic, 2024), `Gemini-1.5-Flash` (DeepMind, 2024), `GLM-4v` (GLM et al., 2024), and the open-source `Qwen2-VL` (Bai et al., 2023a), detailed in Table 7. These advanced models exhibit exceptional image understanding. Some well-known open-source models, such as `Llava-1.6` (Liu et al., 2024a) and `MiniGPT-4` (Zhu et al., 2023), were tested and found to perform poorly. Additionally, their capabilities differ significantly from other models, so they are not discussed in our evaluation. We chose `GPT-4o` as the examiner model for generating image descriptions, questions, and answers due to its strong overall performance. The descriptions were then passed to `Flux-pro` (blackforestlabs, 2024), a text-to-image model known for outstanding image generation. We also experimented with other text-to-image models (Rombach et al., 2022; Podell et al., 2023; Betker et al., 2023). However, their performance was suboptimal.This combination enables effective automated generation of image-based questions, crucial for the evaluation process (Ying et al., 2024; Fu et al., 2024; Xu et al., 2023).

Experimental setting. We set $n = 4$ for the number of general aspects and $m = 6$ for the number of fine-grained aspects, as this configuration yields the highest diversity in the generated aspects as illustrated in Table 1, allowing for a broader range of scenes and elements. We set ω = 10, namely 10 pictures for each fine-grained aspect. Therefore, we evaluate 720 images for each user input (with each user input having three difficulty levels). For easy difficulty, we set the self-validation threshold $\zeta_e = 1$ since the scenes are simpler and contain fewer elements, which justifies a higher threshold. For the medium and hard difficulty levels, the images contain more elements, so we lower the thresholds to $\zeta_m = \zeta_h = 0.8$ to avoid compromising efficiency. The error control mechanism that follows ensures the appropriateness of this threshold.

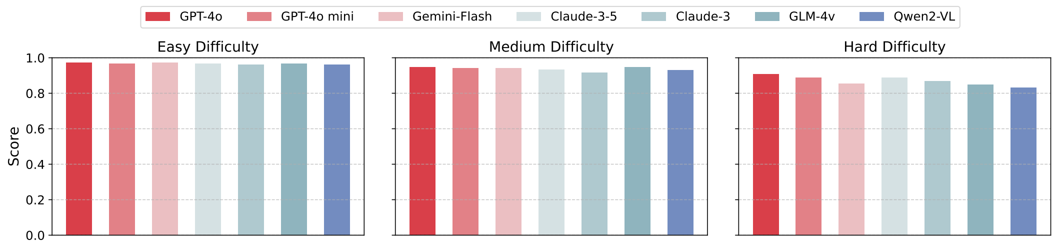


Figure 5: Performance of various LVLMs across three difficulty levels when provided only with image descriptions, without the corresponding images.

4.2 Examiner Priority

To mitigate the potential for answer leakage associated with self-enhancement bias (*e.g.*, when the model being evaluated is also utilized for generating the evaluation cases) in the examiner LLM, we enhance the fairness of the assessment by having AutoBench-V generate questions from image descriptions rather than directly from images. This approach separates the visual information from the generation process, reduceing the risk of self-enhancement bias (Ye et al., 2024) that could occur if questions were derived from the examiner model's (`GPT-4o`) visual capabilities, which might cause unfair comparison. By employing only textual descriptions for generation, we eliminate the influence of `GPT-4o`'s specific visual processing abilities, thereby ensuring a more equitable evaluation.

Table 2: Average performance (Accuracy) of all models at different difficulty levels.

Model	Easy	Medium	Hard
GPT-4o	**90.43%**	**79.81%**	**75.02%**
GPT-4o mini	88.01%	76.98%	70.70%
Gemini-1.5-Flash	88.07%	74.64%	70.85%
Claude-3.5-Sonnet	89.28%	75.49%	63.67%
Claude-3-Haiku	86.82%	73.40%	67.42%
GLM-4v	**90.43%**	77.29%	64.93%
Qwen2-VL	89.57%	79.03%	71.89%

To validate the fairness of this method, we conducted an experiment in which models were presented with image descriptions alongside corresponding questions while avoiding direct access to the images. As demonstrated in Figure 5, the results revealed consistent performance across all models, with minimal variance (0.4% for easy questions and 2.4% for hard questions). This consistency suggests that, in the absence of visual input, models' textual understanding ability is almost equal, which means the benchmark effectively assesses visual comprehension and does not show obvious bias towards the *examiner LVLM* (*i.e.*, `GPT-4o`).

4.3 Main Results

As shown in Figure 6, through the evaluation of various models on AutoBench-V, we can observe several findings that can bring insights for future work. More detailed results are in Table 6.

Model performance decreases as task difficulty increases, with `GPT-4o` showing the strongest average performance across tasks. This trend is consistent across all models, with scores steadily declining as the difficulty increases from easy to hard, as shown in Figure 13. For example, `GPT-4o`'s average score drops from 90.43% at the easy level to 75.02% at the hard level. Despite the overall decline, `GPT-4o` maintains its leading position across all difficulty levels. Additionally, the result highlights that the most notable shift occurs between easy and medium. Although a few samples show improved scores with increased difficulty mildly, the majority trend still experiences a decline, reinforcing the validity of our difficulty grading mechanism.

Table 3: Average accuracy for various user inputs at different difficulty levels.

User Input	Easy	Medium	Hard
Basic.	90.59%	75.90%	63.33%
Spatial.	82.46%	69.14%	63.00%
Seman.	91.28%	79.84%	74.52%
Reason.	86.50%	74.67%	68.97%
Atmos.	94.76%	83.33%	75.66%

As task difficulty increases, the performance disparity between models becomes more pronounced. As illustrated in Figure 7, the performance decline across models varies with increasing

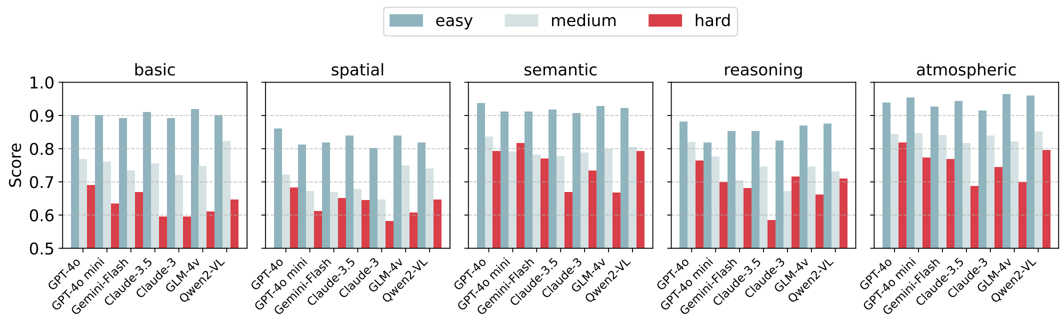


Figure 6: Overall performance of different models at easy, medium, and hard levels.

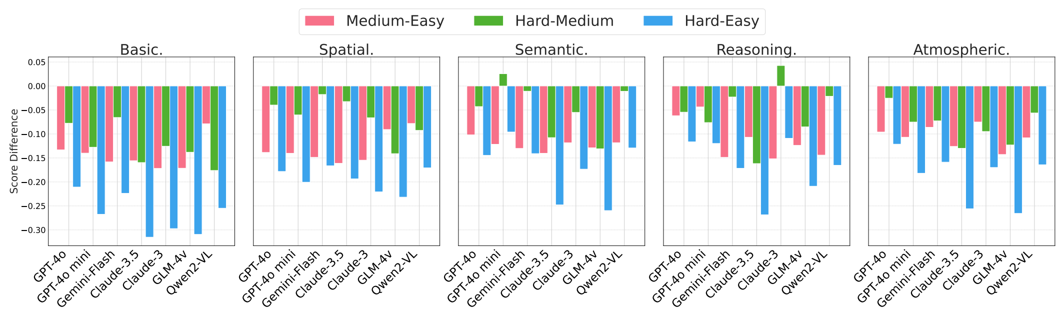


Figure 7: Score variation of models from easy to hard across different user inputs. As task difficulty increases, the performance disparity between models becomes more pronounced.

task difficulty. Models like `Claude-3.5-Sonnet` and `GLM-4v` experience more pronounced drops. In difficult semantic tasks, `GPT-4o` maintains a strong score of 79.36%, compared to `Claude-3.5-Sonnet` and `GLM-4v`, which achieve 66.97% and 66.82%, respectively, highlighting `GPT-4o`'s superior ability to handle complex abstractions. The standard deviations across models at the three difficulty levels are 1.26%, 2.14%, and 3.74%, indicating increased disparity as task difficulty rises. Notably, models like `GPT-4o` and `GLM-4v` show more consistent performance with smaller score variations, suggesting stability across difficulty levels. In contrast, models like `Claude-3.5` and `Qwen2-VL` exhibit greater score fluctuations, indicating higher sensitivity to difficulty changes.

Models demonstrate superior performance in semantic and atmospheric understanding while lagging in spatial and reasoning tasks. As illustrated in Table 3, our results reveal a consistent pattern across difficulty levels. Models excel in semantic and atmospheric understanding, maintaining high accuracy even at the hard level of 74.52% and 75.66% respectively. In contrast, spatial and reasoning prove more challenging, with accuracy dropping to 63.00% at the hard level. These findings indicate that while LVLMs have developed strong capabilities in comprehending semantic content and scene atmosphere, they still struggle with tasks involving spatial relationships and complex visual reasoning. To address these limitations, we suggest future research could explore training strategies that emphasize enhancing spatial reasoning and complex visual problem-solving capabilities in LVLMs.

4.4 Model Rank Overview

Figure 8 reveals distinct performance patterns among different models under various difficulty. Notably, models like `GPT-4o`, while not exhibiting a significant advantage in simpler tasks, demonstrate outstanding performance in more challenging scenarios (*e.g.*,hard questions). Conversely, models such as `GLM-4v` perform well on easier tasks but show diminished capabilities as task difficulty increases. This indicates an imbalance in the model's ca-

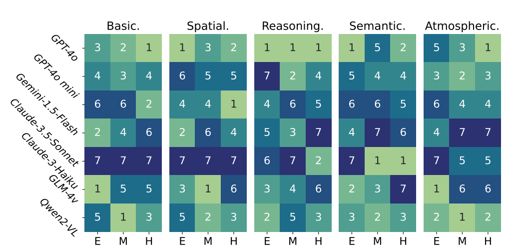


Figure 8: The model performance ranking given five user inputs under different difficulty levels.

pabilities across different difficulty levels, highlighting the importance of cross-difficulty evaluation. It also demonstrates that AUTOBENCH-V is effective in revealing such imbalances.

4.5 HUMAN EVALUATION

We conducted human evaluations in two aspects: the effectiveness of guided description generation and the alignment between questions and reference answers. See the Appendix D for details on the human evaluation. We ultimately represented the results of the human evaluation using the *alignment rate* (the proportion of aligned samples out of the total).

Guided description generation. We developed description generation guidelines for each fine-grained aspect to reduce vagueness in image descriptions, ensuring better alignment with themes and preventing discrepancies. A human evaluation showed that these guidelines significantly improved question-answer alignment, especially in more challenging tasks, as shown in Table 4.

Table 4: Alignment rate of guided description generation.

Task	Easy	Medium	Hard
Before Guide	94.32%	82.30%	77.14%
After Guide	95.20%	88.13%	84.55%
$\Delta_{\uparrow}$	0.88%	5.83%	7.41%

Question-Answer alignment. After implementing the generation guidelines, we conducted a human evaluation to assess the accuracy of the questions and answers generated by the examiner model. As shown in Table 4, the evaluation resulted in high scores, confirming the model's effectiveness in producing well-aligned question-answer pairs for image-based tasks.

4.6 POSITION BIAS

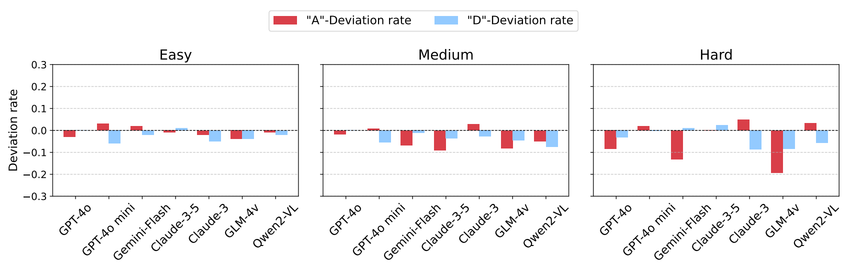


Figure 9: Comparison of answer distribution under position bias conditions. Correct answers at A or D v.s. correct answers are evenly distributed at A,B,C,D.

Since the reference answers generated by LLMs tend to cluster around option A, we manually set the correct options to be evenly distributed. To investigate the necessity of this approach, we conducted experiments to examine potential position bias (Zheng et al., 2023). We evaluated scenarios where all correct answers were placed in either options A and D, comparing the resulting scores with the evenly distributed case (*i.e.*, 25% for each option), as shown in Figure 9. The deviation rate was calculated using the following formula: $R = \frac{S_X - S_U}{S_U}$, where S_X is the model score for condition X (either A or D), and S_U is the score for the scenario when options are evenly distributed.

The position bias becomes more evident with increasing question difficulty. For instance, at the hard level, `GLM-4V` showed a significant bias, with deviation rates of $R_A = -19\%$ and $R_D = -8\%$, suggesting a notable bias when correct answers were concentrated in options A or D, compared to the uniform distribution scenario. Thus, our approach of manually setting an even distribution of answers to avoid position bias is justified and necessary.

5 CONCLUSION

In this work, we introduce AUTOBENCH-V, a fully automated framework designed for benchmarking LVLMs. The framework integrates a series of innovative modules that ensure diversity and reliability in dataset generation, as well as impartiality in model evaluation. Through extensive experiments, we have demonstrated the robustness and unbiased nature of the evaluation process facilitated by AUTOBENCH-V. The insights gleaned from our research provide a solid foundation for future investigations in this field.

REFERENCES

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.

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. Vqa: Visual question answering. *arXiv preprint arXiv:1505.00468*, 2016.

Anthropic. Claude3family. *Anthropic News*, 2024. URL https://www.anthropic.com/news/claude-3-family.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023a.

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, Jiayin Zhang, Juanzi Li, and Lei Hou. Benchmarking foundation models with language-model-as-an-examiner. *arXiv preprint arXiv:2306.04181*, 2023b.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

blackforestlabs. flux_pro. *BlackForestLabs News*, 2024. URL https://blackforestlabs.ai/announcing-black-forest-labs/.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Wenxiao Cai, Yaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. *arXiv preprint arXiv:2406.13642*, 2024.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.

DeepMind. Gemini-1-5-flash. *DeepMind Technologies*, 2024. URL https://deepmind.google/technologies/gemini/flash/.

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

AV Geetha, T Mala, D Priyanka, and E Uma. Multimodal emotion recognition with deep learning: advancements, challenges, and future directions. *Information Fusion*, 105:102218, 2024.

Taraneh Ghandi, Hamidreza Pourreza, and Hamidreza Mahyar. Deep learning approaches on image captioning: A review. *ACM Computing Surveys*, 56(3):1–39, 2023.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

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. *arXiv preprint arXiv:1612.00837*, 2017.

Qiushan Guo, Shalini De Mello, Hongxu Yin, Wonmin Byeon, Ka Chun Cheung, Yizhou Yu, Ping Luo, and Sifei Liu. Regiongpt: Towards region understanding vision language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13796–13806, 2024.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *arXiv preprint arXiv:2303.11897*, 2023.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *arXiv preprint arXiv:1412.2306*, 2015.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench-2: Benchmarking multimodal large language models. *arXiv preprint arXiv:2311.17092*, 2023a.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023b.

Xiang Lisa Li, Evan Zheran Liu, Percy Liang, and Tatsunori Hashimoto. Autobencher: Creating salient, novel, difficult datasets for language models. *arXiv preprint arXiv:2407.08351*, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context. *arXiv preprint arXiv:1405.0312*, 2015.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024a.

Xiao Liu, Xuanyu Lei, Shengyuan Wang, Yue Huang, Zhuoer Feng, Bosi Wen, Jiale Cheng, Pei Ke, Yifan Xu, Weng Lam Tam, et al. Alignbench: Benchmarking chinese alignment of large language models. *arXiv preprint arXiv:2311.18743*, 2024b.

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? *arXiv preprint arXiv:2307.06281*, 2023.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, and Wenqi Shao. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. *arXiv preprint arXiv:2408.02718*, 2024.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*, 2023.

M Ross Quillian. *Semantic memory*. Air Force Cambridge Research Laboratories, Office of Aerospace Research ..., 1966.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *arXiv preprint arXiv:1409.0575*, 2015.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2023.

Siyuan Wu, Yue Huang, Chujie Gao, Dongping Chen, Qihui Zhang, Yao Wan, Tianyi Zhou, Xiangliang Zhang, Jianfeng Gao, Chaowei Xiao, and Lichao Sun. Unigen: A unified framework for textual dataset generation using large language models. *arXiv preprint arXiv:2406.18966*, 2024.

Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2411–2418, 2013.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*, 2024.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*, 2023.

Jian-Ru Xue, Jian-Wu Fang, and Pu Zhang. A survey of scene understanding by event reasoning in autonomous driving. *International Journal of Automation and Computing*, 15(3):249–266, 2018.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Pin-Yu Chen, Nitesh V. Chawla, and Xiangliang Zhang. Justice or prejudice? quantifying biases in llm-as-a-judge. 2024.

Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, et al. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *arXiv preprint arXiv:2306.06687*, 2023.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*, 2024.

Haoxuan You, Rui Sun, Zhecan Wang, Long Chen, Gengyu Wang, Hammad A Ayyubi, Kai-Wei Chang, and Shih-Fu Chang. Idealgpt: Iteratively decomposing vision and language reasoning via large language models. *arXiv preprint arXiv:2305.14985*, 2023.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Weihao Yu, Zhengyuan Yang, Linfeng Ren, Linjie Li, Jianfeng Wang, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Lijuan Wang, and Xinchao Wang. Mm-vet v2: A challenging benchmark to evaluate large multimodal models for integrated capabilities. *arXiv preprint arXiv:2408.00765*, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2024.

Jieyu Zhang, Weikai Huang, Zixian Ma, Oscar Michel, Dong He, Tanmay Gupta, Wei-Chiu Ma, Ali Farhadi, Aniruddha Kembhavi, and Ranjay Krishna. Task me anything. *arXiv preprint arXiv:2406.11775*, 2024a.

Yi-Fan Zhang, Weichen Yu, Qingsong Wen, Xue Wang, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Debiasing multimodal large language models. *arXiv preprint arXiv:2403.05262*, 2024b.

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.

Gengze Zhou, Yicong Hong, and Qi Wu. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7641–7649, 2024.

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

Kaijie Zhu, Jiaao Chen, Jindong Wang, Neil Zhenqiang Gong, Diyi Yang, and Xing Xie. Dyval: Dynamic evaluation of large language models for reasoning tasks. *arXiv preprint arXiv:2309.17167*, 2024a.

Kaijie Zhu, Jindong Wang, Qinlin Zhao, Ruochen Xu, and Xing Xie. Dynamic evaluation of large language models by meta probing agents. *arXiv preprint arXiv:2402.14865*, 2024b.

Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *arXiv preprint arXiv:1905.05055*, 2023.

A Details of Experiment Setting

Model Selection. The details of models selected in our experiments are shown in Table 7.

Computing Resource. For our experiments, all open-source vision-language model inferences were performed locally using NVIDIA GeForce RTX 4090 GPU with 24GB VRAM.

Alignment Evaluation. Inspired by tifa (Hu et al., 2023), we generated consistency tests for images across 12 aspects:*object, human, animal, food, activity, attribute, counting, color, material, spatial, location, shape, other*. For details on the score distribution of the consistency tests without threshold, please refer to Table 5.

Table 5: Alignment score (S_i^d) distribution without setting threshold.

Level	$\leq$ 40%	40%$\sim$60%	60%$\sim$80%	> 80%
Spatial Understanding.				
Easy	0.51%	2.54%	7.61%	89.34%
Medium	0.53%	3.17%	19.05%	77.25%
Hard	1.03%	5.64%	16.92%	76.40%
Basic Understanding.				
Easy	1.94%	2.90%	14.01%	81.16%
Medium	3.38%	5.80%	14.98%	75.85%
Hard	2.86%	8.57%	24.76%	63.81%
Semantic Understanding.				
Easy	0.99%	3.94%	8.37%	86.70%
Medium	0.49%	4.39%	19.02%	76.10%
Hard	1.93%	4.35%	17.39%	76.33%
Reasoning Capacity.				
Easy	1.93%	3.86%	14.49%	79.71%
Medium	2.53%	6.06%	23.23%	68.18%
Hard	6.22%	11.92%	28.50%	53.37%
Atmospheric Understanding.				
Easy	0.48%	2.41%	11.11%	85.99%
Medium	1.38%	2.76%	17.97%	77.88%
Hard	1.79%	6.55%	24.40%	67.26%

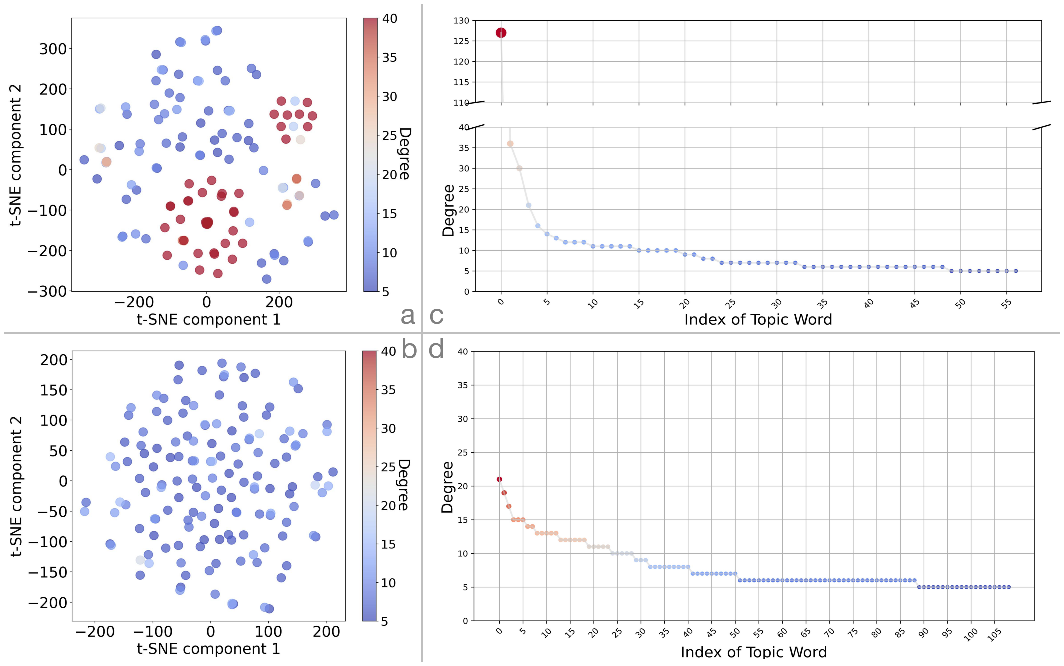


Figure 10: Visualization of image topic words. Topic words are converted into vectors using `bge-large-en-v1.5` (Xiao et al., 2024), then perform dimensionality reduction via t-SNE (Van der Maaten & Hinton, 2008). Topic word distribution without semantic graph (a)(c) and with semantic graph (b)(d). It can be seen that with the semantic graph the diversity of topic words increases and the over-reliance on high-degree words is reduced.

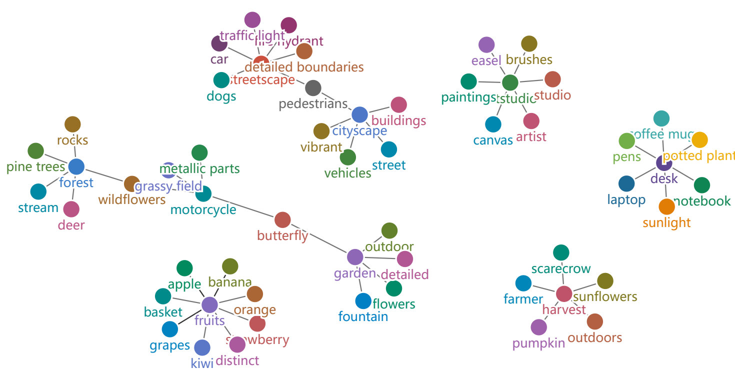


Figure 11: Topic words visualization using semantic graph under basic understanding.

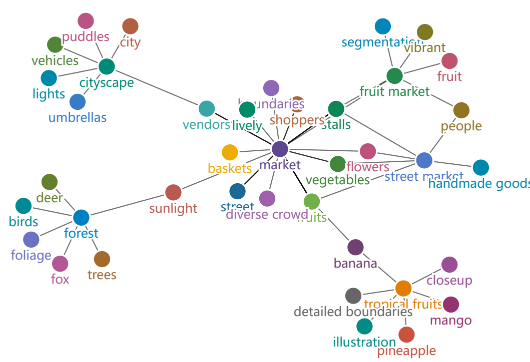


Figure 12: Topic words visulization without using semantic graph under basic undertstanding.

Table 6: Performance (Accuracy) details of all models on five user inputs and three difficulty levels.

Model	User Input↑					
	BASIC.	SPATIAL.	SEMAN.	REASON.	ATMOS.	AVERAGE
Easy						
GPT-4o	90.18%	**86.09%**	**93.81%**	88.13%	93.94%	**90.43%**
GPT-4o mini	90.18%	81.28%	91.24%	81.92%	95.45%	88.01%
Gemini-1.5-Flash	89.29%	81.82%	91.19%	85.31%	92.75%	88.07%
Claude-3.5-Sonnet	91.07%	83.96%	91.75%	85.31%	94.33%	89.28%
Claude-3-Haiku	89.29%	80.21%	90.72%	82.49%	91.41%	86.82%
GLM-4v	**91.96%**	83.96%	87.01%	**92.78%**	**96.45%**	**90.43%**
Qwen2-VL	90.18%	81.82%	92.27%	87.57%	96.00%	89.57%
Medium						
GPT-4o	76.87%	72.25%	**83.64%**	**81.95%**	84.35%	**79.81%**
GPT-4o mini	76.19%	67.26%	79.09%	77.56%	84.78%	76.98%
Gemini-1.5-Flash	73.47%	66.96%	78.18%	70.44%	84.14%	74.64%
Claude-3.5-Sonnet	75.51%	67.84%	77.73%	74.63%	81.74%	75.49%
Claude-3-Haiku	72.11%	64.76%	78.90%	67.31%	83.91%	73.40%
GLM-4v	74.83%	**74.89%**	79.91%	74.63%	82.17%	77.29%
Qwen2-VL	**82.31%**	74.01%	80.45%	73.17%	**85.21%**	79.03%
Hard						
GPT-4o	**69.12%**	**68.28%**	79.36%	**76.50%**	81.82%	**75.02%**
GPT-4o mini	63.43%	61.23%	**81.65%**	69.94%	77.27%	70.70%
Gemini-1.5-Flash	66.91%	65.20%	77.10%	68.16%	76.88%	70.85%
Claude-3.5-Sonnet	59.56%	64.60%	66.97%	58.47%	68.75%	63.67%
Claude-3-Haiku	59.56%	58.15%	73.39%	71.58%	74.43%	67.42%
GLM-4v	61.03%	60.79%	66.82%	66.12%	69.89%	64.93%
Qwen2-VL	64.71%	64.76%	79.36%	71.03%	**79.59%**	71.89%

Table 7: Model names, Creators, whether it is open source, and their purpose.

Model	Creator	Open-Source	Purpose
`GPT-4o`	OpenAI	✗	Examiner&Candidate
`GPT-4o mini`		✗	Candidate
`Gemini-1.5-Flash`	Google	✗	Candidate
`Claude-3.5-sonnet`	Anthropic	✗	Candidate
`Claude-3-haiku`		✗	Candidate
`GLM-4v`	Zhipu AI Inc.	✗	Candidate
`Qwen2-VL`	Alibaba	✓	Candidate
`Flux-pro`	Black Forest Labs	✗	Image generation

Figure 13: We randomly sampled 75 question instances for each user input, with 25 questions from each difficulty level. We then visualized the response patterns of various models to these questions (the red means). The rows in each user input, from top to bottom, represent the response situations of `Claude-3.5-sonnet`, `Claude-3-haiku`, `Gemini-1.5-flash`, `GLM-4v`, `GPT-4o`, `GPT-4o mini`, and `Qwen2-VL`, respectively.

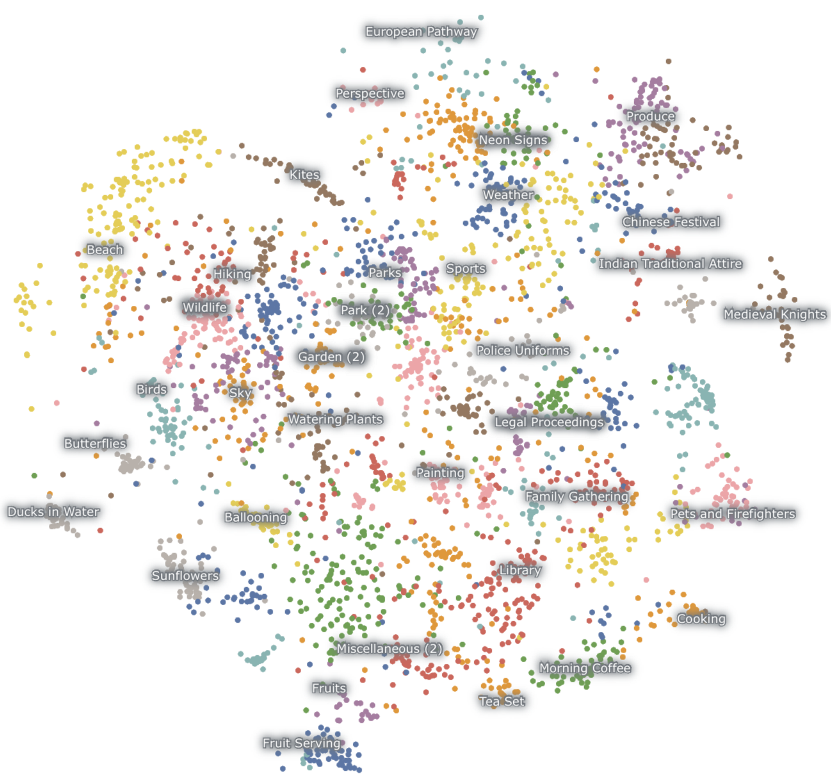


Figure 14: Visualization of all image descriptions in AUTOBENCH-V.

B Details of User Input

In this section, we provide a comprehensive overview of the levels at which we categorize user inputs based on linguistic aspects. Our goal is to offer a comprehensive and broad representation of user requirements for LVLMs. However, as it is challenging to exhaustively cover every aspect, we base our categorization on aspects derived from the literature (Li et al., 2023b; Meng et al., 2024; Liu et al., 2023). These aspects are considered representative and comprehensive examples of the capabilities of LVLMs and other aspects like in (Chen et al., 2024; Xu et al., 2023) can be handled in a similar manner, without requiring additional fine-tuning or adjustments, as our framework is highly extensible, allowing users to propose their own aspects as needed.

B.1 Basic Understanding

Definition and Goal: Basic Understanding refers to the recognition and identification of individual objects, characters, and scenes within an image. The goal is to accurately detect and label relevant elements, providing a foundation for more advanced tasks such as object tracking and scene interpretation (Wu et al., 2013; Xue et al., 2018).

Requirement. This task demands the ability to detect specific objects and differentiate between various types of objects. Additionally, it involves understanding the broader context of the scene and identifying real-life settings to enable accurate interpretation of the image's overall content.

B.2 Spatial Understanding

Definition and Goal. Spatial Understanding refers to the interpretation of the spatial arrangement and positioning of objects within an image (Cai et al., 2024; Guo et al., 2024). The goal is to comprehend both two-dimensional and three-dimensional relationships, determining which objects are in the foreground or background, assessing their relative sizes and orientations, and understanding how they are positioned within the scene.

Requirement. This task demands the ability to perceive depth, estimate distances between objects, and analyze how objects interact within the physical space of the image, providing a more accurate understanding of the spatial structure and context.

B.3 Semantic Understanding

Definition and Goal. Semantic Understanding involves interpreting the higher-level meaning and relationships within an image (Meng et al., 2024). The goal is to move beyond simple object identification to understand the roles and interactions between objects, such as recognizing that a person is riding a bike or that two people are engaged in conversation. This level of understanding aims to capture the context and intent behind the scene, identifying how elements relate to each other to form a coherent narrative or message.

Requirement. This task requires discerning the interactions and relationships between objects, understanding their roles within the scene, and interpreting the overall context to accurately derive the narrative or intended message conveyed by the image.

B.4 Atmospheric Understanding

Definition and Goal. Atmospheric Understanding focuses on grasping the mood, tone, and emotional ambiance conveyed by an image. The goal is to interpret not just what is depicted or how elements are arranged, but also how the scene feels and the emotional resonance it conveys to the viewer. For instance, an image of children laughing under warm sunlight in a lush park combines their expressions, bright colors, and soft lighting to create a joyful and carefree atmosphere.

Requirement. This task requires the ability to capture and interpret subtle emotional cues and tonal qualities of the scene, distinguishing the overall mood and emotional impact of the image from more analytical aspects like semantic or spatial understanding.

B.5 Reasoning Capacity

Definition and Goal. Reasoning Capacity involves interpreting and analyzing the relationships and logical connections between different elements within an image (Zhou et al., 2024; You et al., 2023). The goal is to infer potential outcomes, understand causal relationships, and make predictions about what might happen next based on visual cues. For example, if a person is holding an umbrella and the sky is dark, reasoning capacity would suggest that it might rain soon. This level also includes understanding abstract relationships, such as social dynamics or the intent behind actions, and making judgments about what is likely or possible given the visual information.

Requirement. This task requires the ability to analyze logical connections between elements, infer outcomes, and understand causal relationships, as well as to interpret abstract concepts and make predictions based on the visual context.

C Details of Difficulty Grading

This section describes in detail the difficulty levels for the pictures and questions used in prompts respectively. The following is the instruction guiding the examiner model to generate image descriptions and questions of varying difficulty levels.

C.1 Image Description

Easy difficulty. Generate images with very simple elements, focusing on single, easily recognizable objects placed against a plain or neutral background. The descriptions should be straightforward and unambiguous, *e.g.*, "a red apple on a white background." The focus is on clarity and simplicity, with minimal detail or interaction.

Medium difficulty. Introduce scenes where the required elements interact with their environment naturally but uncomplicatedly. The setting may include multiple common objects and a familiar context, but the composition remains clear and not overly complex, *e.g.*, "a cup on a table in a well-lit kitchen." The background and context are present but not overwhelming, and there are no intricate details or unusual perspectives.

Hard difficulty. Craft descriptions that involve multiple elements interacting with each other, set in a more complex environment. Use varied perspectives, detailed textures, or lighting conditions that add layers of difficulty, *e.g.*, "a reflection of a cat looking out of a rain-soaked window, with a cityscape in the background at dusk." The focus is on creating a rich and intricate scene that challenges the model's ability to render interactions, depth, and subtleties in lighting and perspective.

Moreover, we standardized the description of the observer's perspective in the image description to prevent directional issues from causing confusion. For instance, ambiguities could arise in interpreting relative directions such as left and right, as these can vary significantly depending on the observer's viewpoint.

C.2 Question

Easy difficulty. Focus on questions that require identifying simple, prominent, and explicit details within the image. These questions should be straightforward, relying solely on basic observation without the need for inference or interpretation. For example, you might ask about the color of a specific object, the presence of a single item, or the shape of an easily recognizable feature. The key is to keep the questions direct and simple, ensuring that the answer is obvious and immediately visible in the image.

Medium difficulty. Design questions that necessitate a moderate level of observation and inference. These questions should involve understanding relationships between elements, recognizing interactions, or identifying less prominent features that are still clear but not immediately obvious. Examples could include questions about the relative position of objects, identifying an action taking place, or understanding the context of a scene. The goal is to require some level of thought beyond basic observation, challenging the model to understand the scene's composition or narrative without being overly complex.

Hard difficulty. Create questions that require the model to notice and interpret more detailed aspects of the image. These questions should involve recognizing multiple elements working together, understanding more complex interactions, or identifying details that are present but not immediately obvious. For example, you might ask about the positioning of objects relative to each other in a more crowded scene, subtle changes in lighting or color that affect the appearance of objects, or identifying an element that is not the main focus but is still visible in the background. The aim is to challenge the model to go beyond surface-level details, but without making the task too abstract or overly difficult.

D Human Evaluation

D.1 Details of Human Evaluation

The evaluation was carried out by a panel of five evaluators: three undergraduate students and two PhD students, all possessing professional English skills. Sample annotation screenshots from the human evaluation process are presented in Figure 20 and Figure 21. To ensure unbiased results, each evaluator independently assessed all samples. A sample was considered aligned if it received a majority vote (*i.e.*, more than half of the evaluators agreed on its alignment).

D.2 Human Evaluation Guidelines

In this section, we outline the guidelines followed during human evaluations to ensure reliability and validity.

For **Description Generation Guideline**, the evaluators need to consider the following three points:

▷ **Alignment with Image:** The main criterion is how well the generated description reflects the visual content. Descriptions must accurately correspond to the image, avoiding vague or abstract expressions. Each description should provide clear, specific details that align with the image content and the defined fine-grained aspects.

▷ **Specificity and Clarity:** Ensure that descriptions are specific, directly related to the image, and free from ambiguous or overly generalized language.

▷ **Relevance to Aspects:** Assess whether the description aligns with the corresponding themes and expected content. Descriptions must clearly communicate the intended visual elements and avoid any misalignment between the image and the description.

For **Question-Answer Alignment**, there are two points that the evaluators should consider:

▷ **Clarity and Accuracy:** Each question must be clear, unambiguous, and directly derived from the image. The answers should correspond to observable details or logical inferences from the image, with only one correct answer for each question. There should be no irrelevant or misleading information in the questions or answers.

▷ **Consistency with Image:** Verify that both the question and answer are directly based on the image's content. The evaluation should ensure that there is a logical and clear relationship between the visual cues and the generated question-answer pair, particularly for tasks involving higher difficulty.

E Error Study

Through extensive experimental analysis, we have categorized the common problems encountered by LVLMs in VQA tasks into two main types: *image comprehension errors* and *image reasoning errors*. Regarding the first category, LVLMs often fail to truly understand the details in an image. For instance, in Figure 15, the model failed to notice the *kite* flying in the upper right corner of the image and incorrectly identified a *non-existent action of a squirrel climbing a tree*. In Figure 17, the model mistakenly perceived the firefighter as holding the crowbar with *only one hand instead of two*. In Figure 18, the model failed to recognize that the apple's stem was *slightly tilted to the left*. These errors demonstrate the model's inadequate comprehension of image details.

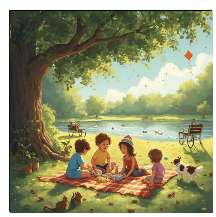
Image reasoning errors occur when models accurately perceive the image content but falter in their reasoning process, leading to incorrect answers. For instance, in Figure 16, the model correctly recognizes that the image does not depict *a cheering crowd*. However, during subsequent reasoning, it convinces itself otherwise, ultimately selecting option C while neglecting to analyze other choices in its explanation. Figure 19 exemplifies a similar issue: the model correctly identifies that *the child is wearing a yellow shirt* but after mentioning *the red of the kite* it erroneously selects B. Red is its final answer. These examples highlight a disconnect between visual perception and logical reasoning in LVLMs, where initial accurate observations can be overridden by flawed deductive processes.

Basic Understanding

Aspect: Multiple Object Identification
Difficulty Level: Hard

Question: Which of the following activities is NOT happening in the generated image?
Choice: "A": "A squirrel climbing a tree", "B": "A kite flying high in the sky", "C": "Children having a picnic on a colorful blanket", "D": "Ducks swimming in a pond"}

Image Description:
please generate a picture from the perspective of an observerA vibrant and bustling outdoor scene featuring a group of children having a picnic in the park. The children are sitting on a large, colorful blanket, surrounded by various snacks and toys. Nearby, a family of squirrels is gathering acorns under a large oak tree, while a kite flies high in the sky. The sunlight filters through the leaves, casting dappled shadows on the grassy ground. Several birds are perched on the branches of the oak tree, and a curious dog is sniffing around the picnic spread. The background includes a pond with ducks swimming and a couple of bicycles leaning against a bench.



Selected Model: Claude-3-Haiku
Answer: "B"
Explanation: The image does not show a kite flying high in the sky. The other activities, such as children having a picnic on a colorful blanket, squirrels climbing a tree, and ducks swimming in a pond, are all visible in the generated image.
Ground Truth: "A"

Figure 15: Error study of Claude-3-Haiku under basic understanding.

Spatial Understanding

Aspect: Foreground vs. Background
Difficulty Level: Hard

Question: In the image, what is positioned immediately to the left of the richly decorated dragon in the foreground?
Choice: "A": "A dancer in a colorful costume", "B": "An illuminated storefront", "C": "A cheering crowd", "D": "A firework display"

Image Description:
Please generate a picture from the perspective of an observer A vibrant carnival parade featuring a richly decorated float with a giant, elaborately detailed dragon in the foreground, its scales shimmering in multicolored lights and its eyes glowing. The dragon is surrounded by dancers dressed in colorful costumes, their movements captured in mid-dance. In the background, a bustling city street lined with illuminated storefronts and cheering crowds, slightly blurred to indicate depth. The night sky is filled with fireworks adding an additional layer of festivity.

Selected Model: Claude-3.5-Sonnet
Answer: "C"
Explanation: In the foreground of the image, to the left of the richly decorated dragon figure, we can see silhouettes of people. While not a cheering crowd per se, these figures represent a group of people watching or participating in what appears to be a festive parade or celebration. The crowd is positioned immediately to the left of the dragon, occupying the foreground space. While there are fireworks in the background and illuminated storefronts visible, these are not positioned immediately to the left of the dragon in the foreground.
Ground Truth: "A"

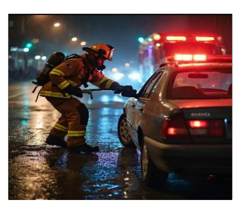
Figure 16: Error study of Claude-3.5-Sonnet under spatial understanding.

Semantic Understanding

Aspect: Object Manipulation
Difficulty Level: Hard

Question: Which hand is the firefighter using to grip the crowbar while prying open the jammed car door?
Choice: "A": "Left hand", "B": "Right hand", "C": "Both hands", "D": "Neither hand"

Image Description:
Please generate a picture from the perspective of an observer. A firefighter in full gear using a large crowbar to pry open a jammed car door on a rain-slicked street at night, the scene illuminated by the flashing red and blue emergency lights, showing intense determination and effort in his stance.



Selected Model: Qwen2-VL
Answer: "B"
Explanation: The firefighter is using the right hand to grip the crowbar while prying open the jammed car door, as indicated by the position of the hand and the direction of the crowbar.
Ground Truth: "C"

Figure 17: Error study of Qwen2-VL under semantic understanding.

Reasoning Capacity

Aspect: Social Interactions
Difficulty Level: Easy

Question: What direction is the stem of the apple tilted towards?
Choice: "A": "Upwards", "B": "Right", "C": "Left", "D": "Downwards"

Image Description:
please generate a picture from the perspective of an observerA single green apple resting upright on a polished white marble countertop. The apple is centered in the frame, positioned to face directly towards the viewer with its stem slightly tilted to the left. The background is plain and white, ensuring the apple is the clear focal point.

Selected Model: GLM-4v
Answer: "A"
Explanation: The stem of an apple is typically the part that connects the fruit to the tree. In the image provided, the stem appears to be pointing upwards, which means it's oriented towards the top of the image. This orientation suggests that if the apple were placed upright on a flat surface, the stem would be the vertical line that points skyward.
Ground Truth: "C"

Figure 18: Error study of GLM-4v under reasoning capacity.

Atmospheric Understanding

Aspect: Color Palette Interpretation
Difficulty Level: Medium

Question: Which color stands out the most in the child's clothing against the background of the green grass and colorful sky?
Choice: "A": "Blue", "B": "Red", "C": "Yellow", "D": "Green"

Image Description:
An image of a child flying a red kite in a lush green park during sunset. The sky is filled with hues of orange, pink, and purple, casting long shadows and creating a warm, nostalgic atmosphere. The child is wearing a bright yellow shirt, which stands out against the green grass and the colorful sky.

Selected Model: Gemini-1.5-Flash
Answer: "B"
Explanation: **The child is wearing a yellow shirt, which stands out against the green grass and the colorful sky. The red of the kite also stands out, but the yellow of the shirt is more prominent since it is closer to the viewer and is a more saturated color.**
Ground Truth: "C"

Figure 19: Error study of Gemini-1.5-Flash under atmospheric understanding.

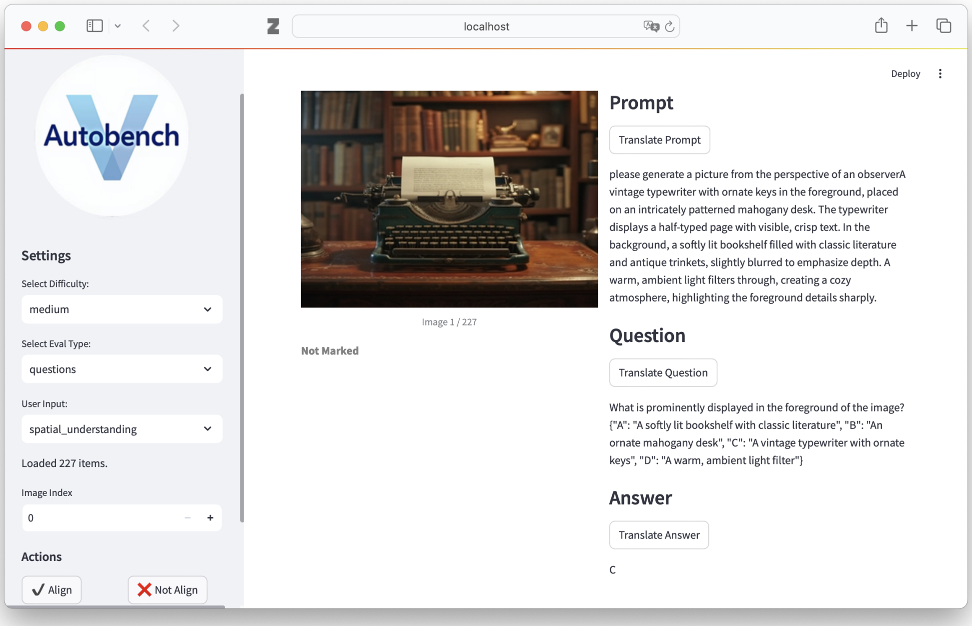


Figure 20: Screenshot of Human Evaluation (Example 1).

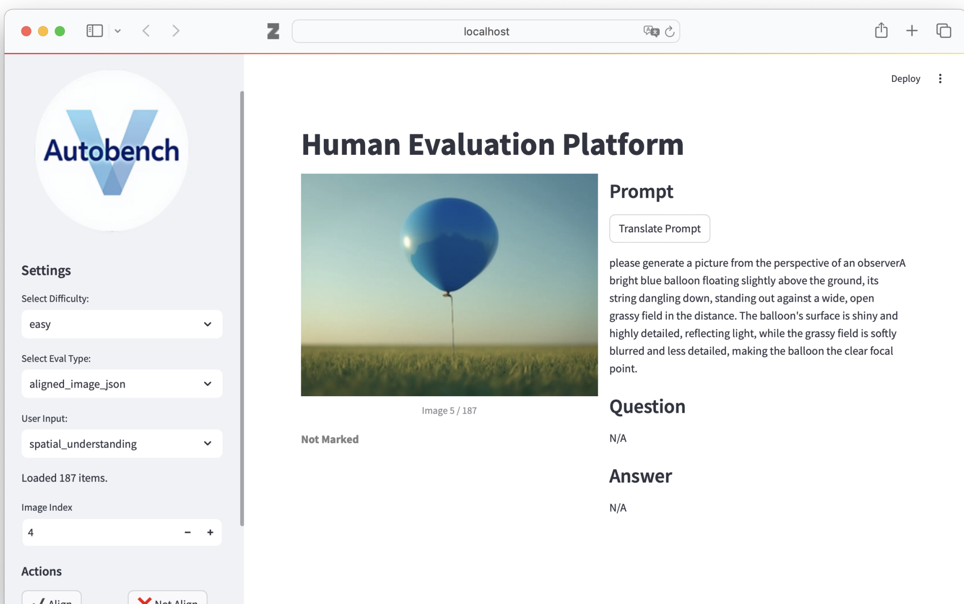


Figure 21: Screenshot of Human Evaluation (Example 2).

F PROMPT TEMPLATE

Prompt Template: Generate Aspects

```
[System]
You are an AI assistant specializing in designing prompts to
test Large Vision-Language Models (LVLMs).  Your task is to
create meticulously {aspect_count} fined-grained aspects that
evaluate LVLMs basic understanding of images.
[Background] Large Vision-Language Models are AI systems
capable of understanding and analyzing images.  Testing these
models across various competencies is crucial for assessing
their performance, limitations, and potential biases.  The
aspects you create will be used to challenge and evaluate
LVMs.
[Instruction]
1.Basic Understanding:  This involves recognizing and
identifying individual objects, characters, and scenes within
an image.  It includes tasks like detecting the presence of
specific items (e.g., cars, trees, people), distinguishing
between different types of objects, and understanding the
general context of the scene (e.g., a park, a city street).
The goal is to accurately label all relevant elements in the
image, providing a foundation for more advanced analysis.
2.Come up with 4 general aspects according to the basic
understanding.
3.Then Create 6 fined-grained aspects within the basic
understanding for each general aspect, do not go beyond.  You
can consider the definition of the basic understanding above.
4.  Please list the aspects without using numbered lists.  5.
Let's think step by step.
[Output Format]
Please strictly respond in the following format:
General Aspect:  [Aspect]
Fined-grained Aspect:  [Aspect]
Introduction:  [Introduction]
```

Figure 22: Prompt template for generating aspects. Here we use the task of basic understanding as an example.

Prompt Template: Generate Guideline

```
[System]
You are an advanced AI simulation assistant specializing in
crafting precise prompts for image generation models.
[Instruction] I will provide you with an aspect for image
generation. Your task is to create a detailed instruction
on how to incorporate this aspect into an image generation
prompt. This instruction should guide an AI to produce a
prompt that will result in a image.
Here are two examples to guide you:
...If the aspect is Foreground vs. Background, then the
prompt should have what's in the foreground and what's in
the background.
...If the aspect is Relative Size Estimation, then you need
to include the size of one thing compared to another.
The aspect we'll focus on is: {aspect}, and the introduction
is: {introduction}
When crafting your instruction, consider the following:
1. Be specific about how the aspect should be represented
visually, similar to the examples provided.
2. Provide clear guidelines on how to balance different
elements within the image.
3. Include tips on avoiding common pitfalls or
misinterpretations related to this aspect.
Remember, the goal is to instruct an AI on creating a
prompt that will generate a single, coherent image. Your
instruction should be comprehensive enough to ensure the
final prompt will produce a high-quality, well-integrated
result.
[Output Format]
Aspect: {aspect}
Introduction: {introduction}
Guidance: [Provide your instruction directly, written in a
clear, authoritative tone. Do not include any explanations,
disclaimers, or additional commentary outside of the
instruction itself. Similar to the example given above, do
not divide it into sections.]
```

Figure 23: Prompt template for generating guideline.

Prompt Template: Generate Image Description

```
[System]
You are an AI assistant tasked with converting user inputs and their
descriptions into suitable prompts for a text-to-image model. These
prompts will generate images to test the capabilities of large
vision language models (LVLMs).
[Background] Large Vision Language Models (LVLMs) are AI systems
proficient in interpreting and analyzing images. Evaluating these
models across different competencies is essential to understanding
their performance, limitations, and potential biases. The prompts
you create will be used to generate images through text-to-image
models, which will then be used to challenge and evaluate LVLMs.
[Instruction]
1.Carefully analyze the given aspect: {aspect}, its introduction:
{introduction} and prompt generation guidance: {guidance}
2.Generate a suitable prompt based on the provided aspect and
introduction for the text-to-image model to create an image.
Ensure that the prompt is composed of simple phrases, avoiding
overly complex descriptions, and is clear enough. If you deem
the description irrelevant to the test content, do not generate a
related prompt.
3.Consider including elements that might be particularly challenging
for LVMs, such as unusual combinations, abstract concepts, or subtle
details.
4.We categorize the difficulty of prompts into easy, medium, and
hard: ...
5.Provide one overarching topic word that encapsulates the essence
of your description.
6.List 4-6 key words that are closely related to your description
and crucial for understanding the image.
7.Avoid using the following words in your new description:
{used_words_str}
8.The required difficulty level is: {level}
9.Please use clear and accurate words, clear logic flow, do not
use too abstract words. Word Choice: Word choice matters. More
specific synonyms work better in many circumstances. Instead of
big, try tiny, huge, gigantic, enormous, or immense. Plural words
and Collective Nouns: Plural words leave a lot to chance. Try
specific numbers. "Three cats" is more specific than "cats."
Collective nouns also work, \flock of birds" instead of "birds."
Focus on What You Want: It is better to describe what you want
instead of what you don't want. If you ask for a party with \no
cake," your image will probably include a cake. Try to be clear
about any context or details that are important to you. Think
about: Subject: person, animal, character, location, object
Medium: photo, painting, illustration, sculpture, doodle, tapestry
Environment: indoors, outdoors, on the moon, underwater, in the
city Lighting: soft, ambient, overcast, neon, studio lights
Color: vibrant, muted, bright, monochromatic, colorful, black and
white, pastel Mood: sedate, calm, raucous, energetic Composition:
portrait, headshot, closeup, birds-eye view But don't write it
directly in colon form, but express it normally in a sentence. ]
[Output Format]
Please strictly respond in the following format:
Aspect: {aspect}
Prompt: [Your detailed image description]
Topic word: [One word that captures the essence of the description]
Key word: [Word1, Word2, Word3,...]
```

Figure 24: Prompt template for generating image descriptions with difficulty grading.

Prompt Template: Generate Q&A

```
[System] You are an AI assistant tasked with converting user inputs
and their descriptions into suitable questions to test the Large
Vision-Language Model's (LVLM) abilities in given aspects.
[Background] Large Vision-Language Models (LVLMs) are AI systems
proficient in interpreting and analyzing images. Evaluating these
models across different competencies is essential to understanding their
performance, limitations, and potential biases. We will provide you with
a prompt to generate an image, which will create a specific image. You
can then formulate questions about this image based on the prompt. The
questions you create will be used to challenge and evaluate LVLMs based on
generated images.
[Instruction]
1.Carefully analyze the given aspect and its Introduction:
Aspect:{aspect}.
2.Generate a suitable question based on the provided image generation
prompt, and aspect to test the LVLM's ability in the given aspect. Ensure
that the question is related to the prompt of the image and is of moderate
difficulty.
3.We categorize the difficulty of questions into easy, medium, and hard.
For easy difficulty, please formulate questions based solely on very
simple details from the image generation prompt, ensuring they adhere to
the scope of the aspect in question. For medium difficulty, ensure the
question is challenging but not overly complex, involving common scenes
and requiring some level of inference or detailed observation. For hard
difficulty, consider incorporating elements that may be particularly
challenging for LVLMs, such as unusual combinations or subtle details,
while keeping the question clear and relevant, and ensure it is more
demanding than the medium level.
4.Avoid using overly complicated language or details unrelated to the
image in the questions.
5. When generating problems of different difficulty, please combine the
current specific aspect.
6.Due to potential discrepancies in image generation, we have detected
the following errors:{elements}. Please avoid referencing these elements
in your questions. If the prompt for generating the image does not
describe in detail what the specific looks like, please do not ask related
questions.
7.The required difficulty level is:{level}
8.Please generate a multiple-choice question, which can either be a
four-option single-choice question or a true/false question. If it is
a true/false question, the options should be A. True B. False.
Image generation prompt:{prompt}
9. The answers in the options need to be differentiated to a certain
extent. There cannot be a situation where multiple options meet the
requirements of the question. There can only be one answer that meets
the question.
Aspect:{aspect}
[Output Format]
Please directly output the generated question in the following JSON
format:
{
"question": "[your question]",
"options": {
"A": "[Option A]",
"B": "[Option B]",
"C": "[Option C]",
"D": "[Option D]"
},
"reference_answer": "A or B or C or D"
}
Without any other information.
```

Figure 25: Prompt template for generating questions.

Prompt Template: Answer

```
[System]
In order to test your ability with pictures, we have a
question about aspect area.  Please answerbased on your
knowledge in this area and your understanding of pictures.
Given the image below, answer the questions:  question based
on the image.
Please give the final answer strictly follow the format [[A]]
(Srtictly add [[ ]] to the choice, and the content in the
brackets should be the choice such as A, B, C, D) and provide
a brief explanation of your answer.  Directly output your
answer in this format and give a brief explanation.
```

Figure 26: Prompt template for LVLM to answer.