V^{2}R-Bench: Holistically Evaluating LVLM Robustness to Fundamental Visual Variations

Anonymous ACL submission

Abstract

Large Vision Language Models (LVLMs) have shown impressive performance on various vision-language tasks. However, while objects in natural scenes inevitably exhibit visual variations in position, scale, orientation, and context due to changes in viewpoint and environment, the robustness of LVLMs to these fundamental visual variations remains largely unexplored. To address this gap, we introduce **V^{2}R-Bench**, a comprehensive benchmark framework for evaluating **V**isual **V**ariation **R**obustness of LVLMs, which encompasses automated evaluation dataset generation and principled metrics for thorough robustness assessment. Through extensive evaluation of 21 LVLMs, we reveal a surprising vulnerability to visual variations, in which even advanced models that excel at complex vision-language tasks significantly underperform on simple tasks such as object recognition. Interestingly, these models exhibit a distinct visual position bias that contradicts theories of effective receptive fields and demonstrate a human-like visual acuity threshold. To identify the source of these vulnerabilities, we present a systematic framework for component-level analysis, featuring a novel visualization approach for aligned visual features. Results show that these vulnerabilities stem from error accumulation in the pipeline architecture and inadequate multimodal alignment. Complementary experiments with synthetic data further demonstrate that these limitations are fundamentally architectural challenges, underscoring the need for architectural innovations in future LVLM designs.[1]

1 Introduction

The rapid development of Large Vision Language Models (LVLMs) (Liu et al., 2023; Lu et al., 2024a) has been driven by two key factors: innovations in model architectures and the availability of high-quality training data. These models have demonstrated impressive results in complex vision-language tasks, achieving human-level performance across various challenges (Fei et al., 2024). To systematically evaluate such capabilities, numerous multimodal benchmarks have been developed to assess models' fundamental knowledge (Fu et al., 2024b), perceptual capability (Wu and Xie, 2023; Zhang et al., 2024b), cognitive understanding (Fu et al., 2024a), and reasoning skills (Lu et al., 2024b; Huang et al., 2024) across various downstream applications.

While current benchmarks extensively evaluate models using images collected in specific scenarios, they overlook a more fundamental and generalizable capability: **the robustness of LVLMs to visual variations**. In input images, objects naturally exhibit diverse variations: spatial positions shift with changes in camera angles and viewpoints; object scales vary depending on viewing distances; orientations deviate from standard poses through rotations and inversions; and objects appear in a range of visual semantic contexts. These visual variations naturally raise several concerns about LVLMs' robustness: whether models maintain consistent perception capabilities across all spatial positions in input images; what visual acuity threshold determines reliable performance; and how changes in orientation and visual context influence model behavior.

Despite the importance of such robustness, current research has primarily focused on model robustness to socio-cultural factors (Ananthram et al., 2024), adversarial prompt attacks (Wu et al., 2024b; Liu et al., 2024b), or corrupted image inputs (Liu et al., 2024a), leaving the impact of natural visual variations largely unexplored. A thorough review of related work is provided in Appendix A.

In this paper, we propose **V^{2}R-Bench**, a compre-

[1] Our code and data will be released in the final version.

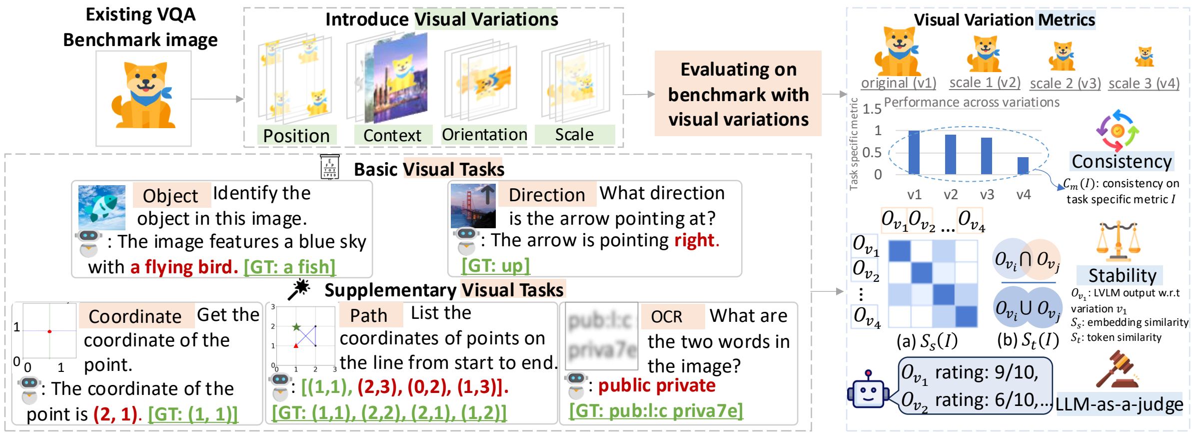

Figure 1: Illustration of V^{2}R-Bench. *Left*: Automated data generation framework that adds visual variations to images, including tasks: basic visual tasks and existing benchmarks, as well as supplementary visual tasks. *Right*: Metrics for robustness testing, including performance consistency, semantic and token-level stability, and LLM-as-a-judge.

hensive benchmark framework to evaluate LVLM robustness against fundamental visual variations. Our framework consists of (1) an automated data generation pipeline and (2) tailored evaluation metrics, which can be readily extended to various VQA tasks. Through extensive evaluation of 21 LVLMs, we uncover surprising findings: despite their excellence in complex multimodal tasks, these models exhibit unexpected vulnerabilities to visual variations, leading to poor performance even in basic tasks such as object recognition. Specifically, we observe: (1) a counter-intuitive position bias where models achieve higher accuracy at image edges rather than the center; (2) a human-like visual acuity threshold where model reliability steadily decreases with object size, reaching and maintaining minimum performance below a critical scale threshold; (3) selective robustness to certain orientations while remaining fragile to others; and (4) a tendency to ground predictions on visual contextual inference rather than direct visual perception.

To identify the root cause of these vulnerabilities, we conduct a systematic analysis of LVLM components. A key innovation in our analysis is a novel visualization method that reconstructs language tokens from aligned visual features, offering insights into how models process and transmit visual information across modalities. Our investigation reveals that inadequate multimodal alignment is the primary bottleneck, as models fail to maintain stable visual representations across variations and struggle to effectively align visual semantics with the language model. To disentangle whether these limitations stem from architectural constraints or data deficiency, we conduct complementary experiments with synthetic training data. The results reveal that these vulnerabilities are fundamentally rooted in architectural design. These findings underscore two critical directions for future LVLM development: stronger multimodal alignment mechanisms to maintain semantic consistency across variations, and unified architectural designs to mitigate error accumulation in current pipeline structures.

Our contributions are summarized as follows:

- We identify and formulate a novel problem: the robustness of LVLMs to visual variations, a fundamental yet overlooked capability essential for reliable vision-language reasoning.
- We propose V^{2}R-Bench, an evaluation framework with automated data generation and tailored metrics, uncovering significant vulnerabilities in current LVLMs through an extensive evaluation of 21 models.
- We develop a systematic component-level analysis with a novel visualization technique for multimodal alignment, revealing the root causes of these vulnerabilities and providing insights for future architectural improvements.

2 Evaluation Framework

In this section, we present our V^{2}R-Bench evaluation framework. 2.1 introduces our automated pipeline for generating diverse visual variations across different task settings. 2.2 describes our evaluation protocols. 2.3 details the construction and characteristics of our evaluation dataset.

2.1 Automated Variations Generation

The automated data generation pipeline incorporates four fundamental visual variations that are prevalent in real-world scenarios:

Position variations investigate whether LVLMs exhibit blind spots[2] in their visual processing, where models may fail to accurately perceive visual information at certain spatial locations.

Scale variations examine the perceptual boundaries of LVLMs when processing objects at different scales, similar to clinical vision tests for humans.

Orientation variations challenge the ability of LVLMs to process objects at different rotational angles, which is crucial for real-world scenarios like robotic navigation where objects rarely appear in canonical orientations.

Context variations test whether model predictions remain consistent across diverse environmental settings, revealing whether LVLMs perform genuine visual perception or rely primarily on contextual cues for inference.

Formally, given an image I, a set of transformed images along these dimensions is generated as:

$$\mathcal{D} = \{T(I, v) | v \in \{P \times S \times R \times C\}\} \quad (1)$$

where $P = \{1, ..., W\} \times \{1, ..., H\}$, $S = [s_{min}, s_{max}]$, $R = [0, 2\pi]$, and $C \in \mathcal{B}$ represent the sets of position, scale, rotation, and context variations respectively. For each question-image pair in the evaluation set, this generation process produces $|P| \times |S| \times |R| \times |C|$ variants in total, enabling a holistic exploration of the entire variation space that leaves no potential vulnerability unexplored.

2.2 Evaluation Protocol

Our evaluation protocol considers two aspects of LVLM robustness: *performance consistency* and *output stability* across variations.

Performance consistency measures whether a model maintains its task-specific metrics across visual variations, which directly quantifies how these variations impact model performance. To quantify this consistency, we define:

$$C_m(I) = 1 - \sqrt{\frac{1}{N} \sum_{v \in \mathcal{V}} (M(I_v) - \overline{M})^2} \quad (2)$$

[2] https://en.wikipedia.org/wiki/Blind_spot_(vision)

where $M(I_v)$ denotes the task-specific metric on variation I_v, $\overline{M}$ represents the mean performance across all variations, and $\mathcal{V}$ is the set of variations. A larger C_m indicates better consistency, with 1 representing that the model is robust enough to be unaffected by visual variations.

Output stability evaluates the generation stability of model outputs at both semantic and token levels, as defined in the following equations:

$$S_s(I) = \frac{1}{|\mathcal{V}|^2} \sum_{v_i \in \mathcal{V}} \sum_{v_j \in \mathcal{V}} sim(E(O_{v_i}), E(O_{v_j})) \quad (3)$$

$$S_t(I) = \frac{1}{|\mathcal{V}|^2} \sum_{v_i \in \mathcal{V}} \sum_{v_j \in \mathcal{V}} \frac{|T(O_{v_i}) \cap T(O_{v_j})|}{|T(O_{v_i}) \cup T(O_{v_j})|} \quad (4)$$

where $E(O)$ denotes the output embedding, $sim(\cdot, \cdot)$ computes the cosine similarity between embeddings, and $T(O)$ represents the set of tokens in output O.

We also employ LLM-as-a-judge (Zheng et al., 2023) to emulate human assessment of LVLM-generated outputs under structured visual variations, providing an additional qualitative perspective on model robustness.

2.3 Dataset Construction

The proposed automated generation pipeline is implemented across two categories of tasks. The first category examines fundamental visual capabilities through object recognition and direction recognition tasks, with target objects and directional indicators being systematically transformed through our visual variations using image processing algorithms and inpainting diffusion models (Corneanu et al., 2024; Lugmayr et al., 2022). The second category extends to existing multimodal benchmarks, focusing on scenarios where variations in position, scale, orientation, and context preserve ground-truth validity, ensuring that any performance changes reflect model robustness rather than ground-truth alteration. The final evaluation datasets contain a total of 428K images. Each category serves a distinct yet complementary purpose in our evaluation: the basic tasks provide controlled, interpretable measures of fundamental capabilities, while the extended benchmarks assess robustness in more naturalistic settings. The detailed implementation of these generation algorithms is provided in Appendix D.

3 Cross-Modal Diagnosis

To understand the underlying mechanisms of LVLM vulnerabilities to visual variations, we propose a systematic analysis framework that examines contributions of each model components to robustness issues. Central to this framework is a novel visualization technique that provides straightforward insights into how visual features extracted by vision encoder are processed through the multimodal alignment module and aligned with language embedding space.

3.1 Component-level Analysis

Here we first formalize the general architecture and process pipeline of modern LVLMs to facilitate subsequent analysis. Given an input question Q and an image I, the vision encoder E_v first extract visual features from the input image:

$$v = E_v(I; \theta_v) \in \mathbb{R}^{N_v \times D_v} \tag{5}$$

These visual features are then projected into the language embedding space through a multimodal alignment module P:

$$h = P(v; \theta_p) \in \mathbb{R}^{N_v \times D_h} \tag{6}$$

Finally, the language model M takes the aligned visual features in conjunction with question embeddings $E_l(Q) \in \mathbb{R}^{N_q \times D_h}$ as input to generate the response in an autoregressive manner:

$$R_t = M(h, E_l(Q), R_{<t}; \theta_m) \quad \text{for } t = 1, \ldots, T \tag{7}$$

As the initial input module of LVLMs, the **vision encoder** E_v fundamentally determines the performance ceiling of the whole model, since the visual information contained in its extracted features represents the upper bound of visual content available to subsequent modules. Given the distinct pre-training paradigms of vision encoders (contrastive learning (Radford et al., 2021) and self-supervised learning (Oquab et al., 2024)), the feature quality assessment differs accordingly: for contrastive-trained encoders, the corresponding text encoder enables zero-shot analysis, while for supervised-trained ones, linear probing (Alain and Bengio, 2018) and clustering analysis serve to assess their extracted features.

The **multimodal projector** P acts as a bridge between the visual feature and language embedding spaces, which raises two questions for its performance analysis: **(RQ1)** Do the projected features h preserve the visual information contained in v?, and **(RQ2)** Do the projected features h align well with the embedding space of the language model M? To answer these questions, two analytical approaches are employed: (1) comparing the performance of pre-projection features v and post-projection features h on visual tasks to quantify potential visual information loss during multimodal alignment, and (2) measuring the discriminability between projected visual features h and the language embeddings of their corresponding captions to assess the quality of modality alignment.

The **language model** M, as the final module for integrating aligned visual features with input questions and generating responses, is inherently affected by potential errors from upstream modules, which complicates the assessment of intrinsic language model robustness. Thus, an evaluation strategy is devised to bypass upstream modules by directly providing visual information as language tokens, which simulates ideal visual feature extraction and multimodal alignment where visual information is perfectly preserved without any loss or misalignment, enabling a controlled setting for analyzing inherent language model capabilities. We construct a set of text-based evaluation datasets that parallel the visual tasks in LVLMs, where visual scenes are simulated through matrix-structured text inputs. The text inputs serve as ideal visual encodings that contain different visual variations. Comparing the performance on these text-based tasks with that of LVLMs reveals whether the vulnerability to visual variations stems from the language model component.

3.2 Visual-Linguistic Feature Analysis

In addition to the aforementioned quantitative analysis of multimodal alignment, our proposed framework also incorporates a novel visualization approach, which reconstructs language tokens from aligned visual features to provide interpretable evidence of the alignment process, an aspect previously unexplored. Specifically, given a single aligned visual feature $h \in \mathbb{R}^{1 \times D_h}$ and token embedding matrix $E \in \mathbb{R}^{|V| \times D_h}$ of the language model M, the feature h can be decoded to a set of language tokens that approximate its semantic meaning:

$$t = \text{topk}\left(\text{softmax}\left(hE^\top\right)\right), \tag{8}$$

where $|V|$ denotes the vocabulary size of the language model and k controls the number of tokens to be selected.

Through these decoded language tokens, the semantic meaning captured by the aligned visual features becomes intuitively understandable, which helps address the following questions: **(RQ3)** how do language models interpret these aligned visual features? **(RQ4)** How robust is the semantics of aligned visual features to visual variations? A detailed investigation of these questions is provided in Section 4.3.

4 Experiments

4.1 Experimental Setup

Detailed experiment settings are provided in Appendix E.

4.2 Impact of Visual Variations

The evaluation results across different LVLM architectures and model scales are presented in Tables 5, and 6. Despite their impressive performance demonstrated on complex visual tasks, these models exhibit surprising vulnerability to simple visual variations, resulting in significantly degraded performance across basic visual tasks. Even proprietary models such as gpt-4o, Claude, and Gemini not only produce incorrect outputs, but also occasionally claim an inability to perceive the visual content entirely, as illustrated in Appendix G. Interestingly, despite claims that some carefully distilled smaller models outperform their larger counterparts on mainstream benchmarks, our analysis reveals that scaling laws still hold for robustness: within the same model architecture, larger models consistently demonstrate better stability across visual variations.

For an in-depth understanding of these vulnerabilities, this study focuses on LLaVA model, which provides full access to its training data, code, and parameters. **Position**: as shown in Figure 8, model performance varies dramatically across different positions. Contrary to the effective receptive field theory (Luo et al., 2017; Raghu et al., 2022) which suggests vision models have better perception of central regions, LVLMs exhibit strong visual position bias with higher accuracy at peripheral regions, raising concerns about their fundamental visual processing mechanisms. Figure 11 presents a more intuitive example using right-pointing arrows as input, illustrating the significant effect of positional variations on prediction outcomes. **Scale**: Figure 3 illustrates a sharp performance decline as object size decreases, stabilizing when the object occupies 1/100 of the image area (equivalent to 1/10 of both width and height). This reveals a potential visual acuity threshold in LVLMs, similar to human vision, that defines a critical boundary below which model outputs become unreliable, serving as a key indicator for deploying LVLMs in fine-grained visual perception tasks. **Orientation**: performance degradation is observed across different orientations, with models exhibiting distinct directional biases: some orientations show robustness, while others lead to significant failures. Interestingly, models like Fuyu and BLIP demonstrate a pronounced predictive tendency, being heavily influenced by the left orientation. **Context**: model predictions vary with different contextual arrangements, correlating with context content, which raises questions about whether the model truly perceives the target objects or infers them from contextual cues.

To further investigate the root causes of these vulnerabilities, we design complementary tasks focusing on fundamental perceptual capabilities. To examine whether position and direction vulnerabilities stem from limitations in spatial and directional perception, we introduce two specialized tasks: coordinate identification, where models directly output position coordinates to test spatial understanding, and path tracking, where models sequentially output coordinates following directional lines to examine continuous directional perception. To verify whether models achieve genuine recognition or merely rely on contextual inference, we develop a modified OCR task where certain letters in fluent text are intentionally replaced with incorrect characters and then blurred at different levels, testing whether models faithfully report the visual content.

As shown in Table 2, all models exhibit poor performance in coordinate recognition and path tracing, indicating these vulnerabilities fundamentally originate from the inability to accurately perceive spatial properties. Figure 12 reveals peak accuracy at initial points followed by monotonically decreasing performance, with marginal recovery at terminal nodes. This indicates not only weak isolated direction recognition but also inadequate performance in visual-following reasoning tasks, suggesting that spatial reasoning capabilities likely derive more from commonsense world knowledge rather than genuine visual understanding. Table 1 results demonstrate that LVLMs tend to overconfidently output contextually inferred conclusions rather than objectively depicted visual con-

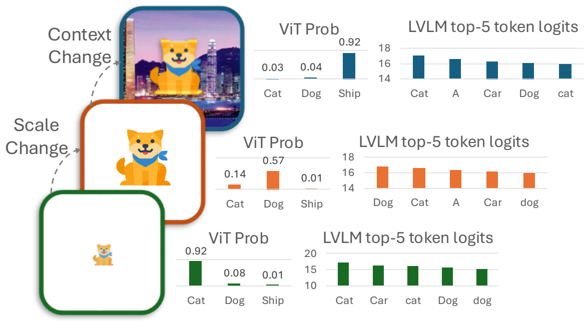

Figure 2: Changes in vision encoder classification probabilities and LVLM token predictions under different visual variations (context and scale). For each variation, we show the ViT's top-3 class probabilities (left) and LVLM's top-5 token logits (right), demonstrating how semantic interpretations shift across visual variations.

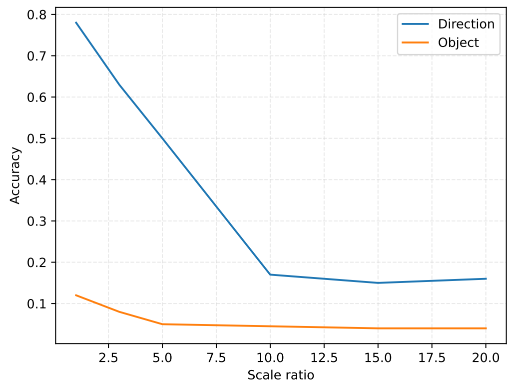

Figure 3: Model performance as a function of relative object scale, where the x-axis represents the ratio between input image dimension (width/height) and object size. A larger ratio indicates the object occupies a smaller portion of the input image.

tent, while LLMs achieve better accuracy under the same task conditions. This contrast proves that the observed limitations stem not from the denoising characteristics (Devlin et al., 2019) inherent to the transformer architecture itself, but rather from the interference introduced by visual modality integration.

4.3 Component Analysis

Vision Encoder. As shown in Figure 8, the LVLM and its vision encoder exhibit similar position bias, achieving higher accuracy in peripheral areas compared to central regions. Moreover, Figure 2 reveals that changes in the prediction label probabilities of the vision encoder align with shifts in the next token logits of the LVLM when scale and context variations are introduced. This behavioral consistency suggests that the LVLM inherits its vulnerability to visual variations from its vision encoder component.

Multimodal Projector. The following analyses address the four research questions posed above.

(RQ1) Visual Information Loss. Linear probing results for pre-projection and post-projection features are presented in Table 4. The significant performance degradation after multimodal projection suggests irrecoverable information loss during the alignment process, potentially contributing to LVLM's vulnerability across visual variations. The poor performance of MM-Projector implies that there is semantic information loss during the modality alignment process, meaning it is the potential cause for the degradation of LVLM robustness across visual variations.

We further compare the fine-tuned vision encoder in LLaVA 1.6 with the original one. While its vision encoder is fine-tuned to adapt image-text features for generative tasks, its linear probing performance underperforms the original vision encoder, revealing a trade-off: task-specific adaptation improves multimodal coherence but erodes the vision encoder's innate spatial representational fidelity. This aligns with LLaVA 1.5's limitations, where only the lightweight projector can be fine-tuned and struggled to mitigate information loss from the frozen CLIP encoder. Our results underscore that while recent advancements in alignment neural networks alleviate alignment bottlenecks, fundamental architectural constraints such as patch-based tokenization and positional bias persist as critical vulnerabilities, necessitating unified approaches to relax the conflicts between visual grounding and multimodal alignment.

(RQ2) Inadequte Multimodal Alignment. Figure 4 illustrates the spatial distribution of visual features, aligned features, and language embeddings. We observe that the distribution of the aligned features is similar to that of image features, meaning that the multimodal process preserves the visual semantics for some extent. Ideally, the aligned features should closely resemble the language embeddings. However, our observation reveals a significant disparity between the aligned feature space and the language embedding space, suggesting a lack of adequate modality alignment. This finding underscores the critical role of the multimodal

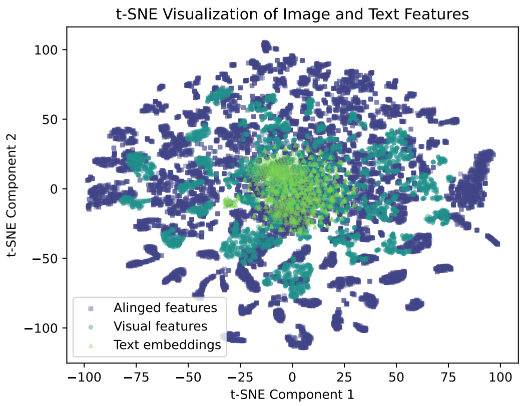

Figure 4: Visualization of different feature representations: aligned visual features projected into the language embedding space, with images from the MSCOCO dataset (Lin et al., 2015) and text embeddings obtained from the language model.

Figure 5: Images with their corresponding word cloud visualization of the decoded aligned features. The dashed red circles highlight selected semantically relevant tokens.

projector, which is a key factor contributing to the LVLM's vulnerability to visual variations.

To explore the image feature representation with regard to visual variations, Figure 10 presents the clustering analysis of image features of the same object at different directions and positions. These variations introduce substantial alterations in image features, highlighting the inherent challenges faced by robust image encoding techniques. These fluctuations can lead to inconsistencies in feature extraction and representation across different instances of the same object or scene. The limited robustness in current encoding methodologies may struggle to effectively capture and encode these variations while maintaining the semantic integrity of the features, potentially compromising the accuracy and reliability of downstream tasks. Addressing this issue requires the development of more resilient encoding strategies that can adapt to diverse visual transformations, enhancing the overall robustness and generalizability of image feature representations in complex visual tasks.

(RQ3) Interpretation of Aligned Features. Figure 5 demonstrates the results of decoding aligned visual features into language tokens. Due to the inherent differences in attention patterns between vision encoder (bidirectional) and language model (autoregressive), these decoded results do not form coherent natural language, with only a subset of tokens being semantically relevant to the image content. These aligned visual features, which reside outside the discrete language embedding space (as shown in Figure 4), can be interpreted as visually conditioned soft prompts (Lester et al., 2021; Gu et al., 2022; Liu et al., 2022), providing implicit cues that contain image-related information to guide model responses to image-related queries. However, prior work (Bailey et al., 2024) has revealed the limitations of such soft prompt approaches, including their instability and potential hidden bugs due to lack of interpretability and the discrepancy from language model embedding space, ultimately leading to the vulnerability to visual variations.

(RQ4) Visual Semantics Vulnerability. Based on the results for RQ3, significant changes in decoded language tokens are observed across visual variations, indicating poor semantic robustness. These inconsistent soft prompt prefixes consequently lead to substantial fluctuations in model outputs, further compromising the model's robustness.

Language Model. It has come to the community's attention that positional biases in natural language widely exist in language models (Wang et al., 2023a; Jung et al., 2019; Zheng et al., 2023; Koo et al., 2024), where models tend to favor information appearing earlier in the text. Similarly, we observe that language models exhibit position bias in matrix-structured text representations of visual information. Furthermore, although object and background information are explicitly represented as language tokens, model predictions are still influenced by context, albeit to a lesser extent than LVLMs, suggesting that the vulnerability primarily stems from the upstream vision encoder and multimodal alignment module, where the aligned visual features fed into the language model are already heavily influenced by visual context. The low accuracy in the coordinate task demonstrates that, similar to LVLMs, language models also lack precise positional awareness of objects. These limitations are rooted in the autoregressive nature of language models, which has the tendency to prioritize sequential dependencies over structural relationships.

4.4 Mitigating Robustness Issues

To determine whether the robustness issues stem from architectural limitations or insufficient training data, we explore improvements through two complementary approaches: First, we conduct controlled experiments on a subset of our test data while maintaining a held-out test set, directly probing the architectural capacity for robust visual understanding. Second, we utilize a general visual instruction tuning dataset injected with visual variations to analyze whether a more diverse training distribution can enhance model robustness. We find that a more diverse dataset offers limited improvement in model robustness, likely due to insufficient data volume (as exhaustively covering these variations requires many more variants). Moreover, while directly training the model on spatial visual tasks improves its performance on position and direction tasks, it does not enhance robustness against position and orientation variations. As Figure 7 suggests, after training, despite improvements in the coordinate task, the path tracing task remains underperforming, due to the lack of sustained visual attention. This reveals that the vulnerability is fundamentally rooted in architectural design choices rather than data limitations. The design of the vision encoder utilizes patch-based tokenization (Radford et al., 2021) and positional embeddings, which may lead to information fragmentation due to arbitrary patch partitioning and position sensitivity induced by explicit positional encoding. Moreover, the cascading pipeline architecture further amplifies vulnerabilities at each component, suggesting the potential benefits of a new architectural approach. Although unified architectures currently underperform on visual understanding tasks due to training stability challenges (Chen et al., 2025a), this direction remains promising.

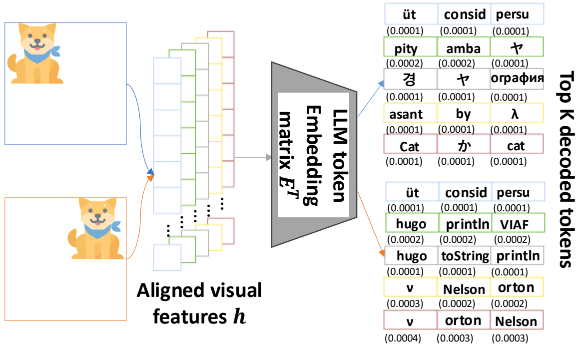

Figure 6: An example of visual-linguistic feature analysis. The decoded language tokens show that the output of MM-Projector does not align well with the text embedding space and suffers from semantic information loss. These aligned features also lack robustness against visual variations.

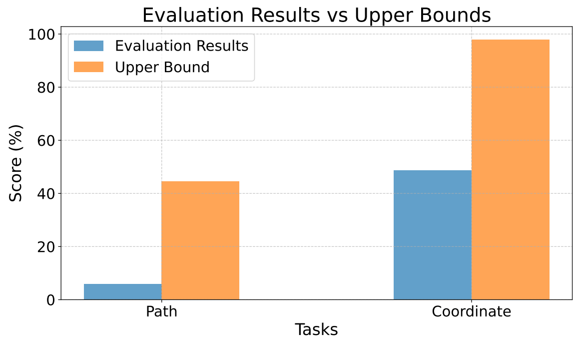

Figure 7: The comparison of benchmark evaluation results and upper bound on coordinate and path tasks, where the upper bound is the testing result of the LVLM fine-tuned on partial benchmark dataset.

5 Conclusion

In this paper, we present V^{2}R-Bench, an evaluation framework designed to assess the robustness of LVLMs against visual variations. Our results show significant vulnerabilities in existing LVLMs and identify their origins as twofold: insufficient multimodal alignment and error accumulation inherent in pipeline model architectures. While synthetic data augmentation seems as a mitigation strategy, fundamental advancement in the field necessitates a shift towards native multimodal architectures rather than the current approach of concatenating separate language and vision modalities. We aim to draw attention to the importance of visual robustness in LVLMs and inspire future research toward more robust architectural paradigms.

Limitations

Aligned with the paper track's focus, this work primarily concentrates on identifying a novel problem, establishing corresponding evaluation methods, and providing initial analytical insights into these vulnerabilities. Regarding solutions to the identified vulnerabilities, only two data synthesis approaches were explored, while potential improvements through architectural modifications, pre-training strategies, or novel modality alignment methods remain unexplored due to their substantial resource requirements. The detailed analysis presented in this paper aims to provide insights for future research addressing these challenges. Additionally, while this study examines four fundamental types of visual variations, it does not exhaust the infinite possibilities of visual transformations, leaving some long-tail cases unexplored.

Ethics Statements

This study reveals a fundamental vulnerability in LVLMs - their lack of robustness to visual variations that naturally occur from camera parameter adjustments and environmental changes. Such robustness deficiency results in significant output inconsistencies across visually similar scenarios, compromising model reliability in real-world deployments.

A critical security implication emerges from these findings: visual variations could serve as a novel attack vector. Through strategic object placement with specific positions, scales, orientations, or contexts, these fundamental visual variations could be exploited to manipulate model behavior. Unlike conventional adversarial attacks requiring sophisticated training procedures, this approach requires no training and generates natural images without artificial artifacts, making such attacks particularly challenging for existing detection mechanisms.

The identified vulnerability underscores the need for increased attention within the research community to these fundamental yet profoundly impactful visual variations, rather than solely pursuing state-of-the-art performance on complex tasks. Enhanced robustness to these variations is crucial for ensuring consistent model performance in natural environments. Moreover, as attacks based on visual variations exploit inherent model vulnerabilities rather than crafted adversarial strategies, addressing this robustness issue becomes essential for improving both the reliability and security of deployed LVLMs.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Guillaume Alain and Yoshua Bengio. 2018. Understanding intermediate layers using linear classifier probes. *Preprint*, arXiv:1610.01644.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. *Preprint*, arXiv:2204.14198.

Amith Ananthram, Elias Stengel-Eskin, Carl Vondrick, Mohit Bansal, and Kathleen McKeown. 2024. See it from my perspective: Diagnosing the western cultural bias of large vision-language models in image understanding. *Preprint*, arXiv:2406.11665.

Federico Arenas. 2020. Blender SDG Icons Generator. https://github.com/federicoarenasl/blender-sdg. Blender add-on for generating 3D Sustainable Development Goals (SDG) icons.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Luke Bailey, Gustaf Ahdritz, Anat Kleiman, Siddharth Swaroop, Finale Doshi-Velez, and Weiwei Pan. 2024. Soft prompting might be a bug, not a feature.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023a. Sharegpt4v: Improving large multi-modal models with better captions. *Preprint*, arXiv:2311.12793.

Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, Benyou Wang, Shaoting Zhang, Bin Fu, Jianfei Cai, Bohan Zhuang, Eric J Seibel, Junjun He, and Yu Qiao. 2024a. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *Preprint*, arXiv:2408.03361.

Shuo Chen, Jindong Gu, Zhen Han, Yunpu Ma, Philip Torr, and Volker Tresp. 2023b. Benchmarking robustness of adaptation methods on pre-trained vision-language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025a. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *Preprint*, arXiv:2501.17811.

Zeren Chen, Zhelun Shi, Xiaoya Lu, Lehan He, Sucheng Qian, Hao Shu Fang, Zhenfei Yin, Wanli Ouyang, Jing Shao, Yu Qiao, Cewu Lu, and Lu Sheng. 2024b. Rh20t-p: A primitive-level robotic dataset towards composable generalization agents. *Preprint*, arXiv:2403.19622.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *Preprint*, arXiv:2412.05271.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, and Chunhua Shen. 2024. Mobilevlm v2: Faster and stronger baseline for vision language model. *Preprint*, arXiv:2402.03766.

Ciprian Corneanu, Raghudeep Gadde, and Aleix M Martinez. 2024. Latentpaint: Image inpainting in latent space with diffusion models. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4322–4331.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. *Preprint*, arXiv:2409.17146.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei

Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Hao Fei, Yuan Yao, Zhuosheng Zhang, Fuxiao Liu, Ao Zhang, and Tat-Seng Chua. 2024. From multimodal LLM to human-level AI: Modality, instruction, reasoning, efficiency and beyond. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries*, pages 1–8, Torino, Italia. ELRA and ICCL.

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. 2024a. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024b. Blink: Multimodal large language models can see but not perceive. *Preprint*, arXiv:2404.12390.

Shaikat Galib, Shanshan Wang, Guanshuo Xu, Pascal Pfeiffer, Ryan Chesler, Mark Landry, and Sri Satish Ambati. 2024. H2ovl-mississippi vision language models technical report. *Preprint*, arXiv:2410.13611.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. 2022. PPT: Pre-trained prompt tuning for few-shot learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8410–8423, Dublin, Ireland. Association for Computational Linguistics.

Phillip Howard, Anahita Bhiwandiwalla, Kathleen C. Fraser, and Svetlana Kiritchenko. 2024. Uncovering bias in large vision-language models with counterfactuals. *Preprint*, arXiv:2404.00166.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. 2024. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. *Preprint*, arXiv:2402.09181.

Kung-Hsiang Huang, Mingyang Zhou, Hou Pong Chan, Yi Fung, Zhenhailong Wang, Lingyu Zhang, Shih-Fu Chang, and Heng Ji. 2024. Do LVLMs understand charts? analyzing and correcting factual errors in chart captioning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 730–749, Bangkok, Thailand. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Taehee Jung, Dongyeop Kang, Lucas Mentch, and Eduard Hovy. 2019. Earlier isn't always better: Subaspect analysis on corpus and system biases in summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3324–3335, Hong Kong, China. Association for Computational Linguistics.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2024. Benchmarking cognitive biases in large language models as evaluators. *Preprint*, arXiv:2309.17012.

Tony Lee, Haoqin Tu, Chi Heem Wong, Wenhao Zheng, Yiyang Zhou, Yifan Mai, Josselin Somerville Roberts, Michihiro Yasunaga, Huaxiu Yao, Cihang Xie, and Percy Liang. 2024. Vhelm: A holistic evaluation of vision language models. *Preprint*, arXiv:2410.07112.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *Preprint*, arXiv:2104.08691.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. *Preprint*, arXiv:2408.03326.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Preprint*, arXiv:2301.12597.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, Lingpeng Kong, and Qi Liu. 2023c. M^3it: A large-scale dataset towards multi-modal multilingual instruction tuning. *Preprint*, arXiv:2306.04387.

Paul Pu Liang, Akshay Goindani, Talha Chafekar, Leena Mathur, Haofei Yu, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2024. Hemm: Holistic evaluation of multimodal foundation models. *Preprint*, arXiv:2407.03418.

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

Daizong Liu, Xiaowen Cai, Pan Zhou, Xiaoye Qu, Xiang Fang, Lichao Sun, and Wei Hu. 2024a. Are large vision-language models robust to adversarial visual transformations?

Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Xiang Fang, Keke Tang, Yao Wan, and Lichao Sun. 2024b. Pandora's box: Towards building universal attackers against real-world large vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024c. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024d. Mmbench: Is your multi-modal model an all-around player? *Preprint*, arXiv:2307.06281.

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. 2024a. Deepseek-vl: Towards real-world vision-language understanding. *Preprint*, arXiv:2403.05525.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024b. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *Preprint*, arXiv:2310.02255.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. *Preprint*, arXiv:2201.09865.

Gen Luo, Xue Yang, Wenhan Dou, Zhaokai Wang, Jiawen Liu, Jifeng Dai, Yu Qiao, and Xizhou Zhu. 2024a. Mono-internvl: Pushing the boundaries of monolithic multimodal large language models with endogenous visual pre-training. *Preprint*, arXiv:2410.08202.

Weidi Luo, Siyuan Ma, Xiaogeng Liu, Xiaoyu Guo, and Chaowei Xiao. 2024b. Jailbreakv: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks. *Preprint*, arXiv:2404.03027.

Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. 2017. Understanding the effective receptive field in deep convolutional neural networks. *Preprint*, arXiv:1701.04128.

Feipeng Ma, Yizhou Zhou, Hebei Li, Zilong He, Siying Wu, Fengyun Rao, Yueyi Zhang, and Xiaoyan Sun. 2024. Ee-mllm: A data-efficient and compute-efficient multimodal large language model. *Preprint*, arXiv:2408.11795.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav

Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. Dinov2: Learning robust visual features without supervision. *Preprint*, arXiv:2304.07193.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. 2022. Do vision transformers see like convolutional neural networks? *Preprint*, arXiv:2108.08810.

Ryan Steed and Aylin Caliskan. 2021. Image representations learned with unsupervised pre-training contain human-like biases. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 701–713, New York, NY, USA. Association for Computing Machinery.

Qwen Team. 2024. Qvq: To see the world with wisdom.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. Is chatgpt a good nlg evaluator? a preliminary study. *Preprint*, arXiv:2303.04048.

Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. 2023b. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *Preprint*, arXiv:2311.07574.

Junyang Wang, Haiyang Xu, Jiabo Ye, Ming Yan, Weizhou Shen, Ji Zhang, Fei Huang, and Jitao Sang. 2024a. Mobile-agent: Autonomous multi-modal mobile device agent with visual perception. *Preprint*, arXiv:2401.16158.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Sibo Wang, Xiangkui Cao, Jie Zhang, Zheng Yuan, Shiguang Shan, Xilin Chen, and Wen Gao. 2024c. Vlbiasbench: A comprehensive benchmark for evaluating bias in large vision-language model. *Preprint*, arXiv:2406.14194.

Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, Xihui Liu, Cewu Lu, Dahua Lin, and Jiangmiao Pang. 2023c. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. *Preprint*, arXiv:2312.16170.

Penghao Wu and Saining Xie. 2023. V*: Guided visual search as a core mechanism in multimodal llms. *Preprint*, arXiv:2312.14135.

Shujin Wu, Yi Fung, Sha Li, Yixin Wan, Kai-Wei Chang, and Heng Ji. 2024a. MACAROON: Training vision-language models to be your engaged partners. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7715–7731, Miami, Florida, USA. Association for Computational Linguistics.

Xiyang Wu, Ruiqi Xian, Tianrui Guan, Jing Liang, Souradip Chakraborty, Fuxiao Liu, Brian M. Sadler, Dinesh Manocha, and Amrit Bedi. 2024b. On the safety concerns of deploying LLMs/VLMs in robotics: Highlighting the risks and vulnerabilities. In *First Vision and Language for Autonomous Driving and Robotics Workshop*.

Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, Wenhao Zheng, Zhaoyang Wang, Xiao Wang, Xuchao Zhang, Chetan Bansal, Marc Niethammer, Junzhou Huang, Hongtu Zhu, Yun Li, Jimeng Sun, Zongyuan Ge, Gang Li, James Zou, and Huaxiu Yao. 2024. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. *Preprint*, arXiv:2406.06007.

Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. 2023. Gpt4tools: Teaching large language model to use tools via self-instruction. *Preprint*, arXiv:2305.18752.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *Preprint*, arXiv:2408.01800.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. Mm-vet: Evaluating large multimodal models for integrated capabilities. *Preprint*, arXiv:2308.02490.

Chi Zhang, Zhao Yang, Jiaxuan Liu, Yucheng Han, Xin Chen, Zebiao Huang, Bin Fu, and Gang Yu. 2023. Appagent: Multimodal agents as smartphone users. *Preprint*, arXiv:2312.13771.

Hao Zhang, Wenqi Shao, Hong Liu, Yongqiang Ma, Ping Luo, Yu Qiao, Nanning Zheng, and Kaipeng Zhang. 2024a. B-avibench: Towards evaluating the robustness of large vision-language model on black-box adversarial visual-instructions. *Preprint*, arXiv:2403.09346.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. 2024b. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *Preprint*, arXiv:2408.13257.

Bo Zhao, Boya Wu, Muyang He, and Tiejun Huang. 2023. Svit: Scaling up visual instruction tuning. *Preprint*, arXiv:2307.04087.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

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.

Yichen Zhu, Minjie Zhu, Ning Liu, Zhicai Ou, Xiaofeng Mou, and Jian Tang. 2024. Llava-phi: Efficient multimodal assistant with small language model. *Preprint*, arXiv:2401.02330.

A Related Work

Large Vision Language Models. The remarkable progress of LVLMs stems from continuous advances in model architectures and the exponential growth of training datasets. On the architectural front, the rapid progress in foundation language models (OpenAI et al., 2024; Touvron et al., 2023; Jiang et al., 2023; Bai et al., 2023a) and vision models (Radford et al., 2021; Oquab et al., 2024), together with increasingly sophisticated multimodal alignment modules (Liu et al., 2023; Li et al., 2023b; Alayrac et al., 2022; Zhu et al., 2023; Wang et al., 2024b), has established the fundamental capabilities of LVLMs in understanding and reasoning across modalities. Building upon these architectural foundations, the emergence of high-quality datasets (Chen et al., 2023a; Zhao et al., 2023; Wang et al., 2023b; Li et al., 2023c) designed for different training stages (e.g., multimodal alignment, visual instruction tuning, preference alignment) has enabled LVLMs to demonstrate exceptional performance across diverse real-world scenarios. More recently, research efforts have focused on optimizing lightweight model architectures and curating datasets tailored to edge scenarios (Ma et al., 2024; Chu et al., 2024; Zhu et al., 2024; Yao et al., 2024), thereby promoting the practical deployment and adoption of LVLMs in real-world applications.

While the pipeline architecture of these LVLMs effectively leverages pretrained domain-specific knowledge from vision and language components, it potentially accumulates errors and vulnerabilities across modules. **However, no systematic investigation has been conducted to attribute model failures to individual architectural components, limiting both model interpretability and understanding of modality alignment.**

LVLM Evaluation and Benchmarking. Recent years have witnessed the emergence of numerous benchmarks for evaluating LVLM capabilities across cognitive and perceptual dimensions (Liang et al., 2024; Fu et al., 2024b,a; Liu et al., 2024d; Yu et al., 2024; Li et al., 2023a; Lee et al., 2024; Wu et al., 2024a), comprehensively assessing various aspects including reasoning skills, understanding abilities, and inherent knowledge. The scope of evaluation has further extended into specialized domains, with an increased emphasis on real-world scenario performance, as researchers develop dedicated benchmarks for embodied intelligence (Wang et al., 2024a; Yang et al., 2023; Zhang et al., 2023), medical image analysis (Xia et al., 2024; Chen et al., 2024a; Hu et al., 2024), and robotic control (Chen et al., 2024b; Wang et al., 2023c). In parallel with capability assessment, researchers have begun investigating LVLM robustness from two critical perspectives: first, examining semantic biases in model responses, particularly those related to societal factors like gender and racial prejudices (Wang et al., 2024c; Howard et al., 2024; Steed and Caliskan, 2021); and second, analyzing adversarial vulnerabilities via carefully crafted visual prompts for assessing model reliability under targeted attacks (Liu et al., 2024a; Luo et al., 2024b; Zhang et al., 2024a; Wang et al., 2024c; Chen et al., 2023b). However, the capacity to withstand basic visual variations, a fundamental aspect of robustness that is widely present in real-world deployment, remains unexplored in current research.

B Discussion

Summary of Empirical Findings Through the evaluation of V^{2}R-Bench, we discover that introducing visual variation causes inconsistent and unstable output. Such vulnerability is due to the error accumulation in the data generation pipeline. We discover that ViT has the tendency to depend on the context during inference rather than truly recognize a target object. Through component analysis, we claim that multimodal projector is the main cause of the vulnerability to visual variations. The inadequate multimodal alignment causes visual information loss and the decoding result of aligned feature does not form coherent natural language.

The Use of Synthetic Data While our framework enables the incorporation of real-world benchmarks, we acknowledge that generated variations may not fully capture the complexity and diversity of real-world visual inputs. To address this, future work can be done to further integrate our framework with diffusion-based contextual blending or 3D synthetic data generation technique such as Blender-SDG (Arenas, 2020).

Future Directions V^{2}R-Bench fills a crucial gap in existing evaluations by systematically testing robustness to fundamental visual variations (position, scale, orientation, context)—ubiquitous in real-world scenarios but overlooked by current benchmarks. For instance, autonomous vehicles require consistent object recognition regardless of camera angles, and medical imaging tools must identify anomalies across scales. By integrating existing benchmarks and evaluating with our synthetic data, our framework directly addresses these needs by enabling models to be stress-tested under realistic conditions. Researchers can also easily expand tests to new variations (e.g., lighting changes) or domains (e.g., CT image).

The benchmark's component-level analysis offers fine-grained diagnostic information, guiding researchers toward targeted improvements. A key future direction lies in enhancing the multimodal alignment mechanisms. Additionally, the identified vulnerabilities are not merely the result of limited data or training strategies but also stem from fundamental architectural constraints requiring further investigation.

C Prompt Template

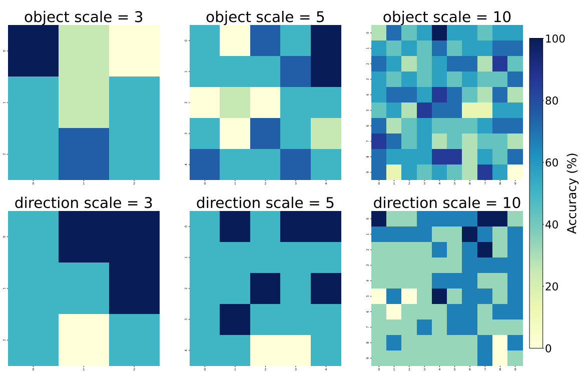

Figure 8: Accuracy heatmaps for object recognition and direction recognition across different object scales and position variations.

Prompt For Evaluation

- **Object:** Identify the object in the image.
- **Direction:** List the direction the arrow is pointing in the image using one of the following: up, down, left, right, top-left, bottom-left, top-right, or bottom-right.
- **Coordinate:** This is a coordinate plot with a single point. Provide the coordinate in the format $(x,)$ for 1D, (x, y) for 2D, or (x, y, z) for 3D.
- **Path:** Describe the coordinates of each point along the line from the start to the end in the format $[(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)]$.

D Experimental Setup

We conduct comprehensive experiments across three distinct datasets to evaluate our model's performance under various conditions. The experimental settings for each dataset are detailed below.

D.1 Coordinate Dataset

For the coordinate dataset, we systematically vary four key parameters to thoroughly assess model performance. The point range parameter defines the spatial extent of the coordinate system, with values spanning from confined spaces ($[-5, 5]$) to broader ranges ($[-10, 10]$, $[0, 10]$, and $[0, 20]$). This variation allows us to evaluate the model's ability to handle different scales of spatial information.

To investigate the impact of visual aids on model performance, we experiment with both the presence and absence of reference lines and grid systems. These binary parameters (True/False) help us understand how additional visual context affects the model's coordinate understanding.

Regarding dimensionality, we focus our analysis on one- and two-dimensional coordinate systems. While three-dimensional coordinates were initially considered, they were ultimately excluded from our final experiments due to consistently poor model performance across preliminary tests.

D.2 Path Dataset

In the path dataset experiments, we explore the model's capability to process and understand connected point sequences. We vary the complexity of paths by adjusting the number of points from 2 to

Model	Visual Dependency					Knowledge Dependency				
	Text	B0	B1	B2	B3	Text	B0	B1	B2	B3
GPT-4o	0.2109	0.0475	0.0586	0.2916	0.4724	0.0821	0.2044	0.2030	0.2004	0.2232
Qwen2-VL	0.2531	0.1169	0.2417	0.3710	0.5344	0.2674	0.2344	0.2284	0.3154	0.4945
Llava-1.6	0.2129	0.5732	1.8404	1.8026	1.4573	0.3743	0.5676	1.8013	1.7919	1.4485

Table 1: Performance Comparison of Models under Different Blur Conditions: LVLMs favor contextual inference over input-output consistency, while LLMs maintain better alignment with inputs. As blur levels increase, LVLMs confidently continue contextual reasoning, implying they inherently operate through inference from ambiguous visual signals rather than direct visual processing.

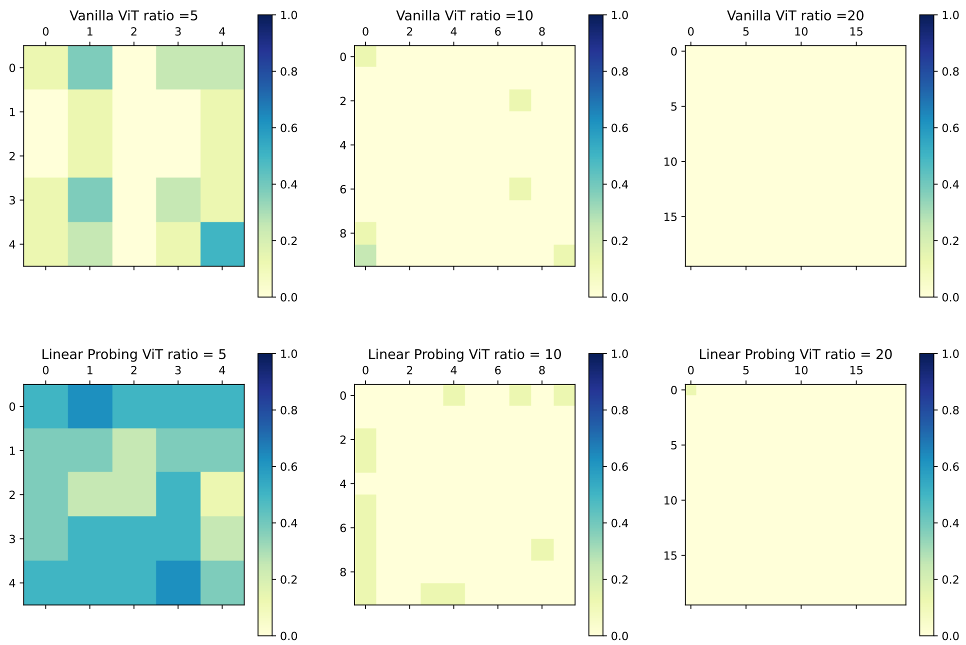

Figure 9: Linear probing results for Vision Encoder showing position bias patterns similar to those observed in LVLMs.

6, creating a progression from simple linear paths to more complex multi-point trajectories.

The spatial distribution of these points is controlled through the same range parameters as the coordinate dataset: $[-5, 5]$, $[-10, 10]$, $[0, 10]$, and $[0, 20]$. For each unique combination of point count and range setting, we generate a substantial set of 100 images, ensuring robust evaluation across different configurations.

We evaluate the performance of path tracing and coordinate recognition through the following metric:

(1) Exact Match Accuracy (EMA) A predicted path is considered correct only if it exactly matches the ground truth answer.

(2) Partial Match Order-Independent Accuracy (PM-IA) In this metric, a point in the predicted path is deemed correct if its coordinates match any point in the ground-truth path. PM-IA is computed as the average accuracy across all positions.

(3) Partial Match Order-Sensitive Accuracy (PM-SA) This metric marks a point in the predicted path as correct only if both its coordinates and position in the path match those of a point in the ground truth path. PM-SA is calculated as the average accuracy across all positions.

Model	EMA	PM-IA	PM-SA	PA
Qwen2-VL-INST	0.060	0.239	0.079	0.487
Molmo-7B-D	0.000	0.160	0.053	0.199
Phi3-Vison-128K-INST	0.062	0.448	0.188	0.707
Phi3.5-Vison-INST	0.005	0.093	0.030	0.123
LLaVA-onevision-Qwen2	0.005	0.074	0.021	0.077
LLaVA-1.6	0.000	0.017	0.005	0.013

Table 2: Path (the first three columns) and coordinate (the last column) task accuracy comparison of 6 selected models across multiple metrics. Metric definitions are in Appendix D.2.

(4) Point Accuracy (PA) This metric is the accuracy of the coordinate recognition.

D.3 Object and Orientation Dataset

The object and orientation dataset is designed to evaluate the model's understanding of object positioning and directionality. We carefully selected ten distinct object categories, including eight animals (shiba dog, cat, bear, eagle, snake, panda, turtle, and fish) and two vehicles (car and plane). This diverse selection allows us to assess the model's performance across varying object morphologies and complexities.

Scale perception is tested through a comprehensive set of object-to-background ratios: 1/2, 1/3, 1/5, 1/10, 1/15, and 1/20. These ratios represent a wide spectrum from prominent objects (1/2) to more subtle presentations (1/20). Furthermore, we evaluate each object against two distinct background types: solid colors for controlled conditions and semantic images for real-world complexity.

The orientation aspect of our experiments encompasses eight distinct directions: the four cardinal directions (up, down, left, right) and their intermediates (top-left, bottom-left, top-right, bottom-right). This comprehensive directional coverage allows us to assess the model's ability to understand and interpret various object orientations.

Through these carefully designed experimental settings, we aim to provide a thorough and systematic evaluation of our model's capabilities across different aspects of visual understanding and spatial reasoning.

D.4 Text Dataset

We create char matrices of size 8*8, 16*16, 24*24, 32*32, 40*40, 64*64. A target word is selected from one of the following, `['dog', 'cat', 'bird', 'lion', 'tiger', 'zebra', 'monkey', 'panda']`, and is positioned within the

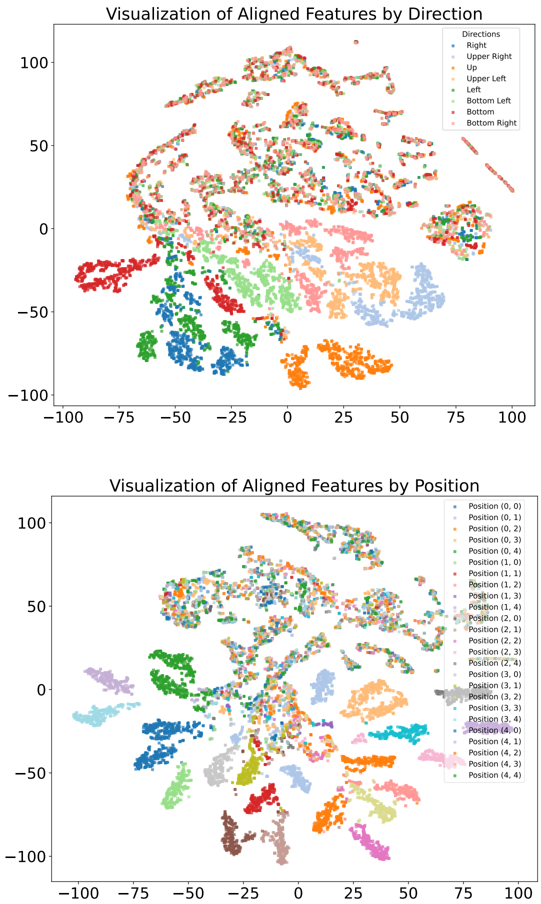

Figure 10: t-SNE visualization of aligned features under directional and positional variations, demonstrating partial sensitivity to visual variations in the feature space.

matrices. Except the object, the rest of the matrices are either asterisks (e.g. corresponding to the w/o BG setting) or random background words. We design 3 tasks to test the robustness of LLM in the cross-component analysis: (1) target word recognition, corresponding to object detection task in image (2) coordinate recognition which corresponds to coordinate recognition and (3) object counting, which is a fundamental skill needed in path tracing. By systematically varying the background contexts while maintaining the target object, we can evaluate whether models truly recognize objects independently or merely rely on contextual associations.

E Experimental Setup

E.1 Evaluated Models

Qwen-VL The Qwen-VL model family represents Alibaba's cutting-edge vision-language model series. The family includes three main

Model	Number		Coordinate		Object	
	w/o BG	w/ BG	w/o BG	w/ BG	w/o BG	w/ BG
Llama-3	0.0688	0.1012	0.0032	0.0024	0.9996	0.8464
Mistral-v0.2	0.2940	0.0776	0.0008	0.000	0.8832	0.8528
Qwen-2	0.3288	0.1364	0.0088	0.0084	0.9952	0.9784

Table 3: Performance of LVLM Language Model Backbones on Text-based Tasks: Comparing scenarios with background tokens represented as asterisks (w/ BG) versus random words (w/o BG).

Model	Zero-shot Classification		Linear Probing	
	Object	Direction	Object	Direction
Original Vision Encoder	0.135	0.151	0.442	0.912
Finetuned Vision Encoder	-	-	0.406	0.826
Multimodal Projector	0.0	0.0	0.032	0.117

Table 4: Accuracy on Object and Direction Tasks: Comparing Original Vision Encoder, Finetuned Vision Encoder (aligned with text features in LLaVA-1.6), and Multimodal Projector.

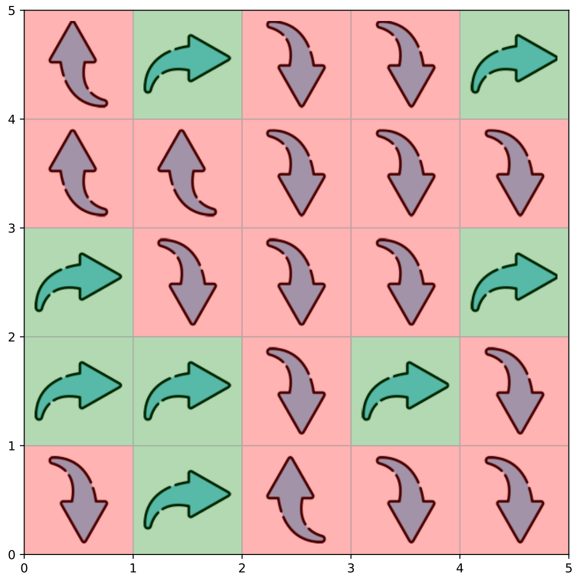

Figure 11: Examples demonstrating position bias effects. Green indicators show correct identifications while red ones represent model predictions, revealing spatial-dependent performance patterns.

variants: the original Qwen-VL (Bai et al., 2023b), which established the foundation for vision-language processing; Qwen2-VL-7B, offering a balanced mid-size option; and Qwen2-VL-72B, the largest and most sophisticated version featuring state-of-the-art visual understanding capabilities and support for videos over 20 minutes long (Wang et al., 2024b). The latest addition, QVQ-72B-Preview, serves as an experimental research model specifically focused on advancing visual reasoning capabilities (Team, 2024).

Molmo Molmo-7B-D (Deitke et al., 2024) is an open-source vision-language model developed by the Allen Institute for AI, built on Qwen2-7B and utilizing OpenAI's CLIP as its vision backbone.

H2OVL The H2OVL-Mississippi (Galib et al., 2024) model family are specifically designed for efficient on-device applications and privacy-focused use cases. The family consists of two specialized models: H2OVL-Mississippi-0.8B, a compact model optimized for text recognition that achieves state-of-the-art performance on OCR-Bench, and H2OVL-Mississippi-2B, a model for general vision-language tasks including image captioning and visual question answering.

Phi-3 Microsoft's Phi-3-vision and Phi-3.5-vision (Abdin et al., 2024) represent a significant advancement in multimodal AI. Phi-3.5-vision, the latest iteration, is a lightweight yet powerful model featuring a 128K token context length and support for both single and multi-image processing.

InternVL The InternVL family includes: Mono-InternVL (Luo et al., 2024a), which established the foundation with its vision-language capabilities; InternVL-2 (Chen et al., 2024c), which expanded the model sizes and improved performance; and InternVL-2.5 (Chen et al., 2025b), the latest iteration that introduces significant architectural and training improvements.

LLaVA LLaVA-1.5 (Liu et al., 2023) uses pretrained CLIP (Radford et al., 2021) and Vicuna language model as the backbone, establishing the foundation with impressive performance across 12 benchmark datasets. LLaVA-OneVision-Qwen2 (Li et al., 2024) pushes performance boundaries across single-image, multi-image, and video scenarios while enabling strong task transfer capabilities. LLaVA-1.6(Liu et al., 2024c), or LLaVA-NeXT, further enhances capabilities with increased input resolution, improved visual reasoning, and enhanced OCR capabilities.

Llama Llama 3.2 Vision models (Dubey et al., 2024) represent Meta's latest advancement in multimodal AI, introducing vision capabilities to the Llama family for the first time. The 11B and 90B parameter versions are specifically designed to handle both text and image inputs, featuring a novel architecture that integrates image encoder representations into the language model.

GLM-4 GLM-4V-9B (GLM et al., 2024) supports high-resolution image processing at 1120*1120 pixels and enables dialogue capabilities in both Chinese and English.

MiniCPM-V MiniCPM-V (Yao et al., 2024) is a series of multimodal large language models (MLLMs) designed specifically for deployment on end-side devices like mobile phones and personal computers.

E.2 Implementation Details

Our experiments are conducted on 8 NVIDIA H100 GPUs. All models maintain their original parameter configurations during inference, with an average processing speed of 525 tokens per second.

To examine spatial reasoning capabilities analogous to those in LVLMs, we design a suite of tasks that evaluate text-based object recognition, counting, and spatial analysis. Using a matrix-structured text input that simulates idealized visual encoding, we assess three fundamental capabilities: (1) target token identification amid background elements, (2) frequency quantification of target tokens, and (3) spatial localization of target tokens within the matrix. The dataset specifications and distributions are detailed in Appendix D.4.

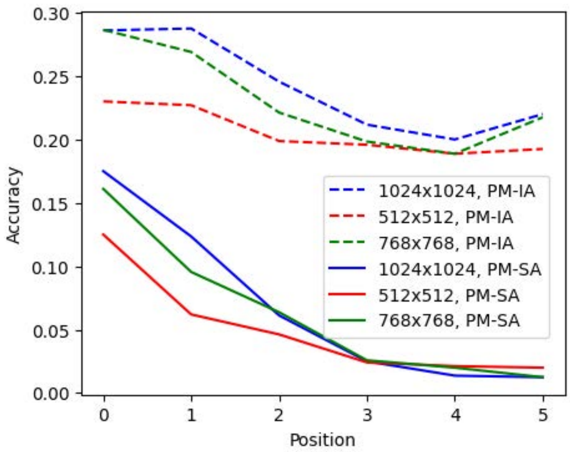

Figure 12: Performance evaluation on a 6-point path tracing task across different positions, where accuracy indicates coordinate prediction precision at each sequential point. Results demonstrate that LVLMs progressively lose accuracy when tracking points along the path.

F Visually Conditioned Soft Prompt

In this section, we explain why aligning visual features to the language model's embedding space through multimodal alignment module can be viewed as a form of visually conditioned soft prompt. Soft prompting prepends learned vectors to the language model's input, optimizing these vectors during training to achieve desired tasks. These vector prefixes learn latent instructions during the tuning process:

$$\mathbf{h}_t = LM([\underbrace{P}_{\text{prompt}}; \mathbf{x}]_{<t}),$$
$$\text{where } P \in \mathbb{R}^{l \times d}, \quad \mathcal{L} = \sum_t -\log p(x_t|\mathbf{h}_t) \tag{9}$$

where P represents the learned soft prompt vectors and $\mathbf{x}$ is the input token embeddings. In LVLM's multimodal alignment process, extracted visual features are mapped to the language model's embedding space through a neural network based multimodal alignment module. The module is trained to transform visual features into embedding space that serve as special task prefixes, guiding model responses to image-related queries::

$$\mathbf{h}_t = LM([\underbrace{f_\theta(E_v(I))}_{\text{visual prompt}}; \mathbf{x}]_{<t}),$$
$$\text{where } \mathcal{L} = \sum_t -\log p(x_t|\mathbf{h}_t, I) \tag{10}$$

where f_θ is the alignment network and $\mathbf{v}$ represents the visual features. Fundamentally, both approaches optimize vector prefixes independent of the discrete language embedding matrix. The key difference lies in their optimization objectives: soft prompts aim to generate task-relevant responses based on the dataset, learning fixed prefixes that remain constant after training. In contrast, while aligned visual features are also trained on the entire dataset, they produce dynamic, image-specific prefixes during inference. Therefore, aligned visual features can be considered a special case of soft prompts - visually conditioned soft prompts - sharing the characteristic of residing in a representation space distinct from the language embedding space.

G Examples.

- **ChatGPT:** Sorry, I don't see the arrow in the image.

- **Claude:** A sleek fish glides through the harbor waters, its scales catching glints of sunlight filtering through the surface. The backdrop features wooden pier posts rising from the water and moored fishing boats gently bobbing in the distance. The water has a deep blue-green tint typical of a working harbor, with small ripples creating patterns of light and shadow. The fish appears to be navigating between the shadows cast by the docks, perhaps searching for food or shelter among the port's underwater structures.

- **Gemini: Reasons:**

  - **Image Information:** The image only shows an animal on the water's surface, without indicating the specific direction of the animal's head, body, or tail.
  - **Lack of Reference Points:** There are no other objects or background elements to help determine the animal's orientation.

  Therefore, it is impossible to determine the animal's orientation from the information provided in the image.

  If you can provide more information about the image, such as the shooting angle or the animal's posture, it may be possible to determine its orientation more accurately.

- **InternVL:** This image shows a pure white background without any recognizable objects or text. It is a simple plane with no complex designs or patterns.

An online anonymous leaderboard is available at https://anonymous.4open.science/r/Visual-Variations-Robustness-EFC3/README.md

Model	Position		Orientation		Scale		Context	
	Object	Direction	Object	Direction	Object	Direction	Object	Direction
Qwen-VL	0.037	0.285	0.023	0.281	0.153	0.477	0.051	0.477
Qwen2-VL-7B	0.059	0.402	0.067	0.427	0.175	0.509	0.067	0.427
Molmo-7B-D	0.260	0.626	0.262	0.646	0.278	0.632	0.262	0.646
h2ovl-2b	0.055	0.398	0.063	0.423	0.172	0.506	0.063	0.423
h2ovl-800m	0.013	0.201	0.011	0.199	0.179	0.332	0.173	0.308
Phi-3-vision	0.015	0.367	0.018	0.379	0.064	0.375	0.018	0.379
Phi-3.5-vision	0.062	0.408	0.070	0.417	0.079	0.430	0.067	0.417
InternVL-Mono	0.057	0.400	0.065	0.425	0.177	0.510	0.065	0.425
InternVL-2	0.061	0.405	0.069	0.430	0.179	0.513	0.069	0.430
InternVL-2.5	0.058	0.403	0.066	0.428	0.176	0.511	0.066	0.418
llava-1.5	0.056	0.399	0.064	0.424	0.174	0.508	0.064	0.424
llava-onevision-qwen2	0.160	0.446	0.173	0.459	0.245	0.474	0.173	0.459
llava-1.6-mistral	0.017	0.190	0.019	0.194	0.044	0.216	0.019	0.194
llava-1.6-vicuna	0.018	0.192	0.020	0.196	0.045	0.218	0.020	0.196
Llama-11B-V	0.057	0.400	0.065	0.425	0.177	0.510	0.065	0.425
Llama-90B-V	0.063	0.412	0.071	0.437	0.182	0.517	0.071	0.437
glm-4v-9b	0.141	0.279	0.093	0.117	0.381	0.257	0.012	0.233
MiniCPM-V-2	0.057	0.392	0.013	0.393	0.156	0.419	0.063	0.428
MiniCPM-V-2.5	0.092	0.448	0.090	0.483	0.210	0.555	0.090	0.463
MiniCPM-V-2.6	0.059	0.405	0.067	0.430	0.178	0.513	0.067	0.430
GPT-4o	0.315	0.782	0.298	0.772	0.319	0.695	0.267	0.799

Table 5: Bias Type and Tasks for Different Models. We present the mean accuracy across variations here as a reference.

Model	Position		Orientation		Scale		Context	
	Object	Direction	Object	Direction	Object	Direction	Object	Direction
Qwen-VL	0.913	0.897	0.924	0.773	0.891	0.899	0.832	0.879
Qwen2-VL-7B	0.945	0.910	0.955	0.664	0.862	0.923	0.937	0.890
Molmo-7B-D	0.911	0.905	0.916	0.739	0.879	0.955	0.882	0.862
h2ovl-2b	0.833	0.910	0.977	1	0.926	1	1	0.979
h2ovl-800m	0.793	0.890	0.921	0.745	0.832	0.926	0.895	0.901
Phi-3-vision	0.968	0.929	1	0.622	0.937	1	1	0.929
Phi-3.5-vision	0.955	0.923	0.945	0.635	1	1	0.937	0.923
InternVL-Mono	0.915	0.905	0.935	0.745	0.915	0.955	0.925	0.915
InternVL-2	0.945	0.925	0.965	0.785	0.945	0.975	0.955	0.945
InternVL-2.5	0.965	0.935	0.985	0.815	0.965	0.985	0.975	0.965
llava-1.5	0.955	0.925	0.975	0.765	0.955	0.985	0.965	0.955
llava-onevision-qwen2	0.929	0.911	0.895	0.712	0.929	1	0.830	0.905
llava-1.6-mistral	0.968	0.937	1	0.788	0.968	0.968	1	0.968
llava-1.6-vicuna	0.965	0.935	0.985	0.775	0.965	0.975	0.985	0.965
Llama-11B-V	0.925	0.913	0.945	0.752	0.925	0.965	0.930	0.925
Llama-90B-V	0.975	0.945	0.986	0.825	0.975	0.984	0.985	0.975
glm-4v-9b glm-4v-9b	0.953	0.925	0.971	0.796	0.959	0.970	0.961	0.955
MiniCPM-V-2	0.932	0.912	0.952	0.762	0.937	0.963	0.948	0.931
MiniCPM-V-2.5	0.953	0.922	0.973	0.782	0.957	0.972	0.968	0.948
MiniCPM-V-2.6	0.962	0.933	0.982	0.793	0.968	0.983	0.977	0.959
GPT-4o	0.983	0.962	1	0.853	0.987	1	1	0.982

Table 6: Comprehensive Evaluation of Model Robustness Across Different Metrics and Test Scenarios.