Towards High Robust Vision-Language Large Models: Benchmark and Method

Supplementary Material

MINYI ZHAO, School of Computer Science
Fudan University, China
YI LIU, ByteDance, China
WENSONG HE, ByteDance, China
BINGZHE YU, School of Computer Science
Fudan University, China
YUXI MI, School of Computer Science
Fudan University, China
SHUIGENG ZHOU*, School of Computer Science
Fudan University, China

Recently, numerous benchmarks have been constructed to evaluate various general capabilities (e.g., perception and reasoning) of Vision-Language Large Models (VLLMs). However, few studies have focused on the robustness of VLLMs when dealing with altered prompts and images. To fill this gap, this paper first constructs a real-world, high-quality, and challenging benchmark, namely **RBench** (i.e., **R**obust **Bench**). Specifically, RBench is human-annotated, with both prompts and images being modified to enrich the difficulty, and cross-validation to ensure data quality. Then, we propose a new method, called **R**obustness **Boost**er (**RBoost** in short), to effectively enhance the robustness of existing VLLMs by automatically generating high-value instruction-tuning training data. Extensive experiments demonstrate the vulnerability of existing VLLMs when handling altered inputs, and the superiority of our RBoost method in improving model robustness. RBench is available at https://github.com/zhaominyiz/RBench.

CCS Concepts: • **Computing methodologies → Knowledge representation and reasoning**.

Additional Key Words and Phrases: Vision-language large model; Robustness; Benchmark; Prompt-tuning.

ACM Reference Format:
Minyi Zhao, Yi Liu, Wensong He, Bingzhe Yu, Yuxi Mi, and Shuigeng Zhou. 2025. Towards High Robust Vision-Language Large Models: Benchmark and Method Supplementary Material. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27-October 31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/xxxxxx.xxxxxx

1 MORE RBENCH CASES

Here, we provide more cases to better demonstrate the diversity and challenge of our collected RBench.

[*] Corresponding author.

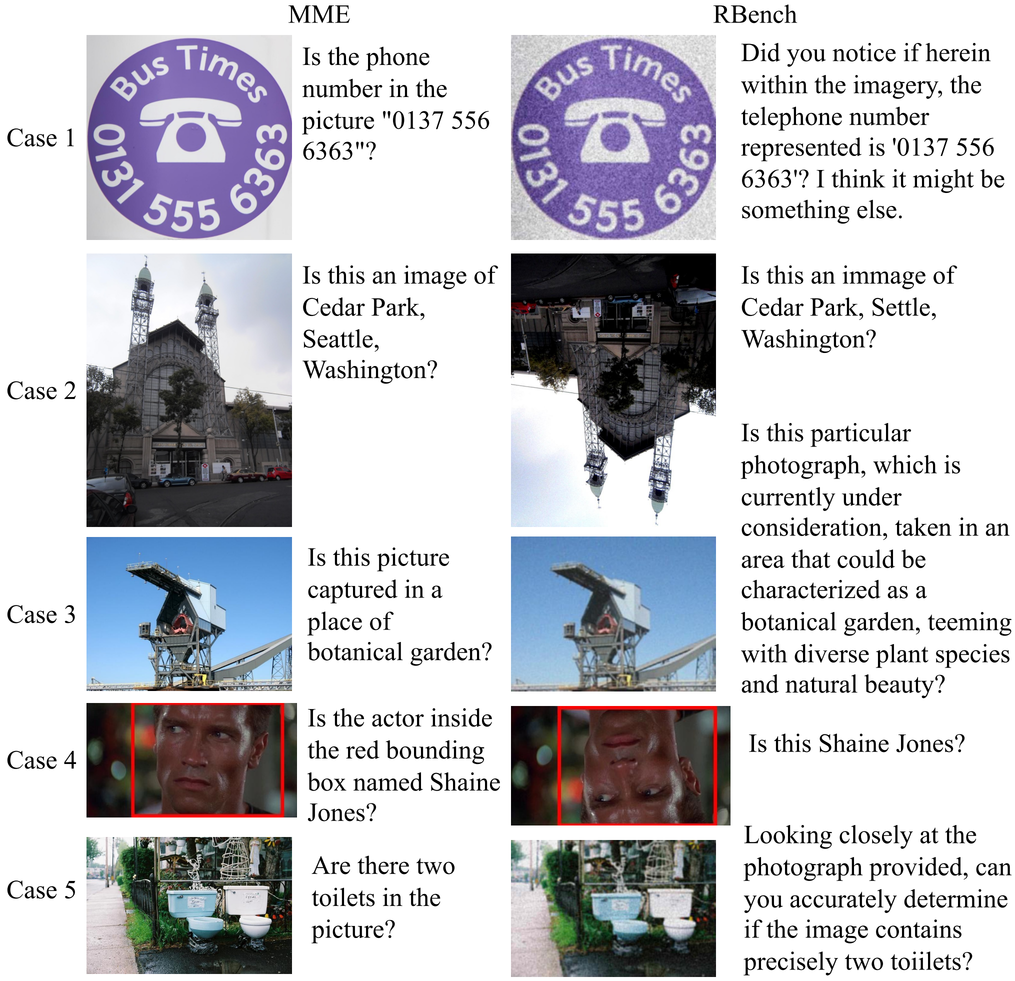
MME
RBench
Case 1
Is the phone number in the picture "0137 556 6363"?
Did you notice if herein within the imagery, the telephone number represented is '0137 556 6363'? I think it might be something else.
Case 2
Is this an image of Cedar Park, Seattle, Washington?
Is this an immage of Cedar Park, Settle, Washington?
Case 3
Is this picture captured in a place of botanical garden?
Is this particular photograph, which is currently under consideration, taken in an area that could be characterized as a botanical garden, teeming with diverse plant species and natural beauty?
Case 4
Is the actor inside the red bounding box named Shaine Jones?
Is this Shaine Jones?
Case 5
Are there two toilets in the picture?
Looking closely at the photograph provided, can you accurately determine if the image contains precisely two toiilets?

Fig. 1. Five representative cases in the RBench dataset.

In Fig. 1, we additionally present five representative cases from RBench. As illustrated, in terms of text modifications, RBench have introduced (not limited to) the following policies: **Misleading prompts** (Cases 1): The prompts introduce ambiguity (*e.g.,* "I think it might be something else" in Case 1), testing the model's resistance to semantic distraction. **Spelling errors** (Cases 2 and 5): RBench includes typos like "immage" (Case 2, misspelled "image") and verbose phrasing (Case 5), simulating real-world input noise. **Prompt length variation** (Case 3: longer, Case 4: shorter): Case 3 extends the prompt to describe a "botanical garden, teeming with diverse plant species," while Case 4 shortens "Is the actor inside the red bounding box named Shaine Jones?" to "Is this Shaine Jones?" – challenging the model's adaptability to input length.

For image alterations: **Blur** (Case 1 RBench image, reducing text clarity), **rotation/flipping** (Case 2 RBench image is upside-down, Case 4 image is vertically flipped), and **content distortion** (e.g., Case 1's blurred phone number, Case 4's flipped face) simulate real-world degradations. These changes force models to handle distorted visual features alongside modified prompts.

Collectively, these cases demonstrate RBench's multidimensional challenge: it not only modifies text (semantic, syntactic, length) and images (quality, orientation, content) but also combines both modalities' perturbations (e.g., Case 2's flipped image + misspelled "immage"). This design ensures RBench evaluates models' robustness under realistic, compounded input variations, unlike MME (which uses clean, concise inputs), making it a rigorous testbed for Vision-Language Large Models.

2 LICENSE AND ACCESSABILITY

RBench is based on the Apache-2.0 license and is open to the community. Researchers can directly access it through the URL in the abstract.