Technical Summary of MM-R^3: On (In-)Consistency of Vision–Language Models

Part I: Problem Formulation, Methods, and Evidence

Kai-Yu Lu

2025/09/10

Bibliographic Information

Title: MM-R3: On (In-)Consistency of Vision-Language Models (VLMs). **Venue and Year:** Findings of ACL 2025. **Authors:** Shih-Han Chou, Shivam Chandhok, James J. Little, Leonid Sigal.

1 Research Problem and Motivation

The study investigates *consistency* in Vision-Language Models (VLMs), defined as the ability to generate semantically equivalent responses when inputs undergo semantically preserved perturbations in language or vision. Prior benchmarking chiefly emphasizes accuracy on tasks such as visual question answering, captioning, and grounding, which leaves an evaluation gap: models can be accurate yet inconsistent across rephrasings or visual restyling. Consistency is a prerequisite for reliability and robustness in deployment settings, motivating a dedicated benchmark and measurement protocol.

2 Related Work

General-purpose multimodal benchmarks (for example, MM-Bench, SEED-Bench, MM-Vet) primarily assess correctness. In the monomodal domain, recent research analyzes LLM consistency and explores lightweight fine-tuning or prompt-adapter methods to stabilize outputs. Building on these lines, the present work extends consistency evaluation to multimodal settings and explores a simple trainable adapter inserted into existing VLMs to reduce inconsistency without overhauling encoders or decoders.

3 Dataset Construction

3.1 Tasks and Sources

The MM-R3 benchmark comprises three tasks designed to preserve semantic intent while perturbing inputs:

- **Question Rephrasing:** Language-only perturbation. Original VQA-style questions are rephrased into multiple semantically equivalent variants. Images and questions come from *InfographicsVQA* and *OKVQA*.
- **Image Restyling:** Vision-only perturbation. Original images are transformed into stylistic variants using four styles (Candy, Mosaic, Udnie, Grayscale). Images are sourced from *Google Landmarks v2* and *Indoor Scene*.

- **Context Reasoning (Masking):** Vision perturbation via partial occlusion. Objects in MSCOCO images are masked with varying types (lines, shapes, colors). The query is standardized as "What kind of object is in the masked region?"

3.2 Scale and Splits

Task	Train #	Test #	Source Datasets
Question Rephrasing	16,894	3,516	InfographicsVQA, OKVQA
Image Restyling	27,226	5,328	Google Landmarks v2, Indoor Scene
Context Reasoning	30,003	4,500	MSCOCO

3.3 Perturbation Protocols

For masking, the masked object area to image area ratio is constrained to $[0.1, 0.25]$. Masks vary by: number of lines $\{1, 3, 5, 7\}$, shape $\{$rectangle, ellipse$\}$, and color $\{$red, blue, green, yellow, white, black$\}$. For image restyling, four fixed styles are applied. For question rephrasing, long questions are rephrased into multiple variants while preserving semantics.

3.4 Quality Control

A human validation on 300 rephrasing pairs and 300 styled images yields 93% and 86% semantic equivalence respectively. Automated validation with a strong external VLM on all rephrasings yields 88% equivalence, providing a conservative lower bound on data quality.

4 Query Protocol and Task Definitions

4.1 Evaluation Settings

Each example is queried multiple times. Two regimes are reported:

- **Sampling:** For each original input, the VLM is queried 4 times with identical conditions to quantify intrinsic stochasticity.
- **All:** Aggregated performance across all perturbed variants for each task.

4.2 Metrics

Let Acc denote accuracy against ground truth via case-insensitive substring matching. Let SGT denote sentence-level semantic similarity with ground truth; it is computed by encoding response and target with a sentence-embedding model and taking the cosine similarity. Let Con (Consistency Accuracy) denote the proportion of response pairs whose pairwise semantic similarity exceeds a fixed threshold. Let SC (Consistency Similarity) denote the average pairwise semantic similarity between responses across variants without thresholding.

4.3 Overall Score

To balance correctness and consistency, an overall score O_{all} uses the harmonic mean of the mean correctness and the mean consistency:

$$\mathrm{Hmean}\big(\mathrm{mean}(\mathrm{Acc}, \mathrm{SGT}),\ \mathrm{mean}(\mathrm{Con}, \mathrm{SC})\big). \tag{1}$$

Symbol meanings and background: Acc is exact-match style correctness using substring matching; SGT is sentence-level semantic similarity against ground truth using pretrained text encoders; Con thresholds pairwise response similarity at 0.7 to count consistent pairs; SC averages pairwise similarities continuously. The harmonic mean is used to mitigate domination when one component is much lower than the other. **Intuition:** a reliable model should be both correct and stable; the harmonic mean penalizes imbalance. **Role in this study:** O_{all} summarizes trade-offs and enables a single-score comparison across models and tasks.

5 Modeling Approach

5.1 Baselines and Hardware

Six open-source VLMs are evaluated: BLIP-2, mPLUG-Owl2, LLaVA 1.5M, MoE-LLaVA, Qwen-VL-Chat, BLIP-3. Three closed-source VLMs are included: Gemini, GPT-4V, GPT-4o. All experiments run on NVIDIA A40 GPUs.

5.2 Adapter for Consistency Improvement

An insertion-only adapter is proposed between the vision-language encoder and the language decoder of a VLM. During fine-tuning, all encoders and the decoder are frozen. The adapter receives encoder outputs, processes them with a bidirectional LSTM followed by max-pooling to obtain a global embedding, and then uses a multi-layer perceptron to project to a prefix-sized vector that is concatenated ahead of the original embeddings for decoding. Only adapter parameters are optimized, which keeps training light-weight and architecture-agnostic.

5.3 Training Data for the Adapter

Adapter training leverages the same data generation pipelines as MM-R3 but uses disjoint samples: 16,894 rephrasing pairs, 27,226 styled images, and 30,003 masked images.

5.4 Hyperparameters

Learning-rate, batch size, and epochs are not specified in the main text. The paper defers further implementation details to the appendix.

6 Empirical Results

6.1 Headline Findings

Across tasks, accuracy is relatively similar across models, while consistency varies substantially. Consistency drops are more pronounced under visual perturbations (restyling and masking) than under linguistic rephrasings. In open-source models, Qwen-VL-Chat leads on Rephrasing accuracy but lags on visual perturbations; BLIP-2 is strongest on Context Reasoning consistency; LLaVA 1.5M and MoE-LLaVA are balanced. Among closed-source models, GPT-4o is strongest overall, but even top models exhibit sizable consistency gaps under perturbations.

6.2 Representative Numbers

Human Upper Bound (Context Reasoning). On 100 examples (300 images), humans achieve Acc/SGT/Con/SC of 66.0/82.0/95.0/97.4, indicating that the task permits consistent human judgments.

Adapter Gains (Table 3). On LLaVA 1.5M for Question Rephrasing, Acc improves from 26.9 to 28.7, while Con and SC improve from 32.5 to 42.5 and from 53.8 to 61.1; O_{all} rises from 43.1 to 49.0. On BLIP-2 for Image Restyling, Acc/SGT/Con/SC improve from 13.0/17.0/38.4/62.8 to 27.0/27.6/49.1/66.7, raising O_{all} from 23.1 to 37.1. On Context Reasoning, LLaVA improves from Acc/SGT/Con/SC of 20.1/28.9/25.9/42.3 to 55.3/71.9/62.2/74.8, increasing O_{all} from 28.5 to 66.0; BLIP-2 improves from 27.9/39.0/82.4/88.8 to 53.5/51.8/88.5/94.0, raising O_{all} from 48.1 to 66.8.

Task-wise Behavior. Rephrasing reveals that models with temperature set to 0 can appear perfectly consistent under repeated sampling yet degrade notably under true rephrasings, reflecting prompt sensitivity. Image Restyling shows stronger dependence on input resolution and style choice; Grayscale perturbs least and yields best consistency among styles, while Mosaic is most challenging. In masking-based Context Reasoning, performance depends more on masked area than mask color; ellipse often yields fewer artifacts than rectangle.

6.3 Innovation vs. Prior Art

The benchmark isolates consistency as a first-class target under both language and vision perturbations and defines a four-metric battery plus the harmonic-mean overall score. The adapter is architectural-minimal, frozen-encoder-friendly, and demonstrably raises consistency across tasks without materially degrading OKVQA accuracy on the base model.

7 Summary

Contributions. A three-task benchmark quantifying VLM consistency under semantically preserving perturbations; four metrics including two for correctness and two for consistency; a harmonic-mean overall score; and a light-weight adapter that consistently improves stability.
Limitations and Future Work. Consistency is operationalized as average behavior under controlled perturbations; alternative definitions may be explored. Metric design currently relies on sentence-level embedding similarity with a fixed threshold; future work may adopt stronger LLM-based judgments and expand scenarios.

Technical Appendices

Metrics Recap

- **Acc** (Accuracy): case-insensitive substring match against ground truth; averaged over the dataset.
- **SGT** (Similarity with Ground Truth): cosine similarity between sentence embeddings of response and ground truth; averaged.
- **Con** (Consistency Accuracy): proportion of response pairs with similarity at least 0.7; averaged.
- **SC** (Consistency Similarity): mean pairwise response similarity without thresholding; averaged.

Experimental Environment

All evaluations are conducted on NVIDIA A40 GPUs. “Sampling” queries each original input four times with identical settings; “All” aggregates over all perturbations per task.