Analyzing Instruction Sensitivity in Vision–Language Representations

Hussein Aveen
ITMO University
Saint Petersburg, Russia
0009-0009-2710-572X

***Abstract*—Vision–language models often produce varied internal representations for instructions with the same semantic meaning when conditioned on the same image. We demonstrate this issue through a controlled experiment on the SmolVLM model, showing measurable divergence between paraphrased instructions. We then explore the hypothesis that instruction consistency can be improved using a contrastive instruction alignment approach with a trainable projection head, and demonstrate on a small scale the potential of this methodology.**

***Index Terms*—component, formatting, style, styling, insert**

I. Literature Review

Fu et al. [1] investigate the extent to which vision–language models (VLMs) effectively utilize their visual encoders when solving vision-centric tasks. By comparing the performance of full VLMs against direct readouts of their underlying vision encoders (e.g., DINO, CLIP), the authors show a dramatic performance drop when tasks are framed through language prompts, often falling from near-ceiling accuracy to near-random chance. Through detailed analysis, they attribute this failure to three factors: degradation of visual representations, brittleness to task prompts, and—most critically—the dominant influence of the language model, which introduces strong linguistic priors that overshadow available visual information. This work highlights a fundamental sensitivity of VLMs to prompt formulation and suggests that linguistic variations can substantially affect model behavior even when visual input remains fixed. However, the analysis primarily focuses on downstream task performance rather than the structure of internal instruction representations.

Dumpala et al. [2] extend this line of investigation by specifically analyzing the sensitivity of generative VLMs to lexical and semantic variations in prompts. Using the SUGARCREPE++ benchmark, they evaluate models such as BLIP, BakLLaVA, and GPT-4o under paraphrased prompts that preserve semantic meaning but alter surface form. Their results show that these generative VLMs are highly sensitive to such variations, leading to significant drops in accuracy and consistency. Moreover, this prompt sensitivity undermines methods designed to improve output consistency, highlighting a fundamental brittleness in current generative VLMs that can compromise downstream applications.

Khosla et al. [3] introduce Supervised Contrastive Learning (SupCon), which extends self-supervised contrastive methods to the fully supervised setting. Unlike standard contrastive losses that only consider a single positive per anchor, SupCon leverages label information to pull all samples of the same class together while simultaneously pushing apart samples from other classes. The authors demonstrate that this approach improves both accuracy and robustness over traditional cross-entropy loss on large-scale image classification tasks. Conceptually, SupCon illustrates that explicitly structuring the embedding space based on known relationships can produce more semantically coherent and consistent representations. This insight is particularly relevant for addressing instruction-level instability in VLMs, where aligning semantically equivalent prompts could improve representation consistency.

Jamal and Mohareri [4] propose a two-stage progressive pre-training strategy for RGB-D datasets using Multi-Modal Contrastive Masked Autoencoders (MM-CMAE). The first stage uses patch-level contrastive learning to align RGB and depth modalities, capturing local cross-modal relationships. The second stage combines masked autoencoding with a denoising objective, encouraging the model to learn both low- and high-frequency features, while also distilling information from the first stage to enhance representation quality. Their method achieves strong performance on semantic segmentation and depth estimation tasks across ScanNet, NYUv2, and SUN RGB-D datasets, including in low-data regimes. This work exemplifies how combining contrastive and reconstruction-based pre-training strategies can yield rich, multi-modal representations, suggesting a potential pathway for improving alignment and stability in instruction-conditioned embeddings.

II. Motivation

Vision–Language Models (VLMs) are increasingly deployed as general-purpose interfaces for perception, reasoning, and control in embodied and robotic systems. In these contexts, natural language serves as a high-level task specification, and it is generally expected that semantically equivalent instructions produce similar internal representations and behaviors. However, recent studies indicate that VLMs are highly sensitive to task prompts: minor lexical or semantic variations can substantially affect outputs, and language priors often dominate over available visual information. Such sensitivity poses risks for downstream applications, particularly in vision–language–action systems, where inconsistent internal representations can lead to unstable or unsafe behavior.

Despite these observations, most evaluations have focused on task-level performance, leaving a gap in understanding the consistency of instruction-conditioned embeddings themselves. In other words, it remains unclear whether semantically equivalent instructions mapped onto the same visual input produce aligned representations within the model. Addressing this question is critical for building robust multimodal systems that can operate reliably under natural language variation, motivating the exploration of methods that explicitly structure and stabilize instruction-level embedding spaces.

III. Problem Statement

Despite their success across a wide range of vision–language tasks, VLMs exhibit significant variability in their internal representations when processing semantically equivalent instructions. This instruction-level inconsistency can undermine the reliability and safety of downstream applications, particularly in settings where precise visual reasoning and consistent action are critical.

In this work, we focus on assessing and improving instruction-conditioned embedding stability in VLMs. Specifically, we investigate the SMOLVLM model, a compact open-source VLM, as a representative testbed for studying this phenomenon. Our goal is to determine whether applying a Contrastive Instruction Alignment Adapter (CIAA) framework, combined with a supervised contrastive loss, can reduce the embedding variance induced by semantically equivalent prompts while preserving the semantic fidelity of the representations.

Formally, given a VLM f and a visual input x, we aim to evaluate whether the embeddings $f(x, i_1), f(x, i_2), \dots, f(x, i_n)$ corresponding to a set of semantically equivalent instructions $\{i_1, i_2, \dots, i_n\}$ are more tightly clustered in the latent space after applying CIAA, compared to the baseline embeddings produced by the unmodified model. The underlying hypothesis is that structured contrastive learning in the CIAA latent space can explicitly align semantically equivalent instructions, reducing representation variance and improving downstream stability.

IV. Metrics to Evaluate the Work

To quantify the effectiveness of the Contrastive Instruction Alignment Adapter (CIAA) in aligning semantically equivalent instructions, we adopt a set of embedding-level similarity metrics. These metrics directly capture how closely instructions with the same meaning cluster in the latent space, and how well they are separated from semantically different instructions.

A. Similarity Matrices

For each model (baseline SMOLVLM and CIAA-enhanced SMOLVLM), we compute pairwise cosine similarities between all instruction embeddings for a given image. This results in a similarity matrix $S \in \mathbb{R}^{n \times n}$, where $S_{ij} = \cos(\mathbf{z}_i, \mathbf{z}_j)$ and $\mathbf{z}_i$ denotes the embedding of the i-th instruction. Visual inspection of these matrices provides qualitative insight into how semantically equivalent instructions cluster together.

B. Within- and Between-Group Similarities

Instructions are grouped according to semantic equivalence. We compute:

- **Within-group similarity:** the average cosine similarity between embeddings of instructions belonging to the same group.
- **Between-group similarity:** the average cosine similarity between embeddings of instructions from different groups.

High within-group similarity indicates that semantically equivalent instructions are well-aligned in the embedding space, while low between-group similarity indicates good separation between semantically different instructions.

C. Summary Ratio

To consolidate these metrics, we define the *within/between ratio* as:

$$\text{Ratio} = \frac{\text{Avg within-group similarity}}{\text{Avg between-group similarity}}.$$

A higher ratio corresponds to tighter clustering of equivalent instructions relative to separation from other instructions, providing a single scalar measure of alignment quality.

All reported results in this work—including the similarity matrices, within- and between-group averages, and the ratio—are computed using the above definitions.

V. Experiments

A. Initial Model Evaluation on SMOLVLM

As a first step, we evaluated the SMOLVLM-256M-Instruct model on individual images with natural language prompts to establish a baseline understanding of its behavior. Using the HuggingFace Transformers library, we loaded both the processor and model, and conditioned the model on a set of user instructions paired with an input image.

For example, when prompted with “Can you describe this image?” alongside a photograph of the Statue of Liberty, the model generated a detailed textual description covering the statue, pedestal, surrounding cityscape, and environmental context. This experiment verified that the model was able to process images and generate coherent responses in natural language.

This step serves as a qualitative baseline, illustrating the model’s current capability and consistency when responding to semantically simple instructions. It also establishes a foundation for subsequent embedding-level analyses and alignment experiments with CIAA.

B. Instruction Embedding Stability Analysis

To quantitatively assess instruction sensitivity, we conducted an experiment to measure how semantically equivalent prompts generate different embeddings in SMOLVLM-256M. Specifically, we used a single image and seven variations of instructions asking about the same visual content, e.g., "What is in this picture?" and "Describe what you see."

We extracted text embeddings from the final hidden layer of the model corresponding to the instruction tokens. Pairwise cosine similarities were computed between these embeddings to measure alignment between responses. The resulting similarity matrix for our experiment is shown in Table I.

TABLE I
PAIRWISE COSINE SIMILARITY BETWEEN INSTRUCTION EMBEDDINGS FOR A SINGLE IMAGE. HIGHER VALUES INDICATE MORE SIMILAR EMBEDDINGS.

Instruction	Instr 1	Instr 2	Instr 3	Instr 4	Instr 5	Instr 6	Instr 7
Instr 1	1.000	0.847	0.836	0.863	0.862	0.906	0.766
Instr 2	0.847	1.000	0.824	0.880	0.903	0.846	0.772
Instr 3	0.836	0.824	1.000	0.818	0.867	0.848	0.842
Instr 4	0.863	0.880	0.818	1.000	0.860	0.851	0.702
Instr 5	0.862	0.903	0.867	0.860	1.000	0.863	0.827
Instr 6	0.906	0.846	0.848	0.851	0.863	1.000	0.789
Instr 7	0.766	0.772	0.842	0.702	0.827	0.789	1.000

The average similarity between different instructions was 0.837, substantially below the expected range of 0.9–1.0 for semantically equivalent queries. This demonstrates that SMOLVLM embeddings are sensitive to minor lexical variations in instructions, confirming the presence of instruction-level instability.

These results provide a quantitative foundation for evaluating approaches such as CIAA to improve instruction consistency. By confirming that embeddings diverge under semantically equivalent prompts, we establish a baseline for subsequent experiments designed to align instruction representations more closely.

C. CIAA Evaluation: Improving Instruction Consistency

To evaluate the effectiveness of our proposed CIAA approach, we conducted experiments on a curated set of instructions divided into semantically similar and dissimilar groups:

- **Group 1:** Semantically similar instructions about the image content.
- **Group 2:** Semantically similar instructions about location information.
- **Group 3:** Dissimilar instructions about metadata (year, photographer, camera).

We computed embeddings for each instruction under two settings: a baseline SMOLVLM-256M model without CIAA, and with CIAA applied to the text encoder. The image encoder was frozen during these experiments to isolate the effect of text-side alignment after initial exploratory experiments showed that jointly updating the image encoder offered minimal additional benefit while increasing instability.

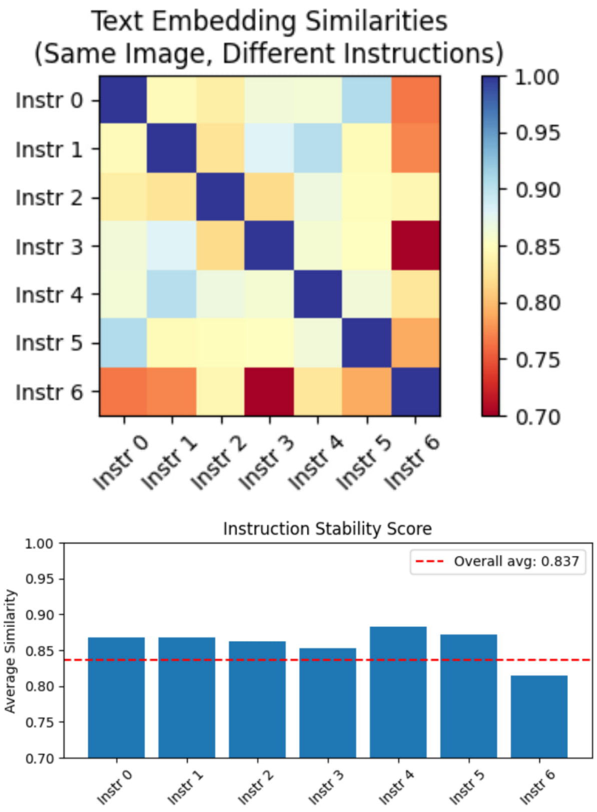

Fig. 1. Left: Heatmap of pairwise cosine similarities between instruction embeddings. Right: Average similarity per instruction with overall mean indicated by a dashed line.

1) Pairwise Similarities: Pairwise cosine similarities were calculated for both baseline and CIAA embeddings. Table II reports within-group and between-group similarities.

TABLE II
WITHIN- AND BETWEEN-GROUP COSINE SIMILARITIES FOR BASELINE AND CIAA EMBEDDINGS.

	Baseline	**CIAA**
Group 1	within: 0.8561, between: 0.8152	within: 0.8839, between: 0.5579
Group 2	within: 0.8614, between: 0.7534	within: 0.9041, between: 0.7545
Group 3	within: 0.7656, between: –	within: 0.7550, between: –

2) Summary Metrics: Overall within-group and between-group similarities were computed to quantify alignment improvements:

TABLE III
SUMMARY METRICS FOR BASELINE AND CIAA EMBEDDINGS.

	Avg within-group	Avg between-group	Ratio (within/between)
Baseline	0.8348	0.7530	1.1086
CIAA	0.8568	0.6142	1.3950
Improvement	+0.0219	-0.1389	+0.2864

These results indicate that CIAA improves semantic grouping by increasing within-group similarity while reducing between-group similarity, yielding a 28.6% improvement in the within/between ratio.

3) Visualization: To qualitatively analyze embedding distributions, we generated similarity heatmaps and UMAP projections. Figures 2 show pairwise similarity matrices for baseline and CIAA embeddings, while Figures 3 illustrate 2D UMAP projections with group labels.

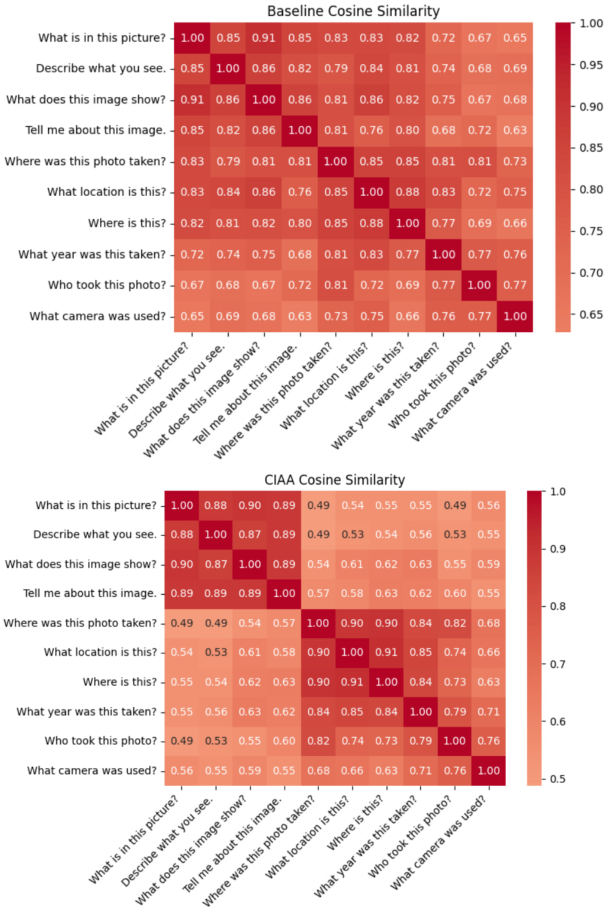

Fig. 2. Pairwise cosine similarity heatmaps. Left: Baseline embeddings. Right: CIAA embeddings.

These visualizations confirm the quantitative findings: CIAA increases cohesion within semantically similar instruction groups while separating dissimilar groups more distinctly. Freezing the image encoder focused training on the text embeddings, which sufficed to achieve stable alignment across prompt variations without destabilizing the visual representation space.

VI. Conclusion

In this work, we identified and quantified a previously underexplored issue in vision–language models: instruction-level embedding instability. Through a targeted experiment on the SMOLVLM-256M model, we showed that semantically equivalent prompts conditioned on the same image can produce measurably different embeddings, which may undermine consistency in downstream tasks. As a proof of concept, we introduced a Contrastive Instruction Alignment Adapter (CIAA) trained with a supervised contrastive loss to align semantically equivalent instruction embeddings. Our evaluation demonstrates that CIAA improves within-group cohesion and increases the ratio of within- to between-group similarity, indicating better semantic alignment.

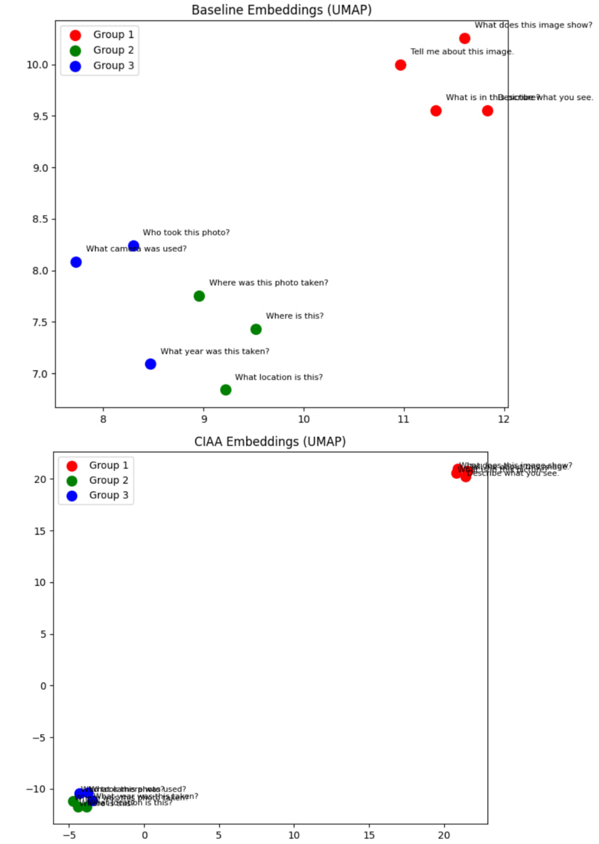

Fig. 3. UMAP projections of instruction embeddings. Left: Baseline. Right: CIAA. Colors indicate instruction groups.

This preliminary study serves as a foundation for future work. Larger instruction sets, additional images, and experiments on larger vision–language models could further validate and extend these findings. The full code for this project is available at https://github.com/aveen007/SMOLEXPT, providing a reproducible starting point for continued investigation into instruction invariance in VLMs.

References

[1] S. Fu, T. Bonnen, D. Guillory, and T. Darrell, “Hidden in Plain Sight: VLMs Overlook Their Visual Representations,” *arXiv preprint arXiv:2506.08008*, 2025.

[2] S. H. Dumpala, A. Jaiswal, C. Sastry, E. Milios, S. Oore, and H. Sajjad, “Sensitivity of Generative VLMs to Semantically and Lexically Altered Prompts,” arXiv preprint arXiv:2410.13030, 2024. [Online]. Available: https://arxiv.org/abs/2410.13030

[3] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, ”Supervised Contrastive Learning,” arXiv:2004.11362, 2021. [Online]. Available: https://arxiv.org/abs/2004.11362

[4] M. A. Jamal and O. Mohareri, “Multi-Modal Contrastive Masked Autoencoders: A Two-Stage Progressive Pre-training Approach for RGBD Datasets,” in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, 2025, pp. 17947–17957, doi: 10.1109/CVPR52734.2025.01672.