Seeing the Blindness: A Quantitative and Visual Analysis of CLIP vs. DINOv2 Embedding Spaces for Subtle Visual Transformations

Daivik Patel, Shrenik Patel

Abstract

Multimodal models such as CLIP have achieved remarkable performance on high-level semantic tasks but exhibit notable limitations in capturing fine-grained visual differences. In this study, we replicate and extend key findings from "Seeing the Blindness" by Liu et al. [**?**], conducting a comparative embedding analysis of CLIP and DINOv2 across visually subtle image transformations (e.g., flipping, zooming, positional changes). By computing standard clustering metrics, visualizing embeddings with t-SNE, and evaluating intra- and inter-cluster distances, we demonstrate that while CLIP forms tighter and more semantically consistent clusters, DINOv2 embeddings better preserve distinctions between fine-grained visual attributes. Our results both confirm and expand upon Liu et al.'s conclusions, providing deeper insights into the structure of embedding spaces and their implications for MLLM visual sensitivity.

1 Introduction

CLIP (Contrastive Language–Image Pre-training) and DINOv2 (Self-supervised Vision Transformer) represent two paradigms in visual representation learning: CLIP aligns images with text in a joint embedding space, while DINOv2 learns visual features without language supervision. While CLIP has dominated as the backbone for Multimodal Large Language Models (MLLMs), Liu et al. (2024) [**?**] highlighted critical failures of CLIP-based models in distinguishing simple visual patterns, such as object orientation and quantity.

This work aims to replicate and expand on these findings. Specifically, we quantitatively evaluate how well CLIP and DINOv2 embeddings separate image groups across 7 transformation types using standard clustering metrics, visualize their 2D structure via t-SNE, and analyze cluster behavior at the individual transformation level. Our methodology is designed to diagnose the extent of CLIP's "visual blindness" and to assess the claimed fine-grained visual sensitivity of DINOv2.

2 Methodology

2.1 Dataset and Transformations

We curated a dataset of approximately 200 base images selected from open-access sources like ImageNet. For each base image, we generated seven visually subtle yet semantically meaningful transformations to test the models' ability to detect fine-grained changes:

- `original` — the unaltered base image
- `flip` — horizontal mirroring of the image
- `zoom` — crop and scale to simulate zoomed-in effect

- `color` — color jittering including hue, brightness, and saturation shifts
- `text` — text overlay added to the image
- `position` — spatial shifting of objects within the frame
- `orientation` — image rotation at random degrees

Each transformed image was saved with a descriptive filename encoding its transformation type. This allowed for consistent automated labeling of transformation categories during processing.

2.2 Model Embeddings

We computed visual embeddings for each image using two pretrained vision models accessed via Hugging Face Transformers:

- **CLIP**: We used the `openai/clip-vit-large-patch14` variant and extracted the image embedding via the `get_image_features` method.
- **DINOv2**: We used `facebook/dinov2-large` and extracted the embedding from the `[CLS]` token of the last hidden state.

Each embedding was L2-normalized to unit length to standardize the representation scale across both models. These high-dimensional embeddings were then used for clustering analysis, distance computation, and 2D visualization with t-SNE.

2.3 Evaluation Metrics

We evaluated the quality of transformation-based grouping using:

- **Silhouette Score** (higher = better separation)
- **Calinski-Harabasz Index** (higher = better cluster separation)
- **Davies-Bouldin Index** (lower = better)
- **Adjusted Rand Index (ARI)** against true labels
- **Intra-cluster distance** (lower = more compact clusters)
- **Inter-cluster distance** (higher = more separated centroids)

2.4 Visualization

We projected high-dimensional embeddings to 2D using t-SNE (perplexity = 30) and generated:

- Overall embedding maps (color-coded by transformation)
- Per-cluster visualizations with centroid distance lines
- Bar plots comparing average intra-cluster distances for each transformation

3 Results

3.1 Clustering Metrics

Metric	CLIP	DINOv2	Better Model	Improvement
Silhouette Score	0.0191	-0.0007	CLIP	-103.6%
Calinski-Harabasz	24.29	4.28	CLIP	-82.4%
Davies-Bouldin	18.14	48.45	CLIP	-167.0%
Adjusted Rand Index	0.2899	0.0583	CLIP	-79.9%
Intra-cluster Distance	11.04	44.69	CLIP	-304.9%
Inter-cluster Distance	4.95	8.42	**DINOv2**	**+70.1%**

Table 1: Comparison of clustering metrics across embedding spaces. CLIP outperforms DINOv2 on 5 of 6 metrics. DINOv2 shows superior separation between transformation categories (inter-cluster distance).

3.2 t-SNE Embedding Visualizations

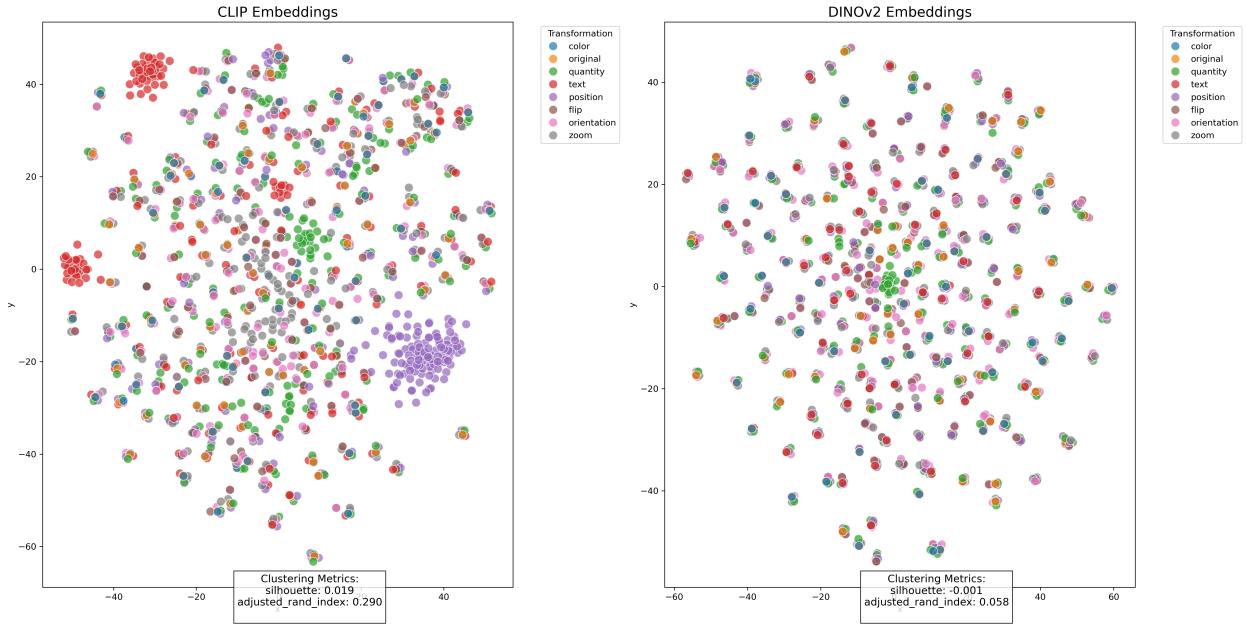

Figure 1: t-SNE projection of CLIP (left) and DINOv2 (right) embeddings. Each point is color-coded by transformation type. CLIP forms tighter semantic clusters; DINOv2 is more dispersed.

3.3 Intra-cluster Distance Visualizations

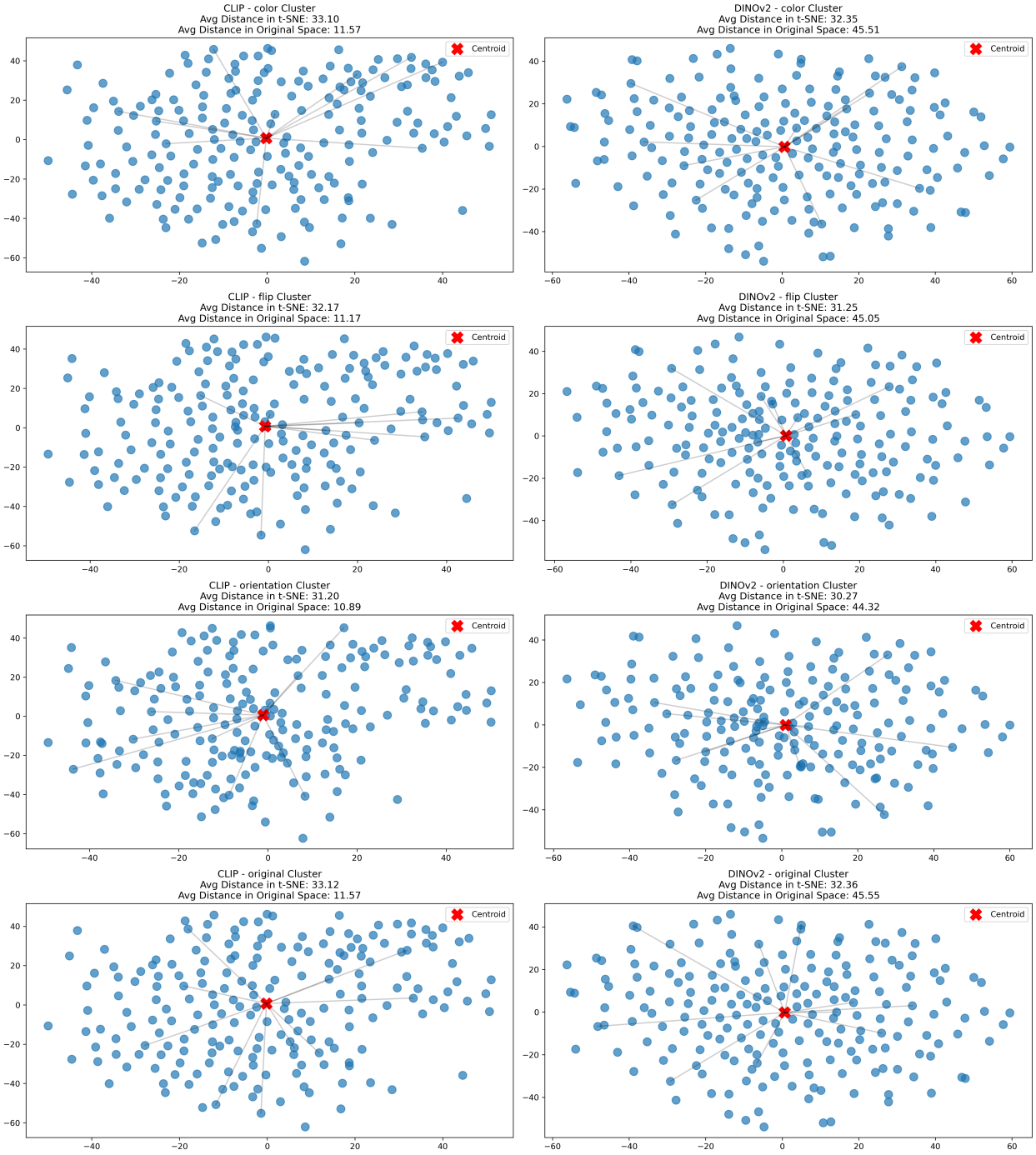

Figure 2: Per-cluster visualization of centroid distance in 2D t-SNE and original embedding space. Each row shows one transformation type; red X denotes centroid. DINOv2 clusters are consistently more dispersed.

3.4 Per-Transformation Clustering Quality

Figure 3: Bar chart comparing average intra-cluster distances (original embedding space) for each transformation. CLIP tends to cluster most transformations more tightly. DINOv2 performs worst on transformations like `flip` and `orientation`, suggesting greater visual sensitivity.

4 Discussion

Our findings replicate and reinforce Liu et al.'s core claim [**?**]: CLIP embeddings are semantically rich but visually coarse. While CLIP produces more compact clusters with higher silhouette and adjusted Rand index scores, this compactness often stems from collapsing visually distinct transformations into semantically similar regions.

In contrast, DINOv2 exhibits higher inter-cluster distances, suggesting it more effectively separates transformation categories. Although its clusters are more dispersed, this likely reflects greater visual sensitivity, precisely the kind of fine-grained detail that CLIP tends to overlook. This trade-off illustrates how CLIP favors semantic abstraction, while DINOv2 prioritizes visual distinction.

Our t-SNE projections and per-transformation analysis reveal that transformations such as `orientation`, `flip`, and `text` are often poorly delineated by CLIP. DINOv2, however, shows clearer separation among these changes. That said, its higher intra-cluster variance indicates less semantic consolidation, which may be a drawback in applications requiring generalization.

These findings support the argument that multimodal models require improved visual grounding. Using DINOv2-like embeddings or blending them with semantically aligned features from models like CLIP can lead to more robust, visually-aware multimodal systems.

5 Contributions

This paper extends prior work in several key ways:

1. We provide a quantitative breakdown of clustering performance using multiple metrics, highlighting specific differences in how CLIP and DINOv2 handle visual variation.

2. Our visual diagnostics like embedding projections and intra-cluster spread charts offer an interpretable view of embedding space structure for each transformation type.

3. We empirically confirm that DINOv2 captures finer-grained visual changes than CLIP, supporting the hypothesis of CLIP's "visual blindness" in alignment-based models.

6 Conclusion

This experiment demonstrates that while CLIP may cluster images more tightly by semantic content, it risks collapsing subtle visual transformations. DINOv2, despite producing looser clusters, encodes meaningful variation across transformation types. For applications that depend on visual grounding rather than pure semantic similarity, our findings suggest a compelling case for incorporating DINO-style representations or hybrid architectures.

References

[1] S. Tong, Z. Liu, Y. Zhai, Y. Ma, Y. LeCun, and S. Xie. Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs. *arXiv preprint arXiv:2401.06209*, 2024.