Many-to-Many, Yet Convergent: Insights into the alignment of Vision and Language Models

Zoe W. He, Sean Trott, and Meenakshi Khosla
(wah016@ucsd.edu, sttrott@ucsd.edu, mkhosla@ucsd.edu)
Department of Cognitive Science, University of California San Diego, La Jolla, CA, USA 92093

Abstract

The "platonic representation" hypothesis holds that vision and language models converge on a shared conceptual space despite being trained on distinct modalities. Yet, much of the evidence for this hypothesis comes from one-to-one image–caption scenarios, where each image is paired with a single descriptive caption. This setup overlooks a fundamental reality: the mapping between images and language is many-to-many, as neither modality uniquely determines the other. In this work, we show that alignment between vision and language models also persists at a finer grain in such many-to-many contexts. Using a forced-choice "Pick-a-Pic" task, we find that human raters' preferences for which of two images better matches a caption are mirrored in the learned embedding space across all vision-language model pairs. This evidence challenges the simplistic view of "one image, one caption" alignment and highlights that models capture finer-grained semantic distinctions akin to human preferences. Moreover, we demonstrate that averaging embeddings across multiple images and multiple captions referring to a shared concept yields significantly stronger alignment than individual image–caption pairs. While one might expect averaging to "blur" representational detail, our results reveal the opposite: aggregating multiple views appears to distill a more universal semantic core. Our findings ultimately reinforce the notion of a shared conceptual space across modalities, underscoring the importance of examining many-to-many correspondences to better understand how such models learn, represent, and unify semantic information.

Keywords: representational alignment; vision; language; cross-modal; meaning

Introduction

The idea of a universal, modality-independent substrate of meaning has intrigued philosophers, cognitive scientists, and neuroscientists. Plato introduced the concept of ideal forms, suggesting that individual percepts derive from an overarching realm of perfect, abstract entities. Similarly, Jerry Fodor's "Language of Thought" hypothesis proposes that minds operate in a universal "code" transcending specific sensory modalities (e.g., vision, audition) and any spoken or written language. Both lines of thought pose a fundamental question: Do putatively distinct cognitive systems—such as vision and language—encode meaning in a shared, abstract space, or are they rooted in modality-specific representations?

Rapid developments in AI—particularly large-scale vision and language models—provide novel tools to explore these ideas computationally. Recent work has shown that despite being trained on distinct modalities, vision and language models often converge onto a shared, abstract semantic space(Huh, Cheung, Wang, & Isola, 2024). However, the majority of empirical evidence for cross-modal alignment derives from experimental setups that assume a simplified one-to-one correspondence between images and captions. These analyses inadvertently mask the complexity of real-world semantics, where an image rarely maps uniquely to a single descriptive phrase, and a given phrase typically corresponds to numerous visual interpretations.

To address this gap, our study employs two complementary analyses that explicitly investigate semantic alignment at a finer granularity using many-to-many mappings. First, using a forced-choice "Pick-a-Pic" task, we show that visual embeddings of human-preferred images align more closely with the language model embeddings of the caption than non-preferred images. Second, using captions selected based on CLIP-scores—a proxy for human preference—we find similar effects. These results indicate that vision and language models converge on a common semantic ground that reflects subtle distinctions aligned with human judgments

In our second analysis, we find that aggregating embeddings across multiple images per caption (and vice versa) consistently improves alignment, suggesting averaging extracts a more stable, modality-independent semantic core. This highlights the importance of examining many-to-many correspondences for understanding cross-modal alignment.

Methods

We compare image representations from Vision Transformers (ViTs) trained on DINOv2 (Oquab et al., 2023) with textual representations of the same images from language models (BLOOM, OpenLLaMA) (BigScience et al., 2022; Geng & Liu, 2023). Multiple model sizes were selected from repositories including Huggingface (Wolf et al., 2019) and PyTorch Image Models (TIMM) (Wightman, 2021). For images, the class token from the penultimate transformer block is used; for language, token activations are averaged from the same layer.

We analyzed 1,000 random samples from each dataset:

- **Pick-A-Pic:** Contains two generated images per caption along with human preference judgments (Kirstain et al., 2023). (Figure 1, A)
- **MS-COCO:** Provides five human-authored captions per image (Lin et al., 2014). (Figure 1, B)

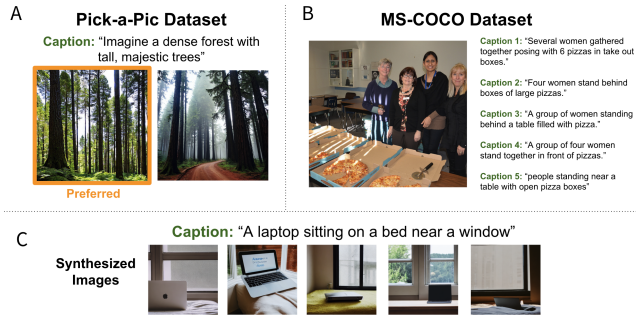

Figure 1: Example data from **(A)** the Pick-A-Pic dataset and **(B)** the MS-COCO dataset. **(C)** Example MS-COCO caption and synthesized images by the stable-diffusion model.

Cross-modal alignment is quantified using linear predictivity—by fitting a Ridge regression (with the regularization parameter optimized over $[10^{-8}, 10^{8}]$ on a log scale) to predict one modality's activations from the other's, with the mean correlation from a 5-fold cross-validation as the alignment score.

Results

Vision–Language Alignment Mirrors Human Preferences

We evaluated whether vision–language alignment reflects human preferences. Images from the "Pick-a-Pic" dataset, which provides two generated images per prompt with human preference judgments, were grouped into high- and low-preference categories. For each group, vision model representations were extracted and linear predictivity scores were computed using the corresponding caption embeddings.

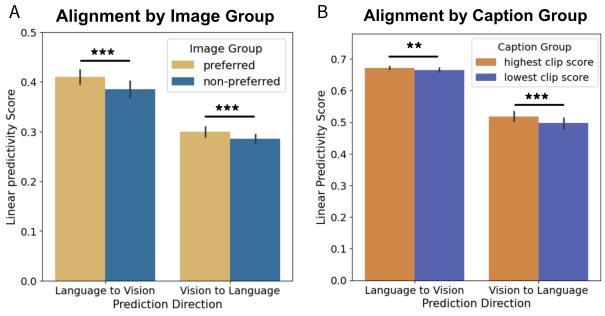

Figure 2: **(A)** Pick-a-Pic linear predictivity scores grouped by image variation (preferred vs. non-preferred) based on human judgments. **(B)** MS-COCO linear predictivity scores grouped by caption variation based on CLIP Scores. Error bars indicating the standard error across all eight model pairs.

Our results indicate that images preferred by human raters exhibit significantly stronger alignment with their associated captions than non-preferred images (paired t-test, language-to-vision: $t(7) = 19.8225$, $p < 0.001$; vision-to-language: $t(7) = 10.2338$, $p < 0.001$; Figure 2A).

To assess alignment from the caption perspective, we computed the CLIP Score (Hessel, Holtzman, Forbes, Le Bras, & Choi, 2021)—a reference-free metric based on the cosine similarity of image–caption embeddings—for all MS-COCO captions, which prior work has shown correlates strongly with human preference (Hessel et al., 2021). Our analysis reveals that captions with higher CLIP Scores are significantly more aligned with their images than those with lower scores (paired t-test, language-to-vision: $t(7) = 3.9231$, $p = 0.0057$; vision-to-language: $t(7) = 17.8350$, $p < 0.001$; Figure 2B).

Together, these findings suggest that the model embeddings capture fine-grained semantic distinctions that mirror human evaluative patterns.

Averaging Embeddings Across Multiple Captions and Images Enhances Alignment

To evaluate whether aggregating multiple caption representations improves vision-language alignment beyond to one-to-one pairings, we computed the alignment score using averaged embeddings across increasing numbers of captions from the MS-COCO dataset. As shown in Figure 3A, the alignment score increases as more caption embeddings are averaged.

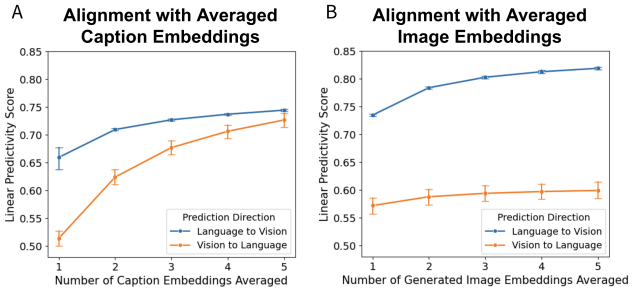

Figure 3: Effect of averaging **(A)** caption embeddings and **(B)** image embeddings on vision–language alignment; error bars show standard error across all eight model pairs.

In a complementary experiment, we synthesized five naturalistic images per caption using the Stable Diffusion model (Figure 1, C) and computed the alignment score based on the averaged embeddings of the generated images (3B). Similar to caption aggregation, increasing the number of aggregated image embeddings further improved the alignment.

Discussion

Our findings show that cross-modal alignment not only persists but also mirrors human preferences in many-to-many image–caption scenarios. The improvement seen with embedding aggregation suggests that combining multiple instances reduces modality-specific noise, thereby distilling a more universal semantic core. These results raise important questions for future work: What types of visual and linguistic information drive this effect, and how does averaging reinforce the features that support a shared conceptual space?

References

BigScience, Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., ... others (2022). *Bloom: A 176b-parameter open-access multilingual language model.* Retrieved from https://arxiv.org/abs/2211.05100

Geng, X., & Liu, H. (2023, May). *Openllama: An open reproduction of llama.* https://github.com/openlm-research/open-llama.

Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., & Choi, Y. (2021). Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the conference on empirical methods in natural language processing (emnlp).* Retrieved from https://arxiv.org/abs/2104.08718

Huh, M., Cheung, B., Wang, T., & Isola, P. (2024). The platonic representation hypothesis. In *Proceedings of the international conference on machine learning (icml).*

Kirstain, A., et al. (2023). *Pick-a-pic: A dataset for evaluating the robustness of vision-language models.* Retrieved from https://arxiv.org/abs/2305.01569

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., ... Zitnick, C. L. (2014). *Microsoft coco: Common objects in context.* Retrieved from https://arxiv.org/abs/1405.0312

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., ... Bojanowski, P. (2023). *Dinov2: Learning robust visual features without supervision.* (Preprint)

Wightman, R. (2021). *Pytorch image models.* https://github.com/rwightman/pytorch-image-models.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... others (2019). *Huggingface's transformers: State-of-the-art natural language processing.* Retrieved from https://arxiv.org/abs/1910.03771