Seeing Through Words, Speaking Through Pixels: Deep Representational Alignment Between Vision and Language Models

Anonymous ACL submission

Abstract

Recent studies have shown that deep vision-only and language-only models—trained on disjoint modalities—nonetheless project their inputs into a partially aligned representational space. Yet we still lack a clear picture of *where* in each network this convergence emerges, *what* visual or linguistic cues support it, and *whether* it endures the many-to-many nature of real image–text relationships. In this work, we systematically investigate these questions. We show that representational alignment emerges most strongly in mid-to-late layers of both vision and language models, suggesting a hierarchical progression from modality-specific to conceptually shared representations. Second, this alignment is robust to appearance-only changes but collapses when semantic content is altered—e.g., object removal in images or word order shuffling that disrupts thematic roles in sentences—highlighting that the shared code is truly semantic rather than form-based. Critically, we move beyond the conventional one-to-one image-caption paradigm to investigate alignment in many-to-many contexts, acknowledging that neither modality uniquely determines the other. Using a forced-choice "Pick-a-Pic" task, we find that human preferences for image-caption matches are mirrored in the learned embedding spaces across all vision-language model pairs. This pattern holds bidirectionally when multiple captions correspond to a single image, demonstrating that models capture fine-grained semantic distinctions similar to human judgments. Surprisingly, aggregating embeddings across multiple images or phrases referring to the same concept amplifies alignment. Rather than "blurring" representational detail, aggregation appears to distill a more universal semantic core. Together, these results demonstrate that vision and language networks converge on a shared semantic code, where the alignment mirrors human judgements, and becomes more pronounced when multiple exemplars of the same concept within a single modality are averaged in representational space. Our work provides compelling evidence for a universal code of meaning that transcends modality, offering critical insights into how neural networks represent and align semantic information across the vision-language divide.

1 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 objects and percepts derive from an overarching realm of perfect, abstract entities. Similarly, Jerry Fodor's "Language of Thought" hypothesis (or mentalese) proposes that minds operate in a universal "code" transcending specific sensory modalities (e.g., vision, audition) and any spoken or written language (Fodor, 1975). Both lines of thought pose a fundamental question: Do putatively distinct cognitive systems—such as vision and language models—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. Large-scale vision-only and language-only models, trained on massive but disjoint corpora, nonetheless exhibit striking representational convergence. Huh et al. coined this phenomenon the "Platonic Representation Hypothesis", showing that increasingly capable LLMs align more tightly with larger vision models. Interestingly, this alignment occurs even without explicit cross-modal training. This "Platonic Representation Hypothesis" is further supported by Maniparambil et al, who demonstrate that this convergence manifests across a range of model architectures and training paradigms (Maniparambil et al., 2024).

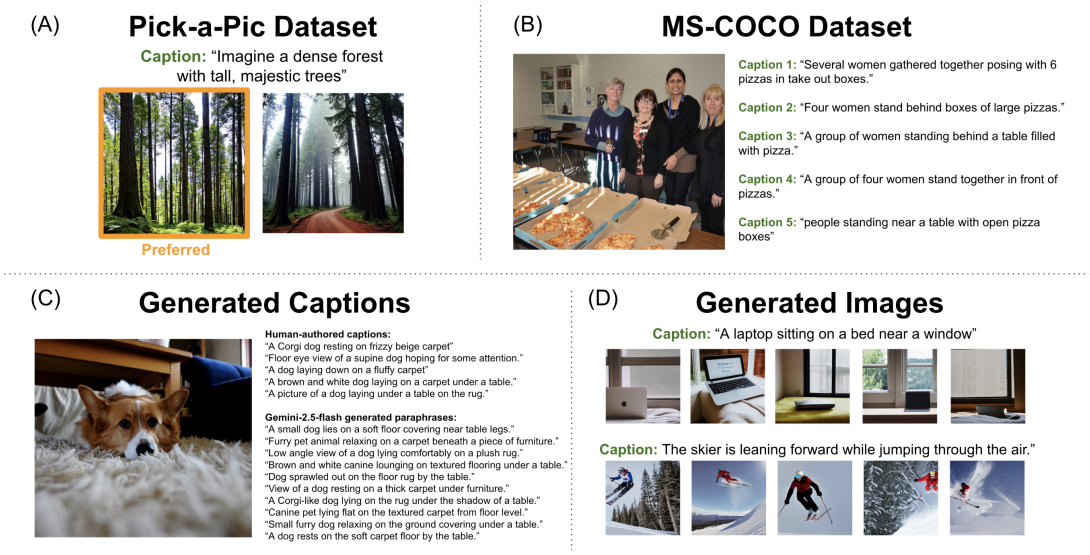

Figure 1: Example data from **(A)** the Pick-A-Pic dataset and **(B)** the MS-COCO dataset. **(C)** Example captions generated by Gemini-2.5-Flash by paraphrasing the human-authored captions in MS-COCO. **(D)** Example MS-COCO captions and synthesized images by the stable-diffusion model.

Critically, cross-modal alignment is not merely correlational. Merullo et al. (2022) show that training just one linear projection is enough to map a frozen vision-transformer's embeddings into the token-embedding space of a frozen language model, letting the stitched system caption images and answer visual questions without any additional multimodal training (Merullo et al., 2022). Similarly, Koh et al. (2023) show analogous gains for the reverse mapping from text to image, showing that a frozen LLM can be visually grounded with a single learned linear map, achieving strong zero-shot performance on tasks such as contextual image retrieval and multimodal dialogue (Koh et al., 2023).

Marjieh et al. (2024) show that even multimodal models like GPT-4 rely predominantly on textual associations rather than direct visual input when predicting human perceptual judgments—highlighting language as a sufficient scaffold for grounding sensory semantics (Marjieh et al., 2024). Bavaresco and Fernández (2025) demonstrate that text alone - when modeled on scale - can implicitly encode rich experiential semantics, echoing Marjieh et al. (2024)'s results on LLMs' ability to recover perceptual hierarchies like the pitch spiral (Bavaresco and Fernández, 2025).

Convergent evidence also emerges from neuroscience. Popham et al. (2021) used within-subject fMRI to chart voxel-wise semantic tuning during silent-movie viewing (purely visual) and narrative listening (purely linguistic) (Popham et al., 2021). They discovered that the two modality-specific maps are topographically contiguous: for every visual category encoded in posterior occipital cortex, a mirror linguistic representation appears immediately anterior to the same cortical border. In other words, visual and linguistic semantics form a single, smoothly joined map that straddles the edge of human visual cortex, implying a tightly aligned cross-modal code rather than two isolated systems. Doerig et al. (2022) asked whether vision already encodes such linguistic semantics. They showed that a vision model trained to translate images directly into sentence embeddings of a language model predicts voxel patterns even better than the embeddings themselves, offering a mechanistic account of how the visual system may recast images into a language-like semantic code by default (Doerig et al., 2022). Saha et al. (2024) went further, finding that off-the-shelf LLM embeddings sometimes outperform dedicated vision models in explaining activity in high-level visual areas. Together, these findings suggest that the cross-modal alignment observed in artificial networks may reflect, or even recapitulate, the brain's own amodal semantic code.

These findings collectively suggest that modern vision and language models, and possibly even brain systems—like Plato's ideal forms—incrementally discard modality-specific details in favor of a shared, amodal semantic code.

Yet critical gaps remain. First, where along the network hierarchy does this alignment emerge, and is it symmetric across modalities? Second, what visual attributes or linguistic properties drive the effect? Third, all previous demonstrations of cross-modal alignment rely on one-to-one image–text pairs. These analyses inadvertently mask the complexity of real-world semantics where no single description exhausts an image's meaning, and the same sentence can fit many images.

In the present study, we fill these gaps through extensive analyses of cross-modal alignment on a broad suite of vision and language encoders. We map alignment layer-by-layer and probe its dependence on targeted manipulations—semantic (object removal, role shuffling) versus appearance-only. Alignment peaks in mid-to-late layers of both modalities, collapses under semantic changes, and is largely unaffected by superficial appearance edits.

To address the third gap about the many-to-many mapping between images and text, 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, for the same image, we analyze pairs of captions selected based on high and low CLIP-scores—previously validated as proxies for human preferences—and observe analogous alignment patterns. 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 investigate the impact of aggregating embeddings across multiple images associated with a single caption and vice versa. Contrary to the intuitive expectation that averaging embeddings would diminish representational specificity, we discover that such aggregation consistently enhances alignment. This suggests that rather than blurring distinctions, averaging distills a more stable, modality-independent semantic core shared across representations. Together, our findings reveal that examining many-to-many correspondences offers richer insights into cross-modal alignment, highlighting a robust convergence toward a shared conceptual space that captures subtle and complex semantic relationships.

2 Methods

We compare image representations from large vision models with textual representations of the same images from large language models. For vision models, we employed Vision Transformers (ViTs) trained via DINOv2 (Oquab et al., 2023) on the LVD-142M dataset. DINOv2 learns rich visual representations by solving a self-distillation task where a student network is trained to match the output distribution of a teacher network (an exponential moving average of the student) while viewing different augmented versions of the same image. For language models, we employed BLOOM (BigScience et al., 2022), a decoder-only transformer-based architecture trained on a massive multilingual corpus, and OpenLLaMA, an open-source reproduction of the LLaMA model trained on publicly available datasets(Geng and 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.

Two datasets are employed:

- **Pick-A-Pic:** An open dataset of over 500,000 human preference judgments on text-to-image outputs, collected from 37 K real-user prompts; each prompt is paired with two generated images and a binary (or tie) preference label (Kirstain et al., 2023). In our experiments, we randomly sample 1,000 prompt–image-pair judgments for analysis (Figure 1A).

- **MS-COCO:** A large-scale image captioning dataset of 123,000 natural photographs depicting complex everyday scenes—people interacting with objects, urban and rural environments, animals, vehicles, indoor and outdoor contexts—each annotated with five human-authored captions (Lin et al., 2014). For our experiments, we randomly sample 1,000 images (and their associated captions) from the official validation split (Figure 1B).

Computing Alignment

To quantify alignment between representations from language and vision models, we use *linear predictivity*. For each pair of representations, $\mathbf{X} \in R^{n \times d_X}$ (e.g., from a vision model) and

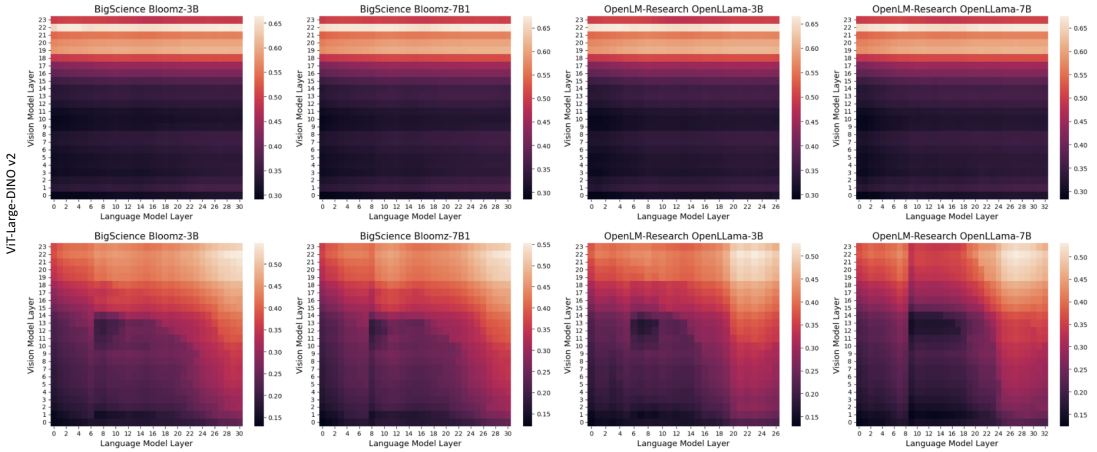

Figure 2: Layer-wise alignment, measured suing linear predictivity score, between one example vision model (ViT-Large-Dinov2) and all language models. Top row: Alignment computed in language-to-vision direction. Bottom row: Alignment computed in vision-to-language direction.

$\mathbf{Y} \in R^{n \times d_Y}$ (e.g., from a language model), we fit a ridge regression from $\mathbf{X}$ to $\mathbf{Y}$:

$$\hat{\mathbf{W}} = \arg\min_{\mathbf{W}} \|\mathbf{X}\mathbf{W} - \mathbf{Y}\|_2^2 + \lambda \|\mathbf{W}\|_F^2, \quad (1)$$

where λ is the regularization parameter, selected via cross-validation over a logarithmically spaced range from 10^{-8} to 10^{8}. We computed the final alignment score by averaging the Pearson correlation between predicted and actual responses across all units and five cross-validation folds.

We treat this as an *asymmetric* similarity measure and report results for both directions: predicting language representations from vision ($\mathbf{X} \to \mathbf{Y}$) and vice versa ($\mathbf{Y} \to \mathbf{X}$). This allows us to disentangle directional differences in information content across modalities.

3 Results

3.1 Layer-Wise Vision-Language Alignment

To pinpoint where vision–language alignment first appears and how it evolves across the network hierarchy, we performed a layer-by-layer mapping between each pair of vision-transformer and language-model embeddings. As shown in Figure 2, both modalities exhibit low cross-modal predictivity in their earliest layers and increase through the mid and later layers. These patterns hold consistently across different vision-language model pairs (see Appendices). These findings demonstrate that both vision and language models transition from modality-bound encoding toward an abstract, shared semantic space as depth increases.

We also observe a clear directional asymmetry in these mappings. When mapping from language to vision, we find that even early language layers can successfully predict later vision layers. In contrast, mapping from vision to language reveals a more graded effect: deeper vision layers progressively yield higher predictivity for deeper language layers. Early vision features poorly predict any language layer, while later vision representations align best with higher language layers. This asymmetry suggests that textual representations abstract away from surface form more rapidly than visual ones, while vision networks require deeper processing to reach a comparable semantic level.

3.2 Semantic content, not surface form, drives cross-modal alignment

We next explore whether the cross-modal correspondence we observe is mainly driven by surface form or by deeper semantic content.

3.2.1 Image manipulations

To dissociate appearance-level similarity from semantic correspondence, we performed four controlled perturbations on each MS-COCO image. Two manipulations altered only the appearance while preserving the full meaning: (i) conversion to grayscale and (ii) 15 degree image rotation. The other two manipulations altered the semantic content with different degrees by exploiting the segmentation masks (Figure 3A) provided with COCO-Stuff (Caesar et al., 2018):

- **Thing-only views** that preserve pixel-perfect instances of the foreground object classes (e.g.

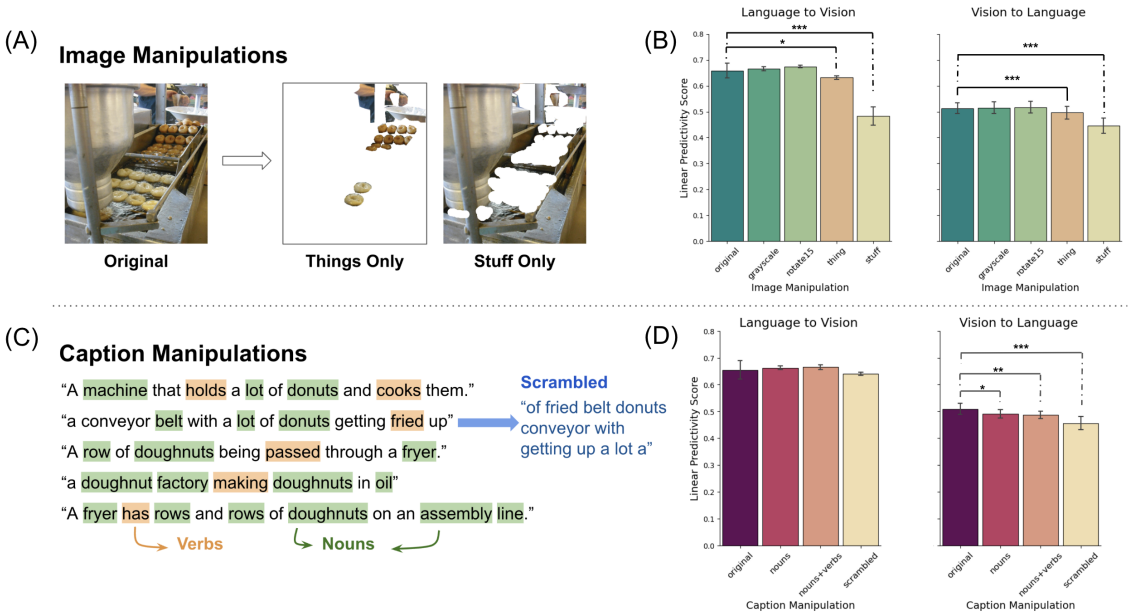

Figure 3: **(A)** Example thing-only and stuff-only images by manipulating the original image using masks from COCO-Stuff. **(B)** Alignment by image manipulations. **(C)** Demonstration of image manipulations: nouns and verbs extraction, and captions scrambling. **(D)** Alignment by caption manipulations. Paired t-tests (n=8 vision-language model pairs per comparison) were conducted separately for each image manipulation, and p-values were adjusted for four comparisons per mapping direction using the Benjamini–Hochberg procedure (FDR).

person, car) but remove the surrounding context to eliminate spatial and contextual relations;

- **Stuff-only views** that retain only the background layout and the scene categories (e.g. grass, wall) while removing the foreground objects.

We find that appearance-only manipulations of image inputs have no notable negative effects on alignment (Figure 3B, grayscale: L→V: $t(7) = -0.8405$, $p = 0.4284$, $q = 0.4284$; V→L: $t(7) = -1.3386$, $p = 0.2226$, $q = 0.2543$; rotation: L→V: $t(7) = -1.7569$, $p = 0.1224$, $q = 0.1631$; V→L: $t(7) = -3.1161$, $p = 0.0169$, $q = 0.0271$). In contrast, deleting semantic content from images results in substantial alignment degradation (Figure 3B). Isolating only the foreground "thing" pixels and removing contextual relations significantly lowered the alignment scores (L→V: $t(7) = 3.4304$, $p = 0.0110$, $q = 0.0220$; V→L: $t(7) = 7.2528$, $p = 0.0002$, $q = 0.0005$). Retaining only the "stuff" background further reduced the alignment (L→V: $t(7) = 10.1267$, $p < 0.0001$, $q = 0.0001$; V→L: $t(7) = 11.7109$, $p < 0.0001$, $q = 0.0001$).

Notably, the decline was systematically steeper in the language-to-vision direction, indicating that mapping from textual embeddings to visual layers depends more heavily on intact visual semantics.

3.2.2 Caption manipulations

To explore the linguistic properties driving the alignment, we separately manipulated the captions in the MS-COCO dataset with different levels of semantic disruption by retaining: (i) nouns only, (ii) nouns and verbs, and (iv) all the words but in scrambled order (Figure 3C).

Interestingly, only in the vision-to-language mapping direction do caption manipulations negatively affect the alignment (Figure 3D, right). Specifically, nouns-only ($t(7) = 3.5956$, $p = 0.0088$, $q = 0.0176$) and nouns+verbs ($t(7) = 5.3561$, $p = 0.0011$, $q = 0.0032$) show similar moderate decreases, while scrambled captions produce the largest drop ($t(7) = 22.8176$, $p < 0.0001$, $q < 0.0001$). This suggests that nouns and verbs carry the primary semantic weight in grounding language to visual content, while word order and the full lexical distribution become even more crucial when projecting from vision to language.

The directional asymmetries we observe—greater sensitivity of language→vision mapping to intact visual semantics and of vision→language mapping to linguistic composition—suggest complementary organizational

principles in how each modality abstracts and transmits meaning across the shared representational space.

3.3 Vision–Language Alignment Mirrors Human Preferences

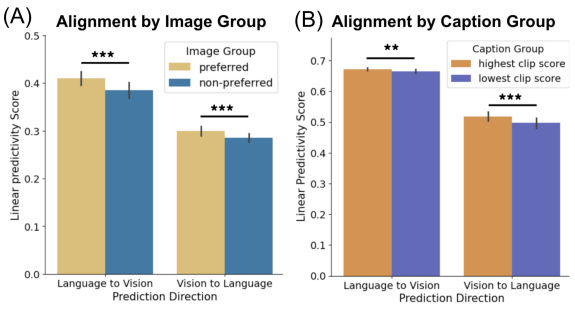

Figure 4: **(A)** Pick-a-Pic dataset Linear predictivity scores grouped by image variation (preferred vs. non-preferred) based on human judgments. **(B)** MS-COCO dataset Linear predictivity scores grouped by caption variation based on CLIP Scores. Error bars indicating the standard error across model pairs.

We next evaluated whether cross-modal alignment tracks fine-grained human preferences. Images from the "Pick-a-Pic" dataset, which provides two generated images for the same 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. This design probes alignment at a finer-grained resolution: can the vision–language mapping replicate the subtle distinctions that lead people to prefer one image over another, even when the linguistic description is identical?

Our results indicate that images preferred by human raters exhibit significantly stronger alignment with their associated captions than non-preferred images across all vision-language model pairs (paired t-test, L→V: $t(7) = 19.8225$, $p < 0.001$; V→L: $t(7) = 10.2338$, $p < 0.001$; Figure 4A). In other words, even when two pictures illustrate the same text, the uni-modal vision and language models collectively "agree" with human raters about which picture is the better semantic fit. This fine-grained sensitivity shows that the cross-modal alignment we measure is not a coarse correlation but captures subtle, human-relevant distinctions within a shared semantic space.

A complementary analysis from the text side reinforces this conclusion. We computed the CLIP Score (Hessel et al., 2021)—a reference-free metric based on the cosine similarity of image–caption embeddings—for all MS-COCO captions, as a reasonable proxy for human preferences (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 4B).

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

3.4 Averaging Embeddings Across Multiple Captions and Images Enhances Alignment

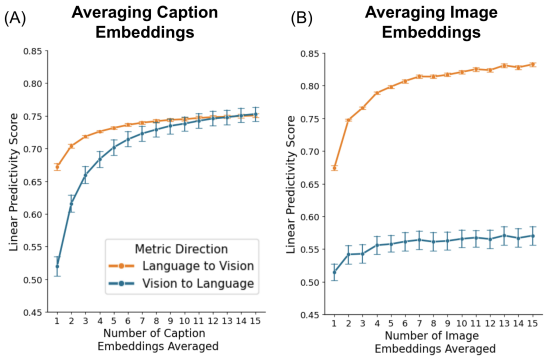

Figure 5: Effect of aggregation on alignment. Cross-modal aggregation: Averaging (**A**) multiple caption embeddings for the same image or (**B**) multiple image embeddings for the same caption steadily increases language–vision and vision–language predictivity. Error bars denote standard error across all model pairs.

To quantify the impact of aggregating caption representations, we progressively averaged embeddings from an increasing number of MS-COCO captions per image and computed cross-modal alignment scores. As shown in Figure 5A, alignment improved monotonically with each additional caption. To locate the point of diminishing returns, we expanded the caption pool by paraphrasing each of the five human-authored captions with Gemini-2.5-Flash (Figure 1C, see Appendix C for prompt), creating up to 15 captions per image. In the vision-to-language mapping, alignment continued to rise until roughly ten captions were included, after which the curve plateaued.

We performed the complementary analysis in the opposite direction by synthesizing up to 15

naturalistic images per caption with Stable Diffusion (Figure 1D). Similar to caption aggregation, increasing the number of aggregated image embeddings further improved the alignment (Figure 5B). The alignment gain is larger when predicting vision from language, and plateaued around seven images.

To confirm that these improvements reflect enhanced semantic information rather than a generic averaging artifact, we repeated both analyses after randomly shuffling the image–caption correspondences (see Figure 8 in Appendix). Under this mismatch baseline, embedding aggregation showed no benefit, demonstrating that the effect depends on semantically matched pairs.

We also observe a clear directional asymmetry both analyses: averaging captions benefitted vision-to-language predictions, whereas averaging images benefitted language-to-vision predictions. This pattern suggests that aggregation may suppress modality-specific noise within the averaged domain, exposing a cleaner semantic signal that is more easily mapped by the other modality.

3.5 Effect of Vision Models on Vision-Language Alignment

To assess the generalizability of our findings, we repeated the analyses on seven ViT backbones that differ in objective (strong AugReg, DINO, large-scale DINOv2, supervised distillation DeiT), data scale (ImageNet-1k vs. ImageNet-21k vs. LVD-142 M), and model size (ViT-B/14, ViT-B/16, ViT-L/14, ViT-L/16).

We observe that the improvement of averaging caption embeddings is generalized across different vision model backbones (Figure 6). Notably, when mapping language features into visual space, the alignment differences scores across ViTs were noticeably larger than in the reverse direction. Furthermore, both training methods and data size appear to affect the alignment. When the model size and data were held constant (ViT-B/16, ImageNet-1k), AugReg produced higher alignment than either DINO or DeiT. Keeping the objective similar but increasing the dataset (DINO-ImageNet1k to DINOv2-LVD142m) improved alignment further. However, a larger dataset did not help the AugReg model: its ImageNet-21k checkpoint aligned worse than its ImageNet-1k counterpart. Our current experiment cannot cleanly disentangle the interaction between objective and data distribution. A systematic experiment would be needed to clarify such interactive effects.

4 Discussion

Our results provide new evidence that purely unimodal vision and language models gravitate toward a common semantic manifold. Alignment (i) peaks in their mid-to-late layers where abstract semantic processing occurs, (ii) reduces when we remove or scramble semantic content but survives appearance-only changes, and (iii) exhibits striking correspondence in fine-grained evaluation scenarios with human judgements (e.g., when comparing alignment scores for multiple candidate images corresponding to the same linguistic expression, the model aligns most strongly with the image humans rate as most semantically congruent with the text, and reciprocally for multiple linguistic descriptions of the same image), and (iv) is markedly enhanced when averaging representations corresponding to the same concept in each modality. Together, these findings refine the emerging "Platonic" view of cross-modal representation: the two modalities do not merely share coarse alignment but capture fine semantic gradients that track human judgments. Our work bridges cognitive science and machine learning by suggesting that a shared code for meaning can emerge implicitly in unimodal systems, even without cross-modal training.

Our work opens several promising avenues for future research. Future studies should investigate how alignment strength varies across different types of visual and linguistic content. Are concrete concepts (e.g., "dog", "chair") more strongly aligned than abstract concepts (e.g., "freedom", "justice")? Understanding these variations could reveal fundamental constraints on cross-modal convergence. Different image types—photographs, illustrations, diagrams, artistic renderings—may exhibit varying degrees of alignment with language. Examining these differences could illuminate how visual style and abstraction influence semantic encoding and cross-modal correspondence.

Our discovery that alignment strengthens when averaging concept-specific representations raises intriguing questions about the geometric properties of these embeddings. Future work should explore whether averaging acts as a denoising mechanism that preserves core semantic content while reducing modality-specific variations. Additionally, it would be interesting to investigate whether averaging techniques applied to paraphrases of the same

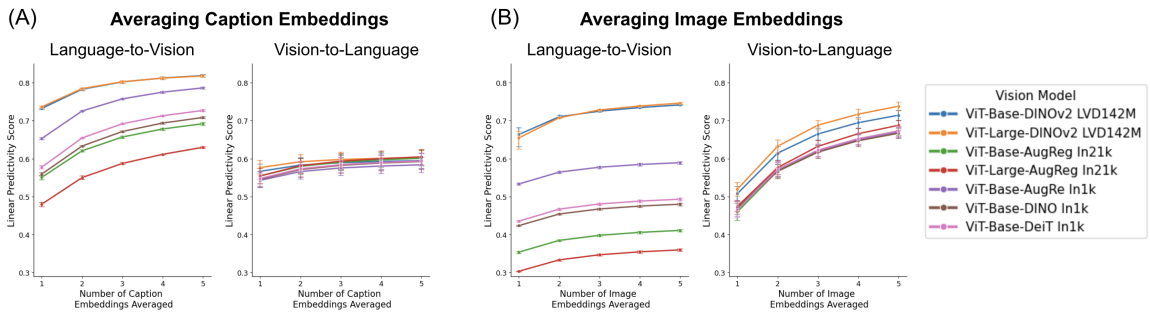

Figure 6: Comparing the alignment of different vision models with language models after averaging **(A)** caption embeddings and **(B)** image embeddings.

linguistic expression could enhance performance on downstream tasks involving natural language inference.

While our study demonstrates alignment at the representation level, identifying which specific features or dimensions drive this alignment remains an open question. Future research should develop techniques to isolate the most aligned dimensions between vision and language models and analyze their semantic properties.

Further, investigating how alignment patterns evolve during training could provide insights into the developmental trajectory of cross-modal correspondence. Do alignment patterns appear early in training and strengthen over time, or do they emerge suddenly after sufficient exposure to domain-specific data? This temporal perspective could reveal fundamental insights about how semantic convergence develops in neural networks trained on different modalities.

Limitations

Our work primarily focuses on linear predictivity as a measure of representational alignment between vision and language models. While this approach offers valuable insights, it represents only one perspective on how these representational spaces may relate to each other. Future work could benefit from employing a broader spectrum of alignment metrics to provide a more complete understanding of vision-language relationships. For instance, more constrained mapping approaches—such as orthogonal transformations in Procrustes analysis (Williams et al., 2021) or permutation-based methods like permutation score and soft matching score (Khosla and Williams, 2024)—might reveal unit-level correspondences between visual and language model representations that linear regression cannot capture. Kernel-based methods (e.g., Representational Similarity Analysis (Kriegeskorte et al., 2008)) would assess population-level relationships between representations, while neighborhood-based approaches (e.g., mutual k-NN) could illuminate local clustering patterns within embedding spaces. These complementary metrics would provide a multi-faceted view of the nature of alignment between vision and language models. Our analysis does not fully reveal which specific features drive the observed alignment between vision-only and language-only models, nor does it identify the scenarios where these models systematically diverge in their representations. Investigating these questions would require more extensive probing of representations across diverse stimuli and large-scale datasets.

The synthetic nature of our image dataset introduces another limitation. While diffusion models generate high-quality images corresponding to text prompts, some generated images may not perfectly capture the semantic content or nuances present in the texts. This potential mismatch between text and generated images could influence our alignment measurements and subsequent interpretations.

Furthermore, our work examines models trained at a specific point in time, with particular architectures and training objectives. As model architectures and training paradigms evolve, the nature of cross-modal alignment may change significantly.

Finally, representational similarity is descriptive. It does not prove shared processing mechanisms or functional interchangeability. Causal interventions are needed to determine whether the aligned dimensions are necessary for each model's downstream behavior.

References

Anna Bavaresco and Raquel Fernández. 2025. Experiential semantic information and brain alignment: Are multimodal models better than language models? *arXiv preprint arXiv:2504.00942*.

BigScience, T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, and 1 others. 2022. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. 2018. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218.

Adrien Doerig, Tim C Kietzmann, Emily Allen, Yihan Wu, Thomas Naselaris, Kendrick Kay, and Ian Charest. 2022. Semantic scene descriptions as an objective of human vision. *arXiv preprint arXiv:2209.11737*, 10.

Jerry A Fodor. 1975. *The language of thought*, volume 5. Harvard university press.

X. Geng and H. Liu. 2023. Openllama: An open reproduction of llama. https://github.com/openlm-research/open-llama.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Meenakshi Khosla and Alex H Williams. 2024. Soft matching distance: A metric on neural representations that captures single-neuron tuning. In *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models*, pages 326–341. PMLR.

A. Kirstain and 1 others. 2023. Pick-a-pic: A dataset for evaluating the robustness of vision-language models. *Preprint*, arXiv:2305.01569.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. 2023. Grounding language models to images for multimodal inputs and outputs. In *International Conference on Machine Learning*, pages 17283–17300. PMLR.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. *Preprint*, arXiv:1405.0312.

Mayug Maniparambil, Raiymbek Akshulakov, Yasser Abdelaziz Dahou Djilali, Mohamed El Amine Seddik, Sanath Narayan, Karttikeya Mangalam, and Noel E O'Connor. 2024. Do vision and language encoders represent the world similarly? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14334–14343.

Raja Marjieh, Ilia Sucholutsky, Pol van Rijn, Nori Jacoby, and Thomas L Griffiths. 2024. Large language models predict human sensory judgments across six modalities. *Scientific Reports*, 14(1):21445.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. 2022. Linearly mapping from image to text space. *arXiv preprint arXiv:2209.15162*.

M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, R. Howes, P.-Y. Huang, H. Xu, V. Sharma, S.-W. Li, W. Galuba, M. Rabbat, M. Assran, N. Ballas, and 7 others. 2023. Dinov2: Learning robust visual features without supervision. Preprint.

Sara F Popham, Alexander G Huth, Natalia Y Bilenko, Fatma Deniz, James S Gao, Anwar O Nunez-Elizalde, and Jack L Gallant. 2021. Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature neuroscience*, 24(11):1628–1636.

Ross Wightman. 2021. Pytorch image models. https://github.com/rwightman/pytorch-image-models.

Alex H Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. 2021. Generalized shape metrics on neural representations. *Advances in Neural Information Processing Systems*, 34:4738–4750.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and 1 others. 2019. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

5 Appendix

A Baseline alignment with shuffled image-caption correspondences.

Under the image-caption mismatch baseline, averaging multiple embeddings does not improve vision-language alignment: the alignment score remains around 0 in both mapping directions (Figure 8).

B Additional Results: Embedding Aggregation Effect on Manipulated Captions.

Given that averaging caption embeddings enhances vision-language alignment, we also explored whether the embeddings of semantically manipulated captions would also benefit from embedding aggregation (Figure 9). Interestingly, the alignment was enhanced even though the embeddings come from manipulated captions.

C MS-COCO caption generation.

Gemini-2.5-Flash Prompt

```
Prompt:

f"""You are an expert image captioner. I'll show you some
       existing captions for an image, and your task is to
       generate 10 NEW captions that:
1. Are similar in style and detail level to the existing
       captions
2. Capture the same meaning but with different wording
3. Are direct, concise descriptions (around 10-15 words each)
4. Are worded differently from each existing caption and from
       each other

Here are the existing captions:
{insert all captions text for the image here}

Generate 10 new captions formatted exactly as:
1. [First new caption]
2. [Second new caption]
3. [Third new caption]
4. [Fourth new caption]
5. [Fifth new caption]
6. [Sixth new caption]
7. [Seventh new caption]
8. [Eighth new caption]
9. [Ninth new caption]
10. [Tenth new caption]"""
```

Table 1: Prompt used for generating new image caption paraphrases from Gemini-2.5-Flash.

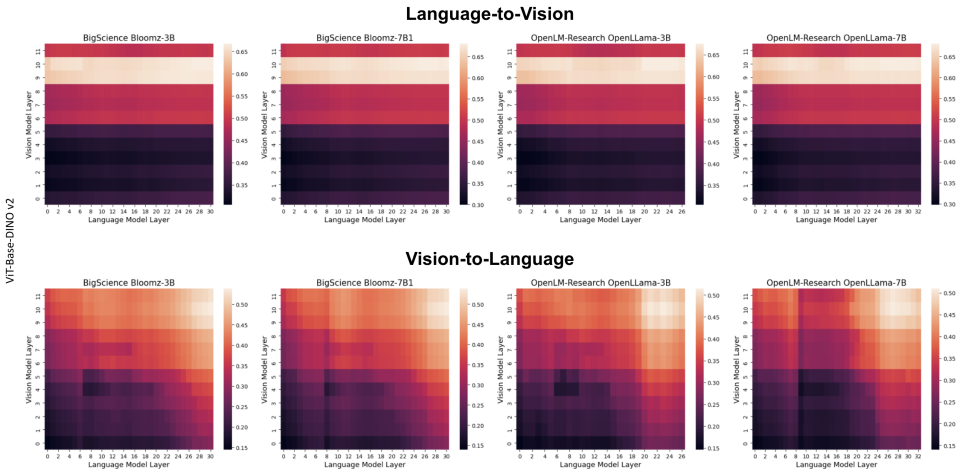

Figure 7: Layer-wise alignment for additional vision-language model pairs (with ViT-Base-DINO v2).

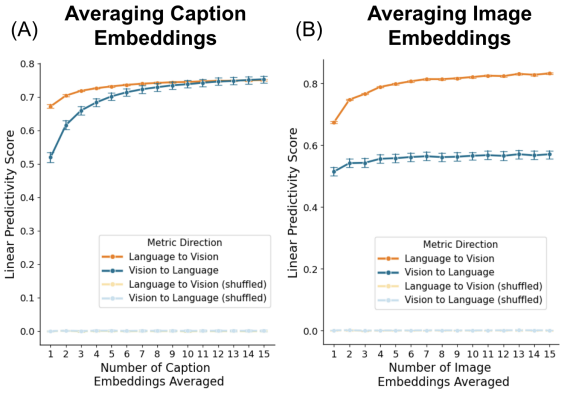

Figure 8: Effect of aggregation on alignment with a mismtach baseline.

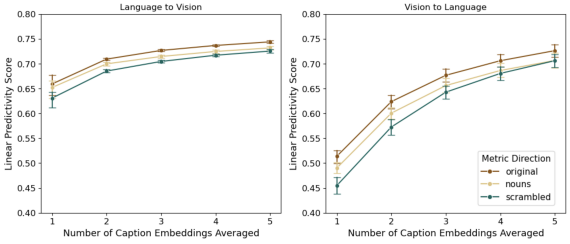

Figure 9: Effect of aggregation on alignment with manipulated captions which either only includes nouns or are scrambled in word order.