# Finding Unsupervised Alignment of Conceptual Systems in Image-Word Representations

**Kexin Luo**[1], **Bei Zhang**[1], **Yajie Xiao**[1], **Brenden M. Lake**[1,2]

[1]Center for Data Science, [2]Department of Psychology, New York University

{kl3108,bz2428,yx1750,brenden}@nyu.edu

## Abstract

Advancements in deep neural networks have led to significant progress in computer vision and natural language processing. These networks, trained on real-world stimuli, develop high-level feature representations of stimuli. It is hypothesized that these representations, stemming from different inputs, should converge into similar conceptual systems, as they reflect various perspectives of the same underlying reality. This paper examines the degree to which different conceptual systems can be aligned in an unsupervised manner, using feature-based representations from deep neural networks. Our investigation centers on the alignment between the image and word representations produced by diverse neural networks, emphasizing those trained via self-supervised learning methods. Subsequently, to probe comparable alignment patterns in human learning, we extend this examination to models trained on developmental headcam data from children. Our findings reveal a more pronounced alignment in models trained through self-supervised learning compared to supervised learning, effectively uncovering higher-level structural connections among categories. However, this alignment was notably absent in models trained with limited developmental headcam data, suggesting more data, more inductive biases, or more supervision are needed to establish alignment from realistic input.

**Keywords:** self-supervised learning; image-word representations; alignment; concept learning; developmental headcam data

## Introduction

Multimodal deep learning, particularly in computer vision and natural language processing, has made substantial advances in a variety of vision-language tasks (OpenAI et al., 2023; J. Li, Li, et al., 2023; Radford et al., 2021; F. Li et al., 2022). This progress, driven by multimodal models trained on aligned input data (e.g. images paired with corresponding captions), has spurred extensive research in cognitive science and machine learning, with a specific focus on the alignment between visual and linguistic modalities. In particular, studies have focused on examining whether the alignment observed in the joint models extends to independently-trained vision and language models. Utilizing representations from distinct visual and linguistic embedding spaces derived from supervised and unsupervised models, these investigations have revealed a high degree of convergence between visual features and linguistic representations across various contexts, including naturalistic object categories and everyday nouns (Sorscher et al., 2022; J. Li, Kementchedjhieva, & Søgaard, 2023), verbs (Y. Zhou et al., 2023), and perceptual adjectives like colors (Abdou et al., 2021). They demonstrate that, despite being trained separately and optimized based on different objectives, visual and language models still exhibit a notable alignment, and a straightforward linear mapping layer is sufficient to effectively transform representations between the visual and textual domains.

Given the impressive alignment achieved through linear mapping, an intriguing question arises: Is it possible to align the conceptual representation in two modalities, without any (even weakly) supervised data or supervised-trained linear projections between them? Roads and Love (2020) recently studied this question and introduced a method for unsupervised alignment. Using embedding spaces derived via unsupervised methods, they found that the alignment across these modalities can occur naturally, driven by inherent correlations among them, and without the need for a linear mapping. This suggests that each concept carries a unique, modality-independent signature, enabling it to be mirrored across different sensory systems and contributing to a substantial alignment naturally. Extending this exploration to humans, Aho, Roads, and Love (2023) demonstrated how such systems alignment could aid in the acquisition of early concepts by children, enhancing the learning of object names in the absence of supervision. This finding reveals that children's early concepts form dense networks ideally suited for systems alignment, underscoring the significance of alignment in multi-modal learning among children.

This study builds on the methodology proposed by Roads and Love (2020) to further investigate the alignment of visual and linguistic systems using unsupervised methods, adopting a paradigm relevant to the noisy, naturalistic, multi-modal input streams observed by children. Our first advancement involves the use of neural network embeddings extracted from raw, pixel-level images, a shift from the reliance on object co-occurrence labels as primarily analyzed by Roads and Love (2020). Furthermore, we extend our analysis across a variety of established vision models and training methodologies, placing a particular emphasis on self-supervised learning methods. The choice of self-supervised learning is key, as such models have not only excelled in vision tasks (Oquab et al., 2023; J. Zhou et al., 2022), but also offer a closer analogy to human learning experiences (Orhan & Lake, 2024; Konkle & Alvarez, 2022). Unlike supervised learning, which relies on synchronous label mapping from images to text, self-supervised learning extracts meaningful patterns from unlabeled data, better reflecting the label-sparse environments typical of children's everyday visual development.



In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.* 

Additionally, to mirror a child's real-life experiences as closely as possible, we test these methods using models trained from scratch on just a single child's egocentric developmental data from Sullivan et al. (2021) (Baby-S from the SAYCam corpus). Previous research has highlighted that unimodal deep models trained on this corpus of children's developmental data can generate meaningful image and text representations (Orhan & Lake, 2024; Wang et al., 2023). This approach probes whether conceptual alignment can emerge organically, without supervision or specific guidance in human learning. It also questions whether the two systems are inherently prepared for alignment or if certain learning prerequisites are essential for such alignment to manifest. This exploration not only enhances our understanding of deep learning models but also offers valuable insights into the developmental processes in humans.

In this study, we present consistent unsupervised alignment between image representations from both supervised and self-supervised computer vision models and common word embeddings. When aligned with the same word embedding space, self-supervised vision models show a higher degree of alignment, potentially due to their capacity to uncover hierarchical structural connections among concepts. Examining models trained on a single child's egocentric developmental data reveals a lower alignment degree compared to established models, suggesting prerequisites for multi-modal alignment from more realistic input.

## Methods

### Established Self-Supervised Models

In this section, we studied the potential for unsupervised alignment between the visual embeddings and word embeddings from pre-trained language and vision models[1].

**Vision Models** We included 7 computer vision models in our analysis, spanning three families of training methods: supervised Vision Transformer (ViT) (Dosovitskiy et al., 2021) and ResNeXt (Xie et al., 2017), self-supervised ViT and ResNet with DINO (Caron et al., 2021) and iBOT (J. Zhou et al., 2022), as well as ViT with CLIP (Radford et al., 2021). All ViTs selected are the 'Base' variant with 16×16 input patch size (with 86M parameters), and both ResNeXt and ResNet included have 50 layers (with 25M parameters). A randomly-initialized ViT is used as a reference of baselines analysis, while the CLIP model, trained on aligned data, is expected to demonstrate the highest alignment and thus serves as the comparative upper limit. Model details are in Table 1.

**Language Models** For our study, to match with the analysis by Roads and Love (2020), we primarily utilized GloVe (Global Vectors for Word Representation) (Pennington, Socher, & Manning, 2014) to obtain word embeddings. This unsupervised learning algorithm provides pre-trained word vectors, which we used for alignment with most of the vision models. We chose single-word embeddings to ensure straightforward concept alignment and to facilitate comparison with models trained on SAYCam-S. However, for models trained using CLIP (Radford et al., 2021), we incorporated the CLIP text encoder to align with the CLIP-based image representations.

| Model | Methods | Training Data |
|---|---|---|
| ViT-B/16 | Random | NA |
| ResNeXt-50 | Supervised | ImageNet (Russakovsky et al., 2015) |
| ViT-B/16 | Supervised | |
| ResNet-50 | SS: DINO | |
| ViT-B/16 | SS: DINO | |
| ViT-B/16 | SS: iBOT | |
| ViT-B/16 | CLIP | 400M Text-image pairs (Radford et al., 2021) |

Table 1: 7 computer vision models used in the analysis.

**Image Stimuli and Representations** For the visual stimuli in our evaluation, we selected images from the ImageNet-1k Dataset (Russakovsky et al., 2015). Our selection process began by defining four major domains for category selection: *Birds, Mammals, Vehicles, Fruits&Vegetables*. To facilitate accurate image-word mapping, we filtered the ImageNet categories to find overlaps with words in the GloVe embedding system. For categories named with multiple words, we considered the last word as the representative label (e.g., *'brown_bear'* as *'bear'*). We then manually assigned each category to its respective domain based on the WordNet hierarchy and selected 20 representative categories per domain.

For each category, we generated visual representations by embedding 100 images using the respective vision model. To verify that the quantity of images sampled did not influence our analysis, we conducted additional tests with larger samples of 500 and 1000 images for each category and found consistent results. ResNet and ResNeXt models produced a single 2048-dimensional vector per image. For ViT models, we used the 768-dimensional CLS token representation from each image. These vectors were then averaged to obtain a representative vector for each category. In the case of ViT-B/16 trained with CLIP, we extracted the CLS states before the projection layer, after passing both the image and text inputs through the model.

**Word Representations** For all models except those trained with CLIP, we employed the pre-trained 300-dimensional word vectors from GloVe, using Wikipedia 2014 and Gigaword 5 corpora (Pennington et al., 2014). Since GloVe provides a single vector for each word, we directly used these embeddings for our analysis. To obtain word embeddings from the CLIP text encoder, we followed the method introduced in Radford et al. (2021) by inputting a full prompt 'A photo of a {category}' together with the corresponding image and replaced {category} with the word. We then extracted the pooled states (EOS token representations), which serve as a summary of the entire input text, for use in our analysis.

[1]Code available here: https://github.com/cindyLuo99/image-word-alignment

### Models Trained Using SAYCam-S

In this section, we introduced the models trained using SAYCam-S, a dataset derived from Child-S's headcam data, and compare the alignment with those analyzed previously.

**Vision Models** We included two vision models: the ViT/B-16 and the ResNeXt-50, both trained using the DINO framework as detailed by Orhan and Lake (2024). These models were trained with visual inputs from Child-S's headcam data and have shown remarkable performance when fine-tuned for various vision tasks.

**Language Model** For the word embedding space, we utilized an LSTM model trained on transcribed child-directed speech from Child-S's headcam data. This model, described in the work by Wang et al. (2023), could generate word embeddings that capture meaningful semantic and syntactic structures.

**Labeled-S Dataset** To ensure the models generate meaningful representations and to mitigate out-of-distribution issues, our alignment evaluation employed the Labeled-S dataset, a curated subset of SAYCam data provided by Orhan and Lake (2024). This dataset includes 22 visual concepts that match with those in the word embedding space. Following the methodology used in the previous section, we obtained image embeddings by averaging the outputs from processing up to 100 images per category through each vision model. The 512-dimensional word embeddings were derived from the LSTM model's embedding space.

### Evaluations of Alignment

**Alignment Correlation** First, a similarity matrix was constructed for each embedding space by computing the cosine similarity between all pairs of concept embeddings within the same modality. Essentially, this matrix captures how closely related each pair of concepts is within its own space. We then quantified the alignment correlation between the two embedding systems using Spearman's rank correlation coefficient ($\rho$). This was done by correlating the upper triangular parts of the similarity matrices (excluding the diagonal). Since Spearman's correlation assesses rank correlation, it is robust against differences in the scales of similarity values.

**Alignment Strength** We used the Alignment Strength metric, as introduced by Roads and Love (2020), to assess how often misaligned mappings show lower correlation compared to the true mapping system. For each actual mapping, we calculated an alignment correlation value, $\rho^*$, using the Spearman correlation method mentioned earlier.

For each level of mapping accuracy, we generated up to 10,000 random mapping systems. We then calculated the alignment correlation for each of these misaligned systems to form a distribution of alignment correlations ($\rho_{\text{misaligned}}$). After running permutations across all levels of mapping accuracy, we determined the percentage of cases where $\rho^*$ exceeded $\rho_{\text{misaligned}}$. This percentage represents the alignment strength between the two systems. The script for this analysis was adapted from Roads and Love (2020).

**Accuracy Correlation** Following the approach of Roads and Love (2020), we also examined the relationship between mapping accuracy and the average alignment correlation of the conditionally sampled 10,000 misaligned systems. Accuracy correlation could reflect whether more accurate mappings lead to higher alignment.

**Recovery Accuracy** In addition, we explored the accuracy of recovering the true mapping by identifying which mapping system exhibited the highest alignment correlation among all systems generated in the previous analysis. This approach tests the effectiveness of using alignment correlation to guide the recovery of the true mapping. To minimize the influence of random seed, we calculated the mean and standard deviation of the recovery accuracy across 50 random seeds from the previous analysis.

## Results

### Alignment Analysis - Established Models

The alignment analysis results, using all 80 categories across four domains, are presented in Table 2. A distinct trend of increasing alignment correlation is evident, beginning with the untrained ViT model, which sets the baseline for our analysis. This is followed by the supervised ViT model. The self-supervised ViT models, utilizing both DINO and iBOT, demonstrated higher alignment correlations compared to their untrained and supervised counterparts. As hypothesized, the highest alignment was observed within the embedding spaces of CLIP, which signals the image-word alignment during the phase of model training. Additionally, variations across network architectures were notable. For instance, the supervised ResNeXt model displayed a relatively high alignment correlation, similar to some self-supervised models, whereas the alignment correlation of the ResNet-DINO was akin to that of the supervised models. This variation could potentially be attributed to differences in model performance in various vision tasks, as both ResNeXt and ViT models have been shown to achieve higher accuracies when employed as backbones for downstream vision tasks.

Alignment strengths closely mirror the observed alignment correlations. Models with higher alignment correlation consistently exhibit greater alignment strength, indicating a reduced likelihood of embedding space mismatches leading to inaccurately high alignment correlations. An exception is noted in the supervised ViT model; despite its lower alignment correlation relative to other models, it achieves a high alignment strength of 0.9994, suggesting robustness of the true mapping against random alternatives. The accuracy correlations also reinforce the presence of alignment, demonstrating that systems with a greater number of correctly mapped pairs tend to show higher alignment correlation.

Finally, the model's recovery accuracy consistently corresponds with its alignment strength. Models demonstrating

| Model | Alignment Correlation | Alignment Strength | Accuracy Correlation | Recovery Accuracy (%) |
|---|---|---|---|---|
| ViT-B/16 - Random | 0.272 | 0.9925 | 0.893 | 89.18 (± 5.06) |
| ResNeXt - Sup | 0.396 | 0.9990 | 0.938 | 94.88 (± 2.20) |
| ViT-B/16 - Sup | 0.312 | 0.9994 | 0.954 | 95.18 (± 1.62) |
| ResNet - DINO | 0.332 | 0.9964 | 0.898 | 91.73 (± 4.19) |
| ViT-B/16 - DINO | 0.478 | 0.9994 | 0.959 | 95.88 (± 1.86) |
| ViT-B/16 - iBOT | 0.486 | 0.9993 | 0.958 | 95.25 (± 1.29) |
| CLIP - GloVe | 0.646 | 0.9997 | 0.963 | 96.53 (± 0.80) |
| CLIP - CLIP | 0.648 | 0.9999 | 0.953 | 96.55 (± 1.29) |

Table 2: Results of the alignment analysis between image and word embedding spaces across all concepts. Except for the last row, all the word embedding spaces used in the analysis were from GloVe.

high alignment correlations and strengths tend to achieve high recovery accuracies, as exemplified by the self-supervised ViT models, each surpassing 95% recovery accuracy. This consistency underscores the effectiveness of high alignment in recovering accurate mapping in practical applications. Additionally, models with lower alignment correlations not only show reduced accuracy in recovering the true mapping but also exhibit greater variability in their effectiveness as a method for identifying the correct mapping system.

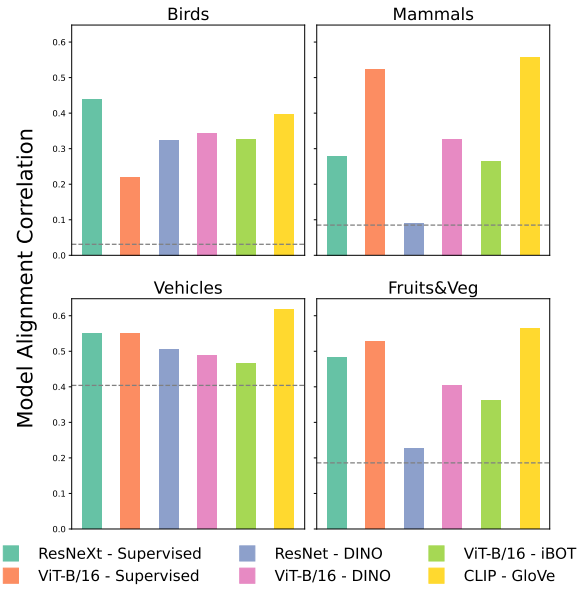



Figure 1: Within domain alignment correlations across models. The gray dotted line represents the alignment correlation obtained from the untrained ViT.

**Within Domain Alignment** Examining the alignment correlation within each domain reveals that the *vehicles* domain consistently exhibits a high degree of alignment across all models, as detailed in Figure 1. Remarkably, this pattern holds true even for the untrained ViT, which shows significant alignment within the *vehicles* domain.

Comparing within-domain alignment to overall concept alignment highlights that CLIP embeddings maintain strong alignment both within individual domains and across the entire concept system. In contrast, self-supervised models, while displaying somewhat weaker within-domain alignment as compared to other models, achieve high overall alignment. This suggests self-supervised models might excel at identifying cross-domain relationships, enhancing their general alignment level.

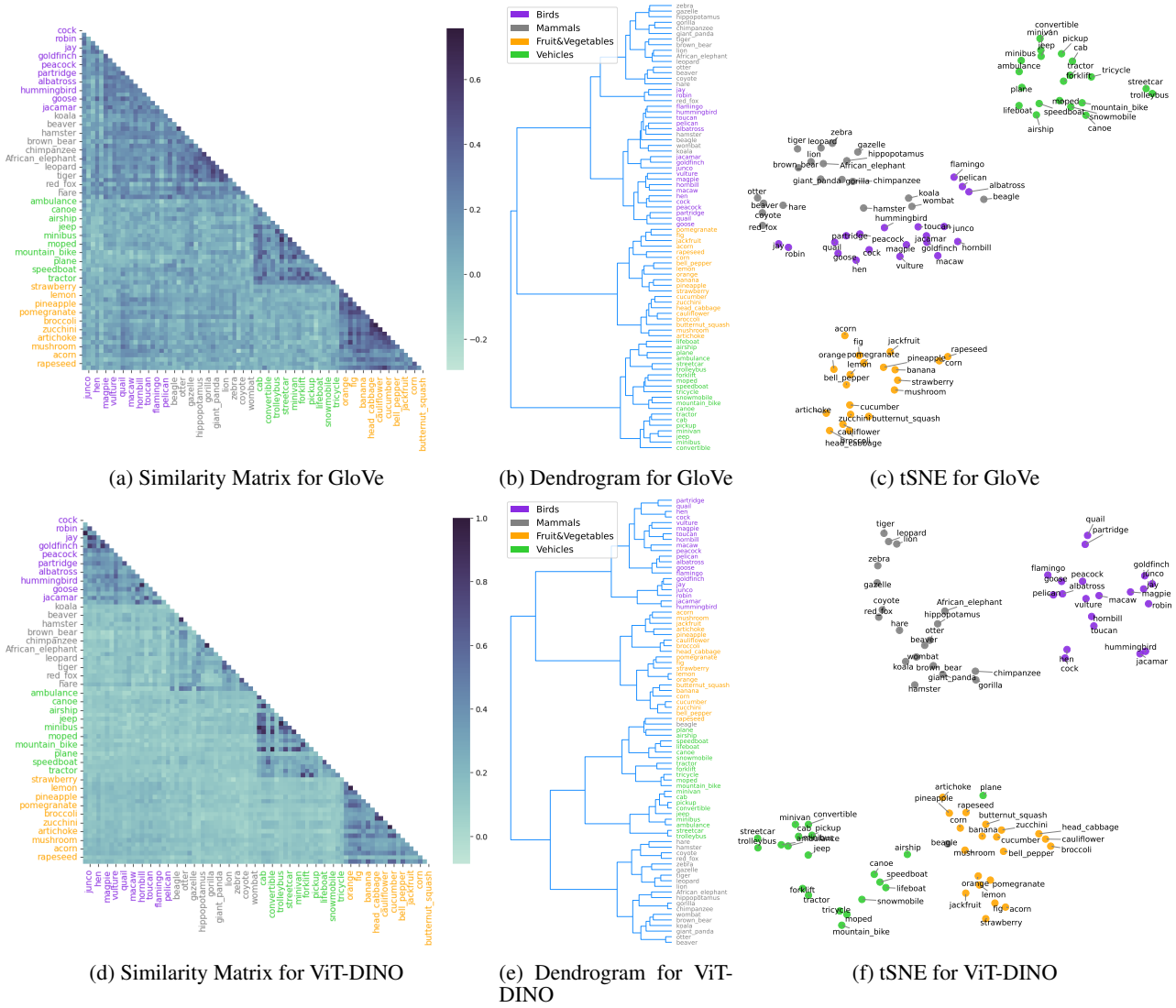
**Qualitative Analysis** As shown in Figure 2, a dendrogram based on GloVe word embeddings revealed structural insights into the word embedding space. The hierarchical clustering indicated four distinct domains (*Birds*, *Mammals*, *Fruits&Vegetables*, *Vehicles*). There was a notable mixing of the *Birds* and *Mammals* clusters (also suggested by the similarity matrix), reflecting their shared animal classification.

The t-SNE visualization of the word embeddings further supported these findings. It depicted intersecting clusters of *Mammals* and *Birds*, while clearly separating the *Vehicles* and *Fruits&Vegetables* domains.

Similar to the word embedding space, the clustering analysis of image embeddings, particularly using the ViT model trained with DINO, also revealed a coherent cluster structure. This may explain the high alignment correlation observed earlier and is indicative of the ability of self-supervised models to discern cross-domain structures. The clusters showed a high level of domain purity, with each primarily containing categories from the same domain. This result also suggests a nuanced understanding of higher-order structures by the vision model, even without explicit domain labels during training, based only on learned visual similarity through self-supervised learning.

## Alignment Analysis - SAYCam-S Models

The alignment correlations derived from image and word embeddings, obtained from models trained from scratch on SAYCam-S, are lower compared to those observed in established pre-trained models. To control for the effect of category size, we randomly sampled 22 categories from the 80-category dataset used in the previous section and conducted the same set of alignment analyses. The alignment correlation and strength levels of the pre-trained models on the 22-category subset were consistent as before, yet accuracy correlation and recovery accuracy dropped by 10-15% compared to the values reported in Table 2.




(a) Similarity Matrix for GloVe (b) Dendrogram for GloVe (c) tSNE for GloVe

(d) Similarity Matrix for ViT-DINO (e) Dendrogram for ViT-DINO (f) tSNE for ViT-DINO

Figure 2: Qualitative analysis of the structures in the image and word embedding spaces, extracted from GloVe and ViT-DINO.

As detailed in Table 3, image embedding spaces from both vision models trained on SAYCam-S each exhibit an alignment correlation of approximately 0.16 with the word embedding space (for comparison, Vong, Wang, Orhan, and Lake (2024)'s model trained on aligned SAYCam-S data achieves 0.37 on this test). This weak correlation, coupled with the stronger correlations for the larger-scale models in the previous sections, highlights the challenges of finding unsupervised alignment from SAYCam. Additionally, as indicated by the low average recovery accuracy and high standard deviation, alignment correlation is unable to guide accurate mapping between the systems.

Interestingly, substituting the LSTM word embeddings for the 22 concepts in the Labeled-S dataset with GloVe word embeddings yields an increase in both alignment correlation and strength, as well as recovery accuracy. These metrics reach levels comparable to those of established models, as outlined in Table 3. This enhancement suggests potential discrepancies between the embedding spaces of LSTM-SAYCam and GloVe. Specifically, it appears that GloVe embeddings encapsulate more intricate structures that align more closely with the visual space extracted from the SAYCam vision models.

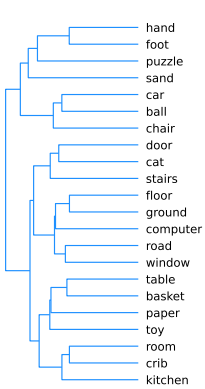
Our qualitative analysis of embedding spaces derived from the SAYCam models reveals some interpretable features within the word embedding space—for example, the close groupings of "hand" with "foot" and "floor" with "ground" as indicated in the dendrogram in Figure 3. However, it is challenging to discern meaningful hierarchical structures, either because of this limited set of categories (which were labeled in Orhan and Lake (2024)) or the inherently nosier representations in models trained from a single child's perspective.

## Discussion

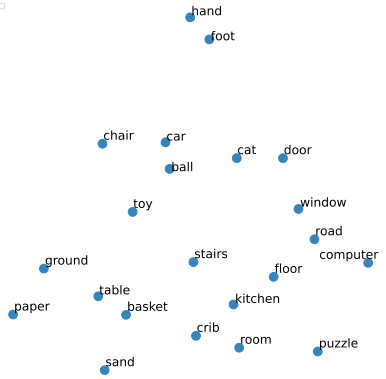
We highlight two contributions of our work. Firstly, by evaluating the unsupervised alignment across a range of estab-

| Vision Model | Language Model | Alignment Correlation | Alignment Strength | Accuracy Correlation | Recovery Accuracy (%) |
|---|---|---|---|---|---|
| ViT-B/16 - SAYCam | LSTM - SAYCam | 0.166 | 0.748 | 0.318 | 26.82 ($\pm$ 15.03) |
| ResNeXt - SAYCam | LSTM - SAYCam | 0.153 | 0.742 | 0.301 | 22.82 ($\pm$ 14.87) |
| ViT-B/16 - SAYCam | GloVe | 0.460 | 0.978 | 0.703 | 60.64 ($\pm$ 12.68) |
| ResNeXt - SAYCam | GloVe | 0.523 | 0.991 | 0.742 | 68.27 ($\pm$ 10.26) |

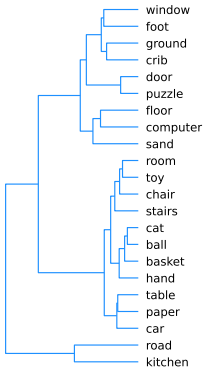
Table 3: Alignment Analysis between the image and word embedding spaces across 22 concepts in Labeled-S. Sources of the embedding spaces are illustrated in the table.



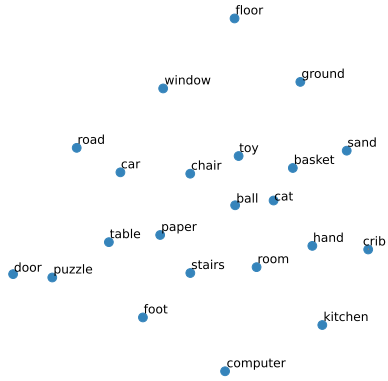

(a) Dendrogram for LSTM - SAYCam




(b) tSNE for LSTM - SAYCam




(c) Dendrogram for ViT - SAYCam




(d) tSNE for ViT - SAYCam

Figure 3: Qualitative analysis of the structures in the image and word embedding spaces, obtained from LSTM-SAYCam and ViT-SAYCam.

lished vision models and training methodologies, we validate the degree of conceptual system alignment between vision and language. Notably, self-supervised models demonstrate a higher alignment correlation compared to supervised models examined, whereas the robustness of such alignment against random alternatives is less dependent on training methods. Our results also demonstrate the effectiveness of high alignment correlation in guiding accurate mapping between systems. Secondly, our exploration of system alignment within the context of human multi-modal learning—using embedding spaces from models trained on a child's everyday experiences (SAYCam-S)—reveals a comparatively weak alignment. This observation suggests that more data, more inductive biases, or more supervision are needed to establish alignment from realistic input.

The observed variance in the degree of multi-modal alignment, particularly between SAYCam models and established models, may partly stem from the evaluation datasets underlying the image embeddings. The inherently noisier nature of the Labeled-S dataset poses challenges in generating representative image embeddings. Unlike ImageNet, where the focal object often dominates the image, Labeled-S images do not necessarily emphasize the labeled object, and SAYCam-trained models are quite sensitive to background features (Orhan & Lake, 2024). Additionally, the categories in Labeled-S are broader than the categories in ImageNet ("car" versus "jeep"), making it hard to form a representative image embedding for each category.

Our results not only highlighted the importance of the dataset characteristics used in alignment analysis, they also underscore the significance of the data used to train the models when forming the embedding spaces on the first place. The weak alignment among SAYCam models could also be related to its limited training data, as compared to the large dataset used to train those established vision and language models, hinting at the necessity for more comprehensive data to support unsupervised multi-modal alignment. Although SAYCam captures everyday learning experiences, it represents only a fraction of our broader experiences, raising the question of whether a more diverse and extensive sampling of such input could bolster alignment.

In summary, our research outlines the necessary conditions for successful unsupervised multi-modal alignment. Given the challenges of unsupervised alignment from a child's egocentric input, a key role for (even weakly and sparsely) supervised learning examples shouldn't be dismissed. This study also showed how many factors can influence alignment success, including type and amount of data, model architecture, and training methodology. We hope these findings will usefully guide future work on multi-modal alignment and how it arises in cognitive development.

## Acknowledgments

We thank Wai Keen Vong, Yanli Zhou, Yulu Qin, Wentao Wang, and Guy Davidson from the Human and Machine Learning Lab at New York University for their helpful feedback on earlier versions of this manuscript and their insightful discussions regarding this project. Additionally, we acknowl-

edge the assistance of Emin Orhan and Wentao Wang for their guidance in accessing their SAYCam-S models (Orhan & Lake, 2024; Wang et al., 2023), which have been crucial to the success of this project. We are also grateful for the SAYCam dataset (Sullivan et al., 2021) which made our work possible.

## References

Abdou, M., Kulmizev, A., Hershcovich, D., Frank, S., Pavlick, E., & Søgaard, A. (2021, November). Can language models encode perceptual structure without grounding? a case study in color. In A. Bisazza & O. Abend (Eds.), *Proceedings of the 25th conference on computational natural language learning* (pp. 109–132). Online: Association for Computational Linguistics. doi: 10.18653/v1/2021.conll-1.9

Aho, K., Roads, B. D., & Love, B. C. (2023). Signatures of cross-modal alignment in children's early concepts. *Proceedings of the National Academy of Sciences*, *120*(42), e2309688120. doi: 10.1073/pnas.2309688120

Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *2021 ieee/cvf international conference on computer vision (iccv)* (p. 9630-9640). doi: 10.1109/ICCV48922.2021.00951

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International conference on learning representations.*

Konkle, T., & Alvarez, G. A. (2022, January). A self-supervised domain-general learning framework for human ventral stream representation. *Nature Communications*, *13*(1), 491. doi: 10.1038/s41467-022-28091-4

Li, F., Zhang, H., Zhang, Y.-F., Liu, S., Guo, J., Ni, L. M., ... Zhang, L. (2022). *Vision-language intelligence: Tasks, representation learning, and large models.*

Li, J., Kementchedjhieva, Y., & Søgaard, A. (2023). *Implications of the convergence of language and vision model geometries.*

Li, J., Li, D., Savarese, S., & Hoi, S. (2023). Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th international conference on machine learning.* JMLR.org.

OpenAI, :, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., ... Zoph, B. (2023). *Gpt-4 technical report.*

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

Orhan, A. E., & Lake, B. M. (2024, March). Learning high-level visual representations from a child's perspective without strong inductive biases. *Nature Machine Intelligence*, *6*(3), 271–283.

Pennington, J., Socher, R., & Manning, C. (2014, October). GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). Doha, Qatar: Association for Computational Linguistics. doi: 10.3115/v1/D14-1162

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... Sutskever, I. (2021). *Learning transferable visual models from natural language supervision.*

Roads, B. D., & Love, B. C. (2020). Learning as the unsupervised alignment of conceptual systems. *Nature Machine Intelligence*, *2*(1), 76–82.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, *115*(3), 211-252. doi: 10.1007/s11263-015-0816-y

Sorscher, B., Ganguli, S., & Sompolinsky, H. (2022). Neural representational geometry underlies few-shot concept learning. *Proceedings of the National Academy of Sciences*, *119*(43), e2200800119. doi: 10.1073/pnas.2200800119

Sullivan, J., Mei, M., Perfors, A., Wojcik, E., & Frank, M. (2021, 03). Saycam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. *Open Mind*, *5*, 1-10. doi: 10.1162/opmi_a_00039

Vong, W. K., Wang, W., Orhan, A. E., & Lake, B. M. (2024). Grounded language acquisition through the eyes and ears of a single child. *Science*, *383*(6682), 504-511. doi: 10.1126/science.adi1374

Wang, W., Vong, W. K., Kim, N., & Lake, B. M. (2023). Finding structure in one child's linguistic experience. *Cognitive Science*, *47*(6), e13305. doi: https://doi.org/10.1111/cogs.13305

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In *2017 ieee conference on computer vision and pattern recognition (cvpr)* (p. 5987-5995). doi: 10.1109/CVPR.2017.634

Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., & Kong, T. (2022). *ibot: Image bert pre-training with online tokenizer.*

Zhou, Y., Tarr, M. J., & Yurovsky, D. (2023). *Quantifying the roles of visual, linguistic, and visual-linguistic complexity in verb acquisition.*