Representation Potentials of Foundation Models for Multimodal Alignment: A Survey

Jianglin Lu[1*], **Hailing Wang**[1], **Yi Xu**[1], **Yizhou Wang**[1], **Kuo Yang**[1], **Yun Fu**[1,2]
[1]Department of Electrical and Computer Engineering, Northeastern University
[2]Khoury College of Computer Science, Northeastern University
Resouce: https://github.com/Jianglin954/Representation-Alignment-Survey

Abstract

Foundation models learn highly transferable representations through large-scale pretraining on diverse data. An increasing body of research indicates that these representations exhibit a remarkable degree of similarity across architectures and modalities. In this survey, we investigate the *representation potentials* of foundation models, defined as the latent capacity of their learned representations to capture task-specific information within a single modality while also providing a transferable basis for alignment and unification across modalities. We begin by reviewing representative foundation models and the key metrics that make alignment measurable. We then synthesize empirical evidence of representation potentials from studies in vision, language, speech, multimodality, and neuroscience. The evidence suggests that foundation models often exhibit structural regularities and semantic consistencies in their representation spaces, positioning them as strong candidates for cross-modal transfer and alignment. We further analyze the key factors that foster representation potentials, discuss open questions, and highlight potential challenges.

1 Introduction

Foundation models, trained through large-scale pretraining on vast and heterogeneous data, have driven remarkable progress and significantly accelerated the pursuit of artificial general intelligence (Bommasani et al., 2021; Cui et al., 2022; Firoozi et al., 2023; Azad et al., 2023; Zhou et al., 2024). By acquiring highly transferable and general-purpose representations, they have become the backbone of a wide spectrum of applications, spanning natural language processing (Liu et al., 2019; He et al., 2020; Rajendran et al., 2024), computer vision (Dosovitskiy et al., 2021; Liu et al., 2022; Woo et al., 2023; Siméoni et al., 2025), speech processing (Belinkov and Glass, 2017; Baevski et al., 2020; Radford et al., 2023), robotics (Brohan et al., 2022; Team et al., 2025), and medical domains (Moor et al., 2023; Huang et al., 2024; Khan et al., 2025).

A growing body of research has shown that the representations learned by foundation models are not only powerful in isolation but also exhibit strong similarity across architectures, training objectives, and even modalities (Wentworth, 2021; Ng et al., 2023; Liu et al., 2023; Sharma et al., 2024; Huh et al., 2024; Maniparambil et al., 2024; Wang et al., 2025). We refer to this capacity as the *representation potential* of foundation models. This perspective carries significant implications: if foundation models naturally converge toward shared representational structures, they may approximate modality-agnostic abstractions and encode common statistical regularities of the world, even without explicit alignment. Understanding these potentials is essential, not only for advancing scientific theories of representation learning but also for enabling practical benefits such as model interoperability, transferability, interpretability, and alignment with biological cognition.

In this survey, we focus on the representation potentials of unimodal foundation models, with the goal of assessing their capacity for alignment. We structure our discussion around four central themes. First, we introduce representative foundation models in vision, language, speech, and multimodality. Second, we review the metrics that make representation alignment measurable, including centered kernel alignment (Kornblith et al., 2019), canonical correlation analysis (Morcos et al., 2018), and mutual nearest neighbors (Haghverdi et al., 2018). Third, we synthesize empirical evidence for representation potentials, drawing from studies in vision, language, speech, cross-modal alignment, and neuroscience. Fourth, we analyze the key factors that drive representation potentials, such as scale, architectural inductive biases, training objectives, and

[*]Corresponding author: jianglinlu@outlook.com.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 16669–16684
November 4-9, 2025 ©2025 Association for Computational Linguistics

task and instruction diversity. Alongside these advances, we highlight pressing open questions: the limits of convergence across modalities, the need for robust evaluation standards, the influence of bias and sociotechnical context, and cases where domain-specific divergence may arise.

The remainder of this survey is organized as follows. In Section 2, we introduce foundation models across modalities. Section 3 reviews major metrics for quantifying representation similarity and alignment. Section 4 presents evidence for representation potentials in vision, language, speech, cross-modal, and neuroscience contexts. Section 5 analyzes the underlying drivers of alignment, including scale, architectures, training paradigms, and tasks. Section 6 discusses open questions and challenges. In Section 7, we conclude with key insights and directions for future research.

2 Foundation Models

This section provides a general definition of foundation models and then presents representative examples across computer vision, natural language processing, speech and multimodal domains.

2.1 Definition

Bommasani et al. (2021) first introduced the term *foundation model* to describe machine learning models trained on vast and diverse datasets, typically with large-scale self-supervision, that can be applied to a broad range of downstream tasks. Three features distinguish foundation models from their earlier predecessors: ① *Broad data*: They are trained on extensive and diverse datasets, often collected at web scale, which provide robust and transferable representations. ② *Self-supervision*: They learn directly from raw, unlabeled data by predicting missing information or inherent structures, thus avoiding the reliance on large volumes of manually annotated data. ③ *Adaptability*: Once trained, they can be fine-tuned, prompted, or otherwise adapted to a wide range of downstream tasks, underscores their general-purpose nature and their ability to serve as a foundation for numerous specialized applications. Based on these characteristics, the following subsections briefly introduce representative foundation models in computer vision, language, speech, and multimodal learning.

2.2 Vision Foundation Models

Vision foundation models (VFMs) (Liu et al., 2022; Siméoni et al., 2025) are large-scale neural architectures designed to learn robust visual representations that transfer across tasks. Canonical examples include ResNet (He et al., 2016), Vision Transformer (ViT) (Dosovitskiy et al., 2021), ConvNeXt (Woo et al., 2023), Dinov2 (Oquab et al., 2023), and the Segment Anything Model (SAM) (Kirillov et al., 2023). VFMs are typically trained on billion-scale image datasets using self-supervised learning, weakly supervised signals, or multimodal objectives. Earlier vision models depended on task-specific annotated datasets, but VFMs provide universal feature embeddings that can be reused or lightly adapted. These features support a wide range of applications, including image classification, object detection, and segmentation, as well as higher-level reasoning tasks such as visual question answering and captioning. Recently, VFMs have also become essential in vision systems such as segmentation-anything frameworks (Kirillov et al., 2023; Ravi et al., 2025), image generative pipelines (Yang et al., 2023; Zhang et al., 2023), and world models (Ha and Schmidhuber, 2018; Zhou et al., 2025). This transition marks a shift in computer vision from narrowly specialized solutions to foundational infrastructures.

2.3 Large Language Models

Large language models (LLMs) (Radford et al., 2019; Touvron et al., 2023; Achiam et al., 2023) are trained on massive text corpora to acquire broad linguistic and semantic knowledge. Representative examples include BERT (Devlin et al., 2019), T5 (Raffel et al., 2020), Qwen (Bai et al., 2023), and the LLaMA-series (Grattafiori et al., 2024), and conversational agents such as ChatGPT (Brown et al., 2020). By predicting masked spans or next tokens, LLMs capture both syntactic structures and semantic relations that generalize across diverse tasks. They can be adapted through fine-tuning, prompting, or in-context learning to applications such as summarization, translation, question answering, reasoning, and dialogue. A defining feature is their scale: empirical studies show that performance improves predictably as the number of parameters, the volume of training data, and the compute budget increase (Kaplan et al., 2020). Beyond higher accuracy, larger models also exhibit emergent abilities that are absent in smaller counterparts (Wei et al., 2022). These findings have shaped both research practices and industrial applications, positioning LLMs as the foundation for multimodal (Yin et al., 2024) and agentic extensions (Li, 2025).

2.4 Speech Foundation Models

Speech foundation models (SFMs) are trained on extensive audio corpora, focusing on speech signals while learning both acoustic-level and linguistic-level abstractions. Representative examples include wav2vec (Schneider et al., 2019; Baevski et al., 2020), HuBERT (Hsu et al., 2021), WavLM (Chen et al., 2022a), Whisper (Radford et al., 2023), and SeamlessM4T (Barrault et al., 2023). These models learn by predicting masked or latent units from raw waveforms, enabling them to serve as universal encoders of speech. SFMs support a wide variety of tasks, ranging from automatic speech recognition and speaker verification to emotion recognition, speech translation, and speech-to-speech generation (Cui et al., 2024; Radford et al., 2023). Beyond supervised fine-tuning, SFMs also demonstrate robust performance in zero-shot and few-shot settings. As a result, SFMs mark a decisive transition from specialized pipelines to broadly adaptable foundation-level architectures.

2.5 Multimodal Foundation Models

Multimodal foundation models (MFMs) integrate signals from multiple modalities, such as vision, language, audio, and video, into a unified architecture. Early MFMs include CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), which use image–text contrastive learning. Later developments such as BLIP (Li et al., 2022) and CoCa (Yu et al., 2022) focus on vision–language generation, while Flamingo (Alayrac et al., 2022) and PaLI (Chen et al., 2022b) advance few-shot multimodal reasoning. More recent large-scale systems such as GPT-4 (Achiam et al., 2023) and Gemini (Team et al., 2023) demonstrate general-purpose multimodal intelligence. Training MFMs typically combines contrastive alignment (for paired modalities such as image and text), cross-modal reconstruction (predicting one modality from another), masked modeling (learning contextual cross-modal embeddings), and instruction tuning (adapting multimodal reasoning to natural language instructions) (Yin et al., 2024). These strategies allow MFMs to generalize across perception tasks such as image captioning and speech-to-text translation, as well as reasoning tasks including visual question answering and multimodal dialogue. Their growing influence signals an important trend: foundation models are evolving from unimodal encoders into integrated multimodal systems that support increasingly rich forms of human–AI interaction.

3 Metrics for Representation Alignment

In this survey, we review the capacity of unimodal foundation models to achieve representation alignment, with a focus on how their learned representations behave across architectures and modalities. The central goal is to uncover potential commonalities and similarities among representations learned by different models and to evaluate the extent to which unimodal foundation models converge in their representation spaces.

Formally, let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \in \mathbb{R}^{n \times d_1}$ and $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\} \in \mathbb{R}^{n \times d_2}$ denote two sets of representations, extracted from distinct neural network layers or from different foundation models. Here, n is the number of samples, and d_1 and d_2 are the feature dimensionalities. The central question is whether $\mathbf{X}$ and $\mathbf{Y}$ encode similar information, possibly up to admissible transformations such as rotation, scaling, or affine mapping. To address this question, we first review representative similarity metrics that have been widely adopted in representation alignment analysis. These metrics provide principled tools for quantifying alignment quality.

3.1 Centered Kernel Alignment

Centered kernel alignment (CKA) (Kornblith et al., 2019; Davari et al., 2023) compares two representation sets by measuring the similarity of their kernel (Gram) matrices, which capture pairwise relationships between samples. We denote $\mathbf{K} = \mathbf{X}\mathbf{X}^\top$ and $\mathbf{L} = \mathbf{Y}\mathbf{Y}^\top$ as the linear Gram (kernel) matrices of $\mathbf{X}$ and $\mathbf{Y}$, which represent inner products between samples in the respective feature spaces. Typically, CKA first centers both kernel matrices to remove the influence of mean offsets:

$$\tilde{\mathbf{K}} = \mathbf{HKH}, \; \tilde{\mathbf{L}} = \mathbf{HLH}. \tag{1}$$

where $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ denotes the centering matrix, $\mathbf{I}_n \in \mathbb{R}^{n\times n}$ is the identity matrix, and $\mathbf{1}_n \in \mathbb{R}^n$ denotes the all-ones column vector. Then, the linear CKA between $\mathbf{X}$ and $\mathbf{Y}$ can be defined as:

$$\mathrm{CKA}(\mathbf{X}, \mathbf{Y}) = \frac{\mathtt{HSIC}(\mathbf{K}, \mathbf{L})}{\sqrt{\mathtt{HSIC}(\mathbf{K}, \mathbf{K})\mathtt{HSIC}(\mathbf{L}, \mathbf{L})}}, \tag{2}$$

where $\mathtt{HSIC}(\cdot, \cdot)$ denotes the Hilbert-Schmidt Independence Criterion (HSIC) that measures the dependence between the two kernel spaces:

$$\mathtt{HSIC}(\mathbf{K}, \mathbf{L}) = \mathrm{tr}(\tilde{\mathbf{K}}\tilde{\mathbf{L}}). \tag{3}$$

CKA normalizes HSIC to produce a scale-invariant similarity measure. It can be interpreted as the cosine of the angle between the centered kernel matrices $\tilde{K}$ and $\tilde{L}$, when viewed as elements in the Hilbert-Schmidt space. It is invariant to isotropic scaling, i.e., $\mathrm{CKA}(c\mathbf{X}, \mathbf{Y}) = \mathrm{CKA}(\mathbf{X}, \mathbf{Y})$ for any scalar $c \neq 0$, and to orthogonal transformations, i.e., $\mathrm{CKA}(\mathbf{XQ}, \mathbf{Y}) = \mathrm{CKA}(\mathbf{X}, \mathbf{Y})$ for any orthogonal matrix $\mathbf{Q}$. The resulting CKA score ranges from 0 to 1, with 1 indicating perfect alignment between $\mathbf{X}$ and $\mathbf{Y}$. Based on CKA, several variants have been proposed to enhance its robustness or adapt it to specific settings, such as unbiased CKA (Song et al., 2007), kernel CKA (Kornblith et al., 2019), and class-conditional CKA (Nguyen et al., 2021). These refinements have made CKA one of the most widely used metrics for comparing neural representations across architectures and training settings.

3.2 Canonical Correlation Analysis

Canonical correlation analysis (CCA) (Hotelling, 1936; Morcos et al., 2018) is a classical statistical technique that identifies linear relationships between two multivariate datasets. It seeks linear projections of two random vectors such that their resulting projected representations are maximally correlated. Assume that both input matrices $\mathbf{X}$ and $\mathbf{Y}$ are centered (i.e., each column has zero mean). CCA aims to find directions $\mathbf{a} \in \mathbb{R}^{d_1}$ and $\mathbf{b} \in \mathbb{R}^{d_2}$ such that the projections $\mathbf{Xa}$ and $\mathbf{Yb}$ are maximally correlated. This is formalized as:

$$\max_{\mathbf{a},\mathbf{b}} \rho = \frac{\mathbf{a}^\top \mathbf{C}_{XY} \mathbf{b}}{\sqrt{\mathbf{a}^\top \mathbf{C}_{XX} \mathbf{a}} \cdot \sqrt{\mathbf{b}^\top \mathbf{C}_{YY} \mathbf{b}}} \tag{4}$$

where $\mathbf{C}_{XX} = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$ and $\mathbf{C}_{YY} = \frac{1}{n}\mathbf{Y}^\top\mathbf{Y}$ are the covariance matrices of $\mathbf{X}$ and $\mathbf{Y}$, respectively, and $\mathbf{C}_{XY} = \frac{1}{n}\mathbf{X}^\top\mathbf{Y}$ is the corresponding cross-covariance matrix. The solution yields the first pair of canonical variates:

$$\mathbf{u}_1 = \mathbf{Xa}_1, \quad \mathbf{v}_1 = \mathbf{Yb}_1, \tag{5}$$

where $\rho_1 = \mathrm{corr}(\mathbf{u}_1, \mathbf{v}_1)$ is the largest (first) canonical correlation. Subsequent pairs $(\mathbf{a}_i, \mathbf{b}_i)$ can be derived in a similar manner, subject to the orthogonality constraints:

$$\mathbf{u}_i^\top \mathbf{u}_j = 0, \quad \mathbf{v}_i^\top \mathbf{v}_j = 0 \quad \text{for} \quad \forall i \neq j. \tag{6}$$

The number of nonzero canonical correlations is at most $r = \min(d_1, d_2)$, and the sequence $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_r \geq 0$ quantifies the strength of the linear relationship between $\mathbf{X}$ and $\mathbf{Y}$.

An important extension is singular vector CCA (SVCCA) (Raghu et al., 2017a; Artetxe et al., 2020), which first reduces both $\mathbf{X}$ and $\mathbf{Y}$ to their dominant subspaces using singular value decomposition and then applies CCA. This improves robustness to noise and has become a standard technique for comparing deep learning representations. Both CCA and SVCCA are invariant to affine transformations, while SVCCA further filters out low-variance directions, improving stability in practice.

3.3 Mutual Nearest Neighbors

Mutual nearest neighbors (MNN) (Haghverdi et al., 2018) define a symmetric relationship between samples from two sets and are commonly used to establish robust correspondences between learned representations. Let $\mathtt{NN}_k(\mathbf{x}_i; \mathbf{Y})$ denote the set of k nearest neighbors of $\mathbf{x}_i \in \mathbf{X}$ in $\mathbf{Y}$, measured under a specific distance metric. A pair $(\mathbf{x}_i, \mathbf{y}_j)$ is said to form a mutual nearest neighbor pair if:

$$\mathbf{y}_j \in \mathtt{NN}_k(\mathbf{x}_i; \mathbf{Y}) \quad \text{and} \quad \mathbf{x}_i \in \mathtt{NN}_k(\mathbf{y}_j; \mathbf{X}). \tag{7}$$

MNN is widely used to reduce false-positive matches, particularly in high-dimensional or noisy representation spaces. Several variants have been developed based on the MNN principle. For example, mutual k-nearest neighbor matching (Huh et al., 2024) has been applied to evaluate cross-modal representation alignment.

The selection of metric depends on the specific aspect of similarity one seeks to capture. In practice, CKA is often preferred for its robustness and interpretability across architectures and tasks, whereas CCA provides useful insights when comparing closely related spaces, and MNN proves valuable when local semantic structures are of interest. Beyond CKA, CCA, and MNN, a number of additional metrics have been developed to capture complementary aspects of representation similarity. Examples include Riemannian distance (Shahbazi et al., 2021), which accounts for the geometry of covariance matrices; similarity-of-similarity matrices (SSM) (Diedrichsen and Kriegeskorte, 2017), which measure agreement in pairwise similarity structures; and rank-based or Jaccard similarity metrics (Wang et al., 2020), which focus on relational consistency. For a systematic overview of these approaches, refer to the paper by Klabunde et al. (2025), which provides a detailed survey of similarity metrics for representation analysis.

4 Representation Potentials of Foundation Models for Alignment

In the following subsections, we review existing works in vision, language, speech, modalities, and neuroscience that explore the representation potentials of foundation models for alignment.

4.1 Representation Alignment in Vision

Within computer vision, a growing body of evidence suggests that models with different architectures, training objectives, and datasets can develop compatible understandings of visual information. Early studies established that shallow features and early convolutional layers behave in similar ways. For example, Lenc and Vedaldi (2015) demonstrated that representations such as histograms of oriented gradients (HOG) and early convolutional filters respond linearly to geometric transformations like warps and flips, revealing that these early features are broadly interchangeable across architectures. Li et al. (2015) showed that independently trained networks often develop neuron clusters with overlapping functions, indicating partial convergence in learned representations. Raghu et al. (2017b) introduced SVCCA and reported that neural representations exhibit strong cross-initialization similarity, with lower layers converging early into compact shared subspaces while higher layers continue to evolve more gradually. Morcos et al. (2018) found that networks with better generalization exhibit higher representation similarity across random initializations, while overfitted networks diverge more. Kornblith et al. (2019) provided systematic evidence that wider models learn more similar representations, early layers converge quickly, and deeper layers often contain redundancies across consecutive layers.

Subsequent works examined alignment under varied training objectives and architectures. Csiszárik et al. (2021) showed that inner representations in deep convolutional networks with identical architectures but different initializations can be closely matched using only a single affine stitching layer. Roeder et al. (2021) proved that a broad class of discriminative and autoregressive models are identifiable in function space up to a linear transformation. Grigg et al. (2021) compared supervised and self-supervised training, finding that intermediate layers are strikingly similar across paradigms, but final layers diverge: supervised models emphasize class-specific structure, whereas self-supervised models emphasize invariance to augmentations. Bansal et al. (2021) introduced the concept of stitching connectivity, showing that identically structured networks trained in different ways can be stitched together at intermediate layers with minimal performance degradation. Caron et al. (2021) highlighted how self-supervised Vision Transformers (ViTs) consistently converge to similar spatial attention patterns and semantic structures, regardless of the specific training setup. Raghu et al. (2021) compared CNNs and ViTs using CKA, finding divergence in early layers but convergence in later ones. Moschella et al. (2023) noted that although absolute coordinates of latent embeddings vary across training runs, relative angular relationships are preserved, reflecting alignment at a relational level.

More recent studies have reinforced and expanded these findings. Shekhar et al. (2023) reported that models trained with the same self-supervised objective tend to learn more similar representations, even when the architectures differ. Oquab et al. (2023) showed that self-supervised ViTs trained on different datasets or initializations learn similar high-level visual structures and that their features are compatible with those of supervised models. Dravid et al. (2023) identified Rosetta neurons, which reliably emerge across model architectures, training paradigms, and tasks. Stoica et al. (2024) demonstrated that independently trained networks can be merged without retraining by aligning and zipping their feature spaces. Sharon and Dar (2025) showed that during training, representation similarity exhibits distinct phases whose clarity depends on both architecture and optimizer: SGD and ViTs exhibit more synchronized, sharply delineated evolution of layer representations, whereas ResNets and Adam yield more gradual or less aligned dynamics. Li et al. (2025) highlighted that bidirectional Transformers serve as strong representation learners, enabling unified modeling of multimodal data distributions through likelihood estimation.

Beyond classification, alignment has also been observed in generative contexts. Yu et al. (2025) argued that the success of diffusion-based generation hinges on learning meaningful representations, and proposed a regularization strategy that aligns noisy denoising states with clean embeddings from pretrained encoders. Moreover, several studies (Balestriero and richard baraniuk, 2018; Kornblith et al., 2019; Roeder et al., 2021; Huh et al.,

2024) converge on a broader trend: representation similarity increases with model scale and performance. In other words, as vision models become larger, more expressive, and more generalizable, their internal representations tend to align more closely, pointing to a fundamental convergence.

4.2 Representation Alignment in Language

LLMs are increasingly demonstrating human-level proficiency across a broad spectrum of natural language processing tasks, including knowledge extraction, reasoning, and dialogue. This widespread improvement suggests a potential convergence in how these models process and represent linguistic information. A growing body of research provides evidence that diverse LLMs, often trained with different architectures and datasets, nonetheless develop aligned internal representations. One line of work highlights consistent structural patterns. Phang et al. (2021) found a block-structured similarity pattern in the hidden representations of fine-tuned RoBERTa (Liu et al., 2019) and ALBERT (Lan et al., 2020), suggesting that training induces stable, repeatable alignment across models. Jiang et al. (2025) similarly observed that representation similarity in transformer models is strongest between adjacent layers, pointing to a layerwise convergence mechanism.

Another direction emphasizes concept-level alignment. Park et al. (2024) formalized what it means for high-level concepts to be linearly represented in LLMs, introduced a causal inner product to capture semantic separability, and showed that high-level concepts in LLaMA2 (Touvron et al., 2023) can be probed or steered as approximate linear directions. Lan et al. (2024) decomposed LLM activations with sparse autoencoders, revealing disentangled features that align closely across different models. Bürger et al. (2024) reported the emergence of a universal two-dimensional truth representation across LLMs of varying sizes and architectures, while Tan et al. (2024) identified strong correlations in both in-distribution and out-of-distribution steerability between LLaMA (Touvron et al., 2023) and Qwen (Bai et al., 2023).

A third body of evidence focuses on transferable features and universal neurons. Del and Fishel (2022) proposed a neuron-wise correlation metric that reveals the "first align, then predict" pattern across languages in multilingual models more faithfully. Gurnee et al. (2024) showed that about one to five percent of neurons in independently seeded GPT2 models are universal, interpretable, and causally relevant for predictions. Oozeer et al. (2025) found that safety-intervention vectors discovered in the activation space of one LLM can be mapped into the activation spaces of other LLMs via learned autoencoder mappings. Chen et al. (2025) find that affine mappings between residual streams allow for effective transfer of learned feature modules, probes, and steering vectors from small to large models. Rinaldi et al. (2025) further demonstrated that task vectors can be transferred from older to newer models without data or retraining by aligning weight structures of the two pretrained models. Lee et al. (2025) showed that token embeddings within a model family share both global and local geometry, enabling cross-model steering despite dimensional differences.

Finally, several studies point to broader universality across model classes. Wang et al. (2025) compared Transformer and Mamba models (Gu and Dao, 2024) trained on the same data and found that they share many internal features and circuits, suggesting substantial but imperfect universality of mechanisms across architectures. Cheng et al. (2025) analyzed intrinsic dimensionality of internal representations in transformer-based language models and identify a high-dimensional abstraction phase within the middle layers. This phase, consistently observed across architectures and datasets, reflects the point where models begin to form abstract, task-relevant representations that generalize well across tasks and models.

4.3 Representation Alignment in Speech

The study of speech models, particularly those trained using self-supervised learning (SSL) (Mohamed et al., 2022; Riera et al., 2023), also reveals emerging evidence of representation alignment. For instance, Ollerenshaw et al. (2021) examined end-to-end automatic speech recognition systems and found that CNN-based models exhibited progressively hierarchical and stable representation similarity as depth increased, whereas LSTM and Transformer architectures displayed less clean or more irregular similarity structures across layers. Chung et al. (2021) reported that the choice of learning objective has a larger impact on how similar representations are across models than architectural choices. Pasad et al. (2023, 2024) provided a detailed analysis of how acoustic, phonetic, and word-level information emerge at different layers, showing that both pretraining objectives and model size determine where key linguistic properties such as identity, pronunciation, syntax, and semantics are encoded. Waheed et al. (2024) showed that certain SFMs achieve strong zero-shot performance on tasks for which they were never explicitly trained, and that this performance correlates with higher-quality underlying representations. Dorszewski et al. (2025) found that Transformer-based speech representation models exhibit a block-structured similarity pattern across layers, with substantial redundancy within blocks. Huo and Dunbar (2025) demonstrated that the contrast between HuBERT (Hsu et al., 2021) and wav2vec2.0 (Baevski et al., 2020) lies not in whether a contrastive or classification objective is used, but in the iterative pseudo-label refinement strategy: multiple clustering iterations yield more aligned representations.

4.4 Representation Alignment Across Modalities

Beyond alignment within individual modalities, an increasing number of studies highlight the potential for foundation model representations to align across different modalities, even when trained independently. One line of studies shows that representations from language and vision models can be brought into correspondence with lightweight mappings. For example, Merullo et al. (2023) investigated the extent to which conceptual representations from frozen text-only and vision-only models align, and found that visual information encoded by image models can be transferred to language models using only a single learned linear projection. Similarly, Koh et al. (2023) proposed a lightweight framework for adapting pretrained text-only LLMs to handle multimodal inputs. Their approach enables LLMs to process interleaved image–text data and generate text interleaved with retrieved images, all without retraining the base model. Maniparambil et al. (2024) examined whether inherent alignment exists between independently trained unimodal vision and language encoders, and found that such models encode semantically similar structures, enabling zero-shot latent space communication without explicit alignment. Zhang et al. (2025) evaluated the degree to which independently trained vision and language models can be aligned and propose an efficient alignment framework for downstream tasks.

Other studies explore alignment between language and auditory representations. Ngo and Kim (2024) showed that auditory and language representations can be approximately aligned through a simple linear transformation, pointing to a shared structural basis across modalities. They further demonstrated that text-only LLMs encode features that align with auditory object representations, such that a contrastive probe can successfully retrieve the correct object label from an audio snippet. Lee et al. (2024) revealed that cross-modal representations, particularly between text and speech, tend to converge in the deeper layers of the models, while the early layers remain modality-specific and specialized for raw input processing. This layer-wise convergence reflects a progressive transformation toward modality-agnostic abstraction as signals propagate through the network.

These results resonate with the Platonic Representation Hypothesis (Huh et al., 2024), which argues that foundation models are converging toward a common statistical model of reality embedded within their representation spaces. All these findings suggest that cross-modal convergence may reflect a deeper tendency of foundation models to discover modality-agnostic abstractions. This emerging evidence points to the possibility that foundation models, even when trained separately, may inhabit overlapping representational manifolds that facilitate alignment, transfer, and integration across modalities.

4.5 Representation Alignment with Neuroscience

While the previous sections focus on alignment within and across artificial modalities, a natural question is whether the representations learned by foundation models also correspond to those observed in biological systems. In this context, representation alignment does not refer to internal consistency within neuroscience, but rather to the similarity and correspondence between model-derived features and neural representations measured in cognitive neuroscience (Pham et al., 2023). This perspective bridges artificial and biological intelligence, providing insights into the extent to which foundation models capture structures present in natural cognition. For example, Chen et al. (2024) showed that both Wav2Vec2.0 and GPT2 predict human auditory cortex responses, with model activations exhibiting strong correlations to brain activity during speech and language perception. Khosla et al. (2024) developed axis-sensitive metrics for alignment and demonstrated that both biological and artificial neural networks exhibit privileged, non-arbitrary axes of representation that converge across systems. Hosseini et al. (2024) examined the representation universality hypothesis, and proposed that artificial neural networks trained on naturalistic data converge toward shared representational structures that also align with the brain. Their findings show that inter-model representational agreement reliably predicts brain alignment, implying that shared tasks and environments naturally drive both artificial and biological systems toward similar representational structures. Doerig et al. (2025) reported that embeddings from LLMs of whole scene captions align closely with high-level visual cortex responses to corresponding images, outperform many image-based models, and that image models trained to predict these caption embeddings can match or exceed vision model alignment to brain activity. Raugel et al. (2025) added that model size, training scale, and the nature of image data all critically drive how well vision transformer models develop representations aligned with human brain activity, with larger, more human-centric models aligning later brain regions and temporal dynamics when given enough training. Feather et al. (2025) proposed the NeuroAI Turing Test, a benchmark that requires models to match both behavior and internal neural representations of brains, arguing that this stronger criterion is necessary for model-brain evaluations in NeuroAI. Taken together, these findings suggest that artificial and biological systems, despite differences in architecture and learning mechanisms, converge toward similar representational frameworks when exposed to comparable sensory inputs and functional objectives. This bridges the gap between artificial intelligence and neuroscience, offering insights into both machine learning and human cognition.

5 Factors Driving Representation Potential for Alignment

A variety of factors contribute to the alignment potential of foundation model representations. Among the most prominent is scale, which encompasses model capacity, dataset size, and computational resources. Kaplan et al. (2020) established scaling laws showing that performance improves predictably with these factors, and subsequent work has demonstrated their impact on representation similarity. For instance, Gammelgaard et al. (2023) showed that as LLMs increase in size and quality, they organize concepts in embedding spaces in ways increasingly similar to the structures of knowledge graphs, suggesting convergence toward human-like conceptual organization derived from text alone. Nguyen et al. (2024) demonstrated that linguistic and cultural diversity in data enhances generalization and robustness. Huh et al. (2024) proposed that across multiple modalities, larger models trained on more diverse datasets naturally develop more aligned representations.

While the Platonic Representation Hypothesis (Huh et al., 2024) suggests that convergence may occur independently of architectural or training objectives, evidence shows that these factors can nonetheless shape the dynamics of alignment. The Transformer architecture (Vaswani et al., 2017), now dominant in foundation models, is argued to inherently facilitate generalizable representations due to its flexible inductive biases (Edelman et al., 2022; Bhattamishra et al., 2023; Geerts et al., 2025). Training paradigms are equally influential. Self-supervision, in particular, has been shown to encourage representations that generalize broadly across tasks and domains. Ciernik et al. (2025) found that self-supervised vision models yield stronger pairwise similarity generalization across datasets compared to models trained with image classification or image–text matching objectives. These results indicate that although convergence may not depend strictly on architecture or objective, both exert important shaping effects on representational outcomes.

The generality of tasks and instructions also affect representation potentials. As foundation models are trained on increasingly diverse task mixtures and fine-tuned with a wide range of instructions, their representation spaces become progressively constrained, encouraging the development of task-agnostic abstractions that potentially support alignment. Sanh et al. (2022) found that prompt-based multitask fine-tuning yields strong zero-shot generalization, frequently exceeding the performance of larger models. Chung et al. (2024) further showed that scaling instruction-finetuning by expanding both the number and variety of finetuning tasks substantially boosts performance across zero-shot, few-shot, reasoning, multilingual, and open-ended benchmarks. Zhang et al. (2024) highlighted that instruction diversity, rather than the sheer number of examples per instruction, is the critical factor for driving generalization.

6 Open Questions

Despite compelling evidence for representation potentials of foundation models for alignment, several limitations and open questions remain. One fundamental limitation arises from differences across modalities. Distinct sensors and perspectives capture complementary aspects of reality, and certain information may be unique to a given sensory channel, which constrains the extent to which their representations can perfectly align. For example, visual data emphasizes spatial and perceptual detail, while language conveys abstract concepts and relationships. Consequently, full convergence to a single, identical representation across modalities is neither achievable nor necessarily desirable. In this sense, alignment should perhaps be understood not as perfect overlap but as the development of partially shared abstractions that remain complementary across modalities. Notably, Lu et al. (2025) recently proposed that foundation models converge toward relational, externally grounded representations, implicitly reflecting a shared relational structure underlying reality. Such representations inherently involve contextualization and mutual reference across samples, and thus offer a principled way of addressing the limitations outlined above.

Another challenge lies in how to evaluate alignment rigorously. While various metrics have been proposed to quantify the similarity between representations, an ongoing debate persists within the research community regarding the effectiveness and interpretability of these measures. It is often unclear whether a given alignment score indicates a strong degree of similarity with only minor discrepancies, or a relatively weak alignment with significant underlying differences that are yet to be fully understood. For example, Wang et al. (2018) introduced a neuron activation subspace match model to define the similarity between networks trained with the same architecture but different random initializations. They showed that convolutional layers often exhibit low similarity across independently trained networks, even with identical architectures. Brown et al. (2023) demonstrated that measures such as model stitching and CKA can reveal internal differences in language models that are invisible to performance-based evaluations. Harvey et al. (2024) provided a decoder-based perspective, arguing that high similarity in metrics like CKA or CCA implies that many features can be decoded similarly, but the reverse does not necessarily hold: models may allow similar decoding for some tasks while still differing geometrically in how information is distributed across dimensions. The absence of a universally accepted standard complicates cross-study comparisons and raises questions about what alignment scores truly capture.

The role of data bias and sociotechnical context also presents a crucial consideration. Training data are often scraped from the internet and thus inherit the biases, cultural norms, and imbalances embedded in human human-generated content. Such biases shape representation spaces and constrain the universality of alignment claims. Beyond data, the broader sociotechnical context, including who builds the models and for what purposes, also influences alignment, raising questions about whether observed consistencies reflects general cognitive structures or artifacts of specific training regimes.

Finally, there are counterexamples and scenarios where representation alignment might not emerge. Highly specialized models, optimized for narrow tasks, may develop unique representations that diverge from general-purpose abstractions. In domains such as robotics, where standardized representations for complex sensorimotor experiences are still under development, the potential for alignment may be constrained by the absence of common frameworks and data formats. In summary, while the evidence for alignment across architectures and modalities is strong, the boundaries of this phenomenon remain poorly understood. Future progress requires both more robust evaluation frameworks and a careful recognition of the contexts and biases that shape representational spaces.

7 Conclusion

In this survey, we synthesize substantial evidence for the representation potential of foundation models, demonstrating their capacity for alignment within individual modalities such as vision, language, and speech, across multimodal combinations, and in relation to representations observed in neuroscience. We further analyze the key factors that foster representation potentials and discuss open questions. Our future work will pursue a deeper theoretical grounding of representation potentials. We believe that continued exploration in this direction will not only drive the development of more interpretable foundation models but also enrich our broader understanding of the principles that underlie both artificial and natural intelligence.

Limitations

This survey has focused on the representation potentials of foundation models across vision, language, speech, multimodality, and neuroscience, emphasizing areas where substantial empirical evidence is available. However, our analysis necessarily excludes domains such as robotics, sensorimotor control, and graphs, where research on representation alignment is still fragmented and publicly available findings are limited. As a result, this survey may not fully reflect the breadth of representational behaviors across all applications of foundation models.

Another limitation arises from the evaluation of alignment itself. While we reviewed a range of commonly used metrics, including CKA, CCA, and MNN, the field still lacks a unified standard for assessing representation similarity. This makes it difficult to integrate results across studies rigorously. Comparisons across modalities, architectures, or training objectives therefore remain partly qualitative, and definitive meta-analyses are constrained by methodological inconsistencies.

Finally, foundation models are rapidly evolving. New models, training paradigms, and evaluation techniques continue to emerge, particularly in cross-modal and neuroscience-inspired settings. Consequently, the conclusions and scope presented here should be regarded as a reflection of the current state of the field rather than a definitive account. We anticipate that many of the questions identified in this survey will be revisited and refined as the field advances.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637.

Bobby Azad, Reza Azad, Sania Eskandari, Afshin Bozorgpour, Amirhossein Kazerouni, Islem Rekik, and Dorit Merhof. 2023. Foundational models in medical imaging: A comprehensive survey and future vision. *arXiv preprint arXiv:2310.18689*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Randall Balestriero and richard baraniuk. 2018. A spline theory of deep learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 374–383. PMLR.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. 2021. Revisiting model stitching to compare neural representations. *Advances in neural information processing systems*, 34:225–236.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, and 1 others. 2023. Seamlessm4t: Massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Yonatan Belinkov and James Glass. 2017. Analyzing hidden representations in end-to-end automatic speech recognition systems. *Advances in Neural Information Processing Systems*, 30.

Satwik Bhattamishra, Arkil Patel, Varun Kanade, and Phil Blunsom. 2023. Simplicity bias in transformers and their ability to learn sparse Boolean functions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5767–5791. Association for Computational Linguistics.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, and 1 others. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, and 1 others. 2022. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*.

Davis Brown, Charles Godfrey, Nicholas Konz, Jonathan Tu, and Henry Kvinge. 2023. Understanding the inner-workings of language models through representation dissimilarity. In *Proceedings of the*

2023 Conference on Empirical Methods in Natural Language Processing, pages 6543–6558. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901.

Lennart Bürger, Fred A Hamprecht, and Boaz Nadler. 2024. Truth is universal: Robust detection of lies in llms. *Advances in Neural Information Processing Systems*, 37.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660.

Alan Chen, Jack Merullo, Alessandro Stolfo, and Ellie Pavlick. 2025. Transferring features across language models with model stitching. *arXiv preprint arXiv:2506.06609*.

Peili Chen, Linyang He, Li Fu, Lu Fan, Edward F Chang, and Yuanning Li. 2024. Do self-supervised speech and language models extract similar representations as human brain? In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2225–2229. IEEE.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022a. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, and 1 others. 2022b. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.

Emily Cheng, Diego Doimo, Corentin Kervadec, Iuri Macocco, Lei Yu, Alessandro Laio, and Marco Baroni. 2025. Emergence of a high-dimensional abstraction phase in language transformers. In *The Thirteenth International Conference on Learning Representations*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Yu-An Chung, Yonatan Belinkov, and James Glass. 2021. Similarity analysis of self-supervised speech representations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3040–3044. IEEE.

Laure Ciernik, Lorenz Linhardt, Marco Morik, Jonas Dippel, Simon Kornblith, and Lukas Muttenthaler. 2025. Objective drives the consistency of representational similarity across datasets. In *Forty-second International Conference on Machine Learning*.

Adrián Csiszárik, Péter Kőrösi-Szabó, Akos Matszangosz, Gergely Papp, and Dániel Varga. 2021. Similarity and matching of neural network representations. *Advances in Neural Information Processing Systems*, 34:5656–5668.

Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Yiwen Guo, and Irwin King. 2024. Recent advances in speech language models: A survey. *arXiv preprint arXiv:2410.03751*.

Yuchen Cui, Scott Niekum, Abhinav Gupta, Vikash Kumar, and Aravind Rajeswaran. 2022. Can foundation models perform zero-shot task specification for robot manipulation? In *Learning for dynamics and control conference*, pages 893–905. PMLR.

MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. 2023. Reliability of CKA as a similarity measure in deep learning. In *The Eleventh International Conference on Learning Representations*.

Maksym Del and Mark Fishel. 2022. Cross-lingual similarity of multilingual representations revisited. *arXiv preprint arXiv:2212.01924*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1*, pages 4171–4186.

Jörn Diedrichsen and Nikolaus Kriegeskorte. 2017. Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS computational biology*, 13(4):e1005508.

Adrien Doerig, {Tim C.} Kietzmann, Emily Allen, Yihan Wu, Thomas Naselaris, Kendrick Kay, and Ian Charest. 2025. High-level visual representations in the human brain are aligned with large language models. *Nature Machine Intelligence*, 7(8):1220–1234.

Teresa Dorszewski, Albert Kjøller Jacobsen, Lenka Tětková, and Lars Kai Hansen. 2025. How redundant is the transformer stack in speech representation models? In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Amil Dravid, Yossi Gandelsman, Alexei A Efros, and Assaf Shocher. 2023. Rosetta neurons: Mining the common units in a model zoo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1934–1943.

Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. 2022. Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning*, pages 5793–5831. PMLR.

Jenelle Feather, Meenakshi Khosla, N Murty, and Aran Nayebi. 2025. Brain-model evaluations need the neuroai turing test. *arXiv preprint arXiv:2502.16238*.

Roya Firoozi, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiyu Liu, Yuke Zhu, Shuran Song, Ashish Kapoor, Karol Hausman, and 1 others. 2023. Foundation models in robotics: Applications, challenges, and the future. *The International Journal of Robotics Research*.

Mathias Lykke Gammelgaard, Jonathan Gabel Christiansen, and Anders Søgaard. 2023. Large language models converge toward human-like concept organization. *arXiv preprint arXiv:2308.15047*.

Jesse Geerts, Stephanie Chan, Claudia Clopath, and Kimberly Stachenfeld. 2025. Relational reasoning and inductive bias in transformers trained on a transitive inference task. *arXiv preprint arXiv:2506.04289*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and Ahmad Al-Dahle et al. 2024. The llama 3 herd of models.

Tom George Grigg, Dan Busbridge, Jason Ramapuram, and Russ Webb. 2021. Do self-supervised and supervised methods learn similar visual representations? *arXiv preprint arXiv:2110.00528*.

Albert Gu and Tri Dao. 2024. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*.

Wes Gurnee, Theo Horsley, Zifan Carl Guo, Tara Rezaei Kheirkhah, Qinyi Sun, Will Hathaway, Neel Nanda, and Dimitris Bertsimas. 2024. Universal neurons in gpt2 language models. *arXiv preprint arXiv:2401.12181*.

David Ha and Jürgen Schmidhuber. 2018. World models. *arXiv preprint arXiv:1803.10122*, 2(3).

Laleh Haghverdi, Aaron TL Lun, Michael D Morgan, and John C Marioni. 2018. Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology*, 36(5):421–427.

Sarah E Harvey, David Lipshutz, and Alex H Williams. 2024. What representational similarity measures imply about decodable information. In *Proceedings of UniReps: the Second Edition of the Workshop on Unifying Representations in Neural Models*, volume 285 of *Proceedings of Machine Learning Research*, pages 140–151. PMLR.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*, pages 770–778.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Eghbal Hosseini, Colton Casto, Noga Zaslavsky, Colin Conwell, Mark Richardson, and Evelina Fedorenko. 2024. Universality of representation in biological and artificial neural networks. *bioRxiv*.

Harold Hotelling. 1936. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Weijian Huang, Cheng Li, Hong-Yu Zhou, Hao Yang, Jiarun Liu, Yong Liang, Hairong Zheng, Shaoting Zhang, and Shanshan Wang. 2024. Enhancing representation in radiography-reports foundation model: A granular alignment algorithm using masked contrastive learning. *Nature Communications*, 15(1):7620.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR.

Robin Huo and Ewan Dunbar. 2025. Iterative refinement, not training objective, makes hubert behave differently from wav2vec 2.0. In *Proc. Interspeech 2025*, pages 261–265.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Jiachen Jiang, Jinxin Zhou, and Zhihui Zhu. 2025. Tracing representation progression: Analyzing and enhancing layer-wise similarity. In *The Thirteenth International Conference on Learning Representations*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Wasif Khan, Seowung Leem, Kyle B See, Joshua K Wong, Shaoting Zhang, and Ruogu Fang. 2025. A comprehensive survey of foundation models in medicine. *IEEE Reviews in Biomedical Engineering*.

Meenakshi Khosla, Alex H Williams, Josh McDermott, and Nancy Kanwisher. 2024. Privileged representational axes in biological and artificial neural networks. *bioRxiv*, pages 2024–06.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, and 1 others. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. 2025. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9):1–52.

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.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR.

Michael Lan, Philip Torr, Austin Meek, Ashkan Khakzar, David Krueger, and Fazl Barez. 2024. Quantifying feature space universality across large language models via sparse autoencoders. *arXiv preprint arXiv:2410.06981*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Andrew Lee, Melanie Weber, Fernanda Viégas, and Martin Wattenberg. 2025. Shared global and local geometry of language model embeddings. In *Second Conference on Language Modeling*.

Hyunji Lee, Danni Liu, Supriti Sinhamahapatra, and Jan Niehues. 2024. How do multimodal foundation models encode text and speech? an analysis of cross-lingual and cross-modal representations. *arXiv preprint arXiv:2411.17666*.

Karel Lenc and Andrea Vedaldi. 2015. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 991–999.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pretraining for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Xinzhe Li. 2025. A review of prominent paradigms for LLM-based agents: Tool use, planning (including RAG), and feedback learning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9760–9779.

Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. 2015. Convergent learning: Do different neural networks learn the same representations? *arXiv preprint arXiv:1511.07543*.

Zijie Li, Henry Li, Yichun Shi, Amir Barati Farimani, Yuval Kluger, Linjie Yang, and Peng Wang. 2025. Dual diffusion for unified image generation and understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2779–2790.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jianglin Lu, Hailing Wang, Kuo Yang, Yitian Zhang, Simon Jenni, and Yun Fu. 2025. The indra representation hypothesis. *Advances in Neural Information Processing Systems*.

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 (CVPR)*.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. 2023. Linearly mapping from image to text space. In *The Eleventh International Conference on Learning Representations*.

Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, and 1 others. 2022. Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1179–1210.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. 2023. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265.

Ari Morcos, Maithra Raghu, and Samy Bengio. 2018. Insights on representational similarity in neural networks with canonical correlation. *Advances in neural information processing systems*, 31.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. 2023. Relative representations enable zero-shot latent space communication. In *The Eleventh International Conference on Learning Representations*.

Evonne Ng, Sanjay Subramanian, Dan Klein, Angjoo Kanazawa, Trevor Darrell, and Shiry Ginosar. 2023. Can language models learn to listen? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10083–10093.

Jerry Ngo and Yoon Kim. 2024. What do language models hear? probing for auditory representations in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 5435–5448.

Thao Nguyen, Maithra Raghu, and Simon Kornblith. 2021. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. In *International Conference on Learning Representations*.

Thao Nguyen, Matthew Wallingford, Sebastin Santy, Wei-Chiu Ma, Sewoong Oh, Ludwig Schmidt, Pang Wei W Koh, and Ranjay Krishna. 2024. Multilingual diversity improves vision-language representations. *Advances in Neural Information Processing Systems*, 37:91430–91459.

Anna Ollerenshaw, Md Asif Jalal, and Thomas Hain. 2021. Insights on neural representations for end-to-end speech recognition. In *Interspeech*.

Narmeen Fatimah Oozeer, Dhruv Nathawani, Nirmalendu Prakash, Michael Lan, Abir HARRASSE, and Amir Abdullah. 2025. Activation space interventions can be transferred between large language models. In *Forty-second International Conference on Machine Learning*.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, and 1 others. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning*, pages 39643–39666. PMLR.

Ankita Pasad, Chung-Ming Chien, Shane Settle, and Karen Livescu. 2024. What do self-supervised speech models know about words? *Transactions of the Association for Computational Linguistics*, 12:372–391.

Ankita Pasad, Bowen Shi, and Karen Livescu. 2023. Comparative layer-wise analysis of self-supervised speech models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Trung Quang Pham, Teppei Matsui, and Junichi Chikazoe. 2023. Evaluation of the hierarchical correspondence between the human brain and artificial neural networks: A review. *Biology*, 12(10).

Jason Phang, Haokun Liu, and Samuel R Bowman. 2021. Fine-tuned transformers show clusters of similar representations across layers. *arXiv preprint arXiv:2109.08406*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017a. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017b. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, volume 30.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. 2021. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128.

Goutham Rajendran, Simon Buchholz, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. 2024. From causal to concept-based representation learning. *Advances in Neural Information Processing Systems*, 37:101250–101296.

Joséphine Raugel, Marc Szafraniec, Huy V Vo, Camille Couprie, Patrick Labatut, Piotr Bojanowski, Valentin Wyart, and Jean-Rémi King. 2025. Disentangling the factors of convergence between brains and computer vision models. *arXiv preprint arXiv:2508.18226*.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. 2025. SAM 2: Segment anything in images and videos. In *The Thirteenth International Conference on Learning Representations*.

Pablo Riera, Manuela Cerdeiro, Leonardo Pepino, and Luciana Ferrer. 2023. Phone and speaker spatial organization in self-supervised speech representations. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5. IEEE.

Filippo Rinaldi, Giacomo Capitani, Lorenzo Bonicelli, Donato Crisostomi, Federico Bolelli, ELISA FICARRA, Emanuele Rodolà, Simone Calderara, and Angelo Porrello. 2025. Update your transformer to the latest release: Re-basin of task vectors. In *Forty-second International Conference on Machine Learning*.

Geoffrey Roeder, Luke Metz, and Durk Kingma. 2021. On linear identifiability of learned representations. In *International Conference on Machine Learning*, pages 9030–9039. PMLR.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, and 21 others. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Mahdiyar Shahbazi, Ali Shirali, Hamid Aghajan, and Hamed Nili. 2021. Using distance on the riemannian manifold to compare representations in brain and in models. *NeuroImage*, 239:118271.

Pratyusha Sharma, Tamar Rott Shaham, Manel Baradad, Stephanie Fu, Adrian Rodriguez-Munoz, Shivam Duggal, Phillip Isola, and Antonio Torralba. 2024. A vision check-up for language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14410–14419.

Yuval Sharon and Yehuda Dar. 2025. How do the architecture and optimizer affect representation learning? on the training dynamics of representations in deep neural networks. *arXiv preprint arXiv:2405.17377*.

Shashank Shekhar, Florian Bordes, Pascal Vincent, and Ari Morcos. 2023. Objectives matter: Understanding the impact of self-supervised objectives on vision transformer representations. *arXiv preprint arXiv:2304.13089*.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, and 7 others. 2025. Dinov3.

Le Song, Alex Smola, Arthur Gretton, Karsten M Borgwardt, and Justin Bedo. 2007. Supervised feature selection via dependence estimation. In *Proceedings of the 24th international conference on Machine learning*, pages 823–830.

G Stoica, D Bolya, J Bjorner, P Ramesh, T Hearn, and J Hoffman. 2024. Zipit! merging models from different tasks without training. In *International Conference on Learning Representations*. International Conference on Learning Representations.

Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. 2024. Analysing the generalisation and reliability of steering vectors. *Advances in Neural Information Processing Systems*, 37.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, and 1 others. 2025. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti

Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Abdul Waheed, Hanin Atwany, Bhiksha Raj, and Rita Singh. 2024. What do speech foundation models not learn about speech? *arXiv preprint arXiv:2410.12948*.

Chenxu Wang, Wei Rao, Wenna Guo, Pinghui Wang, Jun Liu, and Xiaohong Guan. 2020. Towards understanding the instability of network embedding. *IEEE Transactions on Knowledge and Data Engineering*, 34(2):927–941.

Junxuan Wang, Xuyang Ge, Wentao Shu, Qiong Tang, Yunhua Zhou, Zhengfu He, and Xipeng Qiu. 2025. Towards universality: Studying mechanistic similarity across language model architectures. In *The Thirteenth International Conference on Learning Representations*.

Liwei Wang, Lunjia Hu, Jiayuan Gu, Zhiqiang Hu, Yue Wu, Kun He, and John Hopcroft. 2018. Towards understanding learning representations: To what extent do different neural networks learn the same representation. *Advances in neural information processing systems*, 31.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, and 1 others. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

John Wentworth. 2021. Testing the natural abstraction hypothesis. *AI Alignment Forum*.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. 2023. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. 2023. Diffusion models: A comprehensive survey of methods and applications. *ACM computing surveys*, 56(4):1–39.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint*.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. 2025. Representation alignment for generation: Training diffusion transformers is easier than you think. In *The Thirteenth International Conference on Learning Representations*.

Dylan Zhang, Justin Wang, and Francois Charton. 2024. Instruction diversity drives generalization to unseen tasks. *arXiv preprint arXiv:2402.10891*.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. 2023. A tale of two features: Stable diffusion complements DINO for zero-shot semantic correspondence. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Le Zhang, Qian Yang, and Aishwarya Agrawal. 2025. Assessing and learning alignment of unimodal vision and language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14604–14614.

Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, and 1 others. 2024. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *International Journal of Machine Learning and Cybernetics*, pages 1–65.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. 2025. DINO-WM: World models on pretrained visual features enable zero-shot planning. In *Forty-second International Conference on Machine Learning*.