The Indra Representation Hypothesis

Jianglin Lu[1*] **Hailing Wang**[1*] **Kuo Yang**[1] **Yitian Zhang**[1] **Simon Jenni**[3] **Yun Fu**[1,2]
[1]Department of Electrical and Computer Engineering, Northeastern University
[2]Khoury College of Computer Science, Northeastern University
[3]Adobe Research

Abstract

Recent studies have uncovered an interesting phenomenon: unimodal foundation models tend to learn convergent representations, regardless of differences in architecture, training objectives, or data modalities. However, these representations are essentially internal abstractions of samples that characterize samples independently, leading to limited expressiveness. In this paper, we propose *The Indra Representation Hypothesis*, inspired by the philosophical metaphor of Indra's Net. We argue that representations from unimodal foundation models are converging to implicitly reflect a shared relational structure underlying reality, akin to the relational ontology of Indra's Net. We formalize this hypothesis using the $\mathcal{V}$-enriched Yoneda embedding from category theory, defining the Indra representation as a relational profile of each sample with respect to others. This formulation is shown to be unique, complete, and structure-preserving under a given cost function. We instantiate the Indra representation using angular distance and evaluate it in cross-model and cross-modal scenarios involving vision, language, and audio. Extensive experiments demonstrate that Indra representations consistently enhance robustness and alignment across architectures and modalities, providing a theoretically grounded and practical framework for training-free alignment of unimodal foundation models. Our code is available at https://github.com/Jianglin954/Indra.

1 Introduction

Through large-scale pretraining, foundation models have emerged as a transformative paradigm in artificial general intelligence, demonstrating impressive progress across diverse domains, such as natural language processing [12, 26, 46], computer vision [47, 76, 60, 13], and speech processing [66, 3]. These unimodal models are typically trained on web-scale datasets and acquire generalized representations that can be adapted to a broad range of downstream tasks. The representative models, such as BERT [12] for language, ViT [13] for vision, and Wav2Vec [66] for audio, have demonstrated promising performance within their respective domains.

Since real-world information is inherently multimodal (text, images, and audio frequently co-occur and complement each other), relying on a single modality for understanding is typically insufficient. To extend unimodal foundation models for cross-modal tasks, a growing body of research has exploited strong unimodal encoders as the core components to build multimodal systems [62, 42, 61, 63]. They assume that unimodal models are able to provide specialized, high-quality, and high-level representations for each individual modality, which can then be aligned, fused, or bridged to enable multimodal interactions [41, 2, 9]. A common strategy to achieve multimodal understanding is to align unimodal outputs in a shared representation space through cross-modal objectives [62, 41, 45, 55]. This usually relies on external mechanisms, such as alignment losses, fusion modules, and prompt tuning, thus requiring large-scale datasets and extensive retraining for modality alignment.

*Equal Contribution. Corresponding author: JianglinLu@outlook.com.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Interestingly, recent studies [59, 57, 45, 67] suggest that powerful unimodal models may already exhibit latent cross-modal capabilities, as the representations they produce (when grounded on the same physical entity) tend to describe the same underlying semantics from different sensory perspectives. Previous evidence has revealed that adding only a single linear transformation is capable of bridging an auditory model with an LLM [58], integrating a vision model into a large language model (LLM) [55], or conversely, stitching an LLM towards a vision model [36]. Even without retraining, well-pretrained vision encoders exhibit high semantic similarity with language encoders [52]. Further studies [30, 34, 28, 49] have revealed that models trained on different data modalities converge, as different models are all trying to arrive at a representation of reality. Thus, unimodal models may encode modality-agnostic representations in abstract representational space, even without explicit alignment. While conceptually appealing, the specific form of convergent representations and their eventual convergence targets remain elusive and largely unexplored.

In this paper, we posit that the underlying convergent representation is inherently the Indra Representation—a conceptual abstraction inspired by the philosophical metaphor of Indra's Net. Originating in ancient Buddhist philosophy, Indra's Net describes a vast, infinite web of jewels, each reflecting all others. Every jewel is both a part and a reflection of the whole, suggesting that all phenomena are interdependent, mutually defining, and inherently connected. We draw an analogy between this worldview and the notion of representation convergence, and introduce ***The Indra Representation Hypothesis****: well-trained unimodal models tend to produce convergent representations that implicitly reflect a shared relational structure underlying reality, echoing the relational ontology of Indra's Net*. In this view, the representation of each entity is not defined in isolation, but rather emerges from its relational context, i.e., its reflections of all other entities.

To explore this hypothesis, we introduce a theoretical definition of the Indra representation grounded in category theory [33]. Specifically, we define it as the $\mathcal{V}$-enriched Yoneda embedding of a sample within a category enriched over a **Cost**-category. This formulation effectively maps each sample to its covariant Hom-functors in the sample category, thereby encoding its relational profile within the structure of the dataset. We provide theoretical guarantees that the Indra representation is unique, complete, and structure-preserving, offering a principled foundation for its effectiveness. In particular, we prove that it uniquely and completely characterizes each sample within the relational structure induced by a given cost function, while preserving essential properties of that structure.

To instantiate this theory in practice, we adopt angular distance as the cost function, yielding a simple yet powerful realization of the Indra representation. This concrete formulation enables empirical evaluation and allows us to investigate how Indra representations can uncover and support the latent cross-modal capabilities of unimodal models. We validate our approach across a range of scenarios involving cross-model and cross-modal understanding, including single-modality settings, vision-language pairs, and speech-language pairs. Extensive experiments demonstrate the effectiveness and generality of the proposed Indra representation across different architectures and modalities.

2 Preliminaries

2.1 Indra's Net

Indra's Net is a philosophical metaphor originating in ancient Indian and Mahāyāna Buddhist thought, particularly from the Avataṃsaka Sūtra [10, 23]. It is used to symbolize the universe as a web of interdependent connections among all of its members, expressing the concept of interconnectedness, non-duality, and the interpenetration of all phenomena. Francis H. Cook describes it as [11]:

> *Imagine a vast, infinite cosmic net belonging to the god Indra. At each node of the net is a jewel or crystal. Each jewel reflects every other jewel in the net, and in each of those reflections are further reflections of all other jewels, recursively and infinitum.*

The metaphor of Indra's Net resonates deeply with the foundational principles across diverse disciplines. For instance, Gergen et al. [18] posit that identities, thoughts, and actions are not solely products of isolated minds but are co-constructed through interactions and relationships with others. Markus et al. [53] propose the theory of interdependent self-construal, emphasizing that the self is defined relationally through one's social roles, group memberships, and interpersonal obligations. In physics, field theory [14, 54] illustrates a relational structure where the field at any point depends on all sources throughout space, a concept reminiscent of Indra's Net, in which each jewel reflects

and is reflected by all others. Similarly, modern particle physics [19, 5] reveals the properties of an elementary particle through its interactions with other particles. In linguistics, the linguistic principle articulated by J.R. Firth [16], "You shall know a word by the company it keeps", asserts that a word's meaning is derived from its co-occurrence with other words. This principle forms the basis for modern language models like Word2Vec [56] and its successors. Furthermore, Kasulis [32] suggests DNA as a better image of Indra's Net, where every cell contains the blueprint for the whole organism.

2.2 Representation Convergence

Recent studies [30, 75, 21, 15, 74, 40, 49] have revealed a striking phenomenon: unimodal foundation models tend to learn convergent representations, regardless of their architectures, training objectives, or data modalities. For example, Bürger et al. [6] show that a two-dimensional representation of truth emerges universally across LLMs of varying sizes and from different model families. Roeder et al. [65] prove that a wide class of discriminative and autoregressive models are identifiable in function space up to a linear transformation. Tan et al. [68] find strong correlations in both in-distribution and out-of-distribution steerability between LLaMA [69] and Qwen [4]. Huh et al. [30] attribute this convergence to a shared goal: approximating an underlying representation of reality. Khosla et al. [34] argue that both artificial and biological systems converge toward representations that capture the causal structure of the world. Hosseini et al. [28] further observe that high-performing artificial neural networks and biological brains tend to develop similar internal representations under naturalistic training conditions. Additional evidence and analyses on representation convergence can be found in the comprehensive survey [49].

Despite this emerging consensus, it remains unclear how these representations converge and what they ultimately converge to. Notably, prior studies rely primarily on model outputs (embeddings) as proxies for representations, but these representations suffer from ① structural myopia: representations are typically treated as isolated carriers of information, ignoring structural interrelations within the broader data manifold; ② limited expressiveness: unimodal representations often exhibit inferior quality in matching and alignment compared to those from multimodal foundation models; and ③ dimensional incompatibility: representations across models and modalities often differ in dimensionality, thus requiring additional post-processing for cross-modal matching. In light of these challenges, *we argue that representations from model outputs do not reflect the final converged form, but instead serve as the foundation upon which such a form can be built*. In the next section, we introduce a novel representation hypothesis inspired by the metaphor of Indra's Net to hypothesize the concrete structure of convergent representations and illuminate what they ultimately converge to.

3 Methodology

3.1 The Indra Representation Hypothesis

In this paper, we advocate a shift in perspective on representation convergence, inspired by the metaphor of Indra's Net: a sample should be represented not in isolation, but through its pattern of relationships to other samples. In this view, representations emerge from a structure of mutual interdependence. We formalize this perspective through the Indra Representations Hypothesis:

> **The Indra Representation Hypothesis**: *Neural networks, trained with different objectives on different data and modalities, tend to learn convergent representations that implicitly reflect a shared relational structure underlying reality—parallel to the relational ontology of Indra's Net.*

This hypothesis posits that unimodal foundation models, after extensive pretraining, tend to produce representations that converge to capture the inherent relational structure of reality—a structure characterized by interdependence, contextuality, and mutual influence. However, current methods that treat model outputs as final representations fail to reflect this structure. These methods typically emphasize individual embedding information while neglecting the crucial relational patterns between samples. In the next section, we introduce the *Indra Representation*, a novel representation framework inspired by the metaphor of Indra's Net, to reveal the underlying relational structure. In this framework, representations are not independent embeddings but mutually reflective entities woven into a web of interdependent relationships, revealing the deeper relational structure that underpins the data.

3.2 From Metaphor to Theory

Indra's Net is a philosophical metaphor that symbolizes the interconnectedness of the universe. It aligns with the foundational principles in modern science, as mentioned in Section 2.1. To translate this philosophical insight into representation learning, we first introduce the Yoneda Lemma and its corollary, which provide the theoretical foundation for defining our proposed Indra representation.

Lemma 1 (Yoneda Lemma [33, 64])**.** *Let $\mathcal{C}$ be a locally small category, A be an object in $\mathcal{C}$, and $F : \mathcal{C} \to$ **Set** be a functor from $\mathcal{C}$ to the category of sets. Then, there exists a bijection, natural in both A and F, between the set of natural transformations from the hom-functor $h_A = Hom_{\mathcal{C}}(A, -)$ to F, and the set $F(A)$. This bijection is given by:*

$$Nat(h_A, F) \cong F(A). \tag{1}$$

Corollary 1 (Yoneda Embedding [33, 64])**.** *For any two objects A, B in a locally small category $\mathcal{C}$, there is a bijection:*

$$Nat(Hom_{\mathcal{C}}(A, -), Hom_{\mathcal{C}}(B, -)) \cong Hom_{\mathcal{C}}(B, A). \tag{2}$$

This demonstrates that the functor $Y : \mathcal{C}^{op} \to [\mathcal{C}, \textbf{Set}]$, defined by $Y(A) = h_A = Hom_{\mathcal{C}}(A, -)$, is fully faithful. This functor Y is known as the Yoneda embedding.

The Yoneda Lemma provides a profound understanding of how an object in a category is characterized by its relationships (morphisms) with all other objects, rather than by its internal properties. Its corollary further shows that any locally small category $\mathcal{C}$ can be embedded into a category of presheaves on $\mathcal{C}$. To introduce our Indra representations, we give the following definitions:

Definition 1 (Sample Category)**.** *Let $\mathcal{X}$ be a set of samples, possibly infinite. The sample category $\mathcal{C}$ enriched over the **Cost**-category $\mathcal{V} = ([0, \infty], \geq, 0, +)$ consists of: ① Objects: $Ob(\mathcal{C}) = \mathcal{X}$. ② Hom-objects: for every $X_i, X_j \in Ob(\mathcal{C})$, the hom-object $\mathcal{C}(X_i, X_j)$ is given by a cost function $d(X_i, X_j) \in [0, \infty]$, which is an object in $\mathcal{V}$. ③ Identity: for all $X_i \in Ob(\mathcal{C})$, the identity morphism $id_{X_i} : I \to \mathcal{C}(X_i, X_i)$ in $\mathcal{V}$ is $0 \to d(X_i, X_i)$, where $d(X_i, X_i) = 0$. ④ Composition: for all $X_i, X_j, X_k \in Ob(\mathcal{C})$, the composition morphism $M_{X_i,X_j,X_k} : \mathcal{C}(X_j, X_k) \otimes \mathcal{C}(X_i, X_j) \to \mathcal{C}(X_i, X_k)$ in $\mathcal{V}$ is: $d(X_j, X_k) + d(X_i, X_j) \to d(X_i, X_k)$. This morphism exists in $\mathcal{V}$ if and only if $d(X_j, X_k) + d(X_i, X_j) \geq d(X_i, X_k)$, which is precisely the triangle inequality.*

Definition 2 ($\mathcal{V}$-enriched Yoneda embedding)**.** *Let $[\mathcal{C}^{op}, \mathcal{V}]$ be the category of $\mathcal{V}$-presheaves on $\mathcal{C}$. The $\mathcal{V}$-enriched Yoneda embedding is a $\mathcal{V}$-functor $Y : \mathcal{C} \to [\mathcal{C}^{op}, \mathcal{V}]$. For each object $X_i \in Ob(\mathcal{C})$, $Y(X_i)$ is the $\mathcal{V}$-presheaf $h_{X_i} : \mathcal{C}^{op} \to \mathcal{V}$ defined by: $h_{X_i}(X_j) = \mathcal{C}(X_j, X_i) = d(X_j, X_i)$ for any $X_j \in \mathcal{C}^{op}$. For every $X_i, X_j \in Ob(\mathcal{C})$, Y defines a map $Y_{X_i,X_j} : \mathcal{C}(X_i, X_j) \to [\mathcal{C}^{op}, \mathcal{V}](Y(X_i), Y(X_j))$.*

Theorem 1. *The $\mathcal{V}$-enriched Yoneda embedding $Y : \mathcal{C} \to [\mathcal{C}^{op}, \mathcal{V}]$ for the sample category $\mathcal{C}$ enriched over $\mathcal{V} = ([0, \infty], \geq, 0, +)$ with the cost function d is $\mathcal{V}$-fully faithful.*

The sample category $\mathcal{C}$ actually forms a Lawvere metric space [39], and the corresponding $\mathcal{V}$-enriched Yoneda embedding maps each sample X_i to a functor $Y(X_i)$, which captures the cost profile from all other samples to X_i. Theorem 1 shows that the sample category can be fully and faithfully represented within a category of $\mathcal{V}$-presheaf, preserving its entire structure including its metric information. In other words, each sample X_i can be uniquely represented by its cost vector $d(\cdot, X_i)$ with all samples. Based on these, we introduce the Indra representation and state the following theorems:

Definition 3 (**Indra Representation**)**.** *For each sample $X_i \in Ob(\mathcal{C})$, we define its Indra representation as the $\mathcal{V}$-functor $\mathcal{C}(X_i, -)$, i.e., the collection of values obtained by evaluating it under the (covariant) $\mathcal{V}$-enriched Yoneda embedding on all objects of the category $\mathcal{C}$.*

Proposition 1. *If two samples $X_i, X_j \in Ob(\mathcal{C})$ have $\mathcal{V}$-naturally isomorphic Indra representations and the cost function d satisfies the T_0 separation axiom, then $X_i = X_j$.*

Theorem 2. *For any $\mathcal{V}$-functor $P : \mathcal{C} \to \mathcal{V}$, the $\mathcal{V}$-hom-object of $\mathcal{V}$-natural transformations from the Indra representation of sample X_i to P, denoted by $[\mathcal{C}, \mathcal{V}](\mathcal{C}(X_i, -), P)$, is $\mathcal{V}$-isomorphic to $P(X_i)$.*

Corollary 2. *The relational structure among objects in the sample category $\mathcal{C}$ is preserved and reflected in the relationships between their Indra representations.*

The Indra representation of a sample X_i, defined as the $\mathcal{V}$-functor $\mathcal{C}(X_i, -)$, can be interpreted as a relational profile of X_i, that is, the cost from X_i to every other sample in the category. Proposition 1 establishes that the Indra representation is a faithful encoding: no two distinct samples share the same

representation. Theorem 2 further shows that the Indra representation is complete, in the sense that it encapsulates all the information needed to determine how distances from X_i behave under any admissible distance assignment. Furthermore, Corollary 2 demonstrates that the relationships between samples are in one-to-one correspondence with the relationships between their Indra representations.

3.3 Instantiation of Indra Representation

We now demonstrate how to instantiate the Indra representation for a real dataset $\mathcal{X} = \{X_1, \ldots, X_n\}$ consisting of n samples. We define the object set of the enriched category as $\mathrm{Ob}(\mathcal{C}) = \mathcal{X}$ and specify the hom-object $\mathcal{C}(X_i, X_j)$ as the cost $d(X_i, X_j)$ between samples X_i and X_j, $\forall X_i, X_j \in \mathrm{Ob}(\mathcal{C})$. To define a valid Indra representation, the cost function d must satisfy two properties: ① $d(X_i, X_i) = 0$ for $\forall X_i \in \mathcal{X}$; and ② $d(X_i, X_k) \leq d(X_i, X_j) + d(X_i, X_k)$ for $\forall X_i, X_j, X_k \in \mathcal{X}$. Several distance metrics satisfy these conditions. In this work, we adopt a simple yet effective choice by defining $d(X_i, X_j)$ as the angular distance between the model-generated embeddings of X_i and X_j:

$$d(X_i, X_j) := \arccos\left(\frac{f(X_i) \cdot f(X_j)}{\|f(X_i)\|\|f(X_j)\|}\right), \quad \forall X_i, X_j \in \mathrm{Ob}(\mathcal{C}) \tag{3}$$

where $f : \mathcal{X} \to \mathbb{R}^{d^{\mathcal{X}}}$ is a modality-specific foundation model, $f(X_i)$ denotes the internal representation of X_i produced by f, and $d^{\mathcal{X}}$ is the output dimensionality of the model. The use of angular distance ensures that d defines a valid Lawvere cost function. Given this cost function, the Indra representation of each sample $X_i \in \mathcal{X}$ is the covariant Hom-functor:

$$\mathcal{C}(X_i, -) : \mathcal{X} \to [0, \infty], \quad X_j \mapsto d(X_i, X_j). \tag{4}$$

This distance-based embedding forms a principled and interpretable representation. In the finite case, it can be written as $\mathcal{C}(X_i, -) = [d(X_i, X_1), \ldots, d(X_i, X_n)]$, which captures the relational profile of X_i with respect to all other samples.

3.4 Relational Matching across Modalities

Our hypothesis in Section 3.1 posits that unimodal foundation models learn convergent representations that capture the shared relational structure underlying reality. The proposed Indra representations are designed to reflect this inherent structure and can thus be leveraged to improve cross-modal understanding. To demonstrate how Indra representations facilitate relational matching across modalities, we consider a dataset $\mathcal{D} = \{(U_i, Q_i)\}_{i=1}^n$ of n samples, where $U_i \in \mathcal{U}$ and $Q_i \in \mathcal{Q}$ correspond to instances from two distinct modalities, and $\mathcal{D} \subseteq \mathcal{U} \times \mathcal{Q}$. For single-modality scenarios, we define $U_i = Q_i,\ \forall i \in \{1, \ldots, n\}$. We use two pretrained foundation models $f : \mathcal{U} \to \mathbb{R}^{d^{\mathcal{X}}}$ and $g : \mathcal{Q} \to \mathbb{R}^{d^{\mathcal{Y}}}$ to extract modality-specific embeddings, where $d^{\mathcal{X}}$ and $d^{\mathcal{Y}}$ are the embedding dimensionalities of the two models, respectively. Based on these embeddings, we construct the Indra representations $\mathbf{I}^{\mathcal{U}}$ and $\mathbf{I}^{\mathcal{Q}}$ for each modality as follows:

$$\mathbf{I}^{\mathcal{U}}_{ij} = d(U_i, U_j), \quad \mathbf{I}^{\mathcal{Q}}_{ij} = d(Q_i, Q_j), \quad \forall i, j \in \{1, \ldots, n\}, \tag{5}$$

where the cost function d is defined in Equation 3. In practice, we may apply post-processing operations such as sparsification and normalization to enhance robustness:

$$\hat{\mathbf{I}}^{\mathcal{U}} = \texttt{operator}(\mathbf{I}^{\mathcal{U}}), \quad \hat{\mathbf{I}}^{\mathcal{Q}} = \texttt{operator}(\mathbf{I}^{\mathcal{Q}}), \tag{6}$$

where $\texttt{operator}(\cdot)$ denotes the chosen post-processing function. Unlike traditional representation approaches that reflect only internal characteristics of samples, the proposed Indra representations act as external representations, where each vector captures interdependencies by encoding the sample's relative profile within the dataset.

4 Experiments

To comprehensively assess the effectiveness of our Indra representation, we perform evaluations across a range of settings, including unimodal vision, vision–language, and speech–language tasks.

Table 1: Accuracy (%) on CIFAR-10 and CIFAR-100 under different Gaussian noise levels.

CIFAR-10	σ=0.0	σ=3.0	σ=5.0	σ=7.0
`ViT`	93.98	87.75	79.77	68.15
`Indra`	94.84	89.51	80.84	68.71
`Convnext`	97.00	85.89	80.10	65.85
`Indra`	97.21	92.86	81.59	66.64
`Dinov2`	99.19	95.21	85.57	76.54
`Indra`	99.14	96.87	89.73	77.92

CIFAR-100	σ=0.0	σ=3.0	σ=5.0	σ=7.0
`ViT`	79.45	54.69	35.76	27.45
`Indra`	80.09	69.00	51.59	32.74
`Convnext`	85.77	62.79	34.39	21.28
`Indra`	85.64	72.16	51.51	30.25
`Dinov2`	91.97	82.21	63.06	40.16
`Indra`	91.93	84.83	74.29	58.67

Table 2: Accuracy (%) on Office-Home dataset under different Gaussian noise levels.

Art	σ=0.0	σ=3.0	σ=5.0	σ=7.0
`ViT`	80.25	64.40	44.03	22.63
`Indra`	79.63	65.02	43.62	27.57
`Convnext`	89.71	62.76	27.98	12.14
`Indra`	87.86	59.88	28.81	14.20
`Dinov2`	87.65	73.05	46.91	27.78
`Indra`	87.04	70.99	47.53	27.37
Product	σ=0.0	σ=3.0	σ=5.0	σ=7.0
`ViT`	92.34	80.74	61.15	35.25
`Indra`	89.75	81.53	64.08	40.77
`Convnext`	96.62	84.91	44.26	19.37
`Indra`	96.73	85.92	45.61	22.18
`Dinov2`	96.73	93.24	83.33	60.70
`Indra`	96.40	92.79	84.46	60.59

Clipart	σ=0.0	σ=3.0	σ=5.0	σ=7.0
`ViT`	73.20	50.40	28.64	15.23
`Indra`	69.76	54.98	33.10	18.21
`Convnext`	83.62	54.07	20.85	09.74
`Indra`	82.70	57.85	25.09	11.34
`Dinov2`	88.43	75.14	51.09	31.04
`Indra`	87.29	76.63	54.75	33.56
Real	σ=0.0	σ=3.0	σ=5.0	σ=7.0
`ViT`	89.22	82.11	60.09	35.32
`Indra`	87.16	83.49	63.65	40.48
`Convnext`	93.46	82.11	38.30	17.78
`Indra`	93.35	84.63	40.71	19.61
`Dinov2`	92.78	87.39	71.44	48.51
`Indra`	92.89	88.53	73.17	49.89

4.1 Evaluation on Single Modality

Datasets. We first conduct classification tasks on the CIFAR-10 [37], CIFAR-100 [37], and Office-Home [72] datasets. For CIFAR-10 and CIFAR-100, we use the standard data splits provided by `torchvision.datasets` [51]. For Office-Home, we evaluate classification accuracy across four distinct domains: Art, Clipart, Product, and Real-World, using an $80/20$ split for training and testing. Each domain exhibits unique visual styles and distribution shifts, making the dataset a widely used benchmark for evaluating the robustness and generalization of vision models in object recognition tasks. Across all datasets, we adopt logistic regression (i.e., linear probing) to assess the quality of the extracted representations.

Foundation Models. For vision models, we evaluate `ViT` [13], `Convnext` [76], and `Dinov2` [60].

Evaluation Metrics. We assess model classification accuracy (%) using ground-truth labels. To investigate the robustness of Indra representations, we inject Gaussian noise into the features with varying standard deviations $\sigma \in \{0.0, 3.0, 5.0, 7.0\}$. For each noise level, we perturb the features accordingly and train a linear classifier on the noisy representations. This allows us to assess how classification performance degrades as the feature representations are increasingly corrupted by noise.

Analysis. In Tables 1 and 2, we report the classification results on the CIFAR-10, CIFAR-100, and Office-Home datasets. The results clearly show that stronger backbone models (e.g., `Dinov2`) lead to better performance for Indra representations across all noise levels. For instance, on CIFAR-100 with σ=0.0, our Indra representations achieve 91.93% accuracy using `Dinov2` as the backbone, compared to 85.64% with `Convnext` and 80.09% with `ViT`. This performance gap persists and even widens under higher noise: at $\sigma = 7.0$, Indra representations with `Dinov2` maintain 58.67%,

Table 3: Performance on image-text datasets $\mathcal{D}$ using different representations $\mathcal{R}$ (`Orign`: original, `Indra`: Indra representation) with various vision (`Vis-E`) and language (`Lan-E`) models.

$\mathcal{D}$	Vis-E	Lan-E	$\mathcal{R}$	Top-5		Top-10		Top-30		Top-50	
				T→I	I→T	T→I	I→T	T→I	I→T	T→I	I→T
MS-COCO	CLIP-I	CLIP-T	Orign	1.420	1.381	2.734	2.661	7.634	7.470	12.212	11.986
	ViT	BERT	Orign	0.482	0.483	0.967	0.966	2.911	2.905	4.863	4.846
			Indra	0.663	0.832	1.303	1.613	3.787	4.426	6.199	7.036
	ViT	Roberta	Orign	0.486	0.491	0.970	0.981	2.912	2.927	4.853	4.874
			Indra	1.048	0.880	2.065	1.749	5.970	5.149	9.702	8.446
	Convnext	BERT	Orign	0.396	0.474	0.837	0.950	2.603	2.851	4.412	4.755
			Indra	0.612	0.537	1.127	1.022	3.182	2.875	5.242	4.783
	Convnext	Roberta	Orign	0.492	0.480	0.985	0.964	2.962	2.889	4.940	4.824
			Indra	1.005	0.616	1.930	1.217	5.247	3.538	8.267	5.790
	Dinov2	BERT	Orign	0.496	0.473	0.991	0.947	2.969	2.852	4.936	4.760
			Indra	0.540	0.539	1.123	1.022	3.194	2.872	5.277	4.804
	Dinov2	Roberta	Orign	0.468	0.490	0.945	0.982	2.859	2.949	4.779	4.914
			Indra	1.016	0.949	1.978	1.863	5.603	5.370	9.021	8.766
NOCAPS	CLIP-I	CLIP-T	Orign	1.357	1.325	2.556	2.499	6.860	6.717	10.795	10.584
	ViT	BERT	Orign	0.479	0.474	0.956	0.947	2.864	2.844	4.769	4.742
			Indra	0.701	0.667	1.375	1.293	3.960	3.712	6.449	6.069
	ViT	Roberta	Orign	0.484	0.483	0.966	0.963	2.891	2.886	4.814	4.805
			Indra	0.924	0.727	1.792	1.419	5.014	4.102	8.011	6.700
	Convnext	BERT	Orign	0.449	0.451	0.904	0.910	2.754	2.752	4.640	4.604
			Indra	0.415	0.485	0.906	0.971	2.911	2.907	4.821	4.845
	Convnext	Roberta	Orign	0.472	0.462	0.944	0.926	2.833	2.781	4.721	4.647
			Indra	0.764	0.557	1.472	1.102	4.079	3.338	6.494	5.526
	Dinov2	BERT	Orign	0.465	0.439	0.928	0.883	2.701	2.674	4.500	4.485
			Indra	0.566	0.485	1.065	0.971	3.179	2.907	5.238	4.845
	Dinov2	Roberta	Orign	0.497	0.467	0.993	0.936	2.969	2.822	4.938	4.712
			Indra	0.830	0.774	1.604	1.513	4.549	4.324	7.335	7.019

while `Convnext` and `ViT` drop to 30.25% and 32.74%, respectively. In addition, as Gaussian noise increases, our Indra representations consistently retain higher classification accuracy compared to the original representations, highlighting their robustness in the classification tasks. The performance gains of Indra representations hold across multiple backbone architectures (i.e., `ViT`, `Convnext`, and `Dinov2`), indicating the broad applicability of the proposed method.

4.2 Evaluation on Vision & Language Modalities

Datasets. We adopt two widely used image-text datasets: MS-COCO [44] and NOCAPS [1] to evaluate performance on vision and language modalities. MS-COCO serves as a standard benchmark for image captioning and retrieval tasks, while NOCAPS poses a greater challenge due to its focus on novel object categories. We use the validation sets of both datasets for evaluation.

Foundation Models. We use the same vision models as in Section 4.1. For language models, we evaluate `BERT` [12] and `Roberta` [46], both of which are pretrained independently on unimodal data without cross-modal alignment. We include CLIP [62] as the aligned baseline for evaluation.

Evaluation Metrics: We adopt CLIPScore [27] as the evaluation metric, which measures semantic alignment between image and text based on the cosine similarity of their embeddings within the CLIP multimodal space. We report Top-k matching accuracy ($k \in \{5, 10, 30, 50\}$) in both text-to-image (T→I) and image-to-text (I→T) tasks.

Analysis. Table 3 compares the performance of original embeddings versus Indra representations across both datasets using various combinations of vision and language models. The results clearly

Table 4: Performance on audio-text dataset $\mathcal{D}$ using different representations $\mathcal{R}$ (`Orign`: original, `Indra`: Indra representation) with various audio (`Aud-E`) and language (`Lan-E`) models, where `*-b` and `*-l` refer to the base and large versions of model `*`, respectively.

$\mathcal{D}$	Aud-E	Lan-E	$\mathcal{R}$	Top-5		Top-10		Top-30		Top-50	
				T→A	A→T	T→A	A→T	T→A	A→T	T→A	A→T
TIMIT	CLAP-I	CLAP-T	Orign	1.062	1.836	2.046	3.611	5.726	10.146	9.204	16.225
	wav2vec-b	Roberta	Orign	0.319	0.072	0.670	0.276	2.214	1.409	3.655	2.750
			Indra	0.413	0.578	0.819	1.209	2.506	2.953	4.225	4.692
	wav2vec-l	Roberta	Orign	0.418	0.308	0.864	0.634	2.548	2.194	4.200	3.738
			Indra	0.363	0.578	0.908	1.243	2.783	2.956	4.640	4.318
	wavlm-b	Roberta	Orign	0.328	0.328	0.659	0.648	2.080	1.920	3.518	3.208
			Indra	0.436	0.578	0.913	1.241	2.493	2.976	4.227	4.338
	wavlm-l	Roberta	Orign	0.472	0.426	0.912	0.876	2.450	2.518	3.997	4.148
			Indra	0.504	0.578	0.971	1.229	2.693	2.967	4.387	4.335
	hubert-b	Roberta	Orign	0.280	0.431	0.553	0.793	1.978	2.069	3.282	3.240
			Indra	0.322	0.578	0.697	1.232	2.230	2.971	3.880	4.369
	hubert-l	Roberta	Orign	0.449	0.308	0.861	0.638	2.475	1.949	4.119	3.286
			Indra	0.454	0.578	0.878	1.248	2.610	3.003	4.461	4.255

demonstrate that Indra representations lead to consistent performance gains across different architectures and modalities. Significant improvements are observed in both T→I and I→T retrieval, highlighting the effectiveness of our method for cross-modal alignment. These findings suggest that the Indra representation offers a generalizable mechanism to improve vision-language matching, independent of model architecture or dataset. Nonetheless, there remains a noticeable gap in performance compared to the fully aligned CLIP model, indicating further room for improvement.

4.3 Evaluation on Speech & Language Modalities

Datasets. We adopt the TIMIT dataset [17] for audio and language modality experiments. TIMIT contains recordings from 630 speakers representing eight major dialect regions of American English, each reading ten phonetically rich sentences. The dataset provides time-aligned phonetic and word-level transcriptions along with $16k$Hz audio recordings.

Foundation Models. For audio models, we evaluate `wav2vec` [66], `wavlm` [7], and `hubert`[29], using both base and large variants. For the language modality, we use `Roberta` [46]. All models are pretrained independently on unimodal data, without any cross-modal alignment. As an aligned baseline, we include CLAP [77], an audio-language model jointly trained on paired audio-text data.

Evaluation Metrics: Similarly, we adopt CLAPScore [77] as the evaluation metric. We report Top-k matching ($k \in \{5, 10, 30, 50\}$) in both text-to-audio (T→A) and audio-to-text (A→T) tasks.

Analysis. Table 4 presents the audio-text matching results using different audio models. As shown, the Indra representations consistently improve matching performance in both directions across all model configurations. However, compared to the vision-language setting, the improvements in the audio-language modality are relatively modest. This is likely due to the comparatively weaker capacity of the audio models used. Nonetheless, we observe that larger audio models yield better matching accuracy, further supporting the notion that model capacity positively influences cross-modal alignment.

5 Related Work

Instance-Level Representation Learning. In conventional deep learning paradigms, each data instance is independently encoded into a fixed-dimensional vector, optimized using either supervised signals or unsupervised objectives. In supervised learning, representations are shaped by categorical labels [38, 25], whereas in unsupervised settings, objectives such as reconstruction (e.g., autoencoders [73]) or self-prediction (e.g., BERT [12]) guide the learning of instance-level representations. These

approaches primarily focus on learning standalone embeddings and do not explicitly model the relationships between samples, either in training or inference stages.

Contrastive Representation Learning. Contrastive methods such as SimCLR [8], MoCo [24], and BYOL [20] introduce pairwise relational inductive bias during training by encouraging the embeddings of similar (augmented) views to be close, while pushing apart dissimilar ones. Building on this framework, CLIP [62] extends contrastive learning to vision-language pretraining by aligning paired image-text embeddings within a shared multimodal space, enabling strong zero-shot performance. BLIP [42] further integrates contrastive and generative objectives to enhance cross-modal representation learning, achieving state-of-the-art results on several vision-language benchmarks. Despite these advances, the sample representations in contrastive frameworks remain inherently instance-centric during inference. Additionally, contrastive learning approaches typically require large-scale datasets to achieve satisfactory performance, as the training objectives rely heavily on diverse and abundant positive and negative pairs.

Graph-based Representation Learning. Graph neural networks [35] offer a framework for encoding relational information through message passing across graph-structured data. Variants such as GraphSAGE [22] and GAT [71] have improved flexibility and representation ability. However, most graph-based approaches rely on predefined adjacency structures or proximity assumptions [50], which may embed inductive biases that are misaligned with the latent semantics of the data. They typically operate over local k-hop neighborhoods [48], limiting their ability to capture long-range dependencies unless deeply stacked, which can lead to oversmoothing and degraded performance [43].

Attention-based Representation Learning. Transformer-based architectures [70] offer a powerful alternative by leveraging global self-attention to aggregate contextual information. Vision Transformers (ViT) [13] and Perceiver [31] have demonstrated that attention-based architectures can be highly effective even in non-sequential domains. While attention enables global interactions both during training and inference, the resulting token-level representations are determined through dynamic mixing rather than explicitly encoding pairwise or global sample-to-sample relationships, making their geometric interpretation less transparent.

In contrast to the above paradigms, our approach explicitly constructs each representation as a relational profile, specifically, a distance-based reflection of its relationship to all other samples. This design is motivated by the philosophical ontology of Indra's Net, where each entity reflects and is reflected by all others, forming a holistic network of interdependence. More importantly, our Indra representation is formally grounded in the Yoneda Lemma from enriched category theory, offering a theoretically sound and interpretable framework for relational representation learning.

6 Conclusion

In this paper, we present a theoretical and empirical investigation into the convergent behavior of unimodal foundation models. Motivated by the philosophical metaphor of Indra's Net and grounded in enriched category theory, we introduce the Indra representation as a relational encoding that reflects each sample through its relationships with all others. We demonstrate that this representation can be derived via the $\mathcal{V}$-enriched Yoneda embedding and instantiated practically using angular distance. Our theoretical analysis proves that the proposed Indra representations are unique, complete, and structure-preserving, offering a principled basis for training-free alignment of various foundation models. Through extensive experiments across single-modality, vision-language, and speech-language settings, we demonstrate that Indra representations improve robustness and latent cross-modal capabilities of unimodal foundation models. Our findings suggest a new perspective for bridging modalities, which emphasizes the importance of the intrinsic relational structure of data or reality.

Limitations. Constructing exact Indra representations requires a computational complexity of $\mathcal{O}(n^2d)$ and a memory complexity of $\mathcal{O}(n^2)$ for a dataset with n samples and embedding dimension d. This quadratic scaling potentially limits the direct applicability of the exact Indra representations to large-scale datasets. However, the scalability concern is addressable in practice. In the literature, there exists a rich body of work on approximating pairwise distances efficiently. For example, approximate nearest neighbor search (e.g., FAISS, HNSW), landmark-based approximation (e.g., K-means centroids, random subsampling), hashing-based methods (locality sensitive hashing), and sparsified graph constructions. From an application view, these techniques can be readily adapted to approximate the Indra representation at scale without sacrificing its robustness and interpretation.

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019.

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

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

[4] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[5] Juerg Beringer, J-F Arguin, RM Barnett, K Copic, O Dahl, DE Groom, C-J Lin, J Lys, H Murayama, CG Wohl, et al. Review of particle physics. *Physical Review D—Particles, Fields, Gravitation, and Cosmology*, 86(1):010001, 2012.

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

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

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020.

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

[10] Thomas Cleary. *The Flower Ornament Scripture: A Translation of the Avatamsaka Sutra*. Shambhala Publications, Boston, 1993.

[11] F.H. Cook. *Hua-Yen Buddhism: The Jewel Net of Indra*. Iaswr Series. Pennsylvania State University Press, 1977.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, 2019.

[13] 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. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[14] Michael Faraday. *Experimental Researches in Electricity*. Bernard Quaritch, London, 1859. Originally published as a series of papers between 1831 and 1855.

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

[16] John Rupert Firth, editor. *Studies in Linguistic Analysis*. Philological Society, Oxford, 1957. Special volume of the Philological Society.

[17] John S Garofolo, Lori F Lamel, William M Fisher, David S Pallett, Nancy L Dahlgren, Victor Zue, and Jonathan G Fiscus. Timit acoustic-phonetic continuous speech corpus. *(No Title)*, 1993.

[18] Kenneth J. Gergen. *Relational Being: Beyond Self and Community*. Oxford University Press, New York, 2009.

[19] David Griffiths. *Introduction to elementary particles*. John Wiley & Sons, 2020.

[20] Jean-Bastien Grill, Florian Strub, Florent Altmann, Corentin Tallec, Pierre-Henri Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21283, 2020.

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

[22] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, 2017.

[23] Harvard FAS CAMLab. Indra's net, 2022.

[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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

[27] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

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

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

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

[31] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021.

[32] Thomas P. Kasulis. *Engaging Japanese Philosophy: A Short History*. University of Hawai'i Press, Honolulu, 2018.

[33] Gregory Maxwell Kelly. *Basic Concepts of Enriched Category Theory*, volume 64 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, Cambridge, 1982.

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

[35] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

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

[37] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. Technical Report.

[38] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[39] F William Lawvere. Metric spaces, generalized logic, and closed categories. *Rendiconti del seminario matématico e fisico di Milano*, 43:135–166, 1973.

[40] Andrew Lee, Melanie Weber, Fernanda Viégas, and Martin Wattenberg. Shared global and local geometry of language model embeddings. *arXiv preprint arXiv:2503.21073*, 2025.

[41] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

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

[43] Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[44] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.

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

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

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

[48] Jianglin Lu, Yixuan Liu, Yitian Zhang, and Yun Fu. Scale-free graph-language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[49] Jianglin Lu, Hailing Wang, Yi Xu, Yizhou Wang, Kuo Yang, and Yun Fu. Representation potentials of foundation models for multimodal alignment: A survey. *arXiv preprint arXiv:2510.05184*, 2025.

[50] Jianglin Lu, Yi Xu, Huan Wang, Yue Bai, and Yun Fu. Latent graph inference with limited supervision. In *Advances in Neural Information Processing Systems*, 2023.

[51] TorchVision maintainers and contributors. Torchvision: Pytorch's computer vision library. `https://github.com/pytorch/vision`, 2016.

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

[53] Hazel Rose Markus and Shinobu Kitayama. Culture and the self: Implications for cognition, emotion, and motivation. In *College student development and academic life*, pages 264–293. Routledge, 2014.

[54] James Clerk Maxwell. *A Treatise on Electricity and Magnetism*. Clarendon Press, Oxford, first edition, 1873.

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

[56] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

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

[58] Jerry Ngo and Yoon Kim. 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, 2024.

[59] OpenAI. Gpt-4 technical report, 2024.

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

[61] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

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

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

[64] Emily Riehl. *Category theory in context*. Courier Dover Publications, 2017.

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

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

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

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

[69] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[70] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 2017.

[71] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. Graph attention networks. *stat*, 1050(20):10–48550, 2017.

[72] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017.

[73] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

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

[75] John Wentworth. Testing the natural abstraction hypothesis. *AI Alignment Forum*, 2021.

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

[77] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our abstract and introduction clearly showcased our main contributions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of the work in Section 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We provide the details in the supplementary.

   Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We include the implementation details and will release the code for reproducibility.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

Justification: We will release the code for reproducibility.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We conduct experiments following the standard protocol and include the experimental details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provide the details in the supplementary.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the details in the supplementary.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have read and follow the rules.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work has no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have cited the papers that created the datasets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: We do not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: We did not use LLMs in our paper except polishing our presentation.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.