Escaping Plato's Cave: Towards the Alignment of 3D and Text Latent Spaces

Souhail Hadgi[1] Luca Moschella[2,*] Andrea Santilli[2]
Diego Gomez[1] Qixing Huang[4] Emanuele Rodolà[2]
Simone Melzi[3] Maks Ovsjanikov[1]

[1]École polytechnique [2]Sapienza University of Rome
[3]University of Milano-Bicocca [4]The University of Texas at Austin

Abstract

Recent works have shown that, when trained at scale, uni-modal 2D vision and text encoders converge to learned features that share remarkable structural properties, despite arising from different representations. However, the role of 3D encoders with respect to other modalities remains unexplored. Furthermore, existing 3D foundation models that leverage large datasets are typically trained with explicit alignment objectives with respect to frozen encoders from other representations. In this work, we investigate the possibility of a posteriori *alignment of representations obtained from uni-modal 3D encoders compared to text-based feature spaces. We show that naive post-training feature alignment of uni-modal text and 3D encoders results in limited performance. We then focus on extracting subspaces of the corresponding feature spaces and discover that by projecting learned representations onto well-chosen lower-dimensional subspaces the quality of alignment becomes significantly higher, leading to improved accuracy on matching and retrieval tasks. Our analysis further sheds light on the nature of these shared subspaces, which roughly separate between semantic and geometric data representations. Overall, ours is the first work that helps to establish a baseline for post-training alignment of 3D uni-modal and text feature spaces, and helps to highlight both the shared and unique properties of 3D data compared to other representations.*

1. Introduction

Recent advances in multi-modal learning, particularly in vision-language models such as CLIP [47], have sparked interest in extending these successes to the 3D domain. Most current approaches primarily focus on training 3D encoders through triplet-based learning with pre-trained 2D vision and language encoders [32, 64, 69], leveraging new large-scale datasets such as Objaverse [9, 10], and showing promising results in tasks such as zero-shot shape recognition.

While CLIP [47] was trained with explicit alignment objectives between text and image representations, recent work has observed that *even when trained independently*, vision and text encoders tend to exhibit significant similarities [21, 37]. In particular, the learned latent spaces of pure (uni-modal) text and vision encoders have similar proximity structure [21] and can be aligned relatively easily *after training* with a small number of known anchor correspondences [2, 37, 43]. Furthermore, the degree of similarity in learned features across modalities strongly correlates with the quality of performance in various downstream tasks. One prominent interpretation of these results is that given sufficient scale of training data and model complexity, different representations are converging to a shared underlying structure of the physical world. This has given rise to the Platonic Representation Hypothesis [21], where different representations are viewed as projections of reality on particular modalities.

At the same time, as the physical world is inherently (at least) 3-dimensional, a natural question is how the structure of uni-modal vision or text encoders relates to the features learned *directly from* 3D data. This question raises several challenges: first, large-scale 3D datasets have only recently become available [9, 10]; second, most existing 3D "foundation models" are trained with *explicit alignment* objectives with respect to frozen 2D and text encoders, which limits the utility of post-training comparison. Finally, there is a lack of universally agreed-upon architectures and training objectives for 3D data, and many commonly-used architectures tend to have limited generalization power [62].

In this work, we initiate the first study on the relation between 3D and language representations. We formulate the task of post-training alignment between a range of 3D and text encoders and study the accuracy and utility of several alignment strategies.

Our first insight is that when trained on pure 3D data with self-supervised objectives, 3D encoders tend to lead to representations that align only very weakly to text-based representations. We believe that this observation already

*Currently at Apple.

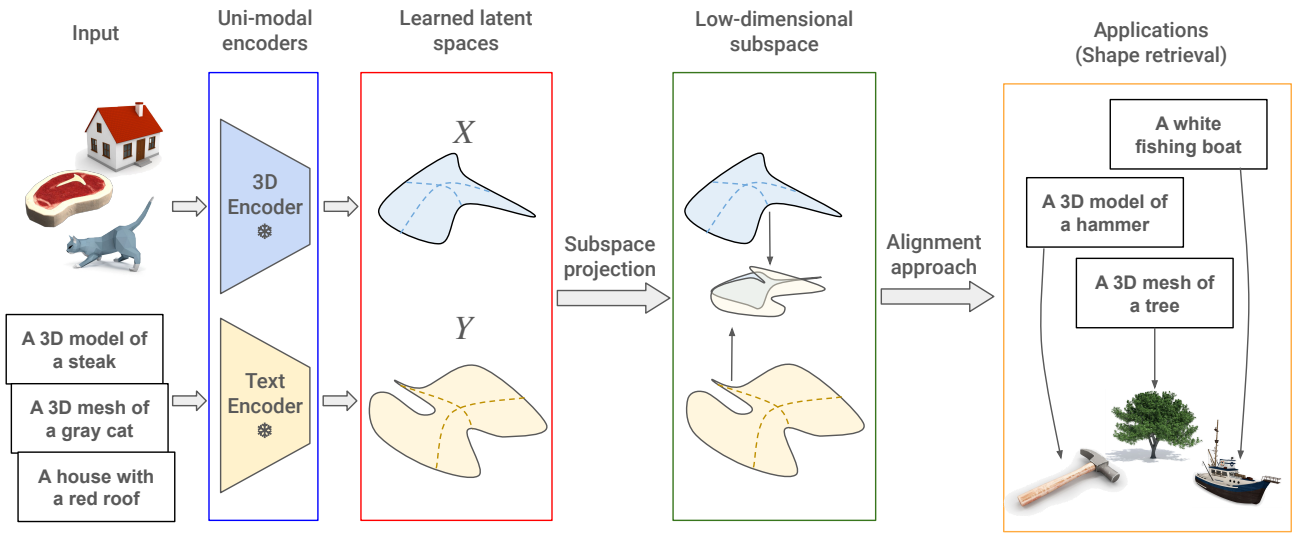

Figure 1. **A visual overview of the proposed approach.** is illustrated. From left to right: We begin with two distinct input collections—one consisting of 3D shapes and the other of textual prompts. In the blue box, independent, frozen uni-modal encoders map each modality into separate, high-dimensional latent spaces, shown in the red box. A dimensionality reduction procedure is applied to project these learned spaces into low-dimensional subspaces, represented in the green box. Finally, an alignment method registers the two low-dimensional subspaces, enabling cross-modal applications such as shape retrieval, with examples depicted in the yellow box.

highlights the difficulty of uni-modal training and sheds light on the scarcity of "pure" 3D foundation models.

Additionally, we reveal that while 3D and text latent spaces are not naturally aligned, effective cross-modal translation can be achieved through *subspace projection and alignment*. Our key insight is that by identifying and operating in correlated subspaces, we can improve the latent space alignment of 3D and text encoders without the need for expensive joint training. To achieve this, we introduce a simple but effective approach that combines Canonical Correlation Analysis (CCA) and existing alignment approaches such as affine translation [36] and local CKA [37]. This approach and our subsequent analysis extend previous works aimed at comparing learned feature spaces, but shows that careful *subspace* alignment can reveal subtle but important similarities, which are otherwise obscured in global comparisons. We provide a visual overview of the proposed approach in Fig. 1.

To summarize, our main contributions are as follows:

1. We extend the analysis of vision-text uni-modal latent space alignment to 3D uni-modal encoders and text, highlighting the limited similarity between these latent spaces and the low efficacy of current alignment approaches.
2. We propose an efficient approach for cross-modal alignment between text and 3D features that combines CCA for subspace selection and existing alignment methods. This approach improves alignment performance with minimal computational overhead, as demonstrated through experiments in matching and retrieval tasks. Our method establishes a baseline for 3D-text cross-modal understanding, providing an alternative to explicit multimodal pre-training for cross-modal tasks.
3. We observe a complementary structure between the spanned subspaces and the original feature spaces, enabling a distinction between geometric and semantic representations.

2. Related Work

Multi-modal representation learning. Multimodal representation learning has surged in recent years, driven by the success of image-language models [25, 30, 47, 52] that enable seamless cross-modal applications. These models serve as the backbone for tasks spanning from text-based image retrieval to generating high-quality visuals from natural language prompts. By aligning visual and linguistic features in a shared latent space, these models have paved the way for advanced interactions between modalities, setting the stage for applications where visual and textual information are jointly processed and understood.

Building on these advancements, vision-language models have recently been adapted for 3D point cloud representation, where 3D-image-text triplets [32, 63, 64, 69] enable contrastive pre-training. These models leverage powerful techniques such as momentum contrast [19] to align representations across modalities, allowing point cloud encoders to be pre-trained in a multimodal context. Additionally, tech-

niques that apply vision-language models directly to 2D projections of point clouds [68, 70] have expanded cross-modal applications to 3D, including zero-shot shape classification, where models trained on 2D-image-text data demonstrate strong performance on 3D-related tasks.

In the domain of multimodal synthesis, recent efforts in 2D and 3D model generation have leveraged diffusion models to tackle complex generation tasks. These models use robust priors, often trained on vast datasets, which allow them to create high-fidelity outputs in scenarios such as novel view synthesis and realistic 3D reconstructions from single RGB images [33, 65]. By combining synthetic data with diffusion-based priors, these frameworks achieve impressive zero-shot generalization and geometry-consistent 3D synthesis, opening new avenues for applications where generating realistic 3D content from limited information is critical.

Representational similarity. The study of representation similarity across neural networks has seen a significant rise in interest, spurred largely by seminal works originating in neuroscience and computational cognitive science. These fields have long been invested in understanding the nature and alignment of cognitive representations, providing a foundational basis that has influenced the current trajectory in machine learning [42, 56]. Based on these insights, various works [1, 31, 40, 41, 43, 44, 59] provide evidence of an intrinsic connection between independently trained networks; the similarity is especially notable among large-scale models, with works such as [39, 40, 55] exploring the phenomenon. In the computer vision and pattern recognition areas, the line of work on similarity-based representations, pioneered by Duin and Pekalska [45], has also been seminal. This line of research, which examines data through the lens of similarity rather than feature attributes, has provided robust frameworks for classification and clustering [5], enabling models to generalize across patterns and variations in complex datasets. Although mostly empirical in nature, these observations find theoretical support in the study of harmonics in neural networks weights [38], Independent Component Analysis [22, 23, 26, 50] and Independent Mechanism Analysis [13, 14, 54]. These works suggest that, when capturing the same underlying data generative factors, deep learning models may converge towards similar structures despite their complexity and non-linearity.

Latent space alignment. More recently, a range of approaches has been developed to *align* latent spaces within the same modality [3, 7, 12], as well as across different modalities [36, 58, 67], particularly between visual and textual domains. Techniques such as Procrustes analysis [60] and several similarity metrics [27, 53, 57, 61], including centered kernel alignment (CKA) [8, 28, 37], have proven instrumental in aligning representations. These methods offer various strategies for quantifying similarity between feature spaces, allowing us to examine cross-modal interactions at a deeper level. HHowever, these methods often focus on aligning entire latent spaces, potentially missing meaningful similarities confined to specific subspaces.

Our approach builds upon CCA [20, 49], a pivotal tool in pattern recognition and multi-view learning applications [16, 51]. This technique identifies maximally correlated subspaces, enabling more refined alignment across modalities by isolating core, mutually relevant components. Leveraging CCA, alignment methods can extend into complex domains, such as connecting 3D and textual latent spaces, as we demonstrate in the following.

3. Method

We compare the similarity between latent (feature) spaces of various 3D and text encoders, introducing a new approach that builds on existing alignment methods to improve their effectiveness. Below, we outline the tools used to measure and align these latent spaces.

3.1. Preliminaries

Centered Kernel Alignment. CKA is a similarity measure frequently used in recent studies [8, 28] to compare representations in neural network feature spaces. Given feature matrices $\mathbf{X} \in \mathbb{R}^{n \times p}$ and $\mathbf{Y} \in \mathbb{R}^{n \times q}$, we apply kernel functions k and l to obtain kernel representations $\mathbf{K} = k(\mathbf{X}, \mathbf{X}) \in \mathbb{R}^{n \times n}$ and $\mathbf{L} = l(\mathbf{Y}, \mathbf{Y}) \in \mathbb{R}^{n \times n}$. CKA is defined as:

$$\mathrm{CKA}(\mathbf{K}, \mathbf{L}) = \frac{\mathrm{HSIC}(\mathbf{K}, \mathbf{L})}{\sqrt{\mathrm{HSIC}(\mathbf{K}, \mathbf{K})\mathrm{HSIC}(\mathbf{L}, \mathbf{L})}} \tag{1}$$

where HSIC is the Hilbert-Schmidt Independence Criterion [15] and can be written as

$$\mathrm{HSIC}(\mathbf{K}, \mathbf{L}) = \frac{1}{(n-1)^2}\mathrm{tr}(\mathbf{KHLH})\,, \tag{2}$$

where $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$.

Canonical Correlation Analysis. CCA [20, 49] is a statistical method that finds linear projections of two datasets, maximizing their correlation in a shared latent space.

Formally, given two sets of zero-centered variables $\mathbf{X} \in \mathbb{R}^{n \times d_1}$ and $\mathbf{Y} \in \mathbb{R}^{n \times d_2}$, CCA seeks two projection matrices $\mathbf{W_X} \in \mathbb{R}^{d_1 \times k}$ and $\mathbf{W_Y} \in \mathbb{R}^{d_2 \times k}$ that map $\mathbf{X}$ and $\mathbf{Y}$ into a common k-dimensional space, maximizing the correlation between the projections $\mathbf{XW_X}$ and $\mathbf{YW_Y}$. The optimization can be expressed as:

$$\max_{\mathbf{W_X}, \mathbf{W_Y}} \mathrm{corr}(\mathbf{XW_X}, \mathbf{YW_Y})\,, \tag{3}$$

where $\mathrm{corr}(\cdot, \cdot)$ represents the correlation between the projected variables.

3.2. Alignment approaches

Recent developments in latent space alignment have introduced various methods. In this work, we examine both the affine transformation approach in [36] and the CKA-based matching approach from [37], observing their limited effectiveness in 3D-text latent space alignment, and propose a method to address these limitations.

Latent Space Translation through Affine Transformation [36]. It is possible to estimate an affine transformation T that maps one latent space $\mathcal{X}$ onto another latent space $\mathcal{Y}$ such that $T(\mathbf{x}) = \mathbf{R}\mathbf{x} + \mathbf{b}, \forall \mathbf{x} \in \mathcal{X}$. To compute T, we assume access to an *anchor* subset consisting of ground-truth paired samples from both latent spaces. These anchors enable us to determine T by optimizing through gradient descent or, if $\mathbf{b} = \mathbf{0}$ by minimizing the least squares error. It assumes that $\mathcal{X}$ and $\mathcal{Y}$ share the same dimensionality and are normalized to zero mean and unit variance. These constraints can be enforced by zero-padding the smaller latent space.

Local CKA-based retrieval and matching [37]. The CKA metric computed between two sets of features is sensitive to ordering and maximized when ground-truth pairs are aligned, this insight can be used to match unseen data by including them in a well-aligned anchors set. Formally, starting from aligned set of features $\mathbf{X}_A$ and $\mathbf{Y}_A$, we compute for a query pair $(\mathbf{x}_q, \mathbf{y}_q)$ its local CKA defined as:

$$\text{localCKA}(\mathbf{x}_q, \mathbf{y}_q) = \text{CKA}(\mathbf{K}_{[\mathbf{X}_A, \mathbf{x}_q]}, \mathbf{K}_{[\mathbf{Y}_A, \mathbf{y}_q]}), \quad (4)$$

where $[\mathbf{X}, \mathbf{x}]$ denotes the column-wise concatenation of matrix $\mathbf{X}$ and vector $\mathbf{x}$. Local CKA calculates similarity in a pairwise manner, accounting for anchor alignment while being sensitive to ordering among query pairs. We introduce all possible query pairs, where correctly matched pairs exhibit the highest local CKA.

3.3. Proposed method.

Our goal is to reliably align the latent spaces of pre-trained 3D and text encoders. Given a dataset of n caption-point cloud pairs, each embedded in the corresponding feature space, and represented as matrices $\mathbf{X} \in \mathbb{R}^{n \times p}$ and $\mathbf{Y} \in \mathbb{R}^{n \times q}$ respectively, we first select a subset of anchor pairs that will guide our translation process, denoted as $\mathbf{X}_A \in \mathbb{R}^{n_A \times p}$ and $\mathbf{Y}_A \in \mathbb{R}^{n_A \times q}$. Building upon the mathematical foundations introduced in Section 3.2, we develop a pipeline that combines dimensionality reduction with the previously introduced alignment methods.

Common Subspace Projection. In our work, we show that 3D and text latent spaces can effectively be aligned in low-dimensional connected subspaces. We begin by applying CCA to the anchor pairs to identify a shared k-dimensional subspace (with $k < p, q$) that connects point cloud and text latent spaces. This yields projection matrices $\mathbf{W}_{\mathbf{X}^A} \in \mathbb{R}^{p \times k}$ and $\mathbf{W}_{\mathbf{Y}^A} \in \mathbb{R}^{q \times k}$. All samples are then projected into this reduced space:

$$\mathbf{X}^r = \mathbf{X}\mathbf{W}_{\mathbf{X}_A}, \quad \mathbf{Y}^r = \mathbf{Y}\mathbf{W}_{\mathbf{Y}_A}. \quad (5)$$

In particular, we refer to the reduced anchors as $\mathbf{X}^r_A \in \mathbb{R}^{n_A \times k}$ and $\mathbf{Y}^r_A \in \mathbb{R}^{n_A \times k}$.

Our experiments show that projecting 3D and text latent spaces into a lower-dimensional, shared subspace improves alignment by isolating features that are highly correlated across modalities

Alignment of projected latent spaces. Given the projected latent spaces, we can apply either of the previously described alignment methods in the reduced space. For the affine transformation approach, we learn a mapping between $\mathbf{X}^r$ and $\mathbf{Y}^r$ using the projected anchor pairs $\mathbf{X}^r_A, \mathbf{Y}^r_A$ to optimize the transformation parameters $\mathbf{R}$ and $\mathbf{b}$:

$$T(\mathbf{X}^r) = \mathbf{R}\mathbf{X}^r + \mathbf{b}, \quad (6)$$

Alternatively, we can employ the local CKA-based matching approach in the projected space. For a query pair $(\mathbf{x}^r_q, \mathbf{y}^r_q)$, we compute its local CKA using the projected anchor sets:

$$\text{localCKA}(\mathbf{x}^r_q, \mathbf{y}^r_q) = \text{CKA}(\mathbf{K}_{[\mathbf{X}^r_A, \mathbf{x}^r_q]}, \mathbf{K}_{[\mathbf{Y}^r_A, \mathbf{y}^r_q]}), \quad (7)$$

4. Experimental setup

4.1. Pre-training Dataset

Prior works in point cloud pre-training, particularly within uni-modal frameworks, have relied heavily on ShapeNet [4], a dataset of 51,300 annotated 3D synthetic shapes spanning 55 categories. ShapeNet has been instrumental in advancing foundational methods, yet it remains limited by the relatively narrow scope of categories. With the release of Objaverse [9], which includes over 800,000 shapes across diverse real-world categories, a new standard for large-scale representation learning in the 3D domain has emerged. Objaverse's extensive shape diversity makes it ideal for both uni-modal and multi-modal learning. Despite these advantages, there is a lack of works that explore uni-modal pre-training specifically on Objaverse.

4.2. Encoders

We explore both multi-modal and uni-modal 3D and text encoders across varying levels of model complexity.

Multi-modal 3D Encoders. For multi-modal pre-training, we use OpenShape, ULIP-2 and Uni3D [32] pre-trained models, trained on point cloud, image, and text triplets of Objaverse with a contrastive pre-text task to align 3D encoders with frozen CLIP encoders. We adopt the Point-BERT-based

Method	3D Encoder	CLIP		RoBERTa		BERT	
		Matching accuracy	Top-5 retrieval	Matching accuracy	Top-5 retrieval	Matching accuracy	Top-5 retrieval
Multi-modal 3D encoder							
Affine + Subspace Projection	OpenShape	67.6	85.6	55.2	75.8	45.0	70.6
Affine + Subspace Projection	ULIP-2	65.6	85.2	47.2	70.6	35.2	59.8
Affine + Subspace Projection	Uni3D	61.4	81.6	45.8	63.8	34.4	47.2
Uni-modal 3D encoder							
Affine	PointBert	15.8	23.6	7.8	15.6	6.4	13.4
Affine	SparseConv	11.0	34.4	6.0	20.0	4.2	16.2
Affine	Pointnet	18.4	21.8	8.0	10.2	9.6	12.2
Affine + Subspace Projection (Ours)	PointBert	**30.8**	42.2	**23.2**	28.4	15.6	18.2
Affine + Subspace Projection (Ours)	SparseConv	21.4	45.6	19.2	16.8	15.8	15.0
Affine + Subspace Projection (Ours)	Pointnet	25.2	36.6	22.4	20.8	**16.6**	16.0
Local CKA	PointBert	5.8	15.2	1.8	1.4	1.0	4.39
Local CKA	SparseConv	3.4	13.6	1.79	1.6	0.6	3.8
Local CKA	Pointnet	6.6	18.0	2.4	2.0	1.0	5.0
Local CKA + Subspace Projection (Ours)	PointBert	29.4	**60.19**	17.0	**42.4**	15.0	**37.2**
Local CKA + Subspace Projection (Ours)	SparseConv	19.0	56.9	15.8	34.0	10.0	30.8
Local CKA + Subspace Projection (Ours)	Pointnet	26.8	53.0	18.6	40.2	14.3	38

Table 1. **Matching and retrieval performance across 3D and text encoders using different alignment approaches.** We use 30,000 anchors for subspace projection and affine transformation approaches, and 1,000 anchors for local CKA. A query set of 500 is uniformly sampled, with results averaged over 3 different seeds. The subspace dimension is fixed at 50. Our approach (Ours) consistently demonstrates improved matching and retrieval performance, with multi-modal 3D encoders setting the upper bound for performance. Additional top-k retrieval metrics are provided in the supplementary.

variant of each model [66]. We primarily focus on OpenShape for simplicity but generalize results to ULIP-2 and Uni3D.

Uni-modal 3D Encoders. For the uni-modal setup, we pre-train PointBERT on Objaverse using its original pretext tasks, which include masked point reconstruction and a uni-modal contrastive loss to encourage robust shape representation. We also explore two additional architectures: a Sparse convolution (MinkowskiNet [6]) model and a simpler architecture in PointNet [46], each pre-trained using a shape-level contrastive learning method [18]. This approach contrasts different partial views of input shapes. Across all 3D encoders in this setup, we fix the latent dimension at 512 to maintain consistency in representation space comparisons.

Text encoders. We use the text encoder from OpenCLIP ViT-bigG-14 [24], chosen to match text encoder used in OpenShape. Additionally, we examine alignment with purely uni-modal text encoders by including BERT [11] and RoBERTa [34], we also evaluate the alignement with T5 [48] in the supplementary.

Across all pre-training setups, parameters are kept consistent to facilitate direct comparisons. We detail the technical details in the supplementary.

4.3. Downstream tasks

We evaluate our alignment approach on Objaverse-LVIS [17], a human-verified test subset of Objaverse that contains 1,156 object categories. This subset is specifically reserved for evaluation and is unseen during pre-training. The captions are generated with Cap3D [35], which provides enhanced descriptive text for each 3D shape. Our evaluation framework consists of two main tasks: matching and retrieval. The matching task involves finding the correct permutation of captions for perfect matching given a shuffled set of query images and their corresponding captions; we utilize the linear sum assignment approach to perform this task [29]. For the retrieval task, the model must identify the correct 3D object from the query set based on a text caption. These tasks are particularly effective in measuring the cross-modal capabilities of encoders, and have been evaluated in prior Vision-Text studies [37]. While our main results emphasize the matching and top-5 retrieval tasks, we provide evaluation of top-1 and top-10 retrieval metrics in the supplementary.

5. Results

5.1. Are 3D and Text Latent Spaces similar ?

To evaluate the inherent similarity between 3D and text feature spaces without alignment, we compute linear CKA scores for both unimodal and multimodal encoders. The results are shown in Fig. 2.

Comparison of 3D-Text and Vision-Text alignment. Prior work [21] reports CKA alignment values ranging from approximately 30% to 48% between different uni-modal vision and text encoders (see Figure 13 in [21]). In stark contrast, we find that the default alignment between 3D and text latent spaces is significantly weaker, with a maximum score of 0.12 observed for the uni-modal PointBERT and CLIP pair. This substantial gap underscores a key insight: unlike vision and text encoders, 3D encoders did not converge to structures similar to text.

Alignment favors multi-modality. We observe that 3D multi-modal encoders demonstrate the highest CKA score with text encoders, particularly with CLIP's text encoder.

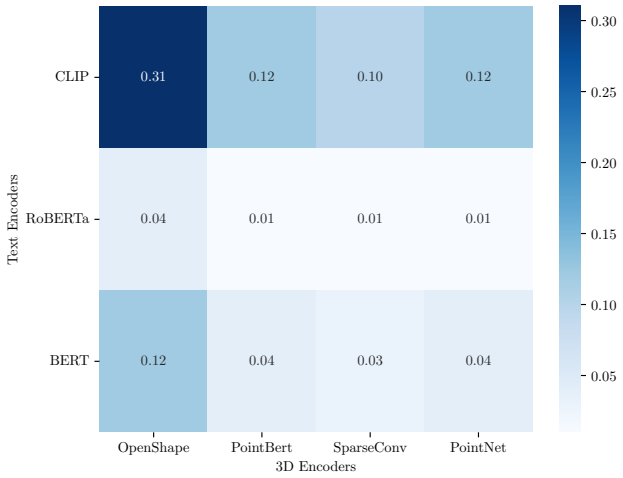

Figure 2. **Linear CKA scores between text and 3D encoders without alignment.** Higher scores reflect stronger alignment between encoder pairs, with the strongest alignment observed between multi-modal 3D encoders (OpenShape) and CLIP text encoder due to their shared training on aligned representations. Uni-modal 3D encoders show significantly lower alignment with text encoders, although slightly higher with the CLIP text encoder.

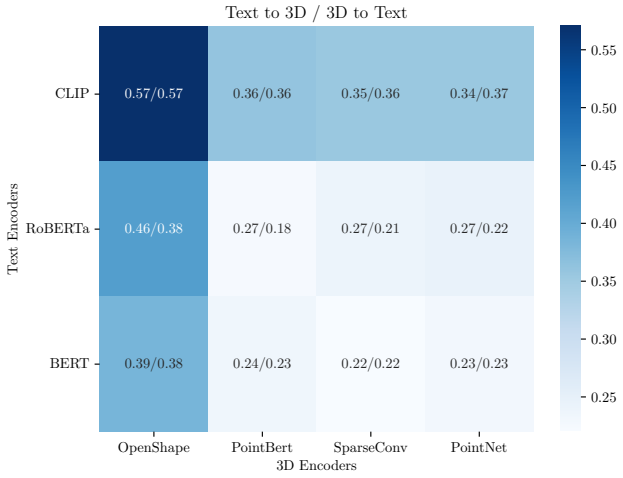

Figure 3. **Linear CKA scores between text and 3D encoders after affine translation.** Affine translation results in a consistent increase in similarity for both 3D-to-text and text-to-3D directions.

This behavior is expected, given that 3D multi-modal encoders are explicitly trained to align with CLIP latent spaces, which share the same textual modality. Even among uni-modal 3D encoders, alignment with the CLIP text encoder is notably higher (e.g. 0.12 for PointBERT-CLIP) compared to alignment with uni-modal text encoders like RoBERTa (e.g., 0.04 for PointBERT-RoBERTa). This suggests that the visual understanding embedded in CLIP's text encoder extends beyond image and text domains to include 3D representations, despite the lack of explicit alignment during pre-training.

5.2. Latent Space Alignment results

We evaluate the performance of the alignment approaches outlined in Sec. 3 using matching and retrieval tasks, and analyze their effectiveness in aligning 3D and text encoders. Unless otherwise specified, we fix the subspace dimension to $d = 50$ and the number of anchors to $30,000$. For downstream tasks, we uniformly sample a query set of size 500 and average results over 3 different seeds. For the affine transformation approach, we present results for the text-to-3D direction, noting that similar performance is achieved in the 3D-to-text direction as seen in Fig. 3.

Existing approaches enable limited alignment We first assess whether previous successful approaches—affine transformation and local CKA—can achieve meaningful 3D-text alignment. As shown in Tab. 1, both methods yield modest alignment improvements over the unaligned baseline, where uni-modal 3D encoders start near zero in matching and retrieval tasks. These results suggest a small alignment shift. However, even with this improvement, alignment remains significantly lower than the alignment achieved in vision-text benchmarks [37]. This finding hints at the limits of uni-modal 3D encoders in achieving similar Vision-Text alignment performance, and might necessitate a different approach to align their latent space with text encoders.

Importance of subspace projection (Ours). While aligning the latent spaces of 3D and text encoders achieved some success with existing methods, results remained well below vision-text alignment benchmarks. Motivated by this limitation, we propose aligning lower-dimensional subspaces, based on the hypothesis that 3D and text representations might intersect within a shared latent subspace. Using a validation set, we report in Fig. 4 the impact of subspace dimension, obtained through our CCA approach (see Sec. 3.3) combined with the affine transformation on the top-5 retrieval accuracy of uni-modal PointBert. While affine alignment alone yields better performance at higher dimensions, our subspace projection approach significantly outperforms it when the subspace dimension is reduced. This finding indicates that alignment quality is optimized within carefully chosen lower-dimensional spaces, reinforcing our hypothesis that 3D-text alignment is more successful within targeted subspaces. The results in Tab. 1 further validate this approach, showing a substantial increase in matching and retrieval performance across all 3D and text encoder pairs, outperforming both alignment approaches when they operate on the original latent space.

Multi-modal encoders as an upper bound of performance. We include 3D multi-modal encoders for two main reasons: (1) As an upper bound for alignment performance, (2) To study how alignment degrades across different text

encoders, including those that were not used during pre-training. In this context, the inclusion of 'Affine + Subspace Projection' multi-modal results in Tab. 1 shows how using our approach significantly narrows the gap between multi-modal and uni-modal performance. However, to illustrate point 1. above, we also measured with cosine similarity the alignment of OpenShape (a multi-modal 3D encoder) and the same CLIP text encoder used during pre-training, achieving 0.94 for top-5 retrieval accuracy, an expected result since 3D and text encoders are explictly aligned during pre-training.

3D encoder complexity's low impact on alignment. PointBERT outperforms the more complex SparseConv among uni-modal 3D encoders, suggesting that increasing model complexity alone does not guarantee improved alignment in 3D-text tasks. Surprisingly, even PointNet—a relatively simple model—achieves similar alignment scores, showing that factors other than model complexity may play a pivotal role in 3D-text alignment. This observation contrasts with vision-text alignment results [21, 37], where complex models typically leverage large datasets more effectively. Our findings thus indicate a distinctive aspect of 3D-text alignment: model simplicity does not impede, and may even aid, the interpretability and compatibility of learned features for cross-modal alignment.

Different alignment techniques have different strengths. Our experiments reveal that different alignment techniques offer complementary strengths across tasks. For instance, affine transformation proves particularly effective for matching tasks across all 3D encoders, while local CKA shows superior performance in top-5 retrieval accuracy. This suggests that while some methods excel in precision tasks, others might better capture broader semantic nuances, making them more suitable for retrieval. Together, these observations imply that a hybrid approach could leverage the unique strengths of each method, opening up promising directions for future cross-modal applications.

Scaling of our approach. In Fig. 5, we explore the scalability of our approach by analyzing how performance responds to increasing the number of anchors. Notably, our subspace projection method scales effectively with anchor count, reaching a plateau before requiring the full dataset (over 800,000 shapes). This scalability highlights the approach's efficiency in learning robust 3D-text mappings with a limited subset of anchor pairs. However, we observe a leveling off in performance gains beyond a certain anchor count, likely due to the constraints imposed by the low-dimensional subspace. This suggests that while our approach is computationally efficient and data-efficient, its reliance on a fixed lower-dimensional subspace could limit its adaptability to larger, more diverse datasets.

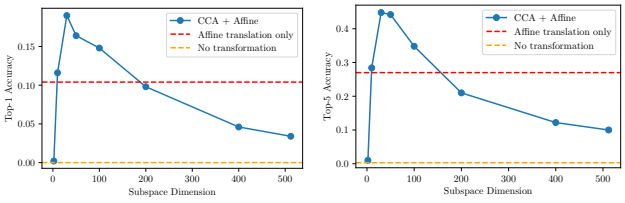

Figure 4. **Impact of subspace dimensionality on retrieval performance.** Comparison of three approaches: our proposed CCA + affine translation method (blue), affine translation without subspace projection (red), and baseline feature space alignment without transformation (orange). Results are shown for the uni-modal PointBERT and CLIP text encoder, with generalizations to other encoders provided in the supplementary.

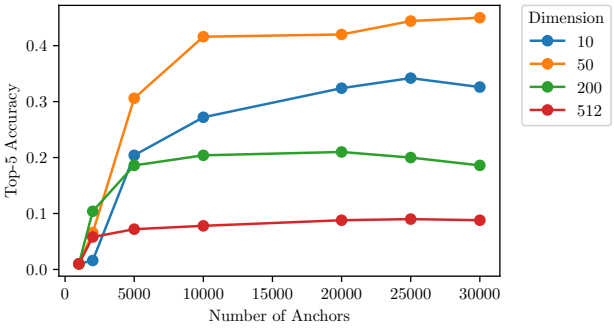

Figure 5. **Effect of anchor set size on retrieval performance.** Validation set results with different subspace dimensions show that retrieval performance improves as the anchor subset size increases but eventually reaches a plateau. Results are shown for the uni-modal PointBERT and CLIP text encoder.

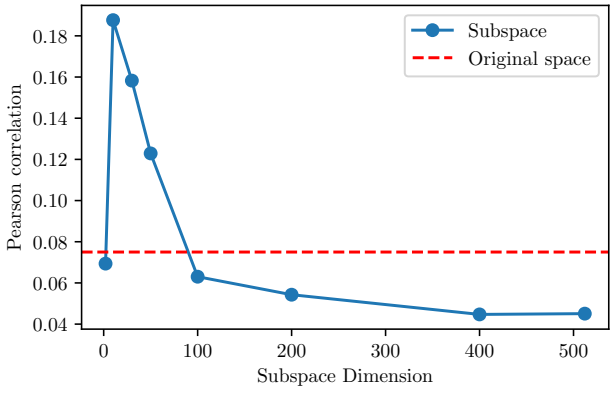

Figure 6. **Pearson correlation between shape query chamfer distances and pairwise distances in the projected text latent subspace.** We observe a higher Pearson correlation in the projected text latent subspace with optimal subspace dimension. Results are shown for the uni-modal PointBERT and CLIP text encoders.

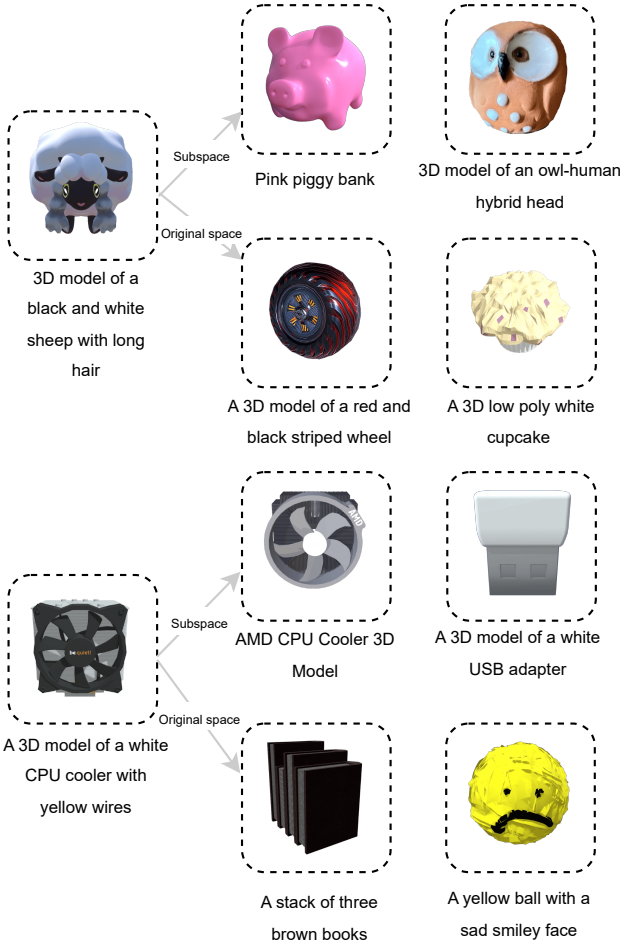

Figure 7. **Shape retrieval comparison between original and reduced latent spaces.** For a given query shape, we retrieve the closest match based on cosine similarity. Results demonstrate a higher semantic understanding in the reduced 3D latent subspace compared to the original latent space. Shown for uni-modal Point-BERT and CLIP.

5.3. Geometries vs. semantics.

Our quantitative evaluation shows that low-dimensional subspace projection significantly improves latent space alignment in the 3D-Text setting. To better understand the characteristics of these subspaces spanned relative to the original spaces, we analyze the increase in semantic and geometric knowledge within the projected spaces.

Increased geometric awareness of the text latent subspace. To quantify the increase in geometric awareness within the projected text subspace, we compute the Pearson correlation between the Chamfer distances of a query set of 500 shapes and the pairwise distances between feature vectors within the text subspace and compare these results to those obtained from the original text latent space. Results in Fig. 6 show that the optimal subspace dimension yields a high correlation with Chamfer distances, signaling an increase in the sensitivity of the text encoder to geometric properties when projected into this specific subspace.

Increased semantic understanding of the 3D latent subspace. To assess the improved semantic capacity of the 3D subspace, we conducted a qualitative analysis of shape retrieval performance in both the reduced and the original latent spaces (Fig. 7). Given a set of query shapes, we observe that the reduced subspace retrieves shapes with stronger semantic similarity (e.g., retrieving animals for a query sheep shape), while the original latent space primarily favors geometric resemblance (e.g retrieving stacked square shapes when querying a *'3D model of a CPU'*). This shift suggests that our subspace projection method enables the 3D encoder to capture semantic relationships that are absent in the full latent space. This effectively enables cross-modal applications and explains its higher matching and retrieval performance.

6. Conclusion, Limitations and Future Work

In this work, we present the first study investigating latent space alignment between 3D and text pre-trained encoders. Building on the hypothesis that these modalities share semantic connections within lower-dimensional subspaces, we propose an effective approach combining CCA projection with affine transformation estimation to translate between modalities' latent spaces. Our empirical results show that optimal cross-modal performance is achieved through low-dimensional subspace projection, and our method successfully improves alignment across diverse 3D and text uni-modal encoders. While CLIP-based multi-modal encoders establish performance upper bounds, we enable significant cross-modal capabilities in uni-modal encoders previously limited to single-modality tasks. We also demonstrate that semantic understanding can be extracted from geometry-aware latent spaces of uni-modal 3D encoders.

Although our work focused on Objaverse, which is the first large-scale 3D dataset, it would be interesting to consider how scaling on Objaverse-XL [10] would affect the alignment quality between 3D and text encoders. Moreover, in this work we do not distinguish object-level vs. scene-level annotations, and decomposing objects or scenes into their composing blocks could shed light onto the compositionality of the learned representations. Finally, the subspace alignment method that we introduce in this work can be broadly applicable to other representations as well. In the future, we plan to use it to investigate the limitations of alignment observed in other representations. In particular, even when trained at significantly higher data scales, images and text representations do not *align perfectly* [21], and it would be interesting to reveal the unique and complementary nature of different modalities via subspace analysis.

Acknowledgements Parts of this work were supported by the ERC Starting Grant 758800 (EXPROTEA), ERC Consolidator Grant 101087347 (VEGA), ANR AI Chair AIGRETTE, as well as gifts from Ansys and Adobe Research. This work was also supported by the Galileo 2022 fellowship from the Università Italo Francese/ Université Franco Italienne (UIF/UFI) within the project G22_4 titled "Multimodal Artificial Intelligence for 3D shape analysis, modeling and applications".

References

[1] Lisa Bonheme and Marek Grzes. How do variational autoencoders learn? insights from representational similarity. *arXiv preprint arXiv:2205.08399*, 2022. 3

[2] Irene Cannistraci, Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, and Emanuele Rodolà. Bootstrapping parallel anchors for relative representations. In *The First Tiny Papers Track at ICLR 2023, Tiny Papers at ICLR 2023*, 2023. 1

[3] Irene Cannistraci, Luca Moschella, Marco Fumero, Valentino Maiorca, and Emanuele Rodolà. From bricks to bridges: Product of invariances to enhance latent space communication. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[4] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 4

[5] Yihua Chen, Eric K. Garcia, Maya R. Gupta, Ali Rahimi, and Luca Cazzanti. Similarity-based classification: Concepts and algorithms. *Journal of Machine Learning Research*, 10(27): 747–776, 2009. 3

[6] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3075–3084, 2019. 5

[7] Donato Crisostomi, Irene Cannistraci, Luca Moschella, Pietro Barbiero, Marco Ciccone, Pietro Lio, and Emanuele Rodolà. From charts to atlas: Merging latent spaces into one. In *NeurIPS 2023 Workshop on Symmetry and Geometry in Neural Representations*, 2023. 3

[8] MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. Reliability of cka as a similarity measure in deep learning. *arXiv preprint arXiv:2210.16156*, 2022. 3

[9] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. 2023 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13142–13153, 2022. 1, 4

[10] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-xl: A universe of 10m+ 3d objects. *arXiv preprint arXiv:2307.05663*, 2023. 1, 8

[11] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 5

[12] Marco Fumero, Marco Pegoraro, Valentino Maiorca, Francesco Locatello, and Emanuele Rodolà. Latent functional maps: a spectral framework for representation alignment. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3

[13] Shubhangi Ghosh, Luigi Gresele, Julius von Kügelgen, Michel Besserve, and Bernhard Schölkopf. Independent mechanism analysis and the manifold hypothesis, 2023. 3

[14] Luigi Gresele, Julius Von Kügelgen, Vincent Stimper, Bernhard Schölkopf, and Michel Besserve. Independent mechanism analysis, a new concept? In *Advances in Neural Information Processing Systems*, 2021. 3

[15] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005. 3

[16] Chenfeng Guo and Dongrui Wu. Canonical correlation analysis (cca) based multi-view learning: An overview. *arXiv preprint arXiv:1907.01693*, 2019. 3

[17] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 5

[18] Souhail Hadgi, Lei Li, and Maks Ovsjanikov. To supervise or not to supervise: Understanding and addressing the key challenges of 3d transfer learning. *arXiv preprint arXiv:2403.17869*, 2024. 5

[19] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. arxiv e-prints, art. *arXiv preprint arXiv:1911.05722*, 2019. 2

[20] Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics: methodology and distribution*, pages 162–190. Springer, 1992. 3

[21] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024. 1, 5, 7, 8

[22] Aapo Hyvarinen and Hiroshi Morioka. Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc. 3

[23] Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ICA Using Auxiliary Variables and Generalized Contrastive Learning. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR. 3

[24] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 5

[25] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2

[26] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational Autoencoders and Nonlinear ICA: A Unifying Framework. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR. 3

[27] Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures, 2024. 3

[28] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR, 2019. 3

[29] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 5

[30] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 2

[31] Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations? *arXiv preprint arXiv:1511.07543*, 2015. 3

[32] Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. Openshape: Scaling up 3d shape representation towards open-world understanding. *Advances in neural information processing systems*, 36, 2024. 1, 2, 4

[33] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. 2023. 3

[34] Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364, 2019. 5

[35] Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3d captioning with pretrained models. *Advances in Neural Information Processing Systems*, 36, 2024. 5

[36] Valentino Maiorca, Luca Moschella, Antonio Norelli, Marco Fumero, Francesco Locatello, and Emanuele Rodolà. Latent space translation via semantic alignment. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 4

[37] 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*, pages 14334–14343, 2024. 1, 2, 3, 4, 5, 6, 7

[38] Giovanni Luca Marchetti and Christopher Hillar. Harmonics of learning: Universal fourier features emerge in invariant networks. https://synthical.com/article/03032df5-f9c5-43a6-8253-d4af1d67e0d4, 2023. 3

[39] Raghav Mehta, Vítor Albiero, Li Chen, Ivan Evtimov, Tamar Glaser, Zhiheng Li, and Tal Hassner. You only need a good embeddings extractor to fix spurious correlations, 2022. 3

[40] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013. 3

[41] Ari S. Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, page 5732–5741, Red Hook, NY, USA, 2018. Curran Associates Inc. 3

[42] Luca Moschella. Latent communication in artificial neural networks, 2024. 3

[43] Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 3

[44] Antonio Norelli, Marco Fumero, Valentino Maiorca, Luca Moschella, Emanuele Rodola, and Francesco Locatello. Asif: Coupled data turns unimodal models to multimodal without training. *Advances in Neural Information Processing Systems*, 36:15303–15319, 2023. 3

[45] E. Pekalska and Robert Duin. Automatic pattern recognition by similarity representations. *Electronics Letters*, 37:159 – 160, 2001. 3

[46] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 5

[47] 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. 1, 2

[48] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 5

[49] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017. 3

[50] Geoffrey Roeder, Luke Metz, and Durk Kingma. On Linear Identifiability of Learned Representations. In *Proceedings of the 38th International Conference on Machine Learning*, pages 9030–9039. PMLR. 3

[51] Jan Rupnik and John Shawe-Taylor. Multi-view canonical correlation analysis. In *Conference on data mining and data warehouses (SiKDD 2010)*, pages 1–4, 2010. 3

[52] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2

[53] Mahdiyar Shahbazi, Ali Shirali, Hamid Aghajan, and Hamed Nili. Using distance on the riemannian manifold to compare representations in brain and in models. *NeuroImage*, 239: 118271, 2021. 3

[54] Joanna Sliwa, Shubhangi Ghosh, Vincent Stimper, Luigi Gresele, and Bernhard Schölkopf. Probing the robustness of independent mechanism analysis for representation learning. In *UAI 2022 Workshop on Causal Representation Learning*, 2022. 3

[55] Gowthami Somepalli, Liam Fowl, Arpit Bansal, Ping Yeh-Chiang, Yehuda Dar, Richard Baraniuk, Micah Goldblum, and Tom Goldstein. Can neural nets learn the same model twice? investigating reproducibility and double descent from the decision boundary perspective. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13689–13698, 2022. 3

[56] Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Erin Grant, Iris Groen, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine L. Hermann, Kerem Oktar, Klaus Greff, Martin N. Hebart, Nori Jacoby, Qiuyi Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas P. O'Connell, Thomas Unterthiner, Andrew K. Lampinen, Klaus-Robert Müller, Mariya Toneva, and Thomas L. Griffiths. Getting aligned on representational alignment, 2023. 3

[57] Shuai Tang, Wesley J. Maddox, Charlie Dickens, Tom Diethe, and Andreas Damianou. Similarity of neural networks with gradients, 2020. 3

[58] Thomas Theodoridis, Theocharis Chatzis, Vassilios Solachidis, Kosmas Dimitropoulos, and Petros Daras. Cross-modal variational alignment of latent spaces. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4127–4136, 2020. 3

[59] Ivan Vulić, Sebastian Ruder, and Anders Søgaard. Are all good word vector spaces isomorphic? *arXiv preprint arXiv:2004.04070*, 2020. 3

[60] Chang Wang and Sridhar Mahadevan. Manifold alignment using procrustes analysis. In *Proceedings of the 25th international conference on Machine learning*, pages 1120–1127, 2008. 3

[61] Alex H Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. Generalized shape metrics on neural representations. In *Advances in Neural Information Processing Systems*, 2021. 3

[62] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 574–591. Springer, 2020. 1

[63] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1179–1189, 2023. 2

[64] Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, et al. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27091–27101, 2024. 1, 2

[65] Yuxuan Xue, Xianghui Xie, Riccardo Marin, and Gerard. Pons-Moll. Human 3Diffusion: Realistic Avatar Creation via Explicit 3D Consistent Diffusion Models. 2024. 3

[66] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, and Jie Zhou. Pre-training 3d point cloud transformers with masked point modeling. 2022 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19291–19300, 2021. 5

[67] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18102–18112, 2022. 3

[68] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8552–8562, 2022. 3

[69] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. *arXiv preprint arXiv:2310.06773*, 2023. 1, 2

[70] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyao Zeng, Shanghang Zhang, and Peng Gao. Pointclip v2: Adapting clip for powerful 3d open-world learning. *arXiv preprint arXiv:2211.11682*, 3(4), 2022. 3