# ISF Research Grant Application 2026: Multimodal Representation Learning from Unpaired Data

Principal Investigator:
Uri Shaham
Department of Computer Science, Bar Ilan University

## 1 Scientific Background

The ability to integrate and reason across multiple data modalities is a central frontier in modern artificial intelligence. As applications increasingly involve diverse sensory or semantic inputs—such as text, images, speech, molecular structures, and biological measurements—there is growing demand for multimodal representation learning (MMRL): learning joint embeddings that capture shared semantics across modalities. This capability is foundational to many recent successes in AI, from vision-language models like CLIP [20] and GPT-4V, to biomedical applications such as protein structure prediction or multi-omics data integration.

However, most advances in MMRL rely heavily on paired supervision: large-scale datasets of aligned samples across modalities (e.g., image–text pairs, audio–video clips). In contrast, many scientific and real-world datasets are unpaired, weakly paired, or noisily aligned. For instance, patient data may include structured clinical tests, free-text notes, and medical images collected asynchronously and with missing links; environmental sensors may capture time series from different locations and modalities with no direct correspondences. The vast majority of multimodal data remains underutilized simply because it lacks perfect alignment.

This project aims to develop a principled and unified framework for multimodal representation learning from unpaired data, grounded in mathematical theory and scalable algorithms. We identify three core challenges that arise in the absence of paired supervision: (i) how to learn shared structure when only partial or noisy correspondences are available; (ii) how to extract statistically correlated components across modalities without access to paired samples; and (iii) how to fuse disparate modality-specific representations into a coherent joint embedding space. These challenges are addressed through three tightly connected technical objectives:

1. **Learning shared representations from weakly-paired data,** leveraging partial or probabilistic alignment to guide robust cross-modal embedding;
2. **Unpaired canonical correlation analysis (CCA),** formulating a new framework for discovering correlated structure from fully unpaired samples;
3. **Unpaired representation fusion,** designing methods to integrate independently trained modality-specific embeddings into a unified representation space.

Each objective tackles a distinct aspect of the unpaired MMRL problem, yet together they contribute to a comprehensive, theoretically grounded approach for learning with multimodal data in real-world settings. The methods developed will build on various mathematical tools from spectral geometry, statistics,

operator theory and optimization, and will be evaluated both empirically and analytically to ensure interpretability, stability, and generalization. While grounded in geometric and spectral methods, this proposal addresses a fundamental challenge in modern AI — the ability to learn unified representations from disjoint, unaligned, or siloed data. Our methods are broadly applicable across science and technology domains where aligned multimodal data is costly or unavailable, making them valuable tools for scalable, data-efficient, and privacy-aware AI systems

### 1.1 Current approaches for representation Learning from unpaired data

Approaches such as CycleGAN [31] and domain-adversarial training [10] aim to align the marginal distributions of different modalities by fooling a discriminator into believing that mapped samples come from a shared domain. While attractive, such approaches come with serious drawbacks, such as instability, mode collapse, and lack of guarantees. Specifically, in scientific applications—where spurious correlations are common and precise interpretation matters—these drawbacks limit the reliability of adversarial approaches.

A second line of work aims at aligning marginal distributions by matching statistical or geometric patterns. Typically, such models implicitly make rigid assumptions, such as structural isomorphism or metric compatibility across modalities, which is problematic when modalities differ in information content, noise, or sampling.

Contrastive methods have seen success in paired settings, and some efforts extend them to unpaired data using heuristic pseudo-pairing strategies. While popular, such methods are mostly heuristic and may introduce noisy or biased pseudo-labels, which may result in embeddings that may capture correlations that do not reflect true cross-modal semantics.

To summarize, across all these approaches, common limitations emerge: A reliance on implicit or fragile alignment signals, lack of generality across domains and modalities, and absence of rigorous theoretical guarantees for shared structure discovery in the unpaired setting. These limitations highlight the need for a new class of methods—mathematically grounded, computationally efficient, and broadly applicable across scientific domains—which this proposal aims to develop.

Multimodal representation learning from unpaired data is rapidly becoming a central research focus, with several fascinating recent contributions that explore representation learning strategies in the absence of direct supervision in various domains, e.g., [16, 29, 30, 11]. This growing body of work highlights both the promise and the complexity of the unpaired setting, motivating the need for new methods that are both mathematically principled and practically effective."

### 1.2 Scientific Potential and AI for Science

Beyond algorithmic innovation, the ability to learn from unpaired multimodal data has transformative potential for AI for science. In scientific domains—such as biology, neuroscience, geophysics, and materials science—data is often multimodal but rarely aligned. Developing methods that can integrate genomics and imaging, or correlate text-based reports with sensor data, without relying on curated pairings, can unlock rich, latent structure in complex systems. Moreover, such methods support key scientific goals: hypothesis generation, data-driven discovery, and interpretable modeling of high-dimensional processes.

In line with emerging trends toward weak supervision, modality fusion, and foundation models, this project aims to establish the mathematical and algorithmic foundations for robust multimodal learning in the absence of explicit labels or pairs—broadening the reach of AI into previously inaccessible or underutilized scientific data regimes.

# 2 Research Objectives and Expected Significance

### Objective 1: Learning Shared Representations from Weakly-Paired Data

This objective aims to develop a theoretical and algorithmic framework for learning shared representations across modalities under weak supervision. A shared representation is a common latent space in which instances from different modalities that convey the same underlying information are mapped to the same—or nearby—points. In many practical settings, such as in science, medicine, and human-centered data, such correspondences are not fully available: data may be only coarsely aligned, sparsely paired, or entirely unpaired. This objective addresses the challenge of learning shared representations in these weakly-paired regimes by exploiting the **universality of embedding geometries**—the observation that meaningful structure in each modality can be captured in a way that is stable, comparable, and aligned across domains. By enabling the discovery of shared latent structure without relying on strong pairing assumptions, this objective contributes to broadening the scope and robustness of multimodal representation learning in real-world, weakly supervised environments.

### Objective 2: Unpaired Canonical Correlation Analysis (CCA)

The second objective is to establish a framework for discovering maximally correlated representations across modalities **without access to paired data**. Canonical Correlation Analysis (CCA) traditionally requires paired samples to identify projections that reveal shared latent structure between two views. In the absence of such pairing, the problem becomes fundamentally ill-posed, as many joint distributions can be consistent with a given pair of marginals. This objective addresses the challenge by introducing a principled criterion for selecting among these: namely, the joint distribution that **maximizes cross-modal correlation**. We show that this joint can be characterized as the solution to an optimal transport problem, augmented with orthogonality constraints to ensure the resulting embeddings behave analogously to classical CCA projections. Beyond the theoretical formulation, this objective also includes the **development of efficient and scalable algorithms** for computing such unpaired CCA embeddings, enabling practical application to large-scale multimodal datasets. This contributes both foundational insights and computational tools to the broader effort of multimodal representation learning from unpaired data.

### Objective 3: Unpaired Representation Fusion

The third objective is to develop a framework for fusing representations across modalities in the absence of pairing, under the assumption that different modalities carry both shared and modality-specific information. In contrast to approaches that focus solely on common latent structure, this objective seeks to learn rich representations that integrate the full informational content of all modalities—capturing both what is shared and what is unique. Achieving this without access to paired data requires novel strategies for aligning and combining modalities. A key innovation in our approach is the use of artificially generated pairs, which serve as anchors for bridging modalities without relying on real correspondences. This departs fundamentally from CycleGAN-style methods, which rely on bidirectional consistency losses and implicitly assume strong information overlap. By relaxing this assumption, our goal is to enable more flexible and expressive multimodal fusion that reflects the complexity of real-world data. This objective advances multimodal representation learning by addressing a central, yet underexplored, challenge: how to integrate complementary signals from unpaired sources into a unified representation space.

### 2.1 Expected significance

Multimodal data is pervasive across science and technology — from medical diagnostics that combine imaging, text, and molecular data, to autonomous systems that process visual, auditory, and spatial signals. Yet in many real-world settings, paired multimodal data is rare or unavailable, severely limiting the applicability of standard multimodal learning approaches. This project addresses this fundamental challenge by developing mathematically grounded and practically effective methods for learning from unpaired multimodal data, thereby expanding the scope and usability of machine learning in real-world contexts.

On the scientific level, the project is expected to make fundamental contributions to the theory of multimodal representation learning. It introduces new frameworks for learning shared and fused representations without supervision, grounded in tools from optimal transport, spectral theory, and statistical dependence. These contributions go beyond heuristic or adversarial approaches by offering a principled understanding of when and how unpaired modalities can be aligned and integrated — filling an important gap in the literature. The project is also expected to yield new algorithmic paradigms that are scalable, robust, and broadly applicable.

From a practical standpoint, the outcomes of this research will be relevant across domains that involve heterogeneous and unaligned data sources. In biomedicine, for example, the ability to integrate genomic, imaging, and clinical text data without requiring aligned patient samples could lead to more holistic diagnostic and prognostic models. In climate science, combining satellite imagery with sensor readings and textual reports can support more comprehensive environmental monitoring. In human-computer interaction, learning from unpaired speech, gesture, and visual input can enable more adaptive and multimodal AI agents.

Moreover, the project aligns with broader trends in AI that prioritize data efficiency, robust generalization, and cross-modal understanding. By enabling flexible and modular representation learning from unpaired data, it supports the development of AI systems that are more adaptable to real-world complexity, including scenarios where supervised data is scarce or privacy constraints prevent alignment.

In summary, this project has the potential to advance both the foundations of machine learning and its practical reach across scientific and technological domains, making multimodal AI more broadly accessible, theoretically principled, and capable of addressing high-impact challenges in science and society.

## 3 Detailed Description of the Proposed Research

### 3.1 Learning Shared Representations from Weakly-Paired Data

Multimodal representation learning aims to construct a common embedding space in which samples from different modalities that convey the same underlying information are mapped to similar representations. In most existing frameworks, this goal is achieved through fully paired supervision, where each sample in one modality is matched with its exact counterpart in the other. However, in many practical settings—such as medicine, scientific research, or human behavior modeling—pairing between modalities is sparse, noisy, or entirely missing. This objective aims to develop a theoretically grounded framework for learning shared representations under weak supervision, leveraging the intrinsic geometry of each modality to guide alignment.

#### 3.1.1 Rationale

**Mathematical Motivation.** Modern pre-trained unimodal foundation models have a proven ability to represent semantics. For example, two given images have close embeddings if their semantic meaning is similar,

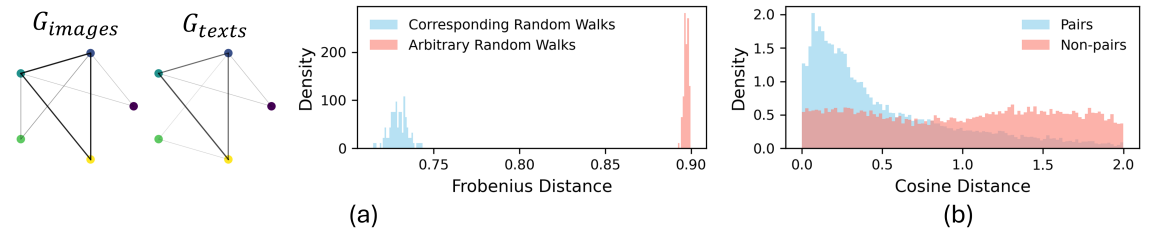



Figure 1: **Empirical demonstration of universality.** (a) Distances between corresponding random walks on image and text graphs from MSCOCO, compared to distances to randomly shuffled (non-matching) walks. Although constructed independently from unimodal features, corresponding walks exhibit significantly greater similarity. (b) Distances between paired and unpaired points in the shared space of aligned 2D spectral embeddings (SEs). Paired points are consistently closer, indicating that the independently learned SEs capture analogous structure across modalities.

and far apart otherwise. These similarities can be captured by a random walk process on the samples' representations. This suggests that a random walk process defined on such unimodal representations should largely correspond to semantic similarity. Therefore, we can expect random walks defined on different unimodal representations that capture semantics well to be highly similar. Random walk processes are finite analogs of diffusion operators. Thereby, the similarity of random walks that are constructed from different, modality-dependent representations implies that the eigenfunctions of the corresponding diffusion operators will have universality properties (i.e., modality-invariance) [6]. Therefore, constructing a spectral embedding (SE) based on the leading eigenvectors of random walks, which are viewed as discrete approximations of the leading eigenfunctions of diffusion operators [2, 23], enables us to take advantage of this concept even in the absence of paired data.

We formalize our assumption as follows. Let $\mathcal{M}$ be a latent, underlying semantic manifold, and let $f, g$ be two transformations, such that $f(\mathcal{M})$ and $g(\mathcal{M})$ represent the two modalities from which we observe samples. There is a body of work specifying conditions under which the spectral properties of $\mathcal{M}$ are preserved under $f, g$. For example, if $f, g$ have bounded distortion and bounded Ricci curvature, the corresponding eigenfunctions of the Laplace-Beltrami operator on $f(\mathcal{M})$ and $g(\mathcal{M})$ are similar in the $L_\infty$ sense [4].

Intuitively, our assumption states that the diffusion operators defined on each modality are relatively similar. This assumption is also empirically supported in recent works [14, 8, 12]. Then, universality is enabled through the eigenfunction preservation properties of the similar diffusion operators. Namely, the eigenfunctions of these operators will be universal, in the sense of modality-invariance (see Figure 1).

In practice, the ability to learn Laplacian eigenfunctions is obtained via SpectralNet [21], a previous work of the PI. While trained to compute the eigenvectors of the graph Laplacian of its training data, being a generalizable parametric map makes it a practical means to compute the eigenfunctions of the Laplacian operator (and thus also of the Diffusion operator), viewing the eigenvectors as a discretization of the eigenfunctions [2, 23]. Crucially, we train SpectralNet on unimodal data only; hence, no paired data is needed to learn the Laplacian eigenfunctions, i.e., our universal embedding functions.

**Overview.** In a recent pre-prent of ours [28], we propose and explore a novel pipeline, named Spectral Universal Embedding (SUE). SUE consists of three steps: SE, CCA and MMD. First, it maps each pre-trained unimodal embedding space into its corresponding eigenspace, to retrieve the global structure of each modality [1, 17, 24]. Using SpectralNet [21], this is done parametrically, allowing generalization to test data. Noteworthy, SE is not unique, as eigenvalues with multiplicity $p$ can yield any basis spanning the $p$-dimensional eigenspace and even single eigenvectors may differ by sign.

To resolve the SE ambiguity and provide additional linear alignment, we use CCA on a minimal number of

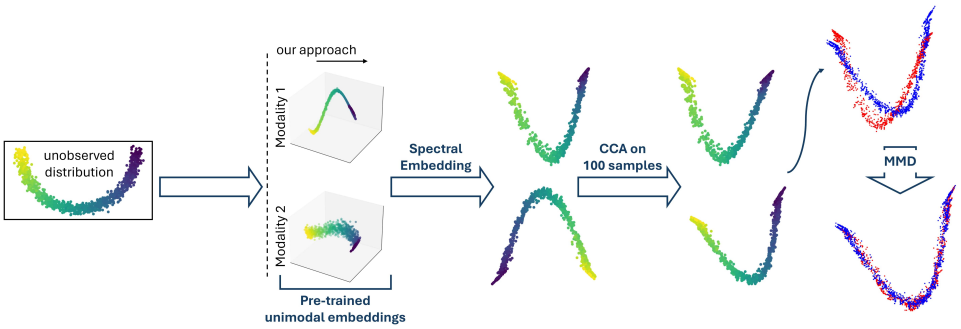



Figure 2: **SUE's overview**. The modalities (represented by their unimodal embeddings) represent an unobserved universal (semantic) distribution; the SE is capable of retrieving this universal structure, up to rotations; CCA on a minimal number of pairs enable linear alignment between the modalities, but not sufficient for a joint universal embedding; the MMD then fixes the misalignment between the modalities, integrating them into the universal embedding space.

paired samples. However, as the CCA purposefully considers a limited number of samples, and the SEs differ by more than an orthogonal transformation, we strengthen the cross-modal alignment using a Maximum Mean Discrepancy (MMD) residual network [22]. This kind of network architecture was originally proposed (by the PI) for batch-effect removal, by minimizing the empirical MMD value of two distributions. Namely, we view the two low-dimensional representations as similar distributions and learn a (close to identity) non-linear shift to align the distributions. The MMD serves as the last step to fine-tune the alignment. Notably, MMD loss does not require paired data, which enables the utilization of the full unpaired dataset. Figure 2 depicts SUE.

### 3.1.2 Uncovering SUE

In this section, we formularize SUE, roughly described in Sec 3.1.1. A summary of the steps of the SUE algorithm is outlined in Algorithm 1.

**Notations.** Throughout this section, we will use the following notations. Let $\mathcal{X} \subseteq \mathbb{R}^{d_1}, \mathcal{Y} \subseteq \mathbb{R}^{d_2}$ be sets of unpaired pre-trained unimodal embeddings of sizes $n_1, n_2$, resp. Accordingly, denote $\mathcal{X}_p = \{x_1, \ldots, x_m\} \subseteq \mathcal{X}, \mathcal{Y}_p = \{y_1, \ldots, y_m\} \subseteq \mathcal{Y}$ to be sets of paired embeddings. Importantly, $m \ll n_1, n_2$. Let $k \geq r$ be two pre-chosen dimensions for the SE and final universal representations.

**Approach.** Given $\mathcal{X}, \mathcal{Y}$, we train two independent SpectralNet models $S_\mathcal{X} : \mathcal{X} \to \mathbb{R}^k$, $S_\mathcal{Y} : \mathcal{Y} \to \mathbb{R}^k$ to approximate the $k$-dimensional SE of each modality. Due to the non-uniqueness of the SE, $S_\mathcal{X}$ and $S_\mathcal{Y}$ might differ by sign and basis of each eigenspace.

To address this ambiguity we utilize $\mathcal{X}_p$ and $\mathcal{Y}_p$. Specifically, we employ CCA on $\big(S_\mathcal{X}(\mathcal{X}_p), S_\mathcal{Y}(\mathcal{Y}_p)\big)$ to obtain the projections $Q_\mathcal{X}, Q_\mathcal{Y} \in \mathbb{R}^{k \times r}$. These projections are used to align $S_\mathcal{X}(\mathcal{X})$ and $S_\mathcal{Y}(\mathcal{Y})$. The linearly aligned SEs approximations can be written as $\tilde{S}_\mathcal{X} := Q_\mathcal{X} \circ S_\mathcal{X}$, $\tilde{S}_\mathcal{Y} := Q_\mathcal{Y} \circ S_\mathcal{Y}$.

Then, we learn a residual neural network $F_\theta : \mathbb{R}^r \to \mathbb{R}^r$ to bring the distribution of the linearly aligned SEs as close as possible. Specifically, we minimize the squared MMD between the two empirical distributions

$$\mathcal{L}_{\text{MMD}} = \frac{1}{m_1^2} \sum_{x_i, x_j \in \mathcal{X}} \kappa(\tilde{x}_i, \tilde{x}_j) - \frac{1}{m_1 m_2} \sum_{x_i \in \mathcal{X}, y_j \in \mathcal{Y}} \kappa(\tilde{x}_i, \tilde{y}_j) + \frac{1}{m_2^2} \sum_{y_i, y_j \in \mathcal{Y}} \kappa(\tilde{y}_i, \tilde{y}_j), \tag{1}$$

where $m_1, m_2$ are the corresponding batch sizes, $\kappa$ is a universal kernel (e.g., RBF kernel), and $\tilde{x}_i = \tilde{S}_\mathcal{X}(x_i)$, $\tilde{y}_i = \tilde{S}_\mathcal{Y}(y_i)$. The final functions can be written as $f_\mathcal{X} := \tilde{S_\mathcal{X}}$, $f_\mathcal{Y} := F_\theta \circ \tilde{S_\mathcal{Y}}$.

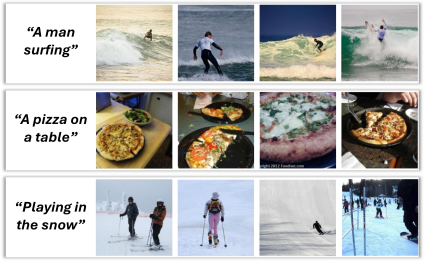



Figure 3: **Almost exclusively unpaired image retrieval.** Retrieved images for custom captions on the MSCOCO dataset, trained with 100 pairs and 10k non-pairs. The retrieved images are highly similar semantically to the text queries, even though almost no pairs were available during training.

Given a new test point $y_t$, sampled from the same distribution as $\mathcal{Y}$, we simply propagate it through $f_{\mathcal{Y}}$, and similarly to a test point sampled from the $\mathcal{X}$ distribution.

**Algorithm 1:** Spectral Universal Embedding (SUE)

**Input:** Unpaired sets of pre-trained unimodal embeddings $\mathcal{X} \in \mathbb{R}^{n_1 \times d_1}$ and $\mathcal{Y} \in \mathbb{R}^{n_2 \times d_2}$, and paired sets $\mathcal{X}_p$ and $\mathcal{Y}_p$ of size $m \geq 0$

**Output:** Maps $f_{\mathcal{X}} : \mathbb{R}^{d_1} \to \mathbb{R}^r, \quad f_{\mathcal{Y}} : \mathbb{R}^{d_2} \to \mathbb{R}^r$ approximating the universal embedding from each modality

1 Train $S_{\mathcal{X}}, S_{\mathcal{Y}}$

2 Perform CCA on $\big(S_{\mathcal{X}}(\mathcal{X}_p), S_{\mathcal{Y}}(\mathcal{Y}_p)\big)$ to obtain projections $Q_{\mathcal{X}}, Q_{\mathcal{Y}} \in \mathbb{R}^{k \times r}$

3 Train a residual neural network $F_\theta : \mathbb{R}^r \to \mathbb{R}^r$ to minimize the MMD loss $\mathcal{L}_{\text{MMD}}$ (Eq. 1)

4 Return the maps:

$$f_{\mathcal{X}} := Q_{\mathcal{X}} \circ S_{\mathcal{X}}, \quad f_{\mathcal{Y}} := F_\theta \circ Q_{\mathcal{Y}} \circ S_{\mathcal{Y}}$$

5 At inference time, propagate the sample $x$ or $y$ through the appropriate map $f_{\mathcal{X}}(x)$ or $f_{\mathcal{Y}}(y)$

### 3.1.3 Preliminary Results

In this section, we provide a demonstration of SUE for vision-language retrieval (Figure 3, Table 1). Additional result demonstrating capabilities in zero-shot classification and image manupulation are not provided, due to space limitations. In addition, Figure 4 demonstrates that SUE is designed to benefit from unpaired data, by analyzing the effects of difference numbers of paired and unpaired instances on the performance of Sue.

**Unpaired samples.** Fig. 4b shows the impact of additional unpaired samples. This experiment is of significant interest, as unpaired samples are usually considered unusable in the multimodal setting for point-to-point matching. The results indicate that additional *unpaired* data significantly enhances retrieval results. This opens the door for a new regime of multimodal learning - using unpaired data with only a minimal number of available pairs.

**Paired samples.** Fig. 4c depicts the results of an analogous experiment examining the effect of the number of paired samples required for the CCA step, with the unpaired samples held constant. As expected, a minimal number of paired samples are required for good results ($\sim$500 in this case of Flickr30k). However, SUE does not rely on additional pairs, as increasing their number above the minimum required is redundant. This outcome highlights the potential for learning significant cross-modal embeddings while focusing on unpaired data, which is much easier to obtain.

Table 1: **Retrieval results.** Results with few paired samples on vision-language (🖼📝) and vision-vision (🖼🖼) datasets from each modality to another: image-to-text (I2T), text-to-image (T2I), edges-to-shoes (E2S), shoes-to-edges (S2E); by SUE and Contrastive. The Imp. column states the relative mean improvement of SUE over Contrastive learning. Using the same small number of pairs, SUE significantly outperforms the popular paired method. **SUE substantially relies on unpaired data.**

| | #paired | | SUE (ours) R@1 | SUE (ours) R@5 | SUE (ours) R@10 | Contrastive R@1 | Contrastive R@5 | Contrastive R@10 | Imp. |
|---|---|---|---|---|---|---|---|---|---|
| MSCOCO 🖼📝 | 100 | I2T | 5.75 | 21.50 | 34.25 | 1.50 | 8.50 | 13.00 | +257.20% |
| | | T2I | 5.25 | 18.25 | 33.25 | 0.80 | 5.80 | 12.20 | |
| Flickr30k 🖼📝 | 500 | I2T | 4.25 | 19.75 | 32.00 | 3.00 | 9.50 | 16.20 | +103.32% |
| | | T2I | 5.75 | 22.00 | 32.75 | 2.50 | 9.80 | 15.00 | |
| Polyvore 🖼📝 | 500 | I2T | 6.00 | 22.75 | 32.25 | 3.20 | 13.8 | 22.5 | +55.67% |
| | | T2I | 4.75 | 20.75 | 32.00 | 4.00 | 11.50 | 23.00 | |
| Edges2Shoes 🖼🖼 | 50 | E2S | 4.00 | 16.00 | 25.25 | 1.0 | 5.50 | 14.00 | +200.51% |
| | | S2E | 3.50 | 17.00 | 27.00 | 0.80 | 6.00 | 12.80 | |

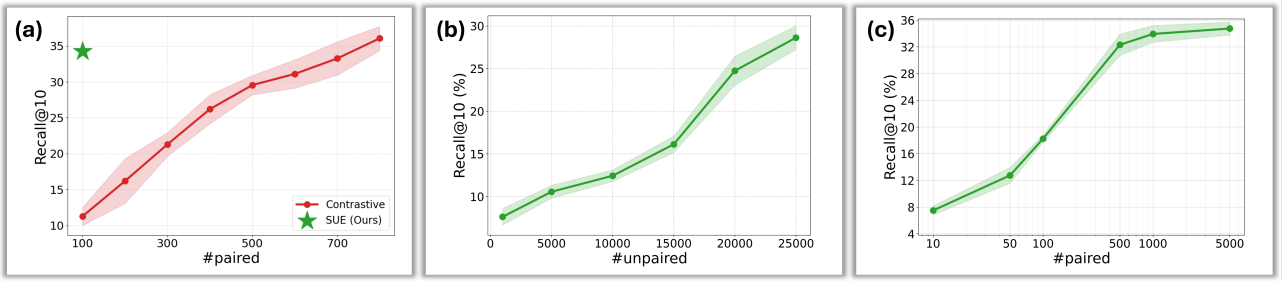



Figure 4: **(a) Contrastive requires an order of magnitude more pairs to achieve similar results as SUE in the weakly-paired regime.** Recall@10 results on MSCOCO by SUE (with 100 pairs) and Contrastive with various numbers of pairs. SUE exploits unpaired data to outperform contrastive learning when limited pairs are available. An order of magnitude more pairs are required to achieve similar results with contrastive learning; (b-c) Effect of #unpaired and #paired samples on Recall@10 results on image retrieval on the Flickr30k dataset. **(b) SUE improves as the amount of unpaired data is increased. (c) SUE relies on non-pairs instead of pairs.** SUE relies minimally on paired data, while substantially on unpaired data, enabling it to enhance its performance with additional unpaired samples, which are much easier to obtain.

### 3.1.4 Future Directions

The proposed method advances multimodal learning by showing that meaningful shared representations can be learned from structure alone, without explicit correspondence. This opens the door to broader deployment of multimodal models in settings where data collection is siloed, incomplete, or privacy-constrained.

As part of this objective, we plan to:

- **Task 1.1**: Formalize conditions under which spectral alignment is provably possible.
- **Task 1.2**: Extend the method to handle multiple modalities, by using multiview CCA machinery
- **Task 1.3**: Apply the method to real-world scientific datasets, such as multi-omics, medical imaging + text, sensor fusion, graphs, and time series
- **Task 1.4**: Investigate robustness to modality-specific distortions and distribution shifts.
- **Task 1.5**: Most importantly, the following objective proposes the development of an unpaired CCA technique. While important in its own right, an immediate application of it would be to turn SUE into a fully unpaired method, as the pairs are used in the SUE pipeline only in CCA.

Ultimately, this objective offers a new paradigm for multimodal learning: instead of relying on dense supervision, we extract and align universal geometric structure, enabling robust, interpretable, and scalable learning in weakly supervised environments.

## 3.2 Objective 2: Unpaired Canonical Correlation Analysis (CCA)

### 3.2.1 Overview of this objective.

Despite recent progress in leveraging unpaired data, no principled extension of CCA to the unpaired setting exists. Our aim is to bridge this gap by establishing a theoretical connection between distributional divergences and correlation, and by formulating a provable equivalence to CCA that holds without access to paired samples. In particular, our theoretical analysis reveals that the Wasserstein distance plays a central role in this equivalence [27]. Specifically, the 2-Wasserstein distance between two marginal distributions $P_X, P_Y$ can be shown to be equivalent to the correlation of their maximally correlated joint distribution, which we denote by $\mathrm{MCJ}(P_X, P_Y)$. This insight leads to propose an approach for unpaired CCA, which we term UCCA, operating by finding linear orthogonal projections for each view, with minimal Wasserstein distance. An important preliminary result of ours (Theorem 3.3) states that, under mild assumptions,

$$\mathrm{UCCA}(P_X, P_Y) = \mathrm{CCA}\big(\mathrm{MCJ}(P_X, P_Y)\big).$$

Intuitively, this means that UCCA recovers the CCA solution of a specific, highly meaningful joint distribution of $P_X, P_Y$.

Building on this theoretical foundation, we aim to develop a practical algorithm that can learn shared representations in fully unpaired settings. The reformulation of correlation maximization as a distribution matching problem enables the application of tools from Riemannian geometry and manifold optimization to the problem of correlation maximization in the unpaired setting. Specifically, we define the following tasks:

- **Task 1.1**: **Theoretical Connection between Wasserstein Distance and CCA:** We aim to prove a formal link between minimizing the Wasserstein distance between two distributions and maximizing the correlation under their maximally correlated joint. This result provides a bridge between optimal transport and classical correlation-based methods.
- **Task 1.2**: **Unpaired Canonical Correlation Analysis (UCCA):** Based on our theoretical insights, we aim to introduce a fully unpaired variant of CCA. This practical tool enables correlation-based learning without any paired data, by connecting Wasserstein distance, correlation, and optimization in a unified framework.
- **Task 1.3**: **Unpaired Nonlinear Shared Representation Learning:** Finally, by integrating our weakly-paired and unpaired techniques, we aim to construct a fully unpaired multimodal learning framework capable of learning nonlinear shared representations.

### 3.2.2 Previous work on Unpaired CCA.

Timilsina et al. [26] proposed a provable framework for unpaired shared component analysis, although its connection to correlation remains unclear. An earlier attempt by Hoshen and Wolf [13] introduced an unpaired variant of CCA; however, their method is unstable and requires multiple runs to obtain satisfactory results, as noted in their own work. Additionally, no implementation is publicly available, limiting its reproducibility and practical use. On a more theoretical front, the concept of a maximally correlated joint distribution has been studied in depth [7, 15, 25], and its connection to optimal transport is well established [27]. However, the link between this joint and the classical CCA algorithm has not been formally drawn.

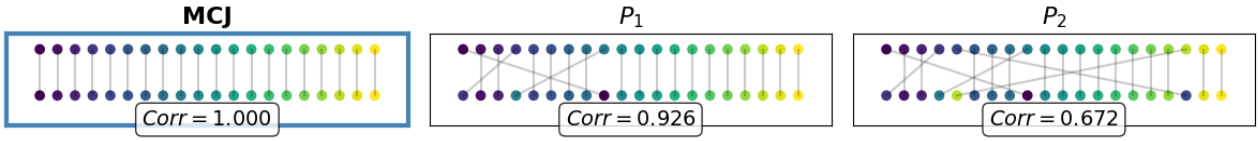



Figure 5: **MCJ.** A demonstration of Maximum Correlation Joint between two uniform distributions. Many joints are possible, but the left joint maximizes the correlation and indeed is the MCJ.

### 3.2.3 Preliminary theoretical results

While in the weakly-paired domain we have developed and assessed a method to capture a universal embedding, in the unpaired domain, our current results are mostly theoretical. To understand our theoretical result, we first need to define a few terms.

**Definition 3.1.** *The Maximum Correlation Joint (MCJ) of two whitened distribution $P_X, P_Y$ is*

$$MCJ(P_X, P_Y) = \arg \sup_{P_{XY} \in \mathcal{J}(P_X, P_Y)} TC(P_{XY})$$

*where $\mathcal{J}(P_X, P_Y)$ is the set of all joint distributions of $P_X$ and $P_Y$, and $TC(P_{XY})$ is the sum of correlations between the corresponding dimensions.*

A demonstration of the MCJ is depicted in Fig. 5. Def. 3.1 lets us reformulate the known connection between Wasserstein distance and correlation, as follows.

**Proposition 3.2.** *Let $P_X, P_Y$ be whitened probability measures, then*

$$TC\big(MCJ(P_X, P_Y)\big) = d - \frac{1}{2} W_2(P_X, P_Y)^2$$

*where $W_2(P_X, P_Y)$ is the 2-Wasserstein distance between $P_X, P_Y$.*

That is, the 2-Wasserstein distance between the marginal distributions $P_X, P_Y$ of two views corresponds to the correlation between the maximally correlated joints of the marginal distributions. For readability, we skip a few formal definitions here, and intuitively define $\text{MCJ}_{\mathcal{F}}(P_X, P_Y)$ as the "best" MCJ of $P_X, P_Y$ in terms of total correlation, over all projections in a function class $\mathcal{F}$. We also denote by UCCA our algorithm for unpaired CCA. That is, minimizing the Wasserstein distance over all orthogonal projections from $\mathbb{R}^d$ to $\mathbb{R}^k$. By that, we can finally state our novel result, which is

**Theorem 3.3.** *Let $P_X, P_Y$ be whitened probability measures. Under mild assumptions,*

$$UCCA(P_X, P_Y) = CCA(MCJ_{V_k(\mathbb{R}^n)}(P_X, P_Y))$$

Intuitively, Thm. 3.3 states that our UCCA algorithm is equivalent to CCA on a specific joint of $P_X, P_Y$ - the best MCJ of their projections.

### 3.2.4 Preliminary empirical results

TODO: COMPLETE RESULTS HERE

## 3.3 Objective 3: Unpaired Representation Fusion

### 3.3.1 Overview of this objective.

The prevailing paradigm in multi-view representation learning, particularly in contrastive self-supervised methods, is to extract only the *shared* information between views while suppressing view-specific information. While this is effective for achieving invariance, it inevitably discards the complementary and unique signals that each modality provides. In contrast, our objective in this case is not merely to align views by eliminating differences, but rather to *fuse* them in a way that leverages both the shared structure and the unique information contained in each view. This richer fusion is critical in settings where each modality contributes distinct yet meaningful aspects of the underlying phenomenon. Crucially, we aim at learning such unified representations across views in the absence of any pairwise correspondences.

Specifically, we plan to achieve this by thinking of each view as a diffusion operator constructed from its data manifold. Using previous methods of the PI for generalizable spectral embeddings [21, 3], we generalize the eigenfunctions of each operator to evaluate across all views, yielding *artificially parallel* diffusion maps. These are then summed into a fused operator that encodes both global and view-specific geometry, serving as a surrogate for true cross-view relationships.

### 3.3.2 Previous work on unpaired cross-domain learning.

In the cross-modal setting, cycle-consistency frameworks such as CycleGAN [31] and StarGAN [5] have been applied to learn mappings between unpaired domains. While successful in some settings, these techniques often struggle to preserve fine-grained structure, are difficult to train, and typically rely on implicit distributional assumptions. Moreover, they do not explicitly model the geometric or spectral structure of the data.

A few recent works address unpaired multi-view scenarios by designing methods for specific tasks such as clustering or classification. These methods typically not designed for learning a unified representation and instead construct task-driven models that operate on cluster level or seek weak correspondences indirectly. While these approaches provide practical solutions in constrained settings, they are not general-purpose multi-view learning frameworks and do not support representation learning that integrates both shared and unique information across modalities. TODO: COMPLETE CITATIONS HERE

### 3.3.3 Mathematical layout.

**Diffusion Operator Fusion via artificial parallelism.** consider unpaired datasets $X^{(v)} \subseteq \mathbb{R}^{n_v \times d_v}$, where $X^{(v)}$ is viewed as a sample from an underlying manifold $\mathcal{M}X^{(v)}$ and where $v$ indexes the views. We begin by constructing view-specific random-walk matrices via

$$W_{ij}^{(v)} = \exp\left(-\frac{\|x_i^{(v)} - x_j^{(v)}\|^2}{2\sigma_v^2}\right) \tag{2}$$

$$D_{ii}^{(v)} = \sum_{j=1}^{n_v} W_{ij}^{(v)} \tag{3}$$

$$P^{(v)} = (D^{(v)})^{-1} W^{(v)}. \tag{4}$$

Using SpectralNet [21, 3] allows us to learn the eigenfunctions $\phi_i^{(v)}$ of the diffusion operators $\mathcal{P}^{(v)}$ whose

finite analogues are the random walk matrices $P^{(v)}$. The operator $\mathcal{P}^{(v)}$ acts on a function $f$ as

$$\mathcal{P}^{(v)} f = \sum_i \lambda_i^{(v)} \langle \phi_i^{(v)}, f \rangle \phi_i^{(v)}. \tag{5}$$

Specifically, taking $f$ to be a Dirac delta function supported on a point $x_j$ enables us to compute an artificial value $P(x_j, x_k)$ for any two points $x_j, x_k \in \bigcup_v \mathcal{X}^{(v)}$ (in particular, ones which do not appear in the original samples) via

$$P^{(v)}(x_j, x_k) = \sum_i \lambda_i^{(v)} \psi_i^{(v)}(x_j) \psi_i^{(v)}(x_k).$$

Importantly, this enables us to obtain artificial parallelism by artificially computing the random walk matrices on parallel data.

**Manifold alignment via functional maps** While each view-specific diffusion operator $P^{(v)}$ admits a spectral decomposition $P^{(v)} f = \sum_i \lambda_i^{(v)} \langle \phi_i^{(v)}, f \rangle \phi_i^{(v)}$, its eigenfunctions $\phi_i^{(v)}$ are intrinsic to the manifold underlying view $v$. Therefore, they cannot be directly evaluated on data from a different view. To fuse the diffusion operator from different views, we need a way to align the eigenfunctions of the different manifolds so that they become comparable. We plan to achieve this with functional maps [19]. A functional map $C_{vw}$ is a linear operator between function spaces over manifolds $\mathcal{M}_v$ and $\mathcal{M}_w$. Specifically, if $f$ is a functional over $\mathcal{M}_v$, with basis expansion

$$f = \sum_{i=1}^{k} a_i^{(v)} \, \phi_i^{(v)} = \mathbf{a}^v \Phi^v$$

and $g$ be its corresponding functional on $\mathcal{M}_w$:

$$g = \sum_{i=1}^{k} a_i^{(w)} \, \phi_i^{(w)} = \mathbf{a}^w \Phi^w$$

the functional map $C_{vw}$ gives a convinient translation between their basis coefficients

$$\tilde{f}_i^{(v \to w)} = \sum_{m=1}^{k} \mathbf{C}[m, i] \, f_m^{(w)}.$$

Typically, the bases that are used are the eigenbases of the corresponding Laplacians. For our purposes, this gives us a machinery to evaluate an eigenfunction $\psi_i^{(v)}(x_j)$ on points belonging to the other manifold $\mathcal{M}_w$. In practice, the matrix $C_{vw}$ is obtained by solving a least squares minimization between known descriptors. In practice, the descriptors often rely on some supervision, and part of our goals in this part of the research will be to design unsupervised descriptors, which will enable us to learn the functional map in the absence of any correspondence between the samples.

We therefore define the following tasks

- **Task 1.1**: **Establish an algorithmic procedure for unpaired representation fusion.** This will be done via the above mathematical layout, by generating artificial parallelism and fusion of diffusion operators. In particular, we will investigate whether such an approach gives benefits (in terms of both downstream tasks accuracy and training efficiency) compared to domain conversion methods such as CycleGAN and StarGAN.

- **Task 1.2**: **Design unsupervised descriptors:** functional maps are typically used in computer graph-

ics. Recently, [9] have also used them for representation learning. However, they report a significant gap between the performance with supervised and unsupervised descriptors. As a by-product of this research objective, we plan to design improved unsupervised descriptors, possibly by leveraging spectral properties.

- **Task 1.3**: **Apply our approach for scientific discovery.** Recent work [18] has shown the applicability of unpaired translation methods for protein data. Inspired by these results, we aim to apply our method to protein and multi-omics data as well, to advance scientific discovery and prediction capabilities.

#### 3.3.4 Preliminary results

TODO: complete this

## 4 Plan of Evaluation

The success of this project will be evaluated through a combination of theoretical analysis, algorithmic development, and empirical validation across synthetic and real-world multimodal datasets. Each of the three objectives will be assessed according to the following criteria:

**Objective 1: Learning Shared Representations from Weakly-Paired Data** We will evaluate the quality of the learned shared representations by measuring cross-modal retrieval performance, alignment consistency, and robustness to pairing noise. Benchmark comparisons will be made against state-of-the-art methods in weakly supervised and semi-supervised multimodal learning. Theoretical evaluation will involve proving conditions under which universality guarantees hold and deriving error bounds on the recovered embeddings.

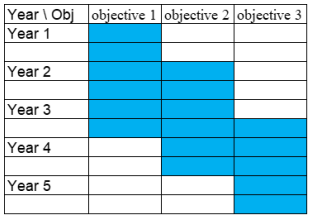
**Objective 2: Unpaired Canonical Correlation Analysis** The effectiveness of the proposed unpaired CCA framework will be assessed through correlation recovery, representation disentanglement, and computational efficiency. Empirical experiments will test the approach on standard unpaired datasets such as cross-lingual word embeddings, image-text pairs, and audio-visual benchmarks. We will also evaluate the practicality of the algorithm under distribution shifts and limited sample regimes.

**Objective 3: Unpaired Representation Fusion** Evaluation will focus on the ability of the model to capture both shared and modality-specific information without supervision. We will design proxy tasks such as zero-shot classification, few-shot transfer, and multimodal completion to quantify the utility of fused representations. Comparisons will include baselines based on CycleGAN-like models, mixture-of-experts, and late fusion methods.

In all cases, evaluation will include ablation studies to isolate the effect of key components and scalability tests on large datasets. Additionally, we will measure generalization to unseen modalities or domains, and validate performance under imperfect or noisy input distributions. The outcomes of the project — including theoretical findings, new algorithms, and empirical benchmarks — will be made available through open-source implementations, peer-reviewed publications, and reproducible research artifacts, allowing the broader community to validate, adopt, and extend the work.

# 5 Work Plan

The work will be performed by the PI, two Ph.D. students, and two M.Sc. students. One Ph.D. student and one M.Sc student will work on objectives 1 and 3, while the other will work on objective 2.




# 6 Broader Impact

This project aims to make foundational contributions to multimodal representation learning under minimal supervision, with broad implications for both the development and responsible deployment of AI systems. By enabling learning from unpaired and weakly aligned data, the proposed research lowers the barrier to applying machine learning in domains where annotation is costly, infeasible, or restricted by privacy — such as healthcare, environmental science, and public policy. These capabilities are especially important for democratizing access to AI in settings where high-quality labeled datasets are not available. Furthermore, the project advances representation learning in a direction that favors modularity, adaptability, and data efficiency, promoting the development of AI systems that are more transparent, robust, and privacy-aware. By reducing reliance on manual supervision and exploiting structure in unpaired data, the proposed methods open opportunities for scientific discovery in fields that increasingly rely on multimodal measurements but lack aligned data — such as genomics, neuroscience, and climate modeling. In doing so, this work contributes to the broader goal of using AI not only to build better models, but also to accelerate progress in science and improve societal outcomes through data integration and cross-modal reasoning.

# 7 Summary

This project develops foundational methods for multimodal representation learning (MMRL) from unpaired data, a setting that reflects the growing prevalence of heterogeneous, weakly aligned information across science and technology. Traditional multimodal learning methods rely on paired supervision, which is often unavailable due to cost, privacy, or measurement constraints. To overcome this, we propose three mathematically grounded objectives: (1) learning shared representations from weakly paired data based on universal spectral embeddings, (2) formulating unpaired canonical correlation analysis (CCA) using optimal transport with orthogonality constraints, and (3) fusing unpaired multimodal data through artificial pairing mechanisms that capture both shared and modality-specific content. These contributions are expected to yield practical, scalable algorithms with broad applicability, from biomedical data integration to cross-modal AI systems. The project will advance the theoretical foundations of MMRL while enabling data-efficient, modular, and privacy-aware machine learning across domains.

**Research team.** I am a statistician by training, with a solid background in mathematics and algorithms and 20 years of experience in machine and deep learning research, both in the industry and academia. As such, I bring a holistic, multi-view perspective, along with a rich toolbox to each of the research objectives. My research

is multi-disciplinary at its core, as it requires knowledge of multiple fields such as machine learning, applied mathematics, computer science, and engineering. Perhaps the best evidence of the multi-disciplinary nature of my research is the papers I publish, which include both rigorous mathematical proofs and practical methods, applied to challenging real-world problems. MY research team currently consists of one Ph.D. student and 14 M.Sc. students. In the past months, four M.Sc. students have graduated, all with publications in major machine learning venues. I credit much of the productivity and creativity of the group to the fruitful discussions and close interactions between the research group members, which I highly encourage, and all the projects described in this research proposal are important elements of my team's research. I also maintain collaborations with several researchers in other departments at Bar Ilan, in other universities in Israel, at also in several US universities, such as Yale and UCSD. I am convinced that both my team at Bar Ilan University and I are well-suited to meet the challenges of this ambitious and fascinating research proposal.

# References


[1] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[2] Mikhail Belkin and Partha Niyogi. Convergence of laplacian eigenmaps. *Advances in neural information processing systems*, 19, 2006.

[3] Nir Ben-Ari, Amitai Yacobi, and Uri Shaham. Generalizable spectral embedding with an application to umap. *arXiv preprint arXiv:2501.11305*, 2025.

[4] Pierre Bérard, Gérard Besson, and Sylvain Gallot. Embedding riemannian manifolds by their heat kernel. *Geometric & Functional Analysis GAFA*, 4:373–398, 1994.

[5] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

[6] Ronald R Coifman. Machine common sense: The darpa perspective. YouTube, 2020. Accessed: 2024-11-11.

[7] DC Dowson and BV666017 Landau. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.

[8] David Fan, Shengbang Tong, Jiachen Zhu, Koustuv Sinha, Zhuang Liu, Xinlei Chen, Michael Rabbat, Nicolas Ballas, Yann LeCun, Amir Bar, et al. Scaling language-free visual representation learning. *arXiv preprint arXiv:2504.01017*, 2025.

[9] Marco Fumero, Marco Pegoraro, Valentino Maiorca, Francesco Locatello, and Emanuele Rodolà. Latent functional maps: a spectral framework for representation alignment. *Advances in Neural Information Processing Systems*, 37:66178–66203, 2024.

[10] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[11] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190, 2023.

[12] Fabian Gröger, Shuo Wen, Huyen Le, and Maria Brbić. With limited data for multimodal alignment, let the structure guide you. *arXiv preprint arXiv:2506.16895*, 2025.

[13] Yedid Hoshen and Lior Wolf. Unsupervised correlation analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3319–3328, 2018.

[14] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Forty-first International Conference on Machine Learning*.

[15] Martin Knott and Cyril S Smith. On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43(1):39–49, 1984.

[16] Shuang Ma, Daniel McDuff, and Yale Song. Unpaired image-to-speech synthesis with multimodal information bottleneck. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7598–7607, 2019.

[17] Boaz Nadler, Stéphane Lafon, Ronald R Coifman, and Ioannis G Kevrekidis. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. *Applied and Computational Harmonic Analysis*, 21(1):113–127, 2006.

[18] Rami Nasser, Leah V Schaffer, Trey Ideker, and Roded Sharan. An adversarial scheme for integrating multi-modal data on protein function. In *International Conference on Research in Computational Molecular Biology*, pages 264–267. Springer, 2025.

[19] Maks Ovsjanikov, Mirela Ben-Chen, Justin Solomon, Adrian Butscher, and Leonidas Guibas. Functional maps: a flexible representation of maps between shapes. *ACM Transactions on Graphics (ToG)*, 31(4):1–11, 2012.

[20] 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.

[21] Uri Shaham, Kelly Stanton, Henry Li, Boaz Nadler, Ronen Basri, and Yuval Kluger. Spectralnet: Spectral clustering using deep neural networks. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

[22] Uri Shaham, Kelly P Stanton, Jun Zhao, Huamin Li, Khadir Raddassi, Ruth Montgomery, and Yuval Kluger. Removal of batch effects using distribution-matching residual networks. *Bioinformatics*, 33(16):2539–2546, 2017.

[23] Zuoqiang Shi. Convergence of laplacian spectra from random samples. *arXiv preprint arXiv:1507.00151*, 2015.

[24] Amit Singer and Ronald R Coifman. Non-linear independent component analysis with diffusion maps. *Applied and Computational Harmonic Analysis*, 25(2):226–239, 2008.

[25] André H Tchen. Inequalities for distributions with given marginals. *The Annals of Probability*, pages 814–827, 1980.

[26] Subash Timilsina, Sagar Shrestha, and Xiao Fu. Identifiable shared component analysis of unpaired multimodal mixtures. *arXiv preprint arXiv:2409.19422*, 2024.

[27] Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2008.

[28] Amitai Yacobi, Nir Ben-Ari, Ronen Talmon, and Uri Shaham. Learning shared representations from unpaired data. *arXiv preprint arXiv:2505.21524*, 2025.

[29] Zhaoyang Zeng, Daniel McDuff, Yale Song, et al. Contrastive learning of global and local video representations. *Advances in Neural Information Processing Systems*, 34:7025–7040, 2021.

[30] Ziqi Zhang, Chengkai Yang, and Xiuwei Zhang. scdart: integrating unmatched scrna-seq and scatac-seq data and learning cross-modality relationship simultaneously. *Genome biology*, 23(1):139, 2022.

[31] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.