# Bridging Latent Spaces Between Modalities

**Yann Baglin-Bunod**
École Polytechnique
yann.baglin-bunod@polytechnique.edu

**Michael Sheroubi**
École Polytechnique
michael.sheroubi@polytechnique.edu

## Abstract

We aim to create a scalable framework for connecting multiple modalities. By training translator functions to connect two modalities together, we can represent modalities as nodes in a graph-based architecture and these translators as edges, enabling flexible cross-modal traversal without requiring direct paired data for all combinations.

***K*eywords** Latent Spaces · Multiple Modalities

## 1 Introduction

We present a generalizable framework for connecting multiple data modalities through a series of bridges between different modality latent spaces. Our work is motivated by recent advances in multimodal learning, such as ImageBind and CLIP, which demonstrate strong zero-shot capabilities and robust alignment between modalities. However, to connect multiple modalities, these approaches are restricted only to modalities with image-paired data. Also, they exhibit a residual "modality gap" where embeddings for different data types cluster separately in latent space. We seek to take advantage of this inherent clustering and create a framework where we can add any modality as long as it can be paired with any another modality, not just images.

In this report, we test our framework on video (images), text, and audio using precomputed embeddings.

## 2 Related Work

### 2.1 CLIP

Contrastive Language-Image Pre-training (CLIP)[1], introduced by OpenAI, is a foundational multimodal model that aligns image and text modalities in a shared latent space. Using a dual-encoder architecture, CLIP trains an image encoder and a text encoder jointly on large-scale data sets of image-text pairs. The model's objective is to maximize the similarity between these image-text pairs while minimizing the similarity between mismatched pairs, using a symmetric cross-entropy loss function.

CLIP demonstrates remarkable zero-shot capabilities across diverse tasks, such as image classification, text-to-image retrieval, and text-guided image manipulation. This is achieved without the need for task-specific fine-tuning, making it highly versatile for open-world applications. By embedding text and images into a shared latent space, CLIP bridges the gap between natural language understanding and visual perception. This alignment enables seamless interaction between modalities, forming the backbone of many vision-language systems.

Despite its success, CLIP faces challenges such as reliance on vast amounts of training data and limited handling of hierarchical or multimodal complexities. It is possible to use data-efficient pretraining strategies that prioritize data quality over quantity to enhance generalization while reducing computational costs.

### 2.2 ImageBind

ImageBind[2] is a state-of-the-art multimodal learning framework designed to unify six distinct modalities into a single joint embedding space.

A key innovation of ImageBind is its ability to extend the capabilities of large-scale vision-language models like CLIP to additional modalities. By leveraging pre-trained image and text encoders alongside self-supervised learning for other modalities, ImageBind achieves emergent (not direct) cross-modal alignment.

ImageBind is a much more scalable approach than other multimodal approaches in that it doesn't require paired datasets across all modalities, it only needs image-paired data to achieve alignment. But while this is a strength in scalability, it is also a big limitation as it only supports modalities with image-paired data.

### 2.3 An Eye For An Ear

The authors[3] address the problem as unsupervised translation between the latent spaces of different modalities. This approach abstracts away the data representation, enabling transfer across modalities and even across distinct types of generative models, such as Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs).

The core methodology involves learning a shared latent space that acts as an interface between pretrained generative models. A variational autoencoder (VAE) serves as the backbone for bridging these models, due to its ability to encode and decode data representations effectively. The shared latent space is further refined by imposing supervised alignment of attributes across domains using a classifier. This ensures semantic consistency and locality preservation during the transfer process. The modular structure of this approach eliminates the need for retraining base generative models, significantly reducing computational overhead.

### 2.4 DMVAE

Disentangled Multi-Modal Variational Autoencoder (DMVAE)[4] is a generative model address challenges in representation learning by disentangling shared and private latent spaces across modalities. Unlike traditional multi-modal VAEs that primarily focus on shared representations, DMVAE explicitly separates the latent space into two distinct parts: a shared space that captures information common across all modalities, and a private space to encodes unique features specific to each modality.

However, the disentanglement process requires careful balancing between shared and private spaces, which can make training computationally intensive and sensitive to hyperparameter tuning. The model relies on paired multi-modal data during training to learn the shared latent space effectively. This requirement may limit its applicability in scenarios with unpaired or missing modalities.

### 2.5 Mind the Gap: Understanding the Modality Gap

CLIP, as explained in the review of its original paper, is a contrastive learning-based approach to training multi-modal models. The authors of *Mind the Gap*[5] study this contrastive learning approach applied to different multi-modal contexts and demonstrate systematic modality gaps across the models. Multi-modal deep neural networks struggle with alignment using contrastive learning methods.

The authors explore different reasons for why this may be true, including the cone effect, which is that neural networks naturally create narrow cones in the embedding space, leading to modality-specific clustering before training. They explain that the contrastive learning objective preserves the gap due to its repulsive structure and does not reduce it. Temperature scaling can affect the size of the gap: a lower temperature value in the contrastive loss increases modality separation.

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

The authors explore post-training alignment between 3D and text latent spaces[6]. Naïve feature alignment does not work well. Unilke vision-text 3D-text embeddings do not align well if they are trained independently. However, they find that projecting the 3D latent space to a subspace helps with alignment to text latent spaces. Canonical Correlation Analysis and affine transformations are used to project features in a lower-dimensional space.

Their method achieves 60% accuracy on the Objaverse dataset, the highest performance in cross-modal alignment without requiring joint pretraining between 3D information and text. Their final best model uses CLIP and Centered

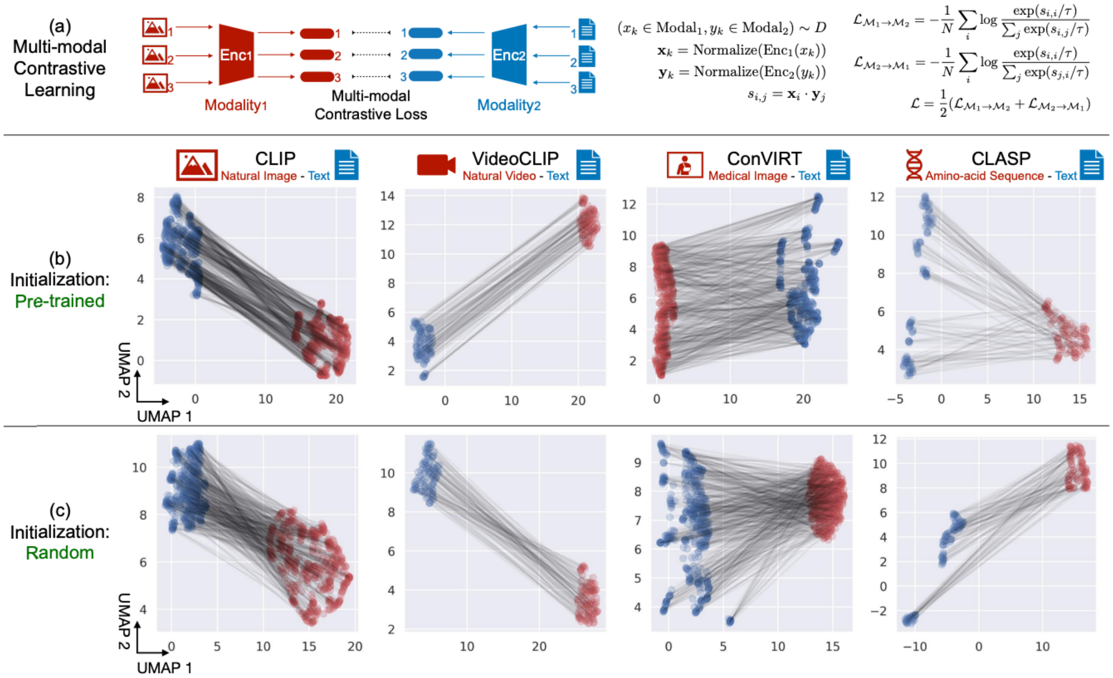

Figure 1: Taken from [5], this figure illustrates the modality gap in multi-modal contrastive learning. (a) Paired inputs are embedded into a shared space using two encoders. (b) UMAP visualization shows a clear separation between modalities in pre-trained models. (c) The gap persists even in randomly initialized models before training.

Kernel Alignment (CKA), a method in which the similarity between latent spaces of 3D and text encoders is calculated and finds a subspace projection that brings 3D and text embeddings closer.

### 2.7 The Platonic Representation Hypothesis

They investigate the idea that a platonic representation of all multi-modal data can be achieved. Across different tasks foundational multi-modal models converge, which may demonstrate that relying on a single universal representation space for multiple modalities is an effective way to bridge them.

## 3 Method

### 3.1 Data Preprocessing Pipeline

The dataset we used performs the following steps to compute embeddings:

1. **Video-to-Embedding:** Video frames are sampled at a fixed rate and fed through a pretrained visual encoder (*e.g.*, ImageBind visual backbone). The resulting latent vectors capture scene semantics.
2. **Audio-to-Embedding:** Audio tracks are normalized and passed into an audio encoder. Similar to the visual pipeline, we obtain a compact embedding that preserves relevant auditory features.
3. **Text-to-Embedding:** Captions, transcripts, and other textual metadata are processed (tokenized, cleaned) and encoded through a text encoder. The output text embeddings align with the same latent space as the video and audio components.

From the dataset, we receive each sample as an `(embedding_video, embedding_audio, embedding_text)` triple alongside some metadata.

### 3.2 Connecting Modalities

Our proposed framework builds upon ImageBind and utilizes multi-layered perceptron (MLP) heads to bridge modality embeddings without loss of information. While the dataset already uses precomputed features in a shared space, our approach does not rely on these embeddings existing in the same space. We mainly rely on the assumption that there exists pre-existing clusters from the cone effect mentioned in "Mind the Gap".

We found that even when each modality resides in the same embedding space, their distributions are not overlapping. This is consistent with previous findings showing that even when trained with contrastive learning objectives, different modalities tend to form distinct clusters in the latent space. These embeddings are learned through separate encoders for each modality, resulting in modality-specific regions within the shared embedding space.

#### 3.2.1 Network Architecture

Our current network consists of an MLP head that maps the input embedding from one modality to the corresponding embedding space in another modality. This network functions as a "translator" between modality embeddings. Unlike approaches that attempt to merge modality representations directly, our bridge preserves the natural clustering of modalities while enabling cross-modal traversal.

The key idea is to use a "reconstruction loss" as the metric for evaluating the quality of these bridges, ensuring that information from one modality can be accurately reconstructed back to itself. This approach allows us to maintain modality-specific information while enabling cross-modal translation.

#### 3.2.2 Creating a Graph

To address the limitations of CLIP and ImageBind, we aim to build a framework that can include new modalities without requiring direct paired data for all possible combinations (CLIP) nor a specific modality (ImageBind),

We propose using a graph framework, where each modality is represented as a node and each translator as an edge connecting these nodes. This allows for cross-modal traversal, where we find the shortest path, the one that preserves the most information. This allows for indirect connections between modalities that might lack paired training data.

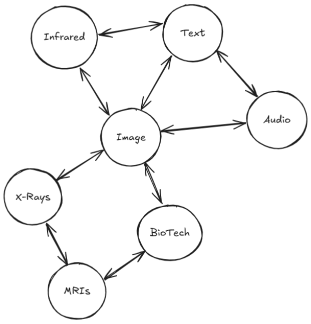

Figure 2: Example graph with multiple modalities

#### 3.2.3 Reconstruction Loss

To evaluate how well our translators preserve information, we try to reconstruct the original embedding by passing it through our from one modality back to itself, we implement a reconstructive loss framework:

1. Take an input embedding from one modality (e.g., image): $\boldsymbol{x}_{img}$.
2. Pass this embedding through a bridge network sequence to obtain a reconstructed embedding: $\hat{\boldsymbol{x}}_{img} = f_{img \to txt}(f_{txt \to img}(\boldsymbol{x}_{img}))$.
3. Compute the cosine similarity between the original and reconstructed embeddings:

$$\mathcal{L}_{\text{recon}} = 1 - \frac{\boldsymbol{x}_{img} \cdot \hat{\boldsymbol{x}}_{img}}{\|\boldsymbol{x}_{img}\| \cdot \|\hat{\boldsymbol{x}}_{img}\|}.$$

4. Use this similarity as the reconstructive loss, which is then optimized during training.

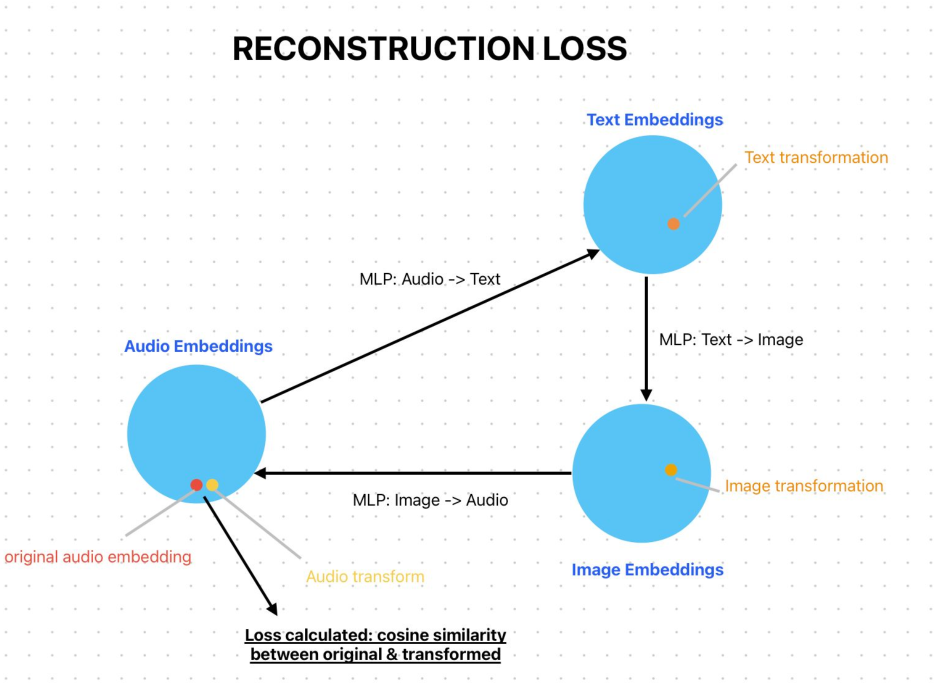

Figure 3: Reconstruction Loss

This loss ensures that traversing between modality spaces preserves semantic content, even when the path involves multiple modality translations.

### 3.3 Creating New Modality Support

To add support for a new modality, we:

1. Create an encoder specifically designed for this modality.
2. Extract features from data of the new modality and map them to the shared latent space used by existing embeddings.
3. Use a contrastive learning module to link the new modality to its shared modality in the latent space.
4. Train an MLP head to map the new modality to its linked modality(s).

### 3.4 Extensions

**Contrastive Fine-Tuning.** We sample matched (video, audio, text) triplets and employ a contrastive loss to pull matched embeddings closer together while pushing apart unrelated samples:

$$\mathcal{L}_{\text{contrastive}} = -\sum \log \frac{\exp(\langle \boldsymbol{v}, \boldsymbol{t} \rangle / \tau)}{\sum_{j \in \text{batch}} \exp(\langle \boldsymbol{v}, \boldsymbol{t}_j \rangle / \tau)},$$

where $\boldsymbol{v}$ and $\boldsymbol{t}$ are embeddings from distinct modalities, and $\tau$ is a temperature hyperparameter. We experiment with multi-modal variants of this objective by incorporating audio embeddings as well.

**Post-Hoc Embedding Alignment.** Even without further fine-tuning, we can post-process embeddings using transformations that align modality-specific subspaces. Techniques include:

- **Centering & Scaling:** Subtract modality-wise means and scale by standard deviation to normalize variance.
- **Procrustes Alignment:** Learn an orthogonal transformation to align centroids across modalities.
- **CCA or SVCCA:** Perform Canonical Correlation Analysis on matched pairs to identify a lower-dimensional subspace with maximal cross-modal correlation.

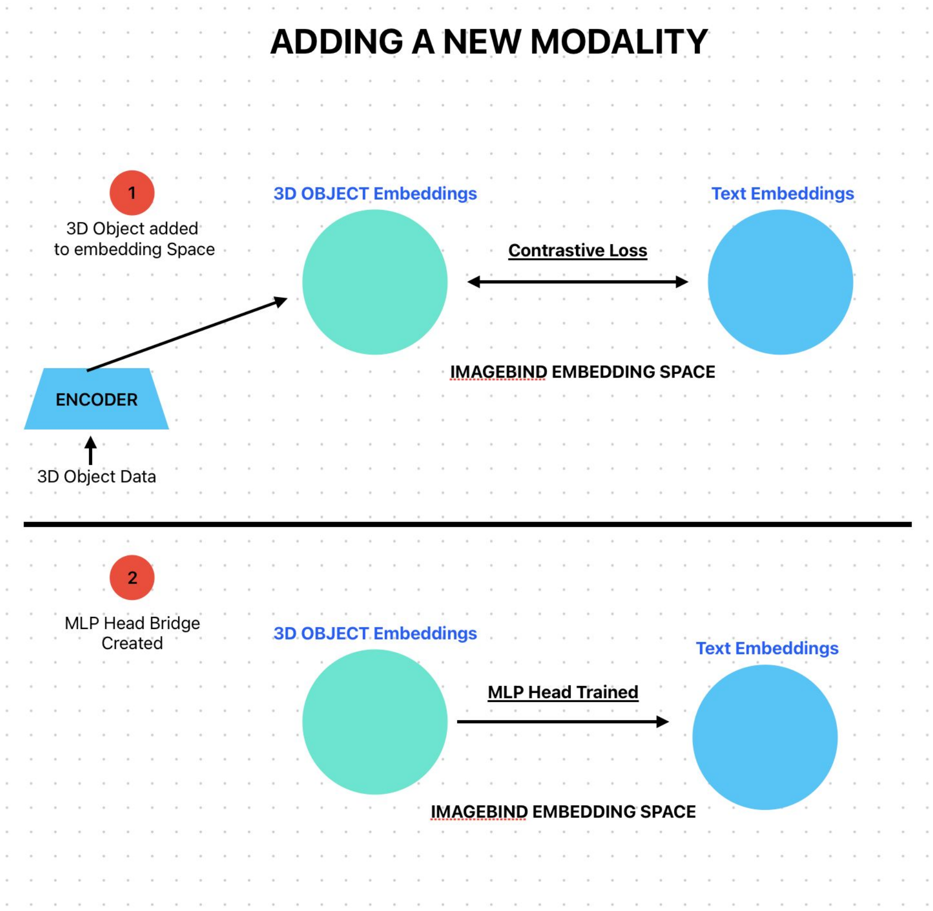

Figure 4: Adding a New Modality

**Disentangled Representations.** To isolate shared from modality-specific features, we optionally train a multimodal VAE variant. Each modality produces a private latent code (unique factors) and a shared code (common factors). By forcing these shared codes to align across data points of the same clip, we close the gap in the shared space.

### 3.5 Progress Analysis and Diagram Creation

Throughout the process, we track metrics such as:

- **Inter-Modality Distance:** Average Euclidean or cosine distances between embeddings of the same clip across modalities.
- **Cluster Separability:** Evaluated via silhouette scores or similar clustering measures.
- **Retrieval Accuracy:** Matching video-to-text or audio-to-text for the same clip.

## 4 Conclusion

This report presents a novel framework for bridging multiple modalities in a shared embedding space, addressing limitations of existing approaches like CLIP and ImageBind. Our method leverages a graph-based architecture where modalities are represented as nodes connected by MLP "translator" edges, enabling flexible cross-modal traversal without requiring direct paired data for all combinations.

Key innovations of our approach include:

1. A reconstruction loss metric to evaluate and optimize the quality of modality bridges

2. A scalable method for incorporating new modalities into the existing framework
3. Extensions like contrastive fine-tuning and post-hoc embedding alignment to further reduce modality gaps

By preserving modality-specific clustering while enabling seamless translation between spaces, our framework offers a promising direction for multimodal AI systems. The ability to indirectly connect modalities lacking paired training data opens up new possibilities for cross-modal applications.

Future work should focus on:

- Empirical evaluation across a wider range of modalities and datasets
- Exploring more sophisticated graph traversal algorithms for optimal cross-modal paths
- Investigating the impact of different network architectures for translator edges
- Developing strategies to mitigate potential information loss during multi-hop translations

As multimodal AI continues to advance, frameworks like ours that can flexibly integrate diverse data types will become increasingly crucial for building more comprehensive and adaptable systems.

## References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[2] 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. *arXiv preprint arXiv:2305.05665*, 2023.

[3] Hugo Malard, Michel Olvera, Stéphane Lathuiliere, and Slim Essid. An eye for an ear: Zero-shot audio description leveraging an image captioner using audiovisual distribution alignment. *arXiv preprint arXiv:2410.05997*, 2024.

[4] Mihee Lee and Vladimir Pavlovic. Private-shared disentangled multimodal vae for learning of latent representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1692–1700, 2021.

[5] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *arXiv preprint arXiv:2203.02053*, 2022.

[6] Souhail Hadgi, Luca Moschella, Andrea Santilli, Diego Gomez, Qixing Huang, Emanuele Rodolà, Simone Melzi, and Maks Ovsjanikov. Escaping plato's cave: Towards the alignment of 3d and text latent spaces. *arXiv preprint arXiv:2503.05283*, 2025.