CLOSING THE MODALITY GAP ENABLES NOVEL MULTIMODAL LEARNING APPLICATIONS

Eleonora Grassucci, Giordano Cicchetti, Danilo Comminiello
Department of Information Engineering, Electronics and Telecommunication
Sapienza University of Rome, Italy

ABSTRACT

In multimodal learning, CLIP has emerged as the *de facto* approach for mapping different modalities into a shared latent space by bringing semantically similar representations closer while pushing apart dissimilar ones. However, CLIP-based contrastive losses exhibit unintended behaviors that negatively impact true semantic alignment, leading to sparse and fragmented latent spaces. This phenomenon, known as the modality gap, has been partially mitigated for standard text and image pairs, but remains unresolved in more complex multimodal settings, especially when integrating three or more modalities. In this work, we propose a modality-agnostic framework that definitively closes the modality gap across multiple modalities, ensuring that semantically related representations are perfectly aligned, regardless of their source modality. Beyond theoretical improvements, we demonstrate that closing the modality gap has profound implications for real-world applications. In semantic communication, our approach enables the transmission of a single compact representation per semantic concept, drastically reducing bandwidth requirements while preserving multimodal reconstruction quality. In medical multimodal learning, our method enhances alignment between radiology images and clinical text, improving cross-modal retrieval and image captioning. We show that our approach not only closes the modality gap permanently but also unlocks new capabilities in downstream applications that were previously limited by poor cross-modal alignment.

1 INTRODUCTION

Multimodal representation learning has emerged as a fundamental paradigm in artificial intelligence, enabling models to process and integrate data from multiple sources such as text, images, and audio. The core assumption behind multimodal learning is that representations carrying similar semantics, regardless of their originating modality, should be mapped close to each other in a shared latent space. However, despite the success of contrastive learning approaches like CLIP Radford et al. (2021), a persistent issue prevents this ideal alignment: the modality gap Liang et al. (2022). When multimodal data are mapped into a shared latent space, samples from the same modality initially cluster together, forming distinct modality-specific groups. Unfortunately, these clusters persist even after training, resulting in a sparse and fragmented latent space Eslami & de Melo (2024); Fahim et al. (2024). In this space, certain modalities are densely packed into small regions, while others are more widely distributed, as shown in Fig. 1. This phenomenon arises because representations from different modalities tend to cluster separately, disrupting semantic coherence and significantly impairing downstream performance.

While previous efforts have attempted to mitigate the modality gap, they have largely focused on specific tasks and settings, leaving many open questions about how to extend these solutions to more complex multimodal settings. Previous works explore the modality gap when only two modalities are present, specifically text and images, and evaluated the performance solely in terms of retrieval tasks Eslami & de Melo (2024); Fahim et al. (2024); Yaras et al. (2024). Such methods do not propose any hypothesis in the case of more modalities, or how the modality gap affects specific downstream tasks and data types.

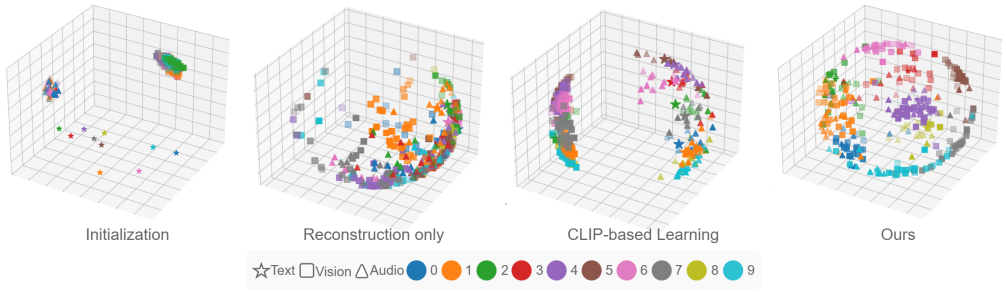

Figure 1: Modality clusters are clearly visible at initialization with image, text, and audio representations grouped and far from each other. After training, when no contrastive learning loss is involved, the space is somewhat randomly organized with no clear semantics (second plot). Then, the conventional CLIP-based models preserve the modality gap (third plot). On the contrary, although the clusters exist at initialization, the proposed method completely closes the gap (last plot), grouping the representations according to the correct textual label. Note that the latent space dimension is set to 3 in this example to directly plot the multimodal latent space.

In this work, we propose a novel, modality-agnostic approach that definitively closes the modality gap across multiple modalities, enabling the creation of a truly unified and structured multimodal latent space. Unlike previous methods that are constrained to specific types of data Ma et al. (2024); Schrodi et al. (2024), our approach generalizes seamlessly across different tasks, applications, and datasets, making it a powerful tool for real-world multimodal learning. Beyond theoretical advancements, closing the modality gap has profound implications for a wide range of real-world applications, spanning diverse fields such as healthcare and wireless communication, as we will explore. Specifically, we focus on two crucial downstream applications.

Semantic Communication. Traditional wireless communication systems focus on transmitting raw data, often leading to significant bandwidth consumption, especially in multimodal scenarios where multiple data streams (e.g., images, text, and audio) must be transmitted simultaneously. The emerging paradigm of semantic communication shifts the focus toward transmitting only the essential semantic content required to reconstruct or interpret the message at the receiver Dai et al. (2021); Xie et al. (2021); Barbarossa et al. (2023). We propose to involve our novel method in semantic communication scenarios to learn a semantically structured latent space where multimodal representations naturally cluster based on meaning rather than modality. Such novel proposed framework enables the transmission of a single compressed representation, rather than modality-specific embeddings, significantly reducing bandwidth consumption while preserving reconstruction performance at the receiver.

Medical Multimodal Alignment. The ability to integrate diverse medical data sources, such as radiology images and clinical notes, is crucial for accurate diagnostics and multimodal information retrieval Wang et al. (2022); Zhang et al. (2023); Chaves et al. (2024); Kumar & Marttinen (2024). Building a misaligned latent space may severely affect the performance of such models to assist clinicians, undermining their trust in learning-based methods for healthcare. Closing the modality gap ensures that heterogeneous data sources contribute meaningfully to a unified representation, improving complex downstream tasks like cross-modal retrieval and data captioning in healthcare. Therefore, in this paper, we will explore how the modality gap affects downstream performance and we propose to involve our novel method to close the gap and improve such performance in radiology-notes scenarios. We show that, by closing the gap, we can indeed improve cross-modal retrieval and image captioning in specific scenarios, possibly leading to enhanced trust of clinicians in AI-assisted diagnostic.

Our main contributions can be summarized as follows:

1. Unlike prior works, we study and resolve the modality gap in scenarios involving more than two modalities. Our approach is modality-agnostic, ensuring consistent performance regardless of the number or nature of modalities.

2. We demonstrate that closing the modality gap profoundly enhances downstream tasks and enables novel applications that where not conceivable with conventional multimodal learning methods.
3. We propose a novel multimodal framework for semantic communication capable of crucially compressing information in a compact semantic representation, saving bandwidth.
4. We study the modality gap in multimodal medical alignment and show that closing it crucially improves the performance in important downstream tasks.

2 Related Work

Multimodal Learning. Starting from the cosine similarity-based CLIP losses Radford et al. (2021) several multimodal models have been developed for two modalities like CLAP Elizalde et al. (2023) or CLIP4Clip Luo et al. (2021). Lately, the same loss has been extended to multiple modalities in ImageBind Girdhar et al. (2023), LanguageBind Zhu et al. (2024), or VAST Chen et al. (2023). More recently, novel approaches have been proposed for multimodal learning to avoid the cosine similarity loss and rethinking multimodal alignment, namely GRAM Cicchetti et al. (2024b) that relies on the volume computation and Symile Saporta et al. (2024), based on total correlation.

Modality Gap. The modality gap has been observed for the first time by Liang et al. (2022), and then studied mainly for the CLIP model Wu et al. (2023); Fahim et al. (2024); Shi et al. (2023) or for generic image and text pairs Yaras et al. (2024); Schrodi et al. (2024); Wu et al. (2023). These works provide theoretical justification for the gap Yaras et al. (2024) and propose to mitigate the gap by fixing the temperature Yaras et al. (2024), by applying post-hoc translations in the latent space Liang et al. (2022); Schrodi et al. (2024), or by sharing the transformer encoder and the projection layer in the vision and language encoders Eslami & de Melo (2024). In any case, each of these methods studied the modality gap in the case of two modalities, without advancing clues on the case of three or more modalities. Moreover, none of these methods closes the gap, as they only mitigate the phenomenon between the two modalities.

3 Closing The Modality Gap

In this Section, we present the limitations of current multimodal models and the proposed novel solutions. In particular, we provide the theoretical definitions to finally build a compact, well-aligned, and semantically meaningful multimodal latent space.

Notation. Given the set of M modalities, M_i is the i-th modality (i.e., text), m_i is a sample from the modality M_i (i.e., *"A dog barking"*), while $\mathbf{m}_i$ is the learned latent representation of the sample m_i.

3.1 Understanding the Modality Gap

Let us consider a set of n samples from M modalities, with $\{(m_{i,1}, m_{i,2}, \ldots, m_{i,M}\}_{i=1}^{n}$ being paired samples from the M modalities. That is, $m_{i,1}$ is the image of a dog, $m_{i,2}$ the caption *"A photo of a dog"*, and $m_{i,3}$ the audio of a dog barking. Multimodal learning aims at training modality-specific encoders $e_M : m_M \rightarrow \mathbf{m}_M, \mathbf{m}_M \in \mathrm{R}^d$ mapping the original data into a shared latent d-dimensional space where representations cluster according to the semantic content, regardless of the original modality. To do so, a huge number of contrastive loss functions have been designed, among which the most common is the one introduced in CLIP Radford et al. (2021) for text and image modalities. The conventional CLIP objective can be expressed in terms of cross-entropy loss function such as:

$$\mathcal{L}_{M_1 \rightarrow M_2} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\mathbf{m}_{1,i}^{\top}\mathbf{m}_{2,i}/\tau)}{\sum_{j=1}^{B}\exp(\mathbf{m}_{1,i}^{\top}\mathbf{m}_{2,j}/\tau)} \tag{1}$$

$$\mathcal{L}_{M_1 \rightarrow M_2} = \frac{1}{B}\sum_{i=0}^{B} H(\mathbb{1}_i, p_i), \tag{2}$$

in which τ is the temperature parameter, H is the cross-entropy function in charge of aligning the one-hot distribution $\mathbb{1}_i$ to the probability density function p_i, whose elements correspond to:

$$p_{i,j} = \frac{\exp(\mathbf{m}_{1,i}^\top \mathbf{m}_{2,i}/\tau)}{\sum_{j=1}^{B} \exp(\mathbf{m}_{1,i}^\top \mathbf{m}_{2,j}/\tau)}. \tag{3}$$

The conventional CLIP loss function is then the average between $\mathcal{L}_{M_1 \to M_2}$ and $\mathcal{L}_{M_2 \to M_1}$ to account for the non-symmetry of the cross-entropy.

The gap between modalities arises at the initializations phase, where different encoders initialized with random weights, represent data in narrow different cones in the shared latent space Liang et al. (2022), as Figure 1 shows. The gap persists even during the entire training phase. Therefore, even though the final learned space is somewhat semantically aligned, positive pairs are decoupled and very distant, as shown in the CLIP-based learning plot in Fig. 1. As demonstrated by Shi et al. (2023); Cicchetti et al. (2024b), the traditional CLIP loss function easily gets stuck in local minima, in which positive pairs are somewhat matched but far from each other. Previous works show that the conventional CLIP loss function is composed of two terms, each with specific objectives Eslami & de Melo (2024): a first term tries to align positive pairs while the second one tries to spread away non-matching pairs. In practice, these two terms provide opposite contributions resulting in balanced and opposite forces. Therefore, models easily end up in local minima, avoiding the gap closure while allowing the representations of the two modalities to align each other in "semantic stripes", a visual clue could be grasped from the left-hand side of Fig. 1. These semantic stripes allow however good performance in retrieval tasks since positive pairs have less cosine similarity with respect to non-matching pairs, even though such similarity is far from the ideal 1.0. Therefore, representations of matching pairs do not lie in the same portion of the latent space and are instead quite far from each other, severely limiting the expressiveness and the real alignment of the multimodal latent space.

3.2 Closing the Modality Gap

We aim to close the modality gap while ensuring consistent alignment among the positive pair distribution. To achieve such scope, we propose two novel losses. The first one, the Align True Pairs loss $\mathcal{L}_{\text{ATP}}$ guarantees the alignment between true pairs. Considering M modalities, among which $\mathbf{a}$ is the anchor one (i.e., the modality to which other modalities are aligned to Girdhar et al. (2023)), the loss is:

$$\mathcal{L}_{\text{ATP}} = \frac{1}{M-1} \sum_{i=0, \mathbf{m}_i \neq \mathbf{a}}^{M} \left(\frac{1}{B} \left(||\mathbf{m}_i - \mathbf{a}||_2^2 \right) \right), \tag{4}$$

where m_i is the i-th modality and B the batch size. The second one, the Centroid Uniformity loss $\mathcal{L}_{\text{CU}}$ ensures uniformity among the modalities in the latent space by:

$$\mathcal{L}_{\text{CU}} = \log \left(\frac{1}{B} \sum_{i=0}^{B} \sum_{j=0, j \neq i}^{B} \exp\left(-2||\mathbf{c}_i - \mathbf{c}_j||_2^2 \right) \right), \tag{5}$$

in which $\mathbf{c}_k$, with $k = i, j$, are the centroids defined as:

$$\mathbf{c}_k = \frac{1}{M} \sum_{k=0}^{M} \mathbf{m}_k, \tag{6}$$

and $\mathbf{c}_k$ is the centroid of the k-th element of the batch built by averaging all the modalities embeddings. The effect of the two losses is complementary. The $\mathcal{L}_{\text{ATP}}$ promotes closeness between positive pairs, effectively enhancing the mean cosine similarity between them. However, involving solely such a loss may produce a side effect: the entire latent space collapses into small portions, placing representations of dissimilar data in the same portion of the latent space. Therefore, the contribution of the $\mathcal{L}_{\text{CU}}$ loss becomes crucial, ensuring the sparsification of the latent space by enforcing uniformity to the centroids while preserving the learned alignment. Indeed, moving the centroids in the space implies moving the modalities representations accordingly, effectively preserving alignment while leveraging the whole space. Without centroids, uniformity should have been applied independently to each modality, disrupting the learned alignment among similar semantic pairs. Additionally, the radial basis function (RBF) kernel in equation 5 is well related to

the uniform distribution on the unit hypersphere where multimodal representations lie Wang & Isola (2020), therefore enforcing the coverage of the whole surface of the hypersphere. Moreover, by directly working on loss functions, the proposed method is modality-agnostic and can be used with any kind of modality. The overall loss function that aims at aligning the true pairs and closing the modality gap is a sum of the two terms:

$$\mathcal{L}_{\text{gap}} = \mathcal{L}_{\text{ATP}} + \mathcal{L}_{\text{CU}}. \tag{7}$$

Such loss should be then combined with the contrastive loss to obtain:

$$\mathcal{L}_{\text{CL}_{\text{gap}}} = \mathcal{L}_{\text{gap}} + \frac{1}{2}\left(\mathcal{L}_{M_1 \to M_2} + \mathcal{L}_{M_2 \to M_1}\right). \tag{8}$$

3.3 Measuring the Alignment

Following Liang et al. (2022), to measure such a gap between two generic modalities M_i and M_j, we measure the effective Euclidean distance between the centroids of each modality:

$$\text{Gap}_{M_i, M_j} = \|\mathbf{c}_{m_i} - \mathbf{c}_{m_j}\|, \tag{9}$$

where $\mathbf{c}_{m_i} = \sum_{\mathbf{m} \in M_i} \mathbf{m}$. Even though the gap is zero, this does not imply that the embeddings are effectively aligned in the latent space. Therefore, we propose to further adopt the mean cosine similarity true pairs metric, defined as:

$$\text{Cos True Pairs}_{M_i, M_j} = \frac{1}{N}\sum_{k=0}^{N}\left(\mathbf{m}_{i,k} \cdot \mathbf{m}_{j,k}^{\top}\right). \tag{10}$$

This metric measures how much the normalized matching pairs are near each other in the latent space. The closer to 1.0, the smaller the angle is between them, and the closer the matching pairs lie in the latent space. In addition, we also propose the mean angular value (AV) metric, computed inside the modality, whose formulation follows:

$$\text{AV}_M = \frac{1}{N^2 - N}\sum_{i=0}^{N}\sum_{j=0, j\neq i}^{N}\left(\mathbf{m}_i \cdot \mathbf{m}_j^{T}\right). \tag{11}$$

This metric measures the intra-modal average cosine similarity. It indicates how much the embeddings of a single modality are spread across the hypersphere. A value of this metric near 1.0 indicates that all the embeddings are very close to each other, while a value of 0 means that the intra-cosine similarity ranges from -1 to 1, indicating a good sparsification of the latent space.

4 Semantic Communication

4.1 Context and Idea

In modern wireless communication systems, the ever-increasing demand for transmitting multimodal data, such as images, text, and audio, has led to significant challenges in terms of bandwidth efficiency and transmission reliability. Traditional approaches focus on compressing and transmitting raw data streams, leading to high communication overhead, especially in multimodal scenarios where multiple representations of the same semantic content must be conveyed over the channel. Semantic communication has recently emerged as a promising paradigm to overcome these limitations by transmitting only the essential semantic information required to reconstruct or interpret the message at the receiver Xie et al. (2021); Barbarossa et al. (2023). However, current semantic communication frameworks mainly focus on transmitting one data type per time Tandon et al. (2023); Grassucci et al. (2023); Guo et al. (2024), thus scaling to sending separate embeddings for each modality in the case of multimodal data Cicchetti et al. (2024a). This requires high bandwidth consumption, sometimes negating the advantages of semantic compression.

In this paper, we propose to rethink the approach to multimodal semantic communication by leveraging the novel proposed method to close the modality gap in multimodal learning. By closing the gap, we can build semantically meaningful clusters in the space, regardless of the original modality.

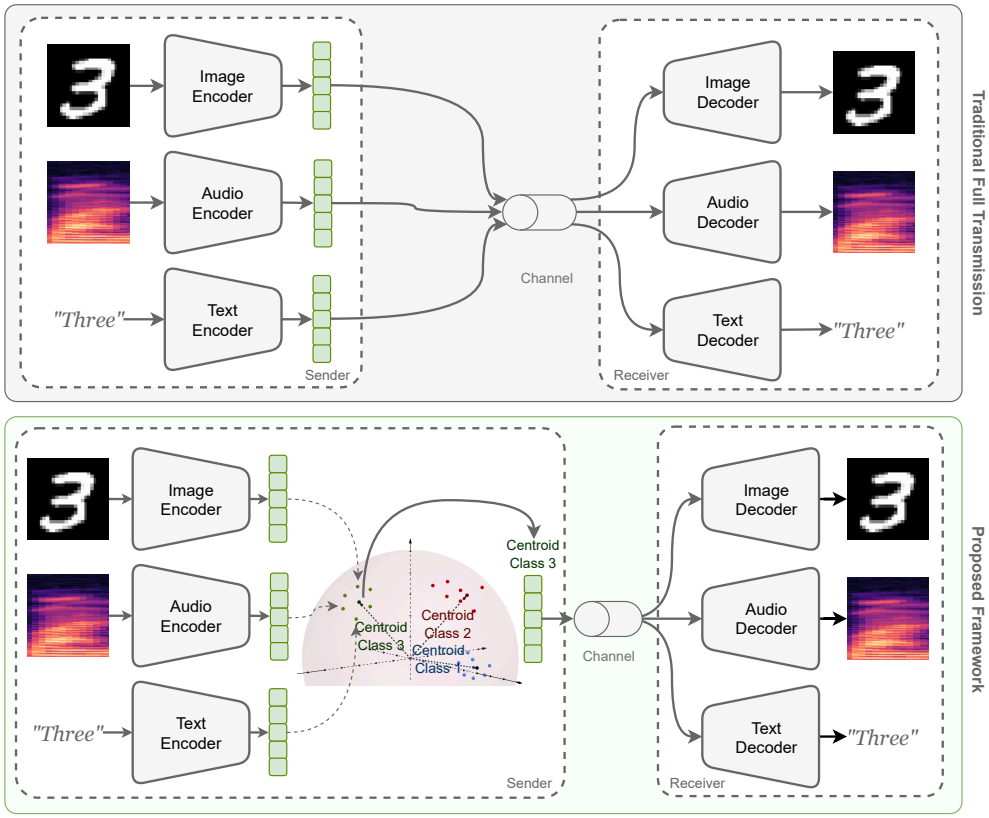

Figure 2: Top: traditional full transmission framework. Bottom: proposed contrastive learning-based semantic communication framework capable of transmitting one embedding only over the channel, saving bandwidth while preserving reconstruction performance.

In this way, data coming from different modalities but with the same semantics have very similar representations in the latent space. The intuition is that, if the multimodal space is so aligned, when a decoder takes in input a latent representation it can decode the information even though such a representation comes from a different modality.

4.2 Proposed Contrastive Learning-based Semantic Communication Framework

We design a multimodal encoder-decoder architecture capable of reconstructing multiple modalities from a single transmitted representation. The core idea is to build a semantically structured latent space where all modalities align perfectly with zero modality gap, thus enabling the transmission of a single latent representation per semantic concept. Therefore and contrary to conventional communication frameworks, at inference time, the sender does not need to transmit separate embeddings for each modality. Instead, given a multimodal input (e.g., an image, an audio clip, and a text description of the same concept), we extract only the centroid of the semantic cluster corresponding to that concept and transmit it over the communication channel, as we show in Fig. 2. This drastically reduces bandwidth consumption while ensuring that receivers have consistent information to reconstruct all original modalities. Indeed, at the receiver side, a set of decoders map the received centroid vector back into its respective modalities (image, text, and audio). Since the training process has fully closed the modality gap, the transmitted centroid contains all the necessary semantic information to allow high-quality reconstruction across modalities, even though only one embedding vector is transmitted.

Table 1: Semantic communication scenario results for embedding dimension equal to 16. CLIP-based learning with learnable temperature (LT) and fixed temperature (FT).

Method	Cos True Pairs (TV) ↑	Cos True Pairs (TA) ↑	Gap ↓	AV(T) ↓	AV(I) ↓	AV(A) ↓
Recon only	0.01	0.01	0.60	0.03	0.13	0.68
CLIP (LT) Radford et al. (2021)	0.15	0.17	0.40	**0.06**	0.07	0.16
CLIP (FT) Yaras et al. (2024)	0.12	0.13	0.53	0.05	0.17	0.33
Ours	**0.37**	**0.40**	**0.12**	0.10	**0.01**	**0.01**

To achieve this result at inference time, we jointly train a set of modality-specific encoders and decoders using an ensemble of loss functions. The first crucial loss is our contrastive loss function designed to close the modality gap. Then, decoder-specific reconstruction losses are applied, such as MSE for image reconstruction, L1 for audio spectrogram reconstruction, and a cross-entropy (CE) loss for text reconstruction.

The total loss function is then:

$$\mathcal{L} = \mathcal{L}_{\text{CL}_{\text{gap}}} + \frac{\lambda}{3}\left(\text{MSE}(m_1, \hat{m}_1) + \text{L1}(m_2, \hat{m}_2) + \text{CE}(m_3, \hat{m}_3)\right), \tag{12}$$

where m_1, m_2, m_3 are image, audio, and text original data, and $\hat{m}_1, \hat{m}_2, \hat{m}_3$ are their reconstructed samples, and λ is the hyperparameter to balance the contributions of the two loss functions.

4.3 Experiments

Settings. We consider a scenario involving three modalities: text, image, and audio. We select two well-known datasets: the MNIST dataset with 60k images, and the Audio-MNIST dataset Becker et al. (2023), which comprises 30k audio samples of spoken digits (0-9) from diverse speakers with different accents. We compute the Mel spectrograms with 128 nmels, fmax at 8000, hop length equal to 512, and 2048 nfft. Text modality is composed of the digit words associated with the images and the audio, i.e., "one" for the digit 1. Keeping it simple, we define three simple encoders, each tailored to process data independently for a specific modality. For both image and audio encoding, we employ a basic CNN with a latent dimension equal to 3 to directly plot the learned spaces in Fig. 1, and equal to 16 to measure the reconstruction performance. As image encoder, we employ a convolutional neural network (CNN) comprising a two-layer convolutional architecture with 32 and 64 filters, respectively, ReLU activation functions, Max Pooling, and a final MLP layer to map the features into the latent space. For the audio modality, we design a three-layer convolutional encoder with 16, 32, and 64 filters, ReLU activations, and an MLP layer to similarly project audio features into the latent space. Finally, we use a Word2Vec Mikolov et al. (2013) architecture for text encoding. By setting the latent dimension to 3, we can easily visualize the embeddings in a three-dimensional space, without using dimensionality reduction methods that approximate the final result by statistical and/or probabilistic methods. We run experiments with the conventional CLIP loss with learnable temperature (LT) Radford et al. (2021), with the recently proposed fixed temperature (FT) by Yaras et al. (2024), and then with the proposed $\mathcal{L}_{\text{CL}_{\text{gap}}}$.

Metrics. To evaluate the performance of the proposed framework, we analyze all the aspects of the framework, ranging from the quality of reconstructions to measures of the semantic alignment of the latent space. For the latter, we employ the metrics introduced in Sec. 3.3. Going through the performance evaluation, we evaluate the audio reconstruction with the mean absolute error (MAE), the image reconstruction with the common mean squared error (MSE), and the text one with the Accuracy (Acc), as we encode this data into one-hot encoding vectors. Finally, to measure the alignment of the latent space, we compute the recall at 1 (R@1) for the task of multimodal video-audio-text retrieval. In addition, we measure the amount of bit-per-pixel (BPP) transmitted over the channel by the conventional communication systems and the proposed one. A lower BPP indicates less bandwidth requirements and a higher compression.

Results. Figure 1 shows the latent three-dimensional multimodal latent space at initialization, without contrastive learning approaches, with conventional CLIP loss, and with the proposed losses. The gap is created at initialization, where modalities are grouped in small potions of the space (first plot), and then preserved by the CLIP-based loss function (third plot). On the contrary, the model trained

Table 2: Semantic communication scenario results for embedding dimension equal to 16. CLIP-based learning with learnable temperature (LT) and fixed temperature (FT).

Method	MSE ↓	MAE ↓	Acc ↑	R@1 ↑	BPP
Full transmission	0.031	0.096	100	100	0.0186
Recon only	0.062	0.136	100	52	-
CLIP (LT) Radford et al. (2021)	<u>0.061</u>	**0.083**	100	100	-
CLIP (FT) Yaras et al. (2024)	0.070	0.115	100	100	-
Ours	**0.042**	<u>0.099</u>	100	100	**0.0062** (-67%)

with the proposed $\mathcal{L}_{\text{CL}_{\text{gap}}}$ completely closes the gap as modalities perfectly overlap. Furthermore, representations cluster according to the correct class, as in the ideal latent space. Table 1 confirms the visual inspections, highlighting that the proposed method has the smallest gap and the highest cosine true pairs. Such space alignment measures are reflected in reconstruction performance shown in Tab. 2, where the proposed method well reconstructs data starting from the sole centroid of the transmitted class. Crucially, with respect to conventional full transmission, our framework has a pivotal lower bit per pixel (BPP), saving the 67% of bandwidth while preserving comparable reconstruction performance.

Importantly, our framework scales naturally to more than three modalities. Since we transmit only one representation per semantic cluster, our method can theoretically achieve a compression ratio of $1/n$, where n is the number of modalities, making it highly suitable for future high-dimensional multimodal communication systems.

5 Medical Data Alignment

Multimodal learning has seen growing adoption in the medical domain, where integrating multiple sources of information, such as radiology images and clinical text, has the potential to improve diagnostic accuracy and clinical decision-making Wang et al. (2022); Zhang et al. (2023); Chaves et al. (2024); Kumar & Marttinen (2024). However, despite the advancements in multimodal learning, to the best of our knowledge, no prior studies have investigated the impact of the modality gap on medical data alignment. The modality gap may introduce severe limitations when learning representations from heterogeneous and content-rich medical data. Indeed, if data is not properly aligned, tasks such as cross-modal retrieval, or captioning may suffer from reduced accuracy and reliability Yaras et al. (2024). Indirectly, a model that inconsistently aligns different modalities or fails to provide coherent predictions across imaging and textual data may crucially undermine the confidence of clinicians in AI-assisted diagnostic tools.

In this Section, we study the impact of the modality gap on multimodal medical data and we propose to leverage our novel loss function $\mathcal{L}_{\text{CL}_{\text{gap}}}$ to definitely close it. Interestingly, we found that the modality gap exists in medical data too, and more seriously, that true pairs are poorly aligned. On average, with the conventional CLIP loss function, true pairs have indeed a cosine similarity of 0.20, corresponding to an angle of 80 degrees. In practice, true pairs are almost orthogonal in the multimodal latent space. Therefore, it is crucial to better align true pairs, close the gap, and build a more aligned latent space to represent multimodal medical data.

5.1 Experiments

Settings. To perform the study on how the modality gap impacts multimodal medical data, we involve the Radiology Object in Context (ROCO) dataset Pelka et al. (2018), containing two modalities. The dataset comprises images from publications available on the PubMed Central Open Access FTP mirror, which were automatically detected as non-compound and either radiology or non-radiology. In this scenario we select only the radiology set comprising of 65420 images for training and 8176 for testing. Each image (no specific body region is selected) is associated with a caption. Captions are very heterogeneous, and comprise examples like *"Showing the subtrochanteric fracture in the porotic bone."* up to like *"A 3-year-old child with visual difficulties. Axial FLAIR image shows a supra-sellar lesion extending to the temporal lobes along the optic tracts (arrows)*

Table 3: Medical data alignment and retrieval results on the ROCO dataset.

Method	Cos True Pairs ↑	Gap ↓	R@1	R@5	R@10
CLIP (LT) Radford et al. (2021)	0.20	0.40	**39.5**	67.4	74.4
CLIP (FT) Yaras et al. (2024)	0.39	0.14	38.3	65.8	75.8
Ours	**0.54**	**0.12**	38.9	**68.8**	**81.8**

Table 4: Medical data captioning results on the ROCO dataset.

Method	Bleu 1 ↑	Bleu 2 ↑	Bleu 3 ↑	Bleu 4 ↑	ROUGE L ↑	CIDER ↑
CLIP (LT) Radford et al. (2021)	16.51	9.82	5.91	3.56	19.61	25.24
CLIP (FT) Yaras et al. (2024)	16.71	**10.07**	6.05	3.59	19.82	**26.05**
Ours	**16.96**	**10.07**	**6.09**	**3.64**	**19.90**	25.22

with moderate mass effect, compatible with optic glioma. FLAIR hyperintensity is also noted in the left mesencephalon from additional tumoral involvement.". To process such data and obtain embeddings we select larger models with respect to the ones in the previous section. As image encoder, we involve EVAClip-ViT-G ($\sim$ 1B parameters), which shows improved performance in zero-shot multimodal scenarios Sun et al. (2023). As text encoder, we involve BERT-B. These two encoder models are already proven to be effective in processing multimodal data Chen et al. (2023).

We conduct retrieval experiments using the conventional CLIP loss function with the learnable temperature parameter Radford et al. (2021). Then, as suggested by Yaras et al. (2024), we fixed the temperature to 0.07 allowing for a partial gap reduction. The last experiments is done using our proposed $\mathcal{L}_{\text{CL}_{\text{gap}}}$. In all the experiments we train the aforementioned framework for 100 epochs using AdamW as optimizer with a fixed learning rate of $1e-4$.

We perform the common image-text retrieval task. Moreover, to further investigate the effectiveness of the proposed loss function in downstream tasks we perform the image captioning task. Following Yan et al. (2022), we involve a decoder serving as the language generator model and we add a specific loss term that, along with our proposed loss functions, trains the text encoder-decoder structure for this specific task. Intuitively, the more aligned the latent space, the better the model will generate the captions from the latent representations. The captioning loss function is:

$$\mathcal{L}_{\text{cap}} = -\sum_{t=1}^{T} \log P_\theta(\mathbf{y}_t|\mathbf{y}_{<t}, \mathbf{x}), \tag{13}$$

where $\mathbf{y}$ is the exact tokenized texts the model aims to learn by maximizing the conditional likelihood under the forward autoregressive factorization. $P_\theta(\mathbf{y}_t|\mathbf{y}_{<t}, \mathbf{x})$ denotes the probability assigned to the token $\mathbf{y}_t$ given as input the past history $\mathbf{y}_{<t}$ and the input features $\mathbf{x}$.

Metrics. Along with standard metrics to measure the semantic alignment of the latent space (i.e., Cos True Pairs, Gap, and Angular Value) introduced in Sec. 3.3, we evaluate the performance in downstream tasks. For the retrieval task, we involve the Recall @1,5,10. For the image captioning task, we consider standard metrics such as BLEU@1, BLEU@2, BLEU@3, BLEU@4, ROUGE-L, and CIDEr.

Results. Tab. 3 and Tab. 4 show the results on the correlation between gap reduction and performance gain. According to the scores in Tab. 3, our introduced loss function can reduce the gap between image and text modalities down to 0.12 while ensuring the closeness of true pairs in the latent space up to 0.54. Interestingly, building such a more aligned space crucially improves the retrieval performance, especially in R@10. The latter result is important, as R@10 measures if the correct result is within the first ten samples in descendent cosine similarity order. Our methods clearly improve this metric up to 7.4 points. This means that the latent space is overall better aligned and that is more likely that the model places the correct data close to the query. Captioning results are instead shown in Tab. 4, and they prove our intuition on the importance of the aligned latent space. According to all the metrics, shaping a better-aligned multimodal latent space considerably enhances the decoder performance in generating the image caption with respect to conventional methods.

6 CONCLUSION

In this paper, we investigated the modality gap in multimodal learning and demonstrated its profound impact on downstream applications, particularly in semantic communication and medical data alignment. While previous studies have focused on mitigating the modality gap for conventional pairs of modalities (i.e., image-text), we proposed a modality-agnostic framework that definitively closes it across multiple modalities, ensuring a fully unified latent space. Our experiments showed that closing the modality gap enables significant advancements in real-world applications. In semantic communication, it allows the transmission of a single compact representation per semantic concept, drastically reducing bandwidth while maintaining reconstruction quality. In medical multimodal learning, we provided the first in-depth study on how the modality gap affects medical data alignment, revealing that misalignment degrades retrieval performance and undermines trust in AI-driven diagnostics. By achieving precise alignment, we improved cross-modal retrieval accuracy and medical captioning, reinforcing the necessity of structured multimodal representations in clinical AI.

ACKNOWLEDGEMENT

This work was partially supported by the Italian Ministry of University and Research (MUR) within the PRIN 2022 Program for the project "EXEGETE: Explainable Generative Deep Learning Methods for Medical Signal and Image Processing", under grant number 2022ENK9LS, CUP B53D23013030006, and by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, Mission 4, Component 2, Investment 1.3, partnership on "Telecommunications of the Future" (PE00000001 - program "RESTART"), and Project PNC 0000001 D3 4 Health, (Digital Driven Diagnostics, prognostics and therapeutics for sustainable Health care) - SPOKE 1 Clinical use cases and new models of care supported by AI/E-Health based solutions - CUP B53C22006120001. Finally, this work has also been partially supported by the "Progetti di Atento Medio" from Sapienza University of Rome with the project titled "SAID: Solving Audio Inverse problems with Diffusion models", under grant RM123188F75F8072.

REFERENCES

Sergio Barbarossa, Danilo Comminiello, Eleonora Grassucci, Francesco Pezone, Stefania Sardellitti, and Paolo Di Lorenzo. Semantic communications based on adaptive generative models and information bottleneck. *IEEE Communications Magazine*, 61(11):36–41, 2023.

Sören Becker, Johanna Vielhaben, Marcel Ackermann, Klaus-Robert Müller, Sebastian Lapuschkin, and Wojciech Samek. AudioMNIST: Exploring explainable artificial intelligence for audio analysis on a simple benchmark. *Journal of the Franklin Institute*, 2023.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, Hany Awadalla, Julia Gong, Houdong Hu, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Yu Gu, Cliff Wong, Mu Wei, Tristan Naumann, Muhao Chen, Matthew P. Lungren, Akshay Chaudhari, Serena Yeung-Levy, Curtis P. Langlotz, Sheng Wang, and Hoifung Poon. Towards a clinically accessible radiology foundation model: open-access and lightweight, with automated evaluation. *ArXiv preprint: arXiv:2403.08002*, 2024.

Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Ming-Ting Sun, Xinxin Zhu, and J. Liu. VAST: A vision-audio-subtitle-text omni-modality foundation model and dataset. In *Neural Information Processing Systems (NeurIPS)*, 2023.

Giordano Cicchetti, Eleonora Grassucci, Jihong Park, Jinho Choi, Sergio Barbarossa, and Danilo Comminiello. Language-oriented semantic latent representation for image transmission. *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, 2024a.

Giordano Cicchetti, Eleonora Grassucci, Luigi Sigillo, and Danilo Comminiello. Gramian multimodal representation learning and alignment. *ArXiv preprint: arXiv:2412.11959*, 2024b.

Jincheng Dai, Ping Zhang, Kai Niu, Sixian Wang, Zhongwei Si, and Xiaoqi Qin. Communication beyond transmitting bits: Semantics-guided source and channel coding. *IEEE Wireless Communications*, 30:170–177, 2021.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP: learning audio concepts from natural language supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Sedigheh Eslami and Gerard de Melo. Mitigate the gap: Investigating approaches for improving cross-modal alignment in clip. *ArXiv preprint: arXiv:2406.17639*, 2024.

Abrar Fahim, Alex Murphy, and Alona Fyshe. It's not a modality gap: Characterizing and addressing the contrastive gap. *ArXiv preprint: arXiv:2405.18570*, 2024.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190, 2023.

Eleonora Grassucci, Sergio Barbarossa, and Danilo Comminiello. Generative semantic communication: Diffusion models beyond bit recovery. *ArXiv preprint: arXiv:2306.04321*, 2023.

Jiangyuan Guo, Wei Chen, Yuxuan Sun, Jia lin Xu, and Bo Ai. VideoQA-SC: Adaptive semantic communication for video question answering. *ArXiv preprint: arXiv:2406.18538*, 2024.

Yogesh Kumar and Pekka Marttinen. Improving medical multi-modal contrastive learning with expert annotations. In *European Conference on Computer Vision*, 2024.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. CLIP4Clip: An empirical study of clip for end to end video clip retrieval. *Neurocomputing*, 508:293–304, 2021.

Xiang Ma, Xuemei Li, Lexin Fang, and Caiming Zhang. Bridging the modality gap: Dimension information alignment and sparse spatial constraint for image-text matching. In *ACM Multimedia*, 2024.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*, 2013.

Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M. Friedrich. Radiology objects in context (roco): A multimodal image dataset. In *MICCAI Workshop on Large-scale Annotation of Biomedical Data and Expert Label Synthesis (LABELS)*, pp. 180–189, 2018.

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. In *International Conference on Machine Learning (ICML)*, 2021.

Adriel Saporta, Aahlad Manas Puli, Mark Goldstein, and Rajesh Ranganath. Contrasting with symile: Simple model-agnostic representation learning for unlimited modalities. In *Neural Information Processing Systems*, 2024.

Simon Schrodi, David T. Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Peiyang Shi, Michael C. Welle, Mårten Björkman, and Danica Kragic. Towards understanding the modality gap in CLIP. In *ICLR 2023 Workshop on Multimodal Representation Learning: Perks and Pitfalls*, 2023.

Quan Sun, Yuxin Fang, Ledell Yu Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: Improved training techniques for clip at scale. *ArXiv preprint: arXiv:2303.15389*, 2023.

Pulkit Tandon, Shubham Chandak, Pat Pataranutaporn, Yimeng Liu, Anesu M. Mapuranga, Pattie Maes, Tsachy Weissman, and Misha Sra. Txt2Vid: Ultra-low bitrate compression of talking-head videos via text. *IEEE Journal on Selected Areas in Communications*, 41(1):107–118, 2023.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. MedCLIP: Contrastive learning from unpaired medical images and text. *Conference on Empirical Methods in Natural Language Processing*, 2022:3876–3887, 2022.

Junkang Wu, Jiawei Chen, Jiancan Wu, Wentao Shi, Xiang Wang, and Xiangnan He. Understanding contrastive learning via distributionally robust optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Huiqiang Xie, Zhijin Qin, Geoffrey Ye Li, and Biing-Hwang Juang. Deep learning enabled semantic communication systems. *IEEE Transactions on Signal Processing*, 69:2663–2675, 2021.

Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and Jiahui Yu. VideoCoCa: Video-text modeling with zero-shot transfer from contrastive captioners, 2022.

Can Yaras, Siyi Chen, Peng Wang, and Qing Qu. Explaining and mitigating the modality gap in contrastive multimodal learning. *ArXiv preprint: arXiv:2412.07909*, 2024.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *ArXiv preprint: arXiv:2303.00915*, 2023.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Hongfa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Liejie Yuan. LanguageBind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *International Conference on Learning Representations (ICLR)*, 2024.