The Modality Gap in Multimodal Semantic Communication

Anonymous submission

Abstract

Multimodal semantic communication (SemCom) is a key paradigm for 6G, yet it faces a critical bandwidth bottleneck. Conventional systems transmit a separate latent vector for each of the M modalities, an approach that scales poorly. In this paper, we argue that this inefficiency is a direct consequence of the modality gap, the persistent structural misalignment in contrastively-trained latent spaces that prevents a single, unified representation of multiple modalities. This paper presents a theoretical framework arguing that reducing the modality gap is the key enabler for efficient multimodal compression. By employing gap-reduction techniques, the M modality-specific embeddings for a single semantic concept collapse into a unified cluster. This alignment enables a novel transmission strategy: sending a single semantic centroid to represent the entire semantic concept, achieving a direct $1/M$ bandwidth reduction. We demonstrate the effectiveness of the proposed solution in different downstream tasks such as classification and reconstruction. This work establishes a clear path toward truly scalable and efficient multimodal semantic communication systems.

Introduction

The relentless growth of multimedia and multimodal content, now dominating internet traffic, imposes extreme pressure on the bandwidth limits of current and future wireless networks. Semantic communication (SemCom) emerges as a transformative paradigm to address this challenge, shifting the communication goal from perfect bit-level replication to the successful conveyance of meaning (Zhang et al. 2024; You et al. 2024; Luo, Chen, and Guo 2022; Huang et al. 2023; Qin et al. 2021). By transmitting only the essential semantic information necessary for a task, SemCom promises orders-of-magnitude reductions in data load, a critical objective for 6G systems (Calvanese Strinati and Barbarossa 2020).

Many modern SemCom frameworks leverage large-scale multimodal models, such as CLIP (Radford et al. 2021), to create a shared latent space (Cicchetti et al. 2025). The objective of such multimodal models is to align different data types, such as images, text, and audio, so that a single concept is represented by a tight cluster in this space. However, the standard contrastive training objective (e.g., InfoNCE (van den Oord, Li, and Vinyals 2018)) used to train these models results in an unintended and persistent artifact known as the modality gap (Liang et al. 2022; Grassucci, Cicchetti, and Comminiello 2025a). This phenomenon is the structural separation in the latent space where embeddings remain clustered by their source modality rather than their semantic content. For example, even after successful training, the vector for an image of a "dog" and the vector for the text "a photo of a dog" will not overlap; instead, they will lie in distinct, parallel regions of the space, separated by a measurable distance (Liang et al. 2022).

Nevertheless, current multimodal semantic communications systems do not face this phenomenon and avoid the limitations of multimodal models at their core by transmitting one representation for each modality (Shen et al. 2025; Zhang et al. 2025). As an example, if they need to transmit a video, with the corresponding audio and textual description, they encode each of these modalities into separated embeddings and then send to the receiver three distinct representation, requiring M transmissions. However, this approach does not scale well when the number of modalities increases (for a grid of sensors, for instance), as it requires increasing the bandwidth proportionally to the number of modalities.

In this paper, we argue that reducing the modality gap is a key enabler for efficient multimodal semantic compression. A latent space with a (near-)zero gap forces embeddings of the same concept to collapse into a single, dense, modality-agnostic cluster. This paper explores the theoretical implications of this alignment. We argue that this structural property allows the transmitter to discard M modality-specific vectors and instead transmit a single, compact representation: the semantic centroid of that cluster. This transition from M representations to one provides a direct $1/M$ reduction in transmission load12, and as we will discuss, this semantically-pure centroid is intrinsically more compressible.

The Modality Gap in Multimodal Learning

The foundational goal of multimodal representation learning is to project data from different sources, such as images $\mathbf{x}^{img} \in \mathcal{M}^{img}$, text $\mathbf{x}^{txt} \in \mathcal{M}^{txt}$, or audio $\mathbf{x}^{audio} \in \mathcal{M}^{audio}$, into a single, shared latent space $\mathbb{R}^d$. This is achieved by training a set of modality-specific encoders, $\mathcal{E}^{img}$, $\mathcal{E}^{txt}$, and $\mathcal{E}^{audio}$, which are typically deep neural networks (e.g., ViT for images, BERT for text, among others). Each encoder $\mathcal{E}^m$ maps a raw input sample $\mathbf{x}^m$ from

its modality $\mathcal{M}^m$ to a d-dimensional latent vector, $\mathbf{z}^m = \mathcal{E}^m(\mathbf{x}^m)$. The primary objective of this training is to ensure that these encoders produce a semantically aligned space. In such a space, representations of the same underlying concept, known as positive pairs, should be mapped to nearby points, regardless of their source modality. For example, the vector $\mathbf{z}^{img}$ for an image of a "3" and the vector $\mathbf{z}^{audio}$ for a spoken audio of "three" should be (near-)identical. To achieve this alignment, the *de facto* standard is contrastive learning, most notably the InfoNCE loss (van den Oord, Li, and Vinyals 2018) popularized by CLIP (Radford et al. 2021). For a batch of B positive pairs, this objective trains the encoders $\mathcal{E}^m$ and $\mathcal{E}^n$ by maximizing the similarity between corresponding pairs $(\mathbf{z}_i^m, \mathbf{z}_i^n)$ while simultaneously minimizing the similarity between all non-corresponding or "negative" pairs $(\mathbf{z}_i^m, \mathbf{z}_j^n)$ where $i \neq j$. This is formally expressed as a symmetric cross-entropy loss. For the $m \to n$ direction, the loss is:

$$\mathcal{L}_{m \to n} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(sim(\mathbf{z}_i^m \mathbf{z}_i^n)/\tau)}{\sum_{j=1}^{B} \exp(sim(\mathbf{z}_i^m, \mathbf{z}_j^n)/\tau)} \quad (1)$$

where $sim(\cdot, \cdot)$ is the cosine similarity and τ is a crucial temperature hyperparameter that scales the distribution of similarities.

However, this training process results in a stable, persistent artifact known as the modality gap (Liang et al. 2022). This gap is a structural misalignment where the learned representations remain partitioned in the latent space according to their source modality rather than according to the semantics. In a space exhibiting this gap, all image embeddings will occupy one region or "cone" (Liang et al. 2022), while all text embeddings will occupy a distinct, parallel region. This separation persists even after training is complete. Consequently, while a positive pair like ($\mathbf{z}^{img}$, $\mathbf{z}^{txt}$) may be closer to each other than to any negative embedding, they remain significantly distant from one another in absolute Euclidean terms. This gap can be empirically quantified by measuring the Euclidean distance between the global centroids of each modality embeddings, $\mathbf{c}^m = \frac{1}{K} \sum_{k=1}^{K} \mathbf{z}_k^m$ by:

$$\text{Gap}_{m,n} = \|\mathbf{c}^m - \mathbf{c}^n\|, \quad (2)$$

where a large, non-zero distance confirms the misalignment. The magnitude of this gap has been shown to be strongly correlated with the choice of the temperature τ (Yaras et al. 2025). Lower temperatures, often used to optimize instance-level retrieval tasks, tend to preserve or widen the gap, whereas higher temperatures tend to reduce it. Nevertheless, it is still largely under debate, as some works state that it does not impact downstream tasks (), while others that it might have mild-positive influence on retrieval tasks ().

Compressing Embeddings by Reducing the Modality Gap

Centroid Representation for SemCom

The structural misalignment of the modality gap, as detailed in Section , poses a direct and fundamental challenge to efficient multimodal semantic communication. In a gap-ridden latent space, the M representations $(\mathbf{z}_i^1, \mathbf{z}_i^2, \ldots, \mathbf{z}_i^M)$ corresponding to a single semantic concept i are spatially distant. Consequently, a transmitter cannot simply select one vector to represent the concept. This forces conventional multimodal SemCom systems to encode and transmit all M latent vectors, resulting in an M-fold transmission load that scales poorly and negates the primary compression benefits of the semantic paradigm.

Our theoretical framework posits that reducing this gap is the key enabler for true multimodal semantic compression. We argue that in a (near-)perfectly aligned space—one where Gap $\to 0$—the embeddings $(\mathbf{z}_i^1, \ldots, \mathbf{z}_i^N)$ for a given concept i would all collapse into a single, dense, modality-agnostic cluster. In this idealized scenario, the individual modality-specific vectors become redundant. This property enables a novel transmission strategy: rather than sending M separate vectors, the transmitter can compute and transmit a single semantic centroid, $\mathbf{c}_i$, to represent the entire concept. This centroid is formally defined as the average of all modality embeddings for a given semantic instance i:

$$\mathbf{c}_i = \frac{1}{M} \sum_{m=1}^{M} \mathbf{z}_i^m \quad (3)$$

This single vector $\mathbf{c}_i$ is then transmitted over the wireless channel. At the receiver, this lone centroid is then involved to accomplish the task at the receiver. As an example, it may be fed as input to all M modality-specific decoders $(\mathcal{D}^1, \ldots, \mathcal{D}^M)$ to reconstruct the data in each of their respective modalities. This approach achieves a direct $1/M$ reduction in the required transmission bandwidth. The central challenge, therefore, becomes how to train the encoders $\mathcal{E}^m$ to produce this idealized, gap-free latent space. We explore two such methods.

How Reducing the Gap

First, the alignment can be implicitly encouraged by modifying the contrastive learning objective itself. As noted in Section , the magnitude of the gap is strongly correlated with the InfoNCE temperature τ from Equation 1. Standard low-temperature ($\tau \leq 0.07$) training optimizes for instance-level retrieval, which inherently preserves the gap. Conversely, training with a sufficiently high, fixed temperature (e.g., $\tau \geq 0.1$) softens the loss distribution (Yaras et al. 2025; Grassucci et al. 2025). This shifts the optimization pressure from discriminating individual instances to reducing the distance from modality-specific clusters, thus reducing the gap. By encouraging all representations of the same semantic concept to group together, this high-temperature training naturally reduces the inter-modality distance, making the semantic centroid $\mathbf{c}_i$ a more faithful representation of the cluster.

Second, a more explicit and robust method is to augment the standard contrastive loss with an objective, $\mathcal{L}_{gap}$, specifically designed to force gap closure. This objective is composed of two complementary terms (Grassucci, Cicchetti, and Comminiello 2025a). The first is the Align True Pairs (ATP) loss, $\mathcal{L}_{ATP}$, which directly minimizes the Euclidean

Table 1: Performance comparison for 3-modality reconstruction. "Ours" transmits only the centroid ($1/M$ TTI). Results adapted from (Grassucci, Cicchetti, and Comminiello 2025c).

Method	Latent dim.	MAE (↓)	MSE (↓)	Acc@1 (↑)	TTI	TTI/TOI
Conventional	3	0.099	0.049	100.0	9	0.002
Ours	3	**0.099**	**0.050**	**100.0**	**3 (-67%)**	**0.001**
Conventional	16	0.078	0.016	100.0	48	0.012
Ours	16	**0.088**	**0.033**	**100.0**	**16 (-67%)**	**0.004**
Conventional	32	0.079	0.014	100.0	96	0.025
Ours	32	**0.091**	**0.038**	**100.0**	**32 (-67%)**	**0.008**

Table 2: Classification results transmitting the centroid $\mathbf{c}_i$ of different temperature τ trained models. Results adapted from (Grassucci et al. 2025).

		RFS 0%		RFS 50%		RFS 90%	
τ	Gap (↓)	Acc@1 (↑)	Acc@5 (↑)	Acc@1 (↑)	Acc@5 (↑)	Acc@1 (↑)	Acc@5 (↑)
0.01	0.604	0.22	0.32	0.07	0.18	0.01	0.02
0.07	0.162	0.50	0.74	0.45	0.63	0.06	0.11
0.1	0.052	**0.51**	0.77	**0.49**	0.70	0.10	0.16
0.2	**0.045**	**0.51**	**0.79**	0.44	**0.76**	**0.14**	**0.19**

distance between all positive pairs. Given a batch of B samples and assuming, without loss of generality, an anchor modality $\mathbf{a}$, this loss is defined as:

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

This term aggressively pulls semantically-related embeddings from different modalities into the same point in the latent space. However, $\mathcal{L}_{ATP}$ alone is insufficient, as it can lead to a representational collapse (Wang and Isola 2020), where all embeddings, regardless of semantics, are mapped to a small region.

To counteract this collapse, the second term, the Centroid Uniformity (CU) loss, $\mathcal{L}_{CU}$, is introduced. This loss functions as a regularizer, promoting sparsity and structural integrity by enforcing a uniform distribution of the semantic centroids ($\mathbf{c}_i$) across the unit hypersphere. This is achieved by maximizing the distance between the centroids of different semantic concepts i and j within the batch:

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

By applying the uniformity constraint to the centroids $\mathbf{c}_i$ rather than to the individual embeddings $\mathbf{z}_i^m$ as in (Wang and Isola 2020), this loss preserves the tight alignment learned by $\mathcal{L}_{ATP}$ while ensuring the overall latent space remains semantically discriminative.

The final gap-closing objective, $\mathcal{L}_{gap} = \mathcal{L}_{ATP} + \mathcal{L}_{CU}$, is then combined with the standard contrastive loss $\mathcal{L}_{CL}$ (the average of $\mathcal{L}_{m\to n}$ and $\mathcal{L}_{n\to m}$ terms) to form the complete training objective:

$$\mathcal{L}_{Total} = \mathcal{L}_{CL} + \lambda \mathcal{L}_{gap} \quad (6)$$

where λ is a hyperparameter balancing the two objectives. This framework trains encoders to produce the aligned, gap-free space required to validate our $1/M$ semantic centroid compression strategy.

Downstream Tasks

The primary benefit of transmitting solely the semantic centroid $\mathbf{c}_i$ instead of multiple embeddings for each modality, enabled by gap reduction, extends beyond simple transmission efficiency. This modality-agnostic representation $\mathbf{c}_i$ serves as a highly efficient, semantically pure input for any task performed at the receiver. This approach avoids the need to store and process M separate, redundant vectors, and, as we will show, the resulting centroid is intrinsically more compressible than its non-aligned counterparts. We demonstrate this efficiency through the downstream tasks of multimodal reconstruction and classification.

Efficient Multimodal Reconstruction

A foundational task for a semantic communication system is the reconstruction of the original multimodal data. In a conventional system, this requires M separate decoders, each receiving its corresponding M latent vectors. In our proposed framework, all M decoders ($\mathcal{D}^1, \ldots, \mathcal{D}^M$) receive the identical semantic centroid $\mathbf{c}_i$ as input (Grassucci, Cicchetti, and Comminiello 2025b,c). Experimental evidence in a three-modality (image, audio, text) scenario demonstrates this capability (Grassucci, Cicchetti, and Comminiello 2025c,b). Despite transmitting only $\mathbf{c}_i$, the receiver-side decoders are able to successfully reconstruct all three modalities, effectively preserving the semantic content (Grassucci, Cicchetti, and Comminiello 2025c). The critical result, highlighted in Table 1, is that this comparable

reconstruction performance is achieved with a 67% reduction in total transmitted bits (TTI) compared to the conventional approach of sending a vector for each modality (Grassucci, Cicchetti, and Comminiello 2025c). This confirms the $1/M$ bandwidth reduction for the reconstruction task.

Efficient Classification and Further Compression

Beyond reconstruction, a key goal of semantic communication is to enable downstream tasks like classification directly from the transmitted latent vector. Here, the semantic centroid $\mathbf{c}_i$ provides a compounded efficiency gain (Grassucci et al. 2025).

First, the centroid $\mathbf{c}_i$ itself serves as a superior, more compact representation for classification than the standard baseline. Conventionally, a receiver would concatenate all M vectors to form a $\mathbf{z}_{concat} = [\mathbf{z}^1, \dots, \mathbf{z}^M]$ embedding. Studies show that using the single centroid $\mathbf{c}_i$, which contains only $1/M$ (e.g., 50%) of the information, achieves comparable or even superior classification accuracy to using the full, concatenated vector (Grassucci et al. 2025). This implies the modality-specific information discarded from the M vectors is largely redundant noise for the classification task, and the centroid $\mathbf{c}_i$ is a cleaner, more semantically-dense representation. Second, this semantically-pure centroid $\mathbf{c}_i$ possesses a lower intrinsic dimensionality. A latent space with a reduced gap, trained at higher temperatures, is inherently more robust to post-training compression techniques such as Random Feature Selection (RFS) (Grassucci et al. 2025). Experimental results for multilabel and single-label classification tasks are stark. Models trained with a large gap ($\tau = 0.01$) suffer a catastrophic performance drop when compressed. Conversely, models trained with a gap-reducing high temperature ($\tau = 0.4$) maintain high accuracy even when 95% of the vector features are randomly dropped (Grassucci et al. 2025). This robustness allows for a second, cascaded layer of compression. Not only do we gain a $1/M$ reduction by using $\mathbf{c}_i$, but $\mathbf{c}_i$ itself can be aggressively compressed. When comparing RFS applied to the baseline "Concat + RFS" versus our "Centroid + RFS", the centroid-based method consistently and significantly outperforms the concatenated baseline at all equivalent compression ratios 2. This demonstrates that gap reduction enables a double-dip in efficiency: an initial $1/M$ compression by transmitting the centroid, followed by substantial further compression (e.g., RFS) with minimal loss of downstream task performance.

Conclusion

In this paper, we have presented a theoretical framework linking multimodal representation alignment with the efficiency of semantic communication. We identified the modality gap, i.e., a persistent structural misalignment in contrastively-trained models, as a fundamental bottleneck that forces conventional systems to transmit M separate latent vectors for M modalities, negating the core promise of semantic compression. We argued that closing this gap is a key theoretical enabler for bandwidth efficiency. By employing methods such as high-temperature training or explicit gap-closing loss functions, a (near-)zero gap latent space can be achieved. This alignment makes modality-specific vectors redundant, enabling a novel and efficient transmission strategy: sending a single semantic centroid to represent the entire multimodal concept. This approach provides a direct $1/M$ reduction in transmission load. Furthermore, we posited that this resulting semantic centroid is not only compact but also intrinsically more compressible. As evidenced by its superior robustness to post-training compression techniques like RFS, this double-dip in efficiency allows for substantial, cascaded compression with minimal impact on downstream task performance, such as reconstruction or classification, at the receiver. This work establishes a clear path toward scalable, efficient, and truly multimodal semantic communication systems.

References

Calvanese Strinati, E.; and Barbarossa, S. 2020. 6G Networks: Beyond Shannon Towards Semantic and Goal-Oriented Communications. *Computer Networks*, 190: 107930.

Cicchetti, G.; Grassucci, E.; Sigillo, L.; and Comminiello, D. 2025. Gramian Multimodal Representation Learning and Alignment. In *Int. Conf. on Learning Representations (ICLR)*.

Grassucci, E.; Cicchetti, G.; and Comminiello, D. 2025a. Closing the gap in multimodal medical representation alignment. In *IEEE Workshop on Machine Learning for Signal Process.*

Grassucci, E.; Cicchetti, G.; and Comminiello, D. 2025b. Closing The Modality Gap Enables Novel Multimodal Learning Applications. In *Second Workshop on Representational Alignment at ICLR 2025*.

Grassucci, E.; Cicchetti, G.; and Comminiello, D. 2025c. Contrastive Learning Enables Low-Bandwidth Semantic Communication. In *AAAI 2025 Workshop on Artificial Intelligence for Wireless Communications and Networking (AI4WCN)*.

Grassucci, E.; Cicchetti, G.; Uncini, A.; and Comminiello, D. 2025. Semantic Compression via Multimodal Representation Learning. *ArXiv preperint: arXiv:2509.24431*.

Huang, J.; Li, D.; Xiu, C. H.; Qin, X.; and Zhang, W. 2023. Joint Task and Data Oriented Semantic Communications: A Deep Separate Source-channel Coding Scheme. *ArXiv preprint: ArXiv:2302.13580*.

Liang, W.; Zhang, Y.; Kwon, Y.; Yeung, S.; and Zou, J. 2022. Mind the Gap: Understanding the Modality Gap in Multimodal Contrastive Representation Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Luo, X.; Chen, H.-H.; and Guo, Q. 2022. Semantic Communications: Overview, Open Issues, and Future Research Directions. *IEEE Wireless Comm.*, 29(1): 210–219.

Qin, Z.; Tao, X.; Lu, J.; and Li, G. Y. 2021. Semantic Communications: Principles and Challenges. *ArXiv preprint: ArXiv:2201.01389*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.;

Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning (ICML)*.

Shen, R.; Wu, H.; Zhang, W.; Hu, J.; and Gunduz, D. 2025. Compression Beyond Pixels: Semantic Compression with Multimodal Foundation Models. In *International Workshop on Machine Learning for Signal Processing (MLSP)*.

van den Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation Learning with Contrastive Predictive Coding. *ArXiv preprint: arXiv:1807.03748*.

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

Yaras, C.; Chen, S.; Wang, P.; and Qu, Q. 2025. Explaining and Mitigating the Modality Gap in Contrastive Multimodal Learning. In *Conference on Parsimony and Learning (CPAL)*.

You, C.; Cai, Y.; Liu, Y.; di Renzo, M.; Duman, T. M.; Yener, A.; and Swindlehurst, A. L. 2024. Next Generation Advanced Transceiver Technologies for 6G. *IEEE Journal on Selected Areas in Communications*.

Zhang, P.; Xu, W.; Liu, Y.; Qin, X.; Niu, K.; Cui, S.; Shi, G.; Qin, Z.; Xu, X.; Wang, F.; Meng, Y.; Dong, C.; Dai, J.; Yang, Q.; Sun, Y.; Gao, D.; Gao, H.; Han, S.; and Song, X. 2024. Intellicise Wireless Networks From Semantic Communications: A Survey, Research Issues, and Challenges. *IEEE Communications Surveys & Tutorials*, 1–1.

Zhang, Y.; Sun, Y.; Guo, L.; Chen, W.; Ai, B.; and Gunduz, D. 2025. Multimodal LLM Integrated Semantic Communications for 6G Immersive Experiences. *2507.04621*.