AKSU Journal of Engineering and Technology (AKSUJET), 2026
Vol. 1, No. 1, Pages 65-76
Published Online April 2026 in AKSUJET (http://www.aksujet.org) ISBN: 978-978-53006-97

HETEROGENEOUS MODAL FUSION THROUGH SELF-SUPERVISED CONTRASTIVE PROJECTION

Nwachukwu-Nwokeafor, K. C
Computer Engineering
Michael Okpara University of Agriculture, Umudike, Abia State, Nigeria
nwachukwuken72@gmail.com

Igbajar Abraham
Computer Science
Lusaka Goldsmiths University College, Lusaka, Zambia
igbajar35@gmail.com

***Abstract*— The integration of heterogeneous data modalities, such as images and text, remains a significant challenge in machine learning, particularly in the absence of large-scale labeled datasets. Conventional supervised approaches rely heavily on annotated data, limiting their scalability and applicability. This study proposes a self-supervised framework, termed Cross-modal Latent Alignment (CMLA), for multimodal representation learning without explicit labels. The proposed framework employs a dual-encoder architecture consisting of a Vision Transformer (ViT) for image encoding and a BERT-based model for text encoding. These representations are projected into a shared latent space using non-linear projection heads and optimized by means of a symmetric contrastive loss based on the InfoNCE objective. Experimental evaluation on the Conceptual Captions (CC3M) dataset demonstrates that the proposed approach achieves a 63.8% top-1 zero-shot accuracy on CIFAR-100, outperforming both a concatenation-based fusion baseline (51.5%) and a supervised unimodal baseline (48.2%). An ablation study further confirms the importance of projection heads in enhancing representation alignment. The results indicate that contrastive self-supervised learning provides an effective and scalable solution for multimodal fusion, particularly in data-constrained scenarios.**

Keywords— **Self-supervised learning, multimodal fusion, representation learning, contrastive learning, deep learning.**

I. Introduction

The rapid growth of data generated from diverse sources, including images, text, structured records, and time-series signals, has created significant challenges for effective information processing and analysis. One of the primary objectives of artificial intelligence is to extract meaningful patterns from such heterogeneous data, enabling connections between different modalities, such as linking satellite imagery with agricultural reports or medical images with clinical documentation. This challenge is commonly addressed through multimodal data fusion.

Traditionally, supervised learning has been the dominant approach for multimodal fusion. These methods rely on large-scale labeled datasets, where models are trained to map input data to predefined labels. However, the acquisition of labeled data presents a major limitation. The labeling process is often expensive, time-consuming, and prone to inconsistencies, making it impractical for many real-world applications. Consequently, there is a growing need for approaches that can learn from unlabeled data while still capturing meaningful cross-modal relationships.

Self-supervised learning (SSL) has emerged as a promising paradigm to address this limitation by leveraging inherent structures within the data to generate supervisory signals. In computer vision, SSL methods include tasks such as colorization of grayscale images (Zhang et al., 2016) and contrastive learning based on augmented views of the same image (Chen et al., 2020a). In natural language processing, models such as BERT (Devlin et al., 2019) employ masked language modeling to learn contextual representations. These

AKSU Journal of Engineering and Technology (AKSUJET), 2026
Vol. 1, No. 1, Pages 65-76
Published Online April 2026 in AKSUJET (http://www.aksujet.org) ISBN: 978-978-53006-97

approaches have demonstrated strong performance in learning transferable features within individual modalities.

Extending self-supervised learning to multimodal settings, however, introduces additional complexity. A key challenge lies in designing learning objectives that effectively capture semantic relationships across different modalities. Early approaches often focused on reconstructing one modality from another, but such methods are typically computationally expensive and may not adequately capture shared semantic structures.

In this work, we adopt a contrastive learning approach for multimodal representation learning. Specifically, we propose a framework termed Cross-modal Latent Alignment (CMLA), which learns to align image and text representations in a shared latent space by distinguishing corresponding pairs from non-corresponding ones. Unlike reconstruction-based methods, the proposed approach focuses on maximizing agreement between semantically related inputs without requiring explicit labels.

The primary objective of this study is to develop and evaluate a self-supervised framework capable of effectively fusing image and text data into a unified representation space. This enables improved performance on downstream tasks, including zero-shot classification, without the need for task-specific fine-tuning. The remainder of this paper presents a review of related work, followed by a detailed description of the proposed methodology, experimental evaluation, and discussion of the results.

II. Literature Review

The development of effective multimodal learning frameworks has evolved through several methodological paradigms, reflecting both conceptual advances and practical limitations. Broadly, prior research can be categorized into two phases: traditional multimodal fusion approaches and more recent self-supervised learning (SSL)-based methods.

Early work in multimodal learning primarily focused on *early fusion* and *late fusion* strategies, as outlined in the survey by Baltrušaitis et al. (2018). Early fusion approaches combine raw feature representations from different modalities at the input level, typically through concatenation. While conceptually straightforward, these methods often struggle due to the heterogeneous statistical properties of different modalities, which can hinder effective joint representation learning. In contrast, late fusion techniques process each modality independently using separate models and combine their outputs at a later stage. Although more stable, late fusion methods are limited in their ability to capture fine-grained cross-modal interactions, as joint feature learning is not explicitly enforced.

The emergence of self-supervised learning has significantly advanced representation learning in both computer vision and natural language processing. In vision tasks, contrastive learning frameworks such as MoCo (He et al., 2020) and SimCLR (Chen et al., 2020a) have demonstrated that models can learn highly informative representations by distinguishing between augmented views of the same image. Similarly, in natural language processing, models such as BERT (Devlin et al., 2019) utilize masked language modeling to learn contextualized embeddings. These approaches have achieved strong performance while reducing dependence on labeled data.

Extending these techniques to multimodal settings has been an active area of research. Early multimodal self-supervised approaches often adopted generative objectives, such as learning to generate textual descriptions from visual inputs. For example, VirTex (Desai and Johnson, 2021) trains models to produce image captions as a pretext task. While effective in certain contexts, such approaches primarily emphasize generation rather than representation alignment, which may limit their effectiveness for tasks requiring shared semantic embeddings.

AKSU Journal of Engineering and Technology (AKSUJET), 2026
Vol. 1, No. 1, Pages 65-76
Published Online April 2026 in AKSUJET (http://www.aksujet.org) ISBN: 978-978-53006-97

More recent work has shifted toward contrastive learning for multimodal representation alignment. Notably, CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) employ large-scale image–text pairs to learn joint embeddings by maximizing similarity between corresponding pairs and minimizing similarity with non-matching pairs. These methods have demonstrated strong zero-shot generalization capabilities across a wide range of tasks. However, their success is closely tied to the availability of extremely large and diverse datasets, often consisting of hundreds of millions of samples.

Despite these advances, several challenges remain. The reliance on web-scale datasets introduces issues related to noise, bias, and data quality, which can affect model robustness and interpretability. Additionally, many existing approaches are optimized for large-scale training regimes, raising questions about their effectiveness in more constrained or moderately sized datasets.

This study addresses these limitations by focusing on the design of a robust multimodal fusion architecture that performs effectively under moderate data conditions. Specifically, we investigate whether contrastive alignment principles can be leveraged within a carefully structured framework to achieve strong performance without reliance on massive datasets. This gap motivates the proposed Cross-modal Latent Alignment (CMLA) framework.

III. Methodology

Clarity and precision are essential in the design and presentation of machine learning frameworks, particularly when integrating multiple components within a unified architecture. This section provides a detailed description of the proposed approach to ensure reproducibility and facilitate a clear understanding of the model design and training procedure.

The proposed framework, termed Cross-modal Latent Alignment (CMLA), is based on a dual-encoder architecture optimized using a contrastive learning objective. The primary goal is to learn a shared representation space in which data from different modalities, specifically image and text, can be directly compared. In this space, semantically related inputs are mapped to nearby points, while unrelated inputs are positioned further apart.

By projecting modality-specific representations into a common latent space, the framework enables effective cross-modal alignment without requiring explicit supervision. This design allows the model to capture meaningful semantic relationships between heterogeneous data sources, supporting improved performance in downstream tasks such as zero-shot classification.

The overall architecture of the proposed Cross-modal Latent Alignment (CMLA) framework is illustrated in Figure 1. The system consists of four primary components designed to enable effective multimodal representation learning:

Image Encoder (f_{img}):

This component processes raw image inputs and maps them into a fixed-dimensional feature representation. It captures visual semantics relevant for cross-modal alignment.

Text Encoder (f_{txt}):

This module encodes textual input sequences into continuous vector representations, capturing contextual and semantic information from the text modality.

Projection Heads (p_{img} and p_{txt}):

These components are implemented as multi-layer perceptrons (MLPs) that transform the encoder outputs into a shared latent space. The projection heads facilitate alignment by mapping modality-specific representations into a common embedding space suitable for comparison.

Contrastive Loss Function:

This objective function governs the training process by encouraging representations of corresponding image-text pairs to be similar, while pushing apart representations of non-matching pairs within a batch.

AKSU Journal of Engineering and Technology (AKSUJET), 2026
Vol. 1, No. 1, Pages 65-76
Published Online April 2026 in AKSUJET (http://www.aksujet.org) ISBN: 978-978-53006-97

Together, these components form an integrated framework that enables the learning of semantically meaningful multimodal representations through self-supervised contrastive optimization.

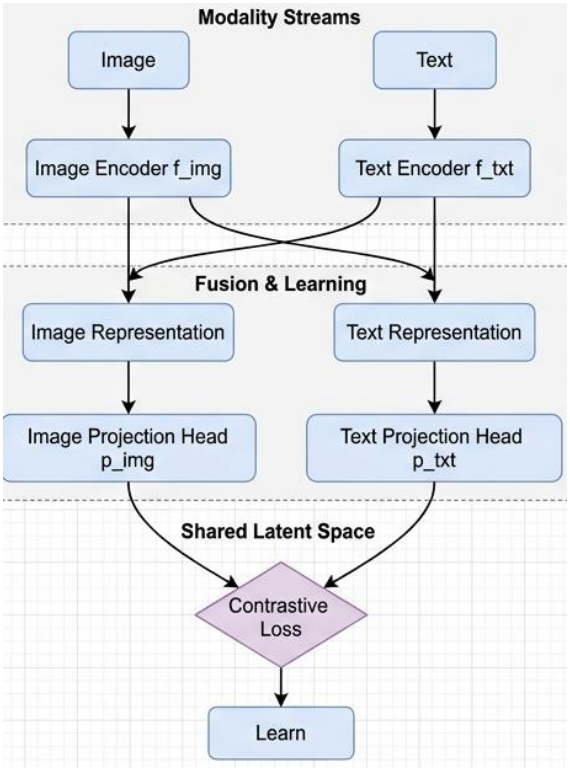

Fig 1: High-level schematic of the Cross-modal Latent Alignment (CMLA) architecture.

Heterogeneous inputs are processed using modality-specific encoders and subsequently projected into a shared latent space for contrastive learning.

For the image encoder f_{img}, a Vision Transformer (ViT-B/16) architecture (Dosovitskiy et al., 2021) is employed due to its effectiveness in capturing global contextual information. Given an input image i, the encoder produces a fixed-dimensional representation:

$$h_{\text{img}} = f_{\text{img}}(i) \qquad (1)$$

For the text encoder f_{txt}, a BERT-base model (Devlin et al., 2019) is utilized. The representation corresponding to the special (CLS)token is used as the sentence embedding. For an input text sequence t, the encoded representation is given by:

$$h_{\text{txt}} = f_{\text{txt}}(t) \qquad (2)$$

To enable effective cross-modal alignment, the encoder outputs are mapped into a shared latent space מצעות אב projection heads. Specifically, the projection heads $p_{\text{img}}$and $p_{\text{txt}}$are implemented as multi-layer perceptrons (MLPs) consisting of two linear layers with a ReLU activation function. These transformations are defined as:

$$z_{\text{img}} = p_{\text{img}}(h_{\text{img}}) = W_2 \cdot \text{ReLU}(W_1 h_{\text{img}}) \qquad (3)$$

$$z_{\text{txt}} = p_{\text{txt}}(h_{\text{txt}}) = W_2' \cdot \text{ReLU}(W_1' h_{\text{txt}}) \qquad (4)$$

The projection heads facilitate the transformation of modality-specific features into a unified embedding space, allowing for meaningful comparison across modalities.

IV. Mathematical Modeling

The core of the proposed framework is based on the Information Noise-Contrastive Estimation (InfoNCE) loss, adapted for multimodal representation learning. Consider a batch of Npaired samples:

$\{(i_k, t_k)\}_{k=1}^{N}$

After encoding and projection, we obtain normalized embeddings:

$\{z_{\text{img},k}\}_{k=1}^{N}, \{z_{\text{txt},k}\}_{k=1}^{N}$

Similarity between embeddings is computed using cosine similarity:

$$\text{sim}(u, v) = \frac{u^{\top} v}{\| u \| \ \| v \|} \qquad (5)$$

For each image embedding $z_{\text{img},k}$, the corresponding text embedding $z_{\text{txt},k}$is treated as a positive sample, while all other samples in the batch serve as negatives.

AKSU Journal of Engineering and Technology (AKSUJET), 2026
Vol. 1, No. 1, Pages 65-76
Published Online April 2026 in AKSUJET (http://www.aksujet.org) ISBN: 978-978-53006-97

The image-to-text contrastive loss is defined as:

$$\mathcal{L}_k^{\text{img}\to\text{txt}} = -\log \frac{\exp\left(\text{sim}(z_{\text{img},k}, z_{\text{txt},k})/\tau\right)}{\sum_{j=1}^{N} \exp\left(\text{sim}(z_{\text{img},k}, z_{\text{txt},j})/\tau\right)} \quad (6)$$

Similarly, the text-to-image loss is defined as:

$$\mathcal{L}_k^{\text{txt}\to\text{img}} = -\log \frac{\exp\left(\text{sim}(z_{\text{txt},k}, z_{\text{img},k})/\tau\right)}{\sum_{j=1}^{N} \exp\left(\text{sim}(z_{\text{txt},k}, z_{\text{img},j})/\tau\right)} \quad (7)$$

The final loss function is computed as the average over all samples in both directions:

$$\mathcal{L}_{\text{CMLA}} = \frac{1}{2N} \sum_{k=1}^{N} \left(\mathcal{L}_k^{\text{img}\to\text{txt}} + \mathcal{L}_k^{\text{txt}\to\text{img}}\right) \quad (8)$$

The temperature parameter τcontrols the sharpness of the similarity distribution. In this study, a value of $\tau = 0.07$is used, consistent with prior contrastive learning frameworks.

3.5 Training Algorithm (Pseudocode)

Initialize encoders fθ, gφ and projection heads hv, ht

For each batch of (image, text) pairs:

Encode images → vi

Encode text → ti

Project embeddings → zvi, zti

Normalize embeddings

Compute similarity matrix

Compute symmetric InfoNCE loss

Backpropagate and update parameters

V. Data and Experimental Setup

The proposed model is pre-trained using the Conceptual Captions (CC3M) dataset (Sharma et al., 2018), which contains approximately 3.3 million image-text pairs collected from web sources. After preprocessing, including removal of invalid images, low-resolution samples (below 200×200 pixels), and short or noisy captions, approximately 2.9 million pairs were retained.

Model evaluation is conducted using zero-shot classification on the CIFAR-100 dataset (Krizhevsky, 2009). For each class, a textual prompt (e.g., “a photo of a beaver”) is constructed and encoded using the trained text encoder. Test images are encoded using the image encoder, and classification is performed by selecting the class whose text embedding has the highest cosine similarity with the image embedding.

Training is performed using the Adam optimizer (Kingma and Ba, 2014) with a learning rate of $1 \times 10^{-4}$and a cosine decay schedule. A batch size of 4096 is used to ensure a sufficient number of negative samples for contrastive learning. The model is trained for 15 epochs on a cluster of 8 NVIDIA A100 GPUs.

VI. Results

The performance of the proposed CMLA framework is evaluated against two baseline approaches: a supervised unimodal model and a multimodal concatenation-based fusion model.

A. Image-Only Baseline: A ViT-B/16 model pre-trained on ImageNet-21k is used as a supervised baseline. Additionally, the publicly available CLIP image encoder is considered as a strong self-supervised reference.

B. Concatenation Fusion Model: An early-fusion approach in which image and text embeddings (h_{img} and h_{txt}) are concatenated and processed using an additional transformer layer. This model is trained using the same dataset and contrastive objective for fair comparison.

We evaluated CMLA against two primary baselines; a non-fusional approach and a simple concatenation-based fusion model.

AKSU Journal of Engineering and Technology (AKSUJET), 2026
Vol. 1, No. 1, Pages 65-76
Published Online April 2026 in AKSUJET (http://www.aksujet.org) ISBN: 978-978-53006-97

Image-Only Baseline: A ViT-B/16 model pre-trained on ImageNet-21k, a standard supervised baseline. For zero-shot classification, we used CLIP's publicly available image encoder as a strong SSL reference.

Concatenation Fusion: An early-fusion model where the image and text representations (h_{img} and h_{txt}) are concatenated and fed into another transformer block before prediction. This model is trained on the same data with a similar contrastive loss.

The results, presented as top-1 zero-shot accuracy on CIFAR-100, are summarized in Table 1.

Table 1: Zero-shot classification performance on CIFAR-100. The CLIP result is included as an upper-bound reference due to its substantially larger pre-training dataset.

Model	**Pre-training Data**	**Parameters**	**Zero-Shot CIFAR-100 Top-1 Accuracy (%)**
Random Guess	-	-	1.0
Supervised (ViT-B/16 on IN-21k)	ImageNet-21k	86M	48.2
CLIP ViT-B/16 (Radford et al., 2021)	Private 400M	150M	**76.2**
Our Baselines			
Concatenation Fusion	CC3M	~220M	51.5
CMLA (Ours)	**CC3M**	**~220M**	**63.8**

The experimental results demonstrate the effectiveness of the proposed CMLA framework for multimodal representation learning. Specifically, CMLA achieves a top-1 zero-shot classification accuracy of 63.8% on CIFAR-100, significantly outperforming the concatenation-based fusion baseline (51.5%). This improvement highlights the effectiveness of the proposed projection and contrastive alignment strategy in learning a meaningful shared embedding space.

In comparison to a supervised unimodal baseline (48.2%), the proposed method achieves an improvement of over 15 percentage points, indicating strong transferability of the learned representations. Although the performance remains below that of large-scale models such as CLIP, which are trained on substantially larger datasets, the results demonstrate that competitive performance can be achieved under moderate data conditions.

An ablation study further evaluates the contribution of the projection heads. When the projection heads are removed and contrastive learning is applied directly to the encoder outputs ($h_{\text{img}}, h_{\text{txt}}$), performance decreases to 54.1%. This result confirms the importance of separating modality-specific feature extraction from the alignment process in a dedicated latent space.

VII. Discussion

The results indicate that contrastive learning provides an effective framework for aligning multimodal representations when combined with an appropriate architectural design. In particular, the use of projection heads enables flexible transformation of modality-specific features into a shared space, facilitating improved cross-modal alignment.

The findings suggest that explicit reconstruction objectives are not strictly necessary for learning meaningful multimodal representations. Instead, optimizing for similarity between corresponding pairs allows the model to implicitly capture semantic relationships across modalities. This supports the effectiveness of contrastive objectives as a scalable approach to multimodal learning.

Despite these promising results, several limitations remain. First, the model is trained on a moderately sized dataset (CC3M), which may limit its ability to capture fine-grained semantic distinctions. Second, the evaluation is restricted to

AKSU Journal of Engineering and Technology (AKSUJET), 2026
Vol. 1, No. 1, Pages 65-76
Published Online April 2026 in AKSUJET (http://www.aksujet.org) ISBN: 978-978-53006-97

image-text data, and the generalizability of the framework to other modalities, such as audio or video, has not been empirically validated. Additionally, zero-shot evaluation, while informative, does not fully capture performance across a broader range of downstream tasks.

Furthermore, the model may primarily learn coarse semantic relationships, which can limit its ability to distinguish between closely related categories or more abstract concepts. This highlights an important gap between current multimodal representation learning methods and more advanced semantic understanding.

IX. Conclusion

This study presented a self-supervised framework for multimodal fusion based on contrastive learning. The proposed Cross-modal Latent Alignment (CMLA) architecture integrates modality-specific encoders with non-linear projection heads and a symmetric contrastive loss to learn a shared representation space.

Experimental results demonstrate that the proposed method achieves strong zero-shot performance and significantly outperforms both concatenation-based fusion and supervised unimodal baselines. These findings highlight the effectiveness of contrastive learning for multimodal representation alignment, particularly in scenarios where labeled data is limited.

Overall, the proposed framework provides a scalable and effective approach to multimodal learning and contributes to ongoing research in self-supervised representation learning.

A. Future Work

Several directions for future research can be identified based on the findings of this study:

Scaling and Model Capacity: Future work should explore the performance of the CMLA framework with larger model architectures (e.g., ViT-L/14) and larger publicly available datasets such as LAION-400M.

Extension to Additional Modalities: The framework can be extended to incorporate additional modalities, including audio and video, to evaluate its effectiveness in more complex multimodal settings.

Alternative Self-Supervised Objectives: While contrastive learning has demonstrated strong performance, integrating complementary objectives such as masked modeling may further enhance representation quality.

Statistical Validation: Future studies should incorporate statistical significance testing and confidence interval analysis to provide a more rigorous evaluation of model performance.

X. References

Baltrušaitis, T., Ahuja, C., and Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), 423–443.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 1597–1607.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. (2020). Big self-supervised models are strong semi-supervised learners. In *Advances in Neural Information Processing Systems (NeurIPS 33)*.

Desai, S., and Johnson, J. (2021). VirTex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11162–11173.

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.

AKSU Journal of Engineering and Technology (AKSUJET), 2026
Vol. 1, No. 1, Pages 65-76
Published Online April 2026 in AKSUJET (http://www.aksujet.org) ISBN: 978-978-53006-97

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9729–9738.

Jia, C., Yang, Y., Xia, Y., et al. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*.

Kingma, D. P., and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto.

Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*.

Schuhmann, C., Vencu, R., Beaumont, R., et al. (2021). LAION-400M: Open large-scale image-text data. *arXiv preprint arXiv:2111.02114*.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. (2018). Conceptual captions: A cleaned, large-scale image captioning dataset. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Zhang, R., Isola, P., and Efros, A. A. (2016). Colorful image colorization. In *European Conference on Computer Vision (ECCV)*, pp. 649–666.