M3CoL: Harnessing Shared Relations via Multimodal Mixup Contrastive Learning for Multimodal Classification

Anonymous authors
Paper under double-blind review

Abstract

Deep multimodal learning has shown remarkable success by leveraging contrastive learning to capture explicit one-to-one relations across modalities. However, real-world data often exhibits shared relations beyond simple pairwise associations. We propose **M3CoL**, a **M**ulti**m**odal **M**ixup **Co**ntrastive **L**earning approach to capture nuanced *shared relations* inherent in multimodal data. Our key contribution is a Mixup-based contrastive loss that learns robust representations by aligning mixed samples from one modality with their corresponding samples from other modalities thereby capturing shared relations between them. For multimodal classification tasks, we introduce a framework that integrates a fusion module with unimodal prediction modules for auxiliary supervision during training, complemented by our proposed Mixup-based contrastive loss. Through extensive experiments on diverse datasets (N24News, ROSMAP, BRCA, and Food-101), we demonstrate that **M3CoL** effectively captures shared multimodal relations and generalizes across domains. It outperforms state-of-the-art methods on N24News, ROSMAP, and BRCA, while achieving comparable performance on Food-101. Our work highlights the significance of learning shared relations for robust multimodal learning, opening up promising avenues for future research.

1 Introduction

The way we perceive the world is shaped by various modalities, such as language, vision, audio, and more. In the era of abundant and accessible multimodal data, it is increasingly crucial to equip artificial intelligence with multimodal capabilities (Baltrušaitis et al., 2018). At the heart of advancements in multimodal learning is contrastive learning, which maximizes similarity for positive pairs and minimizes it for negative pairs, making it practical for multimodal representation learning. CLIP (Radford et al., 2021) is a prominent example that employs contrastive learning to understand the direct link between paired modalities and seamlessly maps images and text into a shared space for cross-modal understanding, which can be later utilized for tasks such as retrieval and classification. However, traditional contrastive learning methods often overlook shared relationships between samples across different modalities, which can result in the learning of representations that are not fully optimized for capturing the underlying connections between diverse data modalities. These methods primarily focus on distinguishing between positive and negative pairs of samples, typically treating each instance as an independent entity. They tend to disregard the rich, shared relational information that could exist between samples within and across modalities. This limited focus can prevent the model from leveraging valuable contextual information, such as semantic similarities or complementary patterns, which can enhance robust representation learning. Consequently, this can lead to suboptimal performance in downstream tasks that require optimized shared representations, such as image-text alignment, cross-modal retrieval, or multimodal fusion tasks.

While traditional contrastive learning methods treat paired modalities as positive samples and non-corresponding ones as negative, they often overlook shared relations between different samples. As shown in the left panel of Figure 1 (Left panel), classical contrastive learning approach assumes perfect one-to-one relations between modalities, which is rare in real-world data. For example, shared elements in images or text can relate even across separate samples, as illustrated by the elements like *"tomato sauce"* and *"basil"* in Figure 1. Our approach, illustrated in the right panel of Figure 1, goes beyond simple pairwise

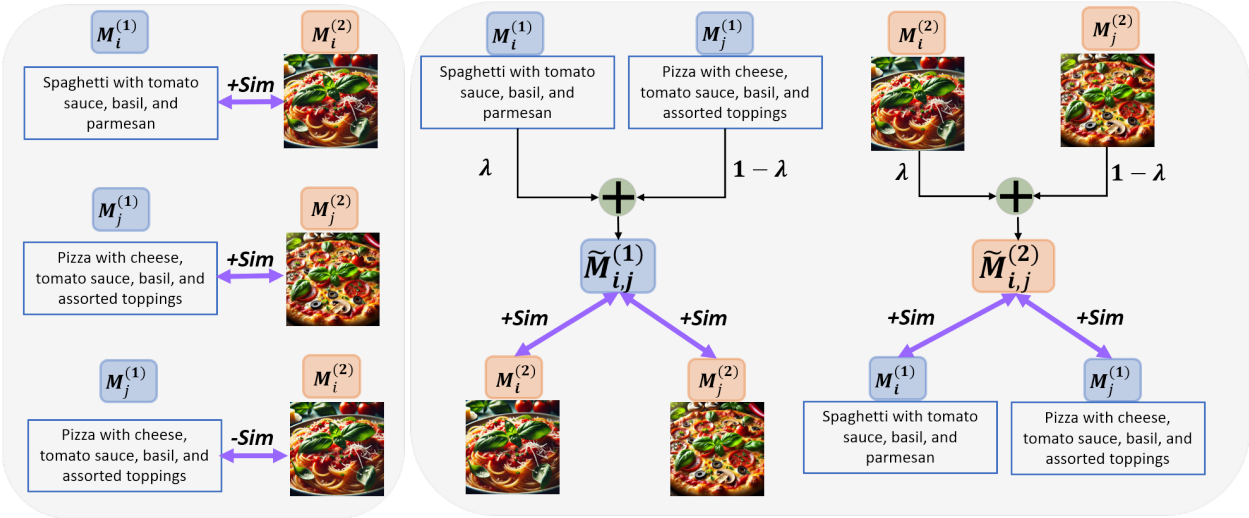


Figure 1: Comparison of traditional contrastive and our proposed M3Co loss. $\mathbf{M}_i^{(1)}$ and $\mathbf{M}_i^{(2)}$ denote representations of the i-th sample from modalities 1 and 2, respectively. Traditional contrastive loss (left panel) aligns corresponding sample representations across modalities. M3Co (right panel) mixes the i-th and j-th samples from modality 1 and enforces the representations of this mixture to align with the representations of the corresponding i-th and j-th samples from modality 2, and vice versa. For the text modality, we mix the text embeddings, while we mix the raw inputs for other modalities. Similarity (Sim) represents type of alignment enforced between the embeddings for all modalities.

alignment by capturing shared relationships across mixed samples. By creating newer data points through convex combinations of data points our method more effectively models complex shared relationships, such as imperfect bijections (Liang et al., 2022a), enhancing multimodal classification performance.

Our approach builds upon the success of data augmentation techniques such as Mixup (Zhang et al., 2017) and their variants (Yun et al., 2019; Cubuk et al., 2019; Hendrycks et al., 2019), which have proven beneficial for enhancing learned feature spaces, improving both robustness and performance. Mixup trains models on synthetic data created through convex combinations of two datapoint-label pairs (Chapelle et al., 2000). These techniques are particularly valuable in low sample settings, as they help prevent overfitting and the learning of ineffective shortcuts (Chen et al., 2020; Robinson et al., 2021), common in contrastive learning. Building on the success of recent Mixup strategies (Shen et al., 2022; Thulasidasan et al., 2019; Verma et al., 2019) and MixCo (Kim et al., 2020), we introduce M3Co, a novel approach that significantly adapts and enhances contrastive learning principles to complex multimodal settings. M3Co modifies the CLIP loss to effectively handle multimodal scenarios, addressing the problem of instance discrimination, where models overly focus on distinguishing individual instances instead of capturing relationships between modalities. By leveraging convex combinations of data for contrastive learning, M3Co eliminates instance discrimination and enhances robust representation learning by capturing shared relations. These combinations serve as structured noise and treated as positive pairs with their corresponding samples from other modalities. Our experimental results demonstrate enhanced ability to capture shared relations enabling improvements in performance and generalization across a range of multimodal classification tasks.

Our key contributions are summarized as follows:

- We propose M3Co, a multimodal contrastive loss (Eq. 8) that utilizes mixed samples to effectively capture shared relationships across different modalities. By going beyond traditional pairwise alignment methods, M3Co makes representations more consistent with the complex, intertwined relationships usually observed in real-world data.
- We introduce a multimodal learning framework (Figure 2) consisting of unimodal prediction modules, a fusion module, and a novel Mixup-based contrastive loss. Our proposed method is

modality-agnostic, allowing for flexible application across various types of data, and continuously updates the representations necessary for accurate and consistent predictions.

- We demonstrate the effectiveness of our methodology by evaluating it on four public multimodal classification benchmark datasets from different domains: two image-text datasets, N24News and Food-101, and two medical datasets, ROSMAP and BRCA (Table 1, 2, 3). Our approach outperforms baseline models, especially on smaller datasets.

2 Methodology

Pipeline Overview: Figure 2 depicts our framework, which comprises of three components: unimodal prediction modules, a fusion module, and a Mixup-based contrastive loss. We obtain latent representations (using learnable modality specific encoders $f^{(1)}$ and $f^{(2)}$) of individual modalities and fuse them (denoted by concatenation symbol '+') to generate a joint multimodal representation, which is optimized using a supervised objective (through classifier 3). The unimodal prediction modules provide additional supervision during training (via classifier 1 and 2). These strategies enable deeper integration of modalities and allow the models to compensate for the weaknesses of one modality with the strengths of another. The Mixup-based contrastive loss (denoted by $\mathcal{L}_{\text{M3Co}}$) continuously updates the representations by capturing shared relations inherent in the multimodal data. This comprehensive approach enhances the understanding of multimodal data, improving accuracy and model robustness.

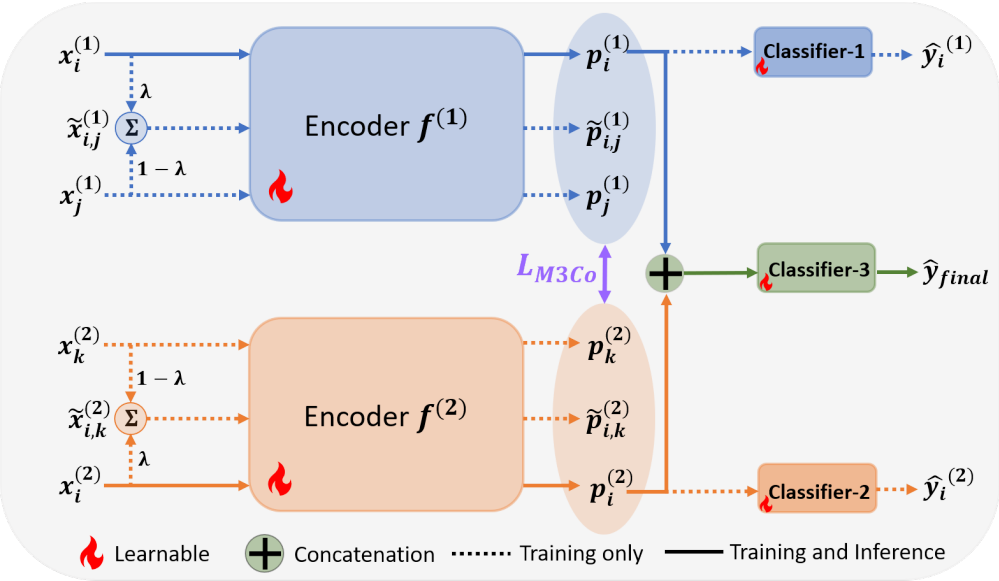


Figure 2: Architecture of our proposed M3CoL model. Samples from modality 1 ($\mathbf{x}_i^{(1)}$, $\mathbf{x}_j^{(1)}$) and modality 2 ($\mathbf{x}_i^{(2)}$, $\mathbf{x}_k^{(2)}$), along with their respective mixed data $\tilde{\mathbf{x}}_{i,j}^{(1)}$ and $\tilde{\mathbf{x}}_{i,k}^{(2)}$, are fed into encoders $f^{(1)}$ and $f^{(2)}$ to generate embeddings. Unimodal embeddings $\mathbf{p}_i^{(1)}$ and $\mathbf{p}_i^{(2)}$ are processed through classifier 1 and 2 to produce predictions $\hat{\mathbf{y}}_i^{(1)}$ and $\hat{\mathbf{y}}_i^{(2)}$ for training supervision only. The unimodal embeddings $\mathbf{p}_i^{(1)}$ and $\mathbf{p}_i^{(2)}$ are concatenated and processed through classifier 3 to yield $\hat{\mathbf{y}}_{\text{final}}$, utilized during training and inference. Additionally, unimodal embeddings $\mathbf{p}_i^{(1)}$, $\mathbf{p}_j^{(1)}$, $\mathbf{p}_i^{(2)}$, $\mathbf{p}_k^{(2)}$, and mixed embeddings $\tilde{\mathbf{p}}_{i,j}^{(1)}$ and $\tilde{\mathbf{p}}_{i,k}^{(2)}$ are utilized by our contrastive loss $\mathcal{L}_{\text{M3Co}}$ for shared alignment.

Multimodal Mixup Contrastive Learning: Given a batch of N multimodal samples, let $\mathbf{x}_i^{(1)}$ and $\mathbf{x}_i^{(2)}$ denote the i-th samples for the first and second modalities, respectively. The modality encoders, $f^{(1)}$ and $f^{(2)}$, generate the corresponding embeddings $\mathbf{p}_i^{(1)}$ and $\mathbf{p}_i^{(2)}$:

$$\mathbf{p}_i^{(1)} = f^{(1)}(\mathbf{x}_i^{(1)}), \quad \mathbf{p}_i^{(2)} = f^{(2)}(\mathbf{x}_i^{(2)}) \tag{1}$$

We generate a mixture, $\tilde{\mathbf{x}}^{(1)}_{i,j}$, of the samples $\mathbf{x}^{(1)}_i$ and $\mathbf{x}^{(1)}_j$ by taking their convex combination. Similarly, we generate a mixture, $\tilde{\mathbf{x}}^{(2)}_{i,k}$, using the convex combination of the samples $\mathbf{x}^{(2)}_i$ and $\mathbf{x}^{(2)}_k$ (Eq. 2). In the case of text modality, instead of directly mixing the raw inputs, we mix the text embeddings (Guo et al., 2019). The mixing indices j, k are drawn arbitrarily, without replacement, from $[1, N]$, for both the modalities. We mix both the modalities using a factor $\lambda \sim \text{Beta}(\alpha, \alpha)$. Based on the findings of (Zhang et al., 2017), which demonstrated enhanced performance for α values between 0.1 and 0.4, we chose $\alpha = 0.15$ after experimenting with several values in this range. The mixtures are fed through the respective encoders to obtain the embeddings: $\tilde{\mathbf{p}}^{(1)}_{i,j}$, and $\tilde{\mathbf{p}}^{(2)}_{i,k}$ (Eq. 3).

$$\tilde{\mathbf{x}}^{(1)}_{i,j} = \lambda_i \cdot \mathbf{x}^{(1)}_i + (1-\lambda_i) \cdot \mathbf{x}^{(1)}_j, \quad \tilde{\mathbf{x}}^{(2)}_{i,k} = \lambda_i \cdot \mathbf{x}^{(2)}_i + (1-\lambda_i) \cdot \mathbf{x}^{(2)}_k \tag{2}$$

$$\tilde{\mathbf{p}}^{(1)}_i = \tilde{\mathbf{p}}^{(1)}_{i,j} = f^{(1)}(\tilde{\mathbf{x}}^{(1)}_{i,j}), \quad \tilde{\mathbf{p}}^{(2)}_i = \tilde{\mathbf{p}}^{(2)}_{i,k} = f^{(2)}(\tilde{\mathbf{x}}^{(2)}_{i,k}) \tag{3}$$

The unidirectional contrastive loss (Sohn, 2016; Chen et al., 2020; Oord et al., 2018; Wu et al., 2018; Zhang et al., 2022) over $\mathbf{p}^{(2)}$ is conventionally defined as:

$$\mathcal{L}_{\text{sim-conv}}(\mathbf{p}^{(1)}, \mathbf{p}^{(2)}) = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(\mathbf{p}^{(1)}_i \cdot \mathbf{p}^{(2)}_i / \tau\right)}{\sum\limits_{j=1}^{N} \exp\left(\mathbf{p}^{(1)}_i \cdot \mathbf{p}^{(2)}_j / \tau\right)} \tag{4}$$

where $\cdot$ indicates dot product and τ is a temperature hyperparameter. While this formulation is needed for computing similarity among aligned samples from different modalities, our loss handles both aligned and non-aligned samples, as this enables to learn a better representation space. To achieve this, we define the unidirectional multimodal contrastive loss between $\mathbf{p}^{(1)}_i$ and $\mathbf{p}^{(2)}_m$ over $\mathbf{p}^{(2)}$ as:

$$\mathcal{L}_{\text{sim}}(\mathbf{p}^{(1)}_i, \mathbf{p}^{(2)}; m) = -\log \frac{\exp\left(\mathbf{p}^{(1)}_i \cdot \mathbf{p}^{(2)}_m / \tau\right)}{\sum\limits_{j=1}^{N} \exp\left(\mathbf{p}^{(1)}_i \cdot \mathbf{p}^{(2)}_j / \tau\right)} \tag{5}$$

where $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$ are $\mathcal{L}^2$ normalized, τ is a temperature hyperparameter, and m is a sample index in $[1, N]$. Although the unidirectional multimodal contrastive loss (Eq. 5) can learn indirect relations, it is insufficient for learning shared semi-positive relations between modalities. Therefore, we introduce a Mixup-based contrastive loss to capture these relations that promotes generalized learning, as this process is more nuanced than simply discriminating positives from negatives. Now, we make our loss bidirectional to encourage improved alignment in the shared representation space and efficient use of training data (Radford et al., 2021; Oord et al., 2018; Sohn, 2016). We define this bidirectional Mixup contrastive loss M3Co for each modality (Eq. 6, 7) and the total M3Co loss (Eq. 8) as:

$$\begin{aligned}\mathcal{L}^{(1)}_{\text{M3Co}} = &\frac{1}{N}\sum_{i=1}^{N}\left[\lambda_i \cdot \mathcal{L}_{\text{sim}}(\tilde{\mathbf{p}}^{(1)}_{i,j}, \mathbf{p}^{(2)}; i) + (1-\lambda_i) \cdot \mathcal{L}_{\text{sim}}(\tilde{\mathbf{p}}^{(1)}_{i,j}, \mathbf{p}^{(2)}; j)\right] \\ &+ \frac{1}{N}\sum_{i=1}^{N}\left\{\lambda_i \cdot \mathcal{L}_{\text{sim}}(\mathbf{p}^{(2)}_i, \tilde{\mathbf{p}}^{(1)}; i) + (1-\lambda_i) \cdot \mathcal{L}_{\text{sim}}(\mathbf{p}^{(2)}_j, \tilde{\mathbf{p}}^{(1)}; i)\right\}\end{aligned} \tag{6}$$

$$\begin{aligned}\mathcal{L}^{(2)}_{\text{M3Co}} = &\frac{1}{N}\sum_{i=1}^{N}\left[\lambda_i \cdot \mathcal{L}_{\text{sim}}(\tilde{\mathbf{p}}^{(2)}_{i,k}, \mathbf{p}^{(1)}; i) + (1-\lambda_i) \cdot \mathcal{L}_{\text{sim}}(\tilde{\mathbf{p}}^{(2)}_{i,k}, \mathbf{p}^{(1)}; k)\right] \\ &+ \frac{1}{N}\sum_{i=1}^{N}\left\{\lambda_i \cdot \mathcal{L}_{\text{sim}}(\mathbf{p}^{(1)}_i, \tilde{\mathbf{p}}^{(2)}; i) + (1-\lambda_i) \cdot \mathcal{L}_{\text{sim}}(\mathbf{p}^{(1)}_k, \tilde{\mathbf{p}}^{(2)}; i)\right\}\end{aligned} \tag{7}$$

$$\mathcal{L}^{(1,2)}_{\text{M3Co}} = \frac{1}{2}\left(\mathcal{L}^{(1)}_{\text{M3Co}} + \mathcal{L}^{(2)}_{\text{M3Co}}\right) \tag{8}$$

where $\mathbf{p}^{(1)}$, $\tilde{\mathbf{p}}^{(1)}$, $\mathbf{p}^{(2)}$, and $\tilde{\mathbf{p}}^{(2)}$ are $\mathcal{L}^2$ normalized. Note that the parts of the loss functions in Eq. (6, 7) inside curly parantheses make them bidirectional. Mixup-based methods enhance generalization by capturing clean patterns in the early training stages but can eventually overfit to noise if continued for larger number of epochs (Liu et al., 2023; Yu et al., 2021; Golatkar et al., 2019). To address this, we implement a schedule that transitions from the Mixup-based M3Co loss to a non-Mixup multimodal contrastive loss. We design this transition so that the non-Mixup loss retains the ability to learn shared or indirect relationships between modalities. By using a bidirectional SoftClip-based loss (Gao et al., 2024; Sohn, 2016; Chen et al., 2020), we relax the rigid one-to-one correspondence, allowing the model to capture many-to-many relations (Gao et al., 2024; 2022). The bidirectional **MultiS**oft**Clip** loss for each modality (Eq. 9, 10) and its combination (Eq. 11) is:

$$\mathcal{L}^{(1)}_{\text{MultiSClip}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{l=1}^{N}\left[\frac{\exp\left(\mathbf{p}^{(1)}_i \cdot \mathbf{p}^{(1)}_l/\tau\right)}{\sum_{t=1}^{N}\exp\left(\mathbf{p}^{(1)}_i \cdot \mathbf{p}^{(1)}_t/\tau\right)} \cdot \left(\mathcal{L}_{\text{sim}}(\mathbf{p}^{(2)}_i, \mathbf{p}^{(1)}; l) + \mathcal{L}_{\text{sim}}(\mathbf{p}^{(1)}_l, \mathbf{p}^{(2)}; i)\right)\right] \tag{9}$$

$$\mathcal{L}^{(2)}_{\text{MultiSClip}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{l=1}^{N}\left[\frac{\exp\left(\mathbf{p}^{(2)}_i \cdot \mathbf{p}^{(2)}_l/\tau\right)}{\sum_{t=1}^{N}\exp\left(\mathbf{p}^{(2)}_i \cdot \mathbf{p}^{(2)}_t/\tau\right)} \cdot \left(\mathcal{L}_{\text{sim}}(\mathbf{p}^{(1)}_i, \mathbf{p}^{(2)}; l) + \mathcal{L}_{\text{sim}}(\mathbf{p}^{(2)}_l, \mathbf{p}^{(1)}; i)\right)\right] \tag{10}$$

$$\mathcal{L}^{(1,2)}_{\text{MultiSClip}} = \frac{1}{2}\left(\mathcal{L}^{(1)}_{\text{MultiSClip}} + \mathcal{L}^{(2)}_{\text{MultiSClip}}\right) \tag{11}$$

where $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$ are $\mathcal{L}^2$ normalized. The M3Co and MultiSClip losses for M modalities is:

$$\mathcal{L}_{\text{M3Co}} = \sum_{i=1}^{M}\sum_{j>i}^{M}\mathcal{L}^{(i,j)}_{\text{M3Co}} \tag{12}$$

$$\mathcal{L}_{\text{MultiSClip}} = \sum_{i=1}^{M}\sum_{j>i}^{M}\mathcal{L}^{(i,j)}_{\text{MultiSClip}} \tag{13}$$

Unimodal Predictions and Fusion: The encoders produce latent representations for each of the M modalities, serving as inputs to individual classifiers that generate modality-specific predictions $\hat{\mathbf{y}}^{(m)}$. These representations are used for modality-specific supervision only during training. The unimodal prediction task involves minimizing the cross-entropy loss $\mathcal{L}_{\text{CE}}$ between these predictions and the corresponding ground truth labels ($\mathbf{y}$), for each modality. The unimodal cross-entropy loss is:

$$\mathcal{L}_{\text{CE-Uni}} = \sum_{m=1}^{M}\mathcal{L}_{\text{CE}}(\mathbf{y}, \hat{\mathbf{y}}^{(m)}) \tag{14}$$

We merge the unimodal latent representations by concatenating them and pass the combined representation to the output classifier. These predictions serve as the final outputs $\hat{\mathbf{y}}_f$ used during inference. The multimodal prediction process aims to minimize the cross-entropy loss between $\hat{\mathbf{y}}_f$ and the corresponding labels. The multimodal cross-entropy loss is:

$$\mathcal{L}_{\text{CE-Multi}} = \mathcal{L}_{\text{CE}}(\mathbf{y}, \hat{\mathbf{y}}_f) \tag{15}$$

Combined Learning Objective: Our overall loss objective utilizes a schedule to combine our M3Co and MultiSClip loss functions weighted by a hyperparamater β, along with the unimodal and multimodal cross-entropy losses. We use M3Co for the first one-third (Liu et al., 2023) part of training, and then transition to MultiSClip as over-training with a Mixup-based loss can potentially harm generalization. The end-to-end loss is defined as:

$$\mathcal{L}_{\text{Total}} = \beta \cdot \mathcal{L}_{\text{M3Co | MultiSClip}} + \mathcal{L}_{\text{CE-Uni}} + \mathcal{L}_{\text{CE-Multi}} \tag{16}$$

3 Experiments

Datasets. We evaluate our approach on four diverse publicly available multimodal classification datasets: N24News (Wang et al., 2022), Food-101 (Wang et al., 2015), ROSMAP (Wang et al., 2021a), and BRCA (Wang et al., 2021a). N24News and Food-101 are both bimodal image-text classification datasets. **Food-101** is a food classification dataset, where each sample is linked with a recipe description gathered from web pages and an associated image. **N24News** is a news classification dataset consisting of four text types (Abstract, Caption, Heading, and Body) along with the corresponding images. Following other works (Zou et al., 2023), we use the first three text types for our experiments. ROSMAP and BRCA are publicly available multimodal medical datasets, each containing three modalities: DNA methylation, miRNA expression, and mRNA expression. **ROSMAP** is an Alzeihmer's diagnosis dataset, while **BRCA** is used for breast invasive carcinoma PAM50 subtype classification. Appendix A.2 provides information about the train-val-test splits.

Evaluation Metrics. The evaluation metric used for N24News and Food-101 is classification accuracy (ACC). For BRCA, we report accuracy (ACC), macro-averaged F1 score (MF1), and weighted F1 score (WF1). For ROSMAP, we use accuracy (ACC), area under the ROC curve (AUC), and F1 score (F1) as the evaluation metrics.

Implementation Details. We use a ViT (pre-trained on the ImageNet-21k dataset) (Dosovitskiy et al., 2020) as the image encoder for N24News and Food-101. For N24News, the text encoder is a pretrained BERT/RoBERTa (Devlin et al., 2018; Zhuang et al., 2021), while we use a pretrained BERT as the text encoder for Food-101. The classifiers for the above two datasets are three layer MLPs with ReLU activations. For ROSMAP and BRCA, which are small datasets, we use two layer MLPs as feature encoders for each modality, and two layer MLPs with ReLU activations as classifiers. The hyperparameter settings and all other details are given in Appendix A.1 .

Baselines. We compare our method with various multimodal classification approaches (Van De Wiel et al., 2016; Wang et al., 2021a; Han et al., 2020; Abavisani et al., 2020; Han et al., 2022; Zou et al., 2023; Liang et al., 2022b; Kiela et al., 2019; 2018; Wang et al., 2022; Vielzeuf et al., 2018; Arevalo et al., 2017; Li et al., 2019; Huang et al., 2020; Kim et al., 2021; Narayana et al., 2019; Liu et al., 2021; Hong et al., 2020; Huang et al., 2021; Singh et al., 2019; Wang et al., 2024). Some methods (Kiela et al., 2019; Vielzeuf et al., 2018; Arevalo et al., 2017) focus on integrating global features from individual modality-specific backbones to enhance classification. Others (Li et al., 2019; Kim et al., 2021; Huang et al., 2020; Narayana et al., 2019) use sophisticated pre-trained architectures fine-tuned for specific tasks. UniS-MMC (Zou et al., 2023), the previous state-of-the-art on Food-101 and N24News, uses contrastive learning to align features across modalities with supervision from unimodal predictions. Similarly, Dynamics (Han et al., 2022), the previous state-of-the-art on ROSMAP and BRCA, applies a dynamic multimodal classification strategy. On Food-101 and N24News, we compare against baseline unimodal networks (ViT and BERT/RoBERTa) and our UniConcat baseline, where pre-trained image and text encoders are fine-tuned independently, and the unimodal representations are simply concatenated for classification. These are typical baselines used in multimodal classification tasks. Detailed baseline descriptions are discussed in Appendix A.7.

4 Results

4.1 Comparison with Baselines

The results are reported as the average and standard deviation over three runs on Food-101/N24News, and five runs on ROSMAP/BRCA. The best score is highlighted in bold, while the second-best score is underlined. The classification accuracy on N24News and Food-101 are displayed in Table 1 and 3 respectively. In the result tables, **ALI** denotes alignment (indicating if the method employs a contrastive component), while **AGG** specifies whether aggregation is early (combining unimodal feature) or late fusion (combining unimodal decisions).

The experimental results from Table 1, 2, 3, reveal the following findings: **(i)** M3CoL consistently outperforms all SOTA methods across all text sources on N24News when using the same encoders, beats SOTA on all evaluation metrics on ROSMAP and BRCA, and also achieves competitive results on Food-101; **(ii)**

contrastive-based methods with any form of alignment demonstrate superior performance compared to other multimodal methods; **(iii)** our proposed M3CoL method, which employs a contrastive-based approach with shared alignment, improves over the traditional contrastive-based models and the latest SOTA multimodal methods. We visualize the unimodal and combined representation distribution of our proposed method using UMAP plots in Figure 7 in Appendix A.5.

Method	Fusion		Backbone		ACC ↑		
	AGG	ALI	Image	Text	Headline	Caption	Abstract
Image-only	-	-	ViT	-	54.1 (*no text source used*)		
Text-only	-	-	-	BERT	72.1	72.7	78.3
UniConcat	Early	✗	ViT	BERT	78.6	76.8	80.8
UniS-MMC	Early	✓	ViT	BERT	80.3	77.5	83.2
M3CoL (Ours)	Early	✓	ViT	BERT	**80.8**$_{\pm 0.05}$	**78.0**$_{\pm 0.03}$	**83.8**$_{\pm 0.06}$
Text-only	-	-	-	RoBERTa	71.8	72.9	79.7
UniConcat	Early	✗	ViT	RoBERTa	78.9	77.9	83.5
N24News	Early	✗	ViT	RoBERTa	79.41	77.45	83.33
UniS-MMC	Early	✓	ViT	RoBERTa	80.3	78.1	84.2
M3CoL (Ours)	Early	✓	ViT	RoBERTa	**80.9**$_{\pm 0.19}$	**79.2**$_{\pm 0.08}$	**84.7**$_{\pm 0.03}$

Table 1: Accuracy (ACC) on N24News on three text sources. AGG denotes early/late modality fusion, ALI indicates presence/absence of alignment. Our method consistently outperforms SOTA across all text sources and backbone combinations. Baseline details are provided in Appendix A.7.

Method	Fusion		ROSMAP			BRCA		
	AGG	ALI	ACC ↑	F1 ↑	AUC ↑	ACC ↑	WF1 ↑	MF1 ↑
GRidge	Early	✗	76.0	76.9	84.1	74.5	72.6	65.6
BPLSDA	Early	✗	74.2	75.5	83.0	64.2	53.4	36.9
BSPLSDA	Early	✗	75.3	76.4	83.8	63.9	52.2	35.1
MOGONET	Late	✗	81.5	82.1	87.4	82.9	82.5	77.4
TMC	Late	✗	82.5	82.3	88.5	84.2	84.4	80.6
CF	Early	✗	78.4	78.8	88.0	81.5	81.5	77.1
GMU	Early	✗	77.6	78.4	86.9	80.0	79.8	74.6
MOSEGCN	Early	✗	83.0	82.7	83.2	86.7	86.8	81.1
DYNAMICS	Early	✗	85.7	86.3	91.1	87.7	88.0	84.5
M3CoL (Ours)	Early	✓	**88.7**$_{\pm 0.94}$	**88.5**$_{\pm 0.94}$	**92.6**$_{\pm 0.59}$	**88.4**$_{\pm 0.57}$	**89.0**$_{\pm 0.42}$	**86.2**$_{\pm 0.54}$

Table 2: Comparison of Accuracy (ACC), Area Under the Curve (AUC), F1 score (F1) on ROSMAP, and Accuracy (ACC), Weighted F1 score (WF1), and Micro F1 score (MF1) on BRCA datasets. AGG denotes early/late modality fusion, ALI indicates presence/absence of alignment. Our method significantly outperforms SOTA across all metrics. Baseline details are provided in Appendix A.7.

4.2 Analysis of Our Method

Effect of Vanilla Mixup. Mixup involves two main components: the random convex combination of raw inputs and the corresponding convex combination of one-hot label encodings. To assess the performance of our M3CoL method in comparison to this Mixup strategy, we conduct experiments on Food-101 and N24News (text source: abstract). We remove the contrastive loss from our framework (Eq. 16) while keeping the rest of the modules unchanged. Table 4 shows that the **Mixup** technique underperforms relative to our proposed M3CoL approach (Testing accuracy curves shown in Figure 6a). The observed accuracy gap can be attributed to excessive noise introduced by label mixing, and the lack of a contrastive approach with

Method	Fusion		Backbone		ACC ↑
	AGG	ALI	Image	Text	
Image-only	-	-	ViT	-	73.1
Text-only	-	-	-	BERT	86.8
UniConcat	Early	✗	ViT	BERT	93.7
MCCE	Early	✗	DenseNet	BERT	91.3
CentralNet	Early	✗	LeNet5	LeNet5	91.5
GMU	Early	✗	RNN	VGG	90.6
ELS-MMC	Early	✗	ResNet-152	BOW features	90.8
MMBT	Early	✗	ResNet-152	BERT	91.7
HUSE	Early	✓	Graph-RISE	BERT	92.3
VisualBERT	✗	✓	FasterRCNN+BERT	BERT	92.3
PixelBERT	Early	✓	ResNet	BERT	92.6
ViLT	Early	✓	ViT	BERT	92.9
CMA-CLIP	Early	✓	ViT	BERT	93.1
ME	Early	✗	DenseNet	BERT	**94.7**
UniS-MMC	Early	✓	ViT	BERT	**94.7**
M3CoL (Ours)	Early	✓	ViT	BERT	$\underline{94.3}_{\pm 0.04}$

Table 3: Accuracy (ACC) comparison on Food-101. AGG denotes early/late modality fusion, ALI indicates presence/absence of alignment. Baseline details are provided in Appendix A.7.

an alignment component. This indicates that the vanilla Mixup strategy introduces additional noise which impairs the model's ability to learn effective representations, while our M3CoL framework benefits from the structured contrastive approach.

Method	ACC ↑			
	ROSMAP	BRCA	Food-101	N24News
Mixup	$84.13_{\pm 0.74}$	$84.52_{\pm 0.46}$	$93.14_{\pm 0.02}$	$81.57_{\pm 0.24}$
M3CoL (No Unimodal Supervision)	$85.14_{\pm 0.85}$	$86.93_{\pm 0.52}$	$94.12_{\pm 0.02}$	$84.26_{\pm 0.11}$
M3CoL (only MultiSClip)	$86.84_{\pm 0.34}$	$87.38_{\pm 0.41}$	$94.23_{\pm 0.01}$	$84.06_{\pm 0.18}$
M3CoL (only M3Co)	$87.42_{\pm 0.63}$	$87.74_{\pm 0.42}$	$94.24_{\pm 0.12}$	$84.57_{\pm 0.08}$
M3CoL (0.33 M3Co + 0.67 MultiSClip)	$\mathbf{88.67}_{\pm 0.94}$	$\mathbf{88.38}_{\pm 0.57}$	$\mathbf{94.27}_{\pm 0.04}$	$\mathbf{84.72}_{\pm 0.03}$

Table 4: Accuracy (ACC) on ROSMAP, BRCA, N24News, and Food-101 datasets under different settings of our method. For N24News, source: abstract and encoder: RoBERTa.

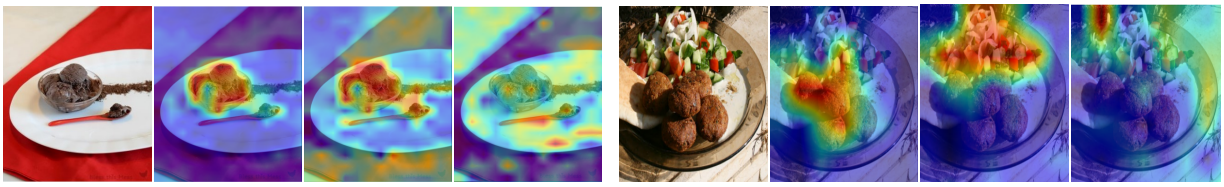
Effect of Loss & Unimodality Supervision. To assess the necessity of each component in the framework, we investigate several design choices: (i) the framework's performance without the supervision of unimodal modules during training, and (ii) the performance differences between using only MultiSClip and only M3Co loss during end-to-end training. The M3CoL (**No Unimodal Supervision**) result indicates that excluding the unimodal prediction module results in a decline in performance as shown in Table 4 and Figure 6a, highlighting its importance as it allows the model to compensate for the weaknesses of one modality with the strengths of another. Additionally, the M3Co loss (**only M3Co**) outperforms the MultiSClip loss (**only MultiSClip**) by learning more robust representations through Mixup-based techniques, which prevent trivial discrimination of positive pairs. Furthermore, using an individual contrastive alignment approach (**only M3Co**) throughout the entire training process without transitioning to the MultiSClip loss results in suboptimal outcomes. This can be attributed to the risk of over-training with Mixup-based loss, which may negatively impact generalization. This demonstrates the necessity of the transition of the contrastive loss during training (**0.33 M3Co + 0.67 MultiSClip**). Figure 6b displays the accuracy plots on the N24News

dataset, for these losses. In addition to the ACC scores presented in Table 4, we also report the performance of other metrics, where available, on the ROSMAP and BRCA datasets under various settings of our method, as shown in Table 5.

Method	ROSMAP		BRCA	
	F1 ↑	AUC ↑	WF1 ↑	MF1 ↑
Mixup	$84.45_{\pm 0.48}$	$88.73_{\pm 0.53}$	$84.66_{\pm 0.48}$	$82.88_{\pm 0.38}$
M3CoL (No Unimodal Supervision)	$86.27_{\pm 0.39}$	$90.20_{\pm 0.82}$	$86.92_{\pm 0.36}$	$85.08_{\pm 0.37}$
M3CoL (only MultiSClip)	$86.76_{\pm 0.58}$	$90.75_{\pm 0.49}$	$87.41_{\pm 0.47}$	$85.48_{\pm 0.35}$
M3CoL (only M3Co)	$87.54_{\pm 0.62}$	$91.41_{\pm 0.54}$	$88.06_{\pm 0.41}$	$85.82_{\pm 0.34}$
M3CoL (0.33 M3Co + 0.67 MultiSClip)	**88.51**$_{\pm 0.94}$	**92.62**$_{\pm 0.59}$	**89.02**$_{\pm 0.42}$	**86.20**$_{\pm 0.54}$

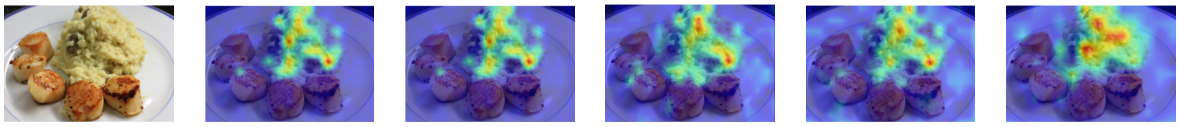
Table 5: Comparison of F1 score (F1), Area Under the Curve (AUC) on ROSMAP, and Weighted F1 score (WF1), Micro F1 score (MF1) on BRCA, under different settings of our method.

Visualization of Attention Heatmaps. The attention heatmaps generated using the embeddings from our trained M3CoL model in Figure 3 and 4 highlight image regions most relevant to the input word. We generate text embeddings for class label words and corresponding image patch embeddings, computing attention scores as their dot product. This visualization aids in understanding the model's focus, decision-making process, and association between class labels and specific image regions. Importantly, it also indicates the correctness of the learned multimodal representations, revealing the model's ability to learn shared relations amongst different modalities, and ground visual concepts to semantically meaningful regions.



(a) Image (b) Ice cream (c) Cream (d) Ice (e) Image (f) Falafel (g) Salad (h) Rice

Figure 3: Text-guided visual grounding with varying input prompts. (a, e) Original images. (b-d) Attention heatmaps for "ice cream" class. (f-h) Heatmaps for "falafel" class. Ice cream example: (b) "Ice cream": Concentrated focus on ice cream, (c) "Cream": Maintained but diffused focus, (d) "Ice": Dispersed attention. Falafel example: (f) "Falafel": Localized focus on falafel, (g) "Salad": Attention shift to salad component, (h) "Rice": Minimal attention (absent in image). Warmer colors indicate higher attention scores.



(a) Risotto (b) Mixup (c) No Unim (d) MultiSClip (e) M3Co (f) M3CoL

Figure 4: Text-guided visual grounding with ablated model variations. (a) Original image. (b-f) Attention heatmaps generated using text embedding (class name: "Risotto") and patch embeddings for different variations of the model. Our proposed M3CoL model (f) demonstrates superior attention localization compared to ablated versions (b-e), corroborating the quantitative results presented in Table 4. Warmer colors indicate higher attention scores. (Here, No Unim: No Unimodal Supervision)

Testing on Random Data and Single-Corrupt Modalities. To showcase the benefits of our framework over traditional contrastive methods, we evaluate the impact of incorporating Mixup-based contrastive loss

(M3Co) during training, highlighting its improvements over standard approaches. It is well-established that deep networks tend to exhibit overconfidence, particularly when making predictions on random or adversarial inputs (Hendrycks & Gimpel, 2016). Previous research has demonstrated that Mixup can mitigate this issue, and our goal is to validate its effectiveness in this context (Thulasidasan et al., 2019). We evaluate

Method	**Modality Corrupted**	
	Image	**Text**
MutliSClip	76.06	46.91
M3CoL	77.24	47.94

Table 6: N24News - Accuracy when tested on data with one corrupt modality.

the confidence scores produced using M3CoL (0.33 M3Co + 0.67 MultiSClip) loss in comparison to only MultiSClip loss when predicting on random noise images and text encoder outputs. Our results show that the model trained with M3CoL exhibits lower confidence in its predictions when both modalities are replaced with random inputs. This demonstrates that incorporating M3CoL enhances the reliability of predictions, especially in the presence of corrupted or random inputs.

To evaluate the robustness of our approach, we conduct experiments where one input modality was corrupted with random noise. Table 6 compares the performance of M3CoL (0.33 M3Co + 0.67 MultiSClip) against only MultiSClip under these conditions. Our M3CoL method demonstrates superior robustness to modality corruption, consistently outperforming MultiSClip. For image corruption, we substituted the original images

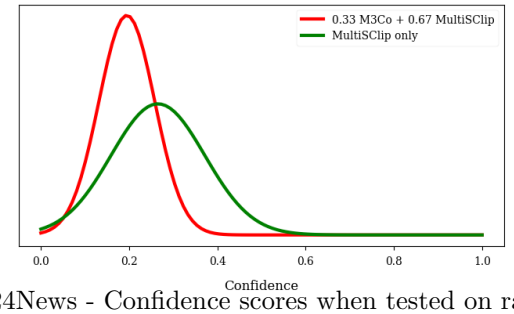


Figure 5: N24News - Confidence scores when tested on random inputs.

with random noise sampled from a Gaussian distribution, parameterized to match the mean and variance of the training set. Similarly, for text corruption, we replaced the original text embeddings with random outputs from the text encoder, again following a Gaussian distribution with statistics matching the training data. For both the above experiments, we use the N24News dataset, with the abstract as the text source and a RoBERTa-based text encoder.

Error Analysis. To evaluate the efficacy of our multimodal approach in integrating and leveraging image and text features, we performed a comprehensive error analysis, comparing it with image-only (ViT) and text-only (RoBERTa) models using the N24News dataset (see Table 9 in Appendix A.4). The analysis reveals that our method excels when both modalities are correctly classified (42.71:0.03 correct-to-incorrect ratio). This demonstrates that our model can learn valuable insights from the fusion of image and text features, which may not be discovered when processing them separately. In cases where only one modality is correctly classified, our model effectively leverages the accurate modality (27.77+8.11=35.88):(1.29+3.25=4.54) correct-to-incorrect ratio. This demonstrates our method's robustness and its ability to outperform unimodal approaches.

5 Related Work

Contrastive Learning. Contrastive learning has driven significant progress in unimodal and multimodal representation learning by distinguishing between similar (positive) and dissimilar (negative) pairs. In multimodal contexts, cross-modal contrastive techniques align representations from different modalities

(Radford et al., 2021; Jia et al., 2021; Kamath et al., 2021), with approaches like CrossCLR (Zolfaghari et al., 2021) and GMC (Poklukar et al., 2022) focusing on global and modality-specific representations. Contrastive learning approaches for paired image-text data, such as CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), and BASIC (Pham et al., 2023), have demonstrated remarkable success across diverse vision-language tasks. Subsequent works have aimed to enhance the efficacy and data efficiency of CLIP training, incorporating self-supervised techniques (SLIP (Mu et al., 2022), DeCLIP (Li et al., 2021)) and fine-grained alignment (FILIP (Li et al., 2023)). The CLIP framework relies on data augmentations to prevent overfitting and the learning of ineffective shortcuts (Chen et al., 2020; Robinson et al., 2021), a common practice in contrastive learning.

Unimodal and Multimodal Data Augmentation. Data augmentation has been integral to the success of deep learning, especially for small training sets. In computer vision, techniques have evolved from basic transformations to advanced methods like Cutout (DeVries & Taylor, 2017), Mixup (Zhang et al., 2017), CutMix (Yun et al., 2019), and automated approaches (Cubuk et al., 2019; 2020). NLP augmentation includes paraphrasing, token replacement (Zhang et al., 2015; Jiao et al., 2019), and noise injection (Yan et al., 2019). Multimodal data augmentation, primarily focused on vision-text tasks, has seen limited exploration, with approaches including back-translation for visual question answering (Tang et al., 2020), text generation from images (Wang et al., 2021b), and external knowledge querying for cross-modal retrieval (Gur et al., 2021). MixGen (Hao et al., 2023) generates new image-text pairs through image interpolation and text concatenation. In contrast, our proposed augmentation technique focusing on the early training phase is fully automatic, applicable to arbitrary modalities, and designed to leverage inherent shared relations in multimodal data.

Relation to Mixup. Mixup (Zhang et al., 2017), a pivotal regularization strategy, enhances model robustness and generalization by generating samples through convex combinations of existing data points. Originally introduced for computer vision, it has been adapted to NLP by applying the technique to text embeddings (Guo et al., 2019). Our proposed augmentation differs from Mixup in several key aspects: it is designed for multimodal data, takes inputs from different modalities, and does not rely on one-hot label encodings. By extending the Mixup paradigm to complex, multimodal scenarios and focusing on the early training phase, our method broadens its applicability while leveraging inherent shared relations in multimodal data.

6 Discussion and Limitations

Discussion and Conlusions. Aligning representations across modalities presents significant challenges due to the complex, often non-bijective relationships in real-world multimodal data (Liang et al., 2022a). These relationships can involve many-to-many mappings or even lack clear associations, as exemplified by linguistic ambiguities and synonymy in vision-language tasks. We propose M3Co, a novel contrastive-based alignment method that captures shared relations beyond explicit pairwise associations by aligning mixed samples from one modality with corresponding samples from others. Our approach incorporates Mixup-based contrastive learning, introducing controlled noise that mirrors the inherent variability in multimodal data, thus enhancing robustness and generalizability. The M3Co loss, combined with an architecture leveraging unimodal and fusion modules, enables continuous updating of representations necessary for accurate predictions and deeper integration of modalities. This method generalizes across diverse domains, including image-text, high-dimensional multi-omics, and data with more than two modalities. Experiments on four public multimodal classification datasets demonstrate the effectiveness of our approach in learning robust representations that surpass traditional multimodal alignment techniques.

Limitations and Future Work. M3CoL demonstrates promising results, yet faces optimization challenges due to the inherent limitations of multimodal frameworks, particularly extended training times on large-scale datasets like Food-101. The method's modality-agnostic nature and effective use of mixup augmentation suggest its potential adaptability to various multimodal tasks, especially where data augmentation and learning real-world inter-modal relationships are crucial. Future work should focus on investigating domain adaptation strategies, validating M3CoL's utility on downstream tasks such as visual question answering and information retrieval, and enhancing model interpretability through explainable AI techniques. These advancements, coupled with comprehensive hyperparameter tuning, will likely broaden M3CoL's impact in multimodal research.

References

Mahdi Abavisani, Liwei Wu, Shengli Hu, Joel Tetreault, and Alejandro Jaimes. Multimodal categorization of crisis events in social media. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14679–14689, 2020.

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

Olivier Chapelle, Jason Weston, Léon Bottou, and Vladimir Vapnik. Vicinal risk minimization. *Advances in neural information processing systems*, 13, 2000.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 113–123, 2019.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 702–703, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, Rongrong Ji, and Chunhua Shen. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *Advances in neural information processing systems*, 35:35959–35970, 2022.

Yuting Gao, Jinfeng Liu, Zihan Xu, Tong Wu, Enwei Zhang, Ke Li, Jie Yang, Wei Liu, and Xing Sun. Softclip: Softer cross-modal alignment makes clip stronger. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1860–1868, 2024.

Aditya Sharad Golatkar, Alessandro Achille, and Stefano Soatto. Time matters in regularizing deep networks: Weight decay and data augmentation affect early learning dynamics, matter little near convergence. *Advances in Neural Information Processing Systems*, 32, 2019.

Hongyu Guo, Yongyi Mao, and Richong Zhang. Augmenting data with mixup for sentence classification: An empirical study. *arXiv preprint arXiv:1905.08941*, 2019.

Shir Gur, Natalia Neverova, Chris Stauffer, Ser-Nam Lim, Douwe Kiela, and Austin Reiter. Cross-modal retrieval augmentation for multi-modal classification. *arXiv preprint arXiv:2104.08108*, 2021.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. In *International Conference on Learning Representations*, 2020.

Zongbo Han, Fan Yang, Junzhou Huang, Changqing Zhang, and Jianhua Yao. Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20707–20717, 2022.

Xiaoshuai Hao, Yi Zhu, Srikar Appalaraju, Aston Zhang, Wanqian Zhang, Bo Li, and Mu Li. Mixgen: A new multi-modal data augmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 379–389, 2023.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019.

Danfeng Hong, Lianru Gao, Naoto Yokoya, Jing Yao, Jocelyn Chanussot, Qian Du, and Bing Zhang. More diverse means better: Multimodal deep learning meets remote-sensing imagery classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5):4340–4354, 2020.

Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34:10944–10956, 2021.

Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*, 2019.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1780–1790, 2021.

Douwe Kiela, Edouard Grave, Armand Joulin, and Tomas Mikolov. Efficient large-scale multi-modal classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*, 2019.

Sungnyun Kim, Gihun Lee, Sangmin Bae, and Se-Young Yun. Mixco: Mix-up contrastive learning for visual representation. *arXiv preprint arXiv:2010.06300*, 2020.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pp. 5583–5594. PMLR, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021.

Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23390–23400, 2023.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and trends in multimodal machine learning: Principles, challenges, and open questions. *arXiv preprint arXiv:2209.03430*, 2022a.

Tao Liang, Guosheng Lin, Mingyang Wan, Tianrui Li, Guojun Ma, and Fengmao Lv. Expanding large pre-trained unimodal models with multimodal information injection for image-text multimodal classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15492–15501, 2022b.

Huidong Liu, Shaoyuan Xu, Jinmiao Fu, Yang Liu, Ning Xie, Chien-Chih Wang, Bryan Wang, and Yi Sun. Cma-clip: Cross-modality attention clip for image-text classification. *arXiv preprint arXiv:2112.03562*, 2021.

Zixuan Liu, Ziqiao Wang, Hongyu Guo, and Yongyi Mao. Over-training with mixup may hurt generalization. *arXiv preprint arXiv:2303.01475*, 2023.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European conference on computer vision*, pp. 529–544. Springer, 2022.

Pradyumna Narayana, Aniket Pednekar, Abishek Krishnamoorthy, Kazoo Sone, and Sugato Basu. Huse: Hierarchical universal semantic embeddings. *arXiv preprint arXiv:1911.05978*, 2019.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, et al. Combined scaling for zero-shot transfer learning. *Neurocomputing*, 555:126658, 2023.

Petra Poklukar, Miguel Vasco, Hang Yin, Francisco S Melo, Ana Paiva, and Danica Kragic. Geometric multimodal contrastive representation learning. In *International Conference on Machine Learning*, pp. 17782–17800. PMLR, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Joshua Robinson, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, and Suvrit Sra. Can contrastive learning avoid shortcut solutions? *Advances in neural information processing systems*, 34:4974–4986, 2021.

Zhiqiang Shen, Zechun Liu, Zhuang Liu, Marios Savvides, Trevor Darrell, and Eric Xing. Un-mix: Rethinking image mixtures for unsupervised visual representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2216–2224, 2022.

Amrit Singh, Casey P Shannon, Benoît Gautier, Florian Rohart, Michaël Vacher, Scott J Tebbutt, and Kim-Anh Lê Cao. Diablo: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics*, 35(17):3055–3062, 2019.

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016.

Ruixue Tang, Chao Ma, Wei Emma Zhang, Qi Wu, and Xiaokang Yang. Semantic equivalent adversarial data augmentation for visual question answering. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, pp. 437–453. Springer, 2020.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in neural information processing systems*, 32, 2019.

Mark A Van De Wiel, Tonje G Lien, Wina Verlaat, Wessel N van Wieringen, and Saskia M Wilting. Better prediction by use of co-data: adaptive group-regularized ridge regression. *Statistics in medicine*, 35(3): 368–381, 2016.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International conference on machine learning*, pp. 6438–6447. PMLR, 2019.

Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. Centralnet: a multilayer approach for multimodal fusion. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 0–0, 2018.

Jiahui Wang, Nanqing Liao, Xiaofei Du, Qingfeng Chen, and Bizhong Wei. A semi-supervised approach for the integration of multi-omics data based on transformer multi-head self-attention mechanism and graph convolutional networks. *BMC genomics*, 25(1):86, 2024.

Tongxin Wang, Wei Shao, Zhi Huang, Haixu Tang, Jie Zhang, Zhengming Ding, and Kun Huang. Mogonet integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nature communications*, 12(1):3445, 2021a.

Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frederic Precioso. Recipe recognition with large multimodal food dataset. In *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 1–6. IEEE, 2015.

Zhen Wang, Xu Shan, Xiangxie Zhang, and Jie Yang. N24news: A new dataset for multimodal news classification. In *Proceedings of the Language Resources and Evaluation Conference*, pp. 6768–6775, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.729.

Zixu Wang, Yishu Miao, and Lucia Specia. Cross-modal generative augmentation for visual question answering. *arXiv preprint arXiv:2105.04780*, 2021b.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018.

Ge Yan, Yu Li, Shu Zhang, and Zhenyu Chen. Data augmentation for deep learning of judgment documents. In *Intelligence Science and Big Data Engineering. Big Data and Machine Learning: 9th International Conference, IScIDE 2019, Nanjing, China, October 17–20, 2019, Proceedings, Part II 9*, pp. 232–242. Springer, 2019.

Hao Yu, Huanyu Wang, and Jianxin Wu. Mixup without hesitation. In *Image and Graphics: 11th International Conference, ICIG 2021, Haikou, China, August 6–8, 2021, Proceedings, Part II 11*, pp. 143–154. Springer, 2021.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6023–6032, 2019.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pp. 2–25. PMLR, 2022.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. A robustly optimized bert pre-training approach with post-training. In *Proceedings of the 20th chinese national conference on computational linguistics*, pp. 1218–1227, 2021.

Mohammadreza Zolfaghari, Yi Zhu, Peter Gehler, and Thomas Brox. Crossclr: Cross-modal contrastive learning for multi-modal video representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1450–1459, 2021.

Heqing Zou, Meng Shen, Chen Chen, Yuchen Hu, Deepu Rajan, and Eng Siong Chng. Unis-mmc: Multimodal classification via unimodality-supervised multimodal contrastive learning. *arXiv preprint arXiv:2305.09299*, 2023.

A Appendix

A.1 Experimental Details

The models were trained on either an NVIDIA RTX A6000 or an NVIDIA A100-SXM4-80GB GPU. The results are reported as the average and standard deviation over three runs on Food-101 and N24News, and five runs on ROSMAP and BRCA. We use a grid search on the validation set to search for optimal hyperparameters. The temperature parameter for the M3Co and MultiSClip losses is set to 0.1. The corresponding loss coefficient β is 0.1 to keep the loss value in the same range as the other losses. We use the Adam optimizer (Kingma & Ba, 2014) for all datasets. For Food-101 and N24News, the learning rate scheduler is ReduceLROnPlateau with validation accuracy as the monitored metric, lr factor of 0.2, and lr patience of 2. For ROSMAP and BRCA, we use the StepLR scheduler with a step size of 250. For Food-101 and N24News, the maximum token length of the text input for the BERT/RoBERTa encoders is 512. Other hyperparameter details are provided in Table 7.

Hyperparameter	**N24News**	**Food-101**	**ROSMAP**	**BRCA**
Embedding dimension	768	768	1000	768
Classifier dimension	256	256	1000	768
Learning rate	10^{-4}	10^{-4}	$5 \cdot 10^{-3}$	$5 \cdot 10^{-3}$
Weight decay	10^{-4}	10^{-4}	10^{-3}	10^{-3}
Batch size	32	32	-	-
Batch gradient	128	128	-	-
Dropout (classifier)	0	0	0.5	0.5
Epochs	50	50	500	500

Table 7: Experimental hyperparameter values for our proposed model across all the four datasets.

A.2 Dataset Information and Splits

The datasets used in our experiments can be downloaded from the following sources: Food-101 from https://visiir.isir.upmc.fr, N24News from https://github.com/billywzh717/N24News, and BRCA and ROSMAP from https://github.com/txWang/MOGONET.

To ensure a fair comparison with previous works, we adopt the default split method detailed in Table 8. As the Food-101 dataset does not include a validation set, we partition 5,000 samples from the training set to create one, which is conistent with other baselines.

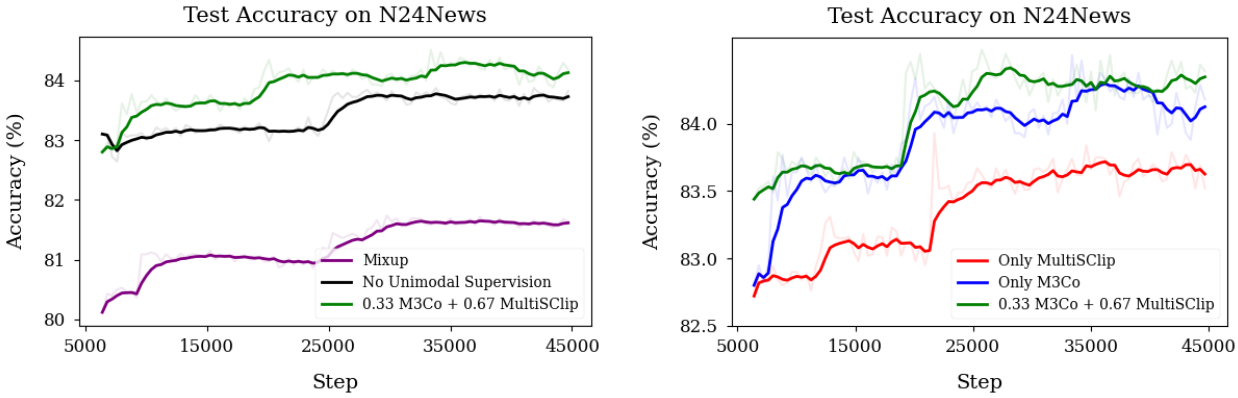
Dataset	**Modalities**	**Modality Types**	**Train**	**Validation**	**Test**	**Classes**
Food-101	2	Image, text	60101	5000	21695	101
N24News	2	Image, text	48988	6123	6124	24
ROSMAP	3	mRNA, miRNA, DNA	245	-	106	2
BRCA	3	mRNA, miRNA, DNA	612	-	263	5

Table 8: Statistics for the four datasets: Food-101, N24News, ROSMAP, and BRCA. Note: miRNA stands for microRNA, and mRNA stands for messenger RNA.

A.3 Ablation Studies on the N24News Dataset

The accuracy plots for the N24News dataset (text source: abstract, text encoder: RoBERTa) are used to compare our method and its variants. Our proposed M3CoL approach outperforms the Mixup technique, as

shown in Figure 6a. Ablating the unimodal supervision in M3CoL leads to a performance decline, indicating the importance of the unimodal prediction module, as shown in Figure 6a. Furthermore, the M3Co loss achieves better results than the MultiSClip loss. Training solely with either the M3Co loss or the MultiSClip loss alignment approach yields suboptimal performance when compared to their strategic combination, as shown in Figure 6b. The quantitative results are given in Table 4.



(a) Comparison of M3CoL and its variants using Mixup and No Unimodal Supervision.

(b) Comparison of M3CoL and its variants using only M3Co and only MultiSoftClip loss.

Figure 6: Test accuracy plots showing comparison of M3CoL and its variants on the N24News dataset (text source: abstract, text encoder: RoBERTa).

A.4 Error Analysis

We provide an in-depth error analysis on the N24News dataset in Section 4.2. As shown in Table 9, our method excels not only when both modalities are correctly classified but also demonstrates strong performance even when both are misclassified, indicating effective feature fusion. Additionally, it successfully leverages information when only one modality is classified correctly. This analysis demonstrates our method's robustness and highlights its superiority over unimodal approaches.

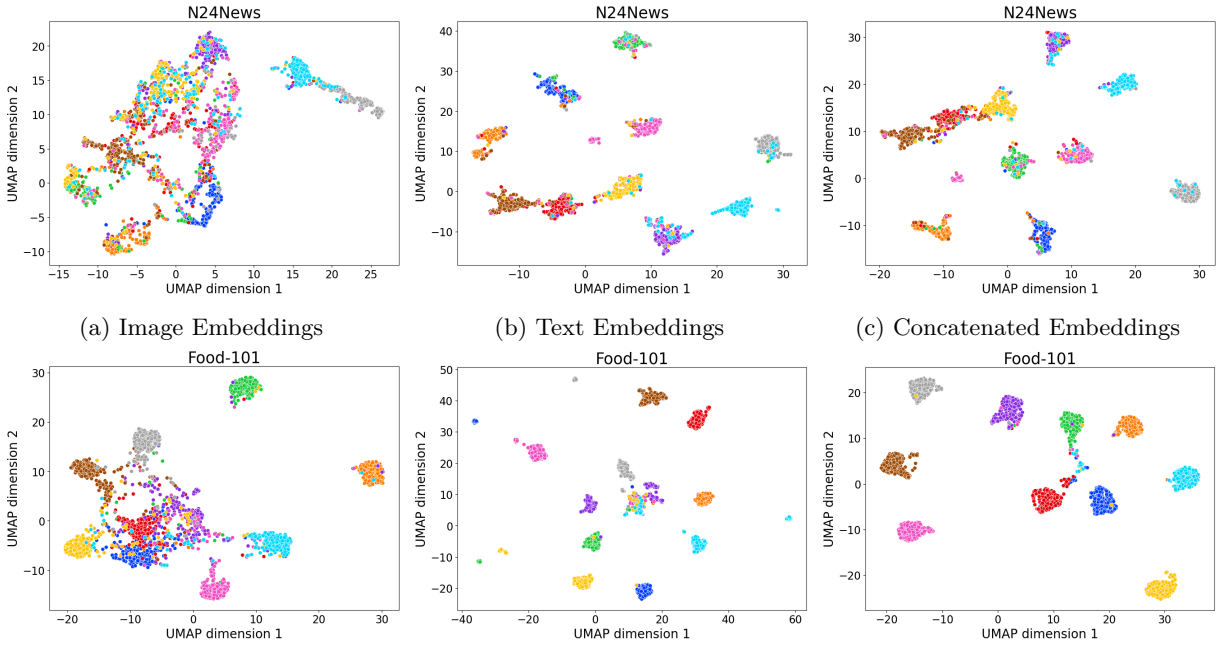
Unimodal Prediction		**Multimodal Prediction %**	
Text	**Image**	**Correct**	**Incorrect**
True	True	42.71	0.03
True	False	27.77	1.29
False	True	8.11	3.25
False	False	2.31	14.53

Table 9: Error analysis on N24News with text encoder RoBERTa and text source "headline". True and False denote the correctness of the unimodal predictions. Multimodal Prediction % shows the resulting test set ratio of the final predictions.

A.5 UMAP Plots

We generate UMAP plots on embeddings derived from the N24News and Food-101 datasets to visualize the clustering performance of our M3CoL model. For each dataset, we randomly select 10 classes and generate the corresponding embeddings from the image encoder, text encoder, and their concatenated multimodal representatio, using our trained M3CoL model. Figure 7 shows that the image embeddings depict less distinct clusters, indicating less effective inter-cluster separation compared to the text encoder embeddings. The concatenated embeddings, however, result in the best-defined clusters, suggesting that the final multimodal

representations preserve and potentially enhance class-distinguishing features. These observations align with our quantitative results presented in Table 1 and 3.



(a) Image Embeddings (b) Text Embeddings (c) Concatenated Embeddings

(d) Image Embeddings (e) Text Embeddings (f) Concatenated Embeddings

Figure 7: UMAP plots of embeddings from the N24News (source: abstract and encoder: RoBERTa) and Food-101 datasets. We generate UMAP plots for the representations generated by the image encoder, text encoder, and their concatenated multimodal representations, using our trained M3CoL model. Concatenated embeddings exhibit superior clustering, while text embeddings outperform image embeddings. Consistent patterns across datasets demonstrate M3CoL's effectiveness in fusing multimodal information and enhancing semantic representations.

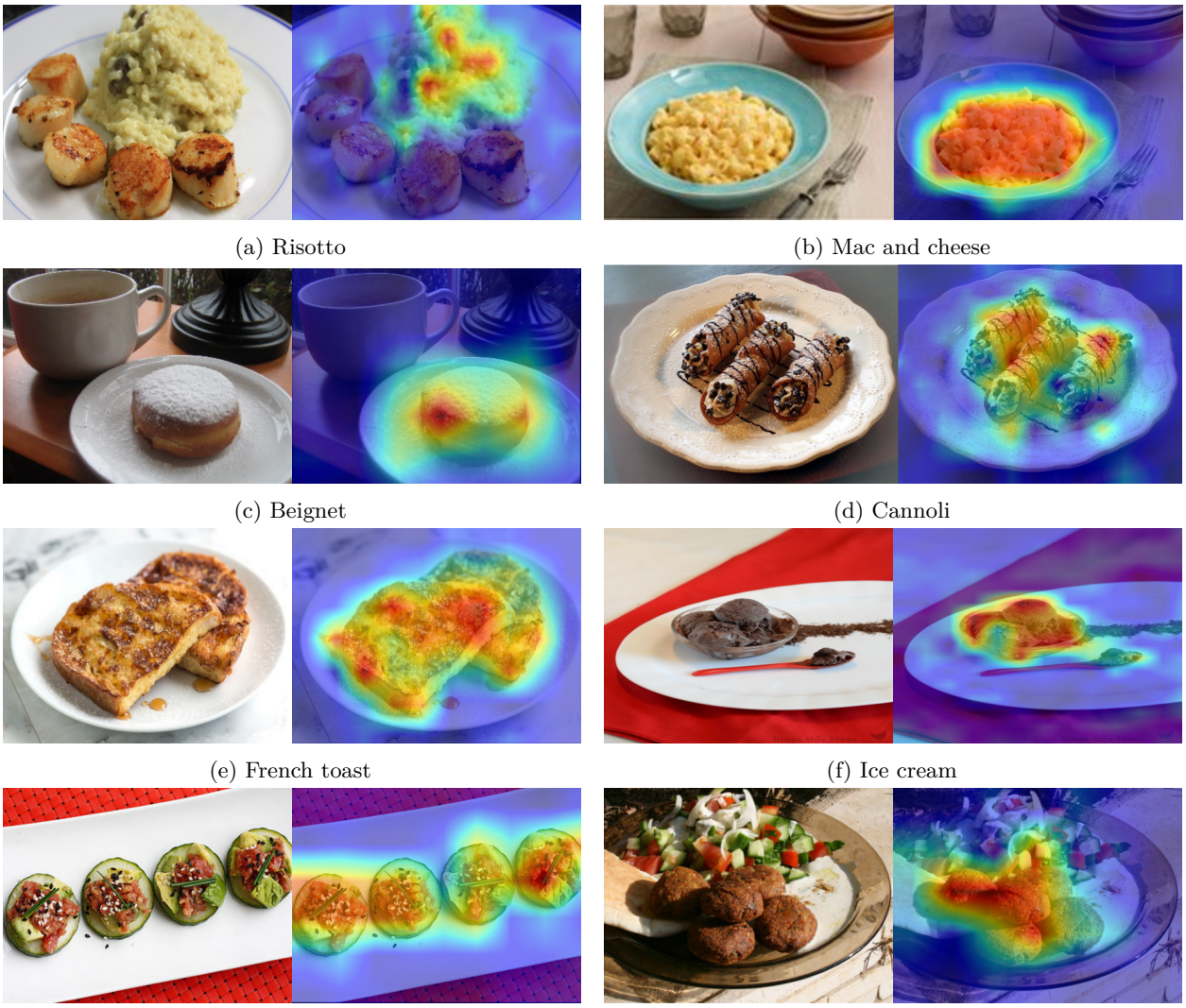
A.6 Additional Visualization Attention Heatmaps

Following Section 4.2, we generate heatmaps using class embeddings and patch embeddings for some more examples in Food-101. These are depicted in Figure 8.

A.7 Baseline Details

The baselines used in our comparsions are described in details as follows:

- **GRidge** (Van De Wiel et al., 2016) dynamically incorporates multimodal data to adjust regularization penalties, improving predictive accuracy in genomic classification scenarios.
- **BPLSDA** (Block partial least squares discriminant analysis) (Singh et al., 2019) analyzes multimodal data in latent space and **BSPLSDA** (Block sparse partial least squares discriminant analysis) (Singh et al., 2019) adds sparsity constraints to BPLSDA to extract relevant features.
- **MOGONET** (Wang et al., 2021a) integrates GCNs with a View Correlation Discovery Network (VCDN) to process multi-omics data. The initial predictions from each omics-specific GCN are consolidated using the VCDN, which identifies and leverages cross-omics label correlations to improve prediction accuracy.
- **CF** (Concatenation of Final Multimodal Representations) (Hong et al., 2020; Huang et al., 2021) creates representations by combining late stage representations of multiple modalities.



(a) Risotto (b) Mac and cheese (c) Beignet (d) Cannoli (e) French toast (f) Ice cream (g) Tuna tartare (h) Falafel

Figure 8: Attention heatmaps demonstrating text-guided visual grounding for samples in the Food-101 dataset. Warmer colors indicate higher attention scores.

- **TMC** (Han et al., 2020) enhances decision-making by dynamically integrating multiple views based on confidence levels, for robust and reliable fusion.

- **MCCE** (Abavisani et al., 2020) uses DenseNet and BERT for feature extraction then applying stochastic transitions between multi-modal embeddings during training to enhance generalization and to handle sparse data effectively.

- **Dynamics** (Han et al., 2022) presents an approach for trustworthy multi-modal classification, specifically designed for high-stakes environments like medical diagnosis. The model dynamically assesses both feature-level and modality-level informativeness, using a sparse gating mechanism to filter and integrate the most relevant features and modalities per sample.

- **UniS-MMC** (Zou et al., 2023) uses a contrastive learning approach that relies on making unimodal predictions, evaluating the agreement or discrepancy between these predictions and the ground truth, and using this insight to align feature vectors across various modalities through a contrastive loss.

- **ME** (Liang et al., 2022b) leverages cross-modal information by transforming features between modalities. It achieves this by integrating a Multimodal Information Injection Plug-in (MI2P) with pre-trained models, enabling them to process image-text pairs without structural modifications.
- **MMBT** (Kiela et al., 2019) leverages the strengths of pre-trained text and image encoders, effectively combining them within a BERT-like framework by mapping image embeddings into the textual token space.
- **ELS-MMC** (Kiela et al., 2018) investigates various multimodal fusion techniques for integrating discrete (text) and continuous (visual) modalities to enhance classification tasks in a resource-efficient manner.
- **N24News** (Wang et al., 2022) presents a novel dataset from The New York Times with text and image data across 24 categories. It utilizes a multitask multimodal strategy, employing ViT for image processing and RoBERTa for text analysis, with features concatenated for final classification.
- **CentralNet** (Vielzeuf et al., 2018) uses separate convolutional networks for each modality, linked via a central network that generates a unified feature representation and also applies multi-task learning to refine and regulate these features.
- **GMU** (Arevalo et al., 2017) employs multiplicative gates that dynamically adjust the influence of each modality on its activation thereby deriving a sophisticated intermediate representation tailored for specific applications.
- **VisualBERT** (Li et al., 2019) employs a series of transformer layers to align textual elements and corresponding image regions through self-attention mechanisms.
- **PixelBert** (Huang et al., 2020) directly aligns image pixels with textual descriptions using a deep multi-modal transformer and establishes a direct semantic connection at the pixel and text level.
- **ViLT** (Kim et al., 2021) implements a BERT-like transformer model that processes visual data in a convolution-free manner, similar to textual data, thereby simplifying input feature extraction and reducing computational demands.
- **HUSE** (Narayana et al., 2019) constructs a shared latent space that aligns image and text embeddings based on their semantic similarity, enhancing cross-modal representation.
- **CMA-CLIP** (Liu et al., 2021) enhances CLIP (Radford et al., 2021) by integrating two cross-modality attention mechanisms: sequence-wise and modality-wise attention. These attention modules refine the relationships between image patches and text tokens, allowing the model to focus on relevant modalities for specific tasks.
- **MOSEGCN** (Wang et al., 2024) utilizes transformer multi-head self-attention and Similarity Network Fusion (SNF) to learn correlations within and among different omics. This information is then fed into a self-ensembling Graph Convolutional Network (SEGCN) for semi-supervised training and testing.

A.8 Use of Generative AI Models

In this work, we use the following generative AI model:

- Gemini 1.0 Pro (Team et al., 2023) to generate food item images and captions as displayed in Figure 1, which serve as sample representations from the Food-101 dataset.