Neurocomputing 647 (2025) 130502

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Borrowing treasures from neighbors: In-context learning for multimodal learning with missing modalities and data scarcity

Zhuo Zhi [a],*, Ziquan Liu [b], Moe Elbadawi [c], Adam Daneshmend [d], Mine Orlu [e], Abdul Basit [e], Andreas Demosthenous [a], Miguel Rodrigues [a]

[a] Department of Electronic and Electrical Engineering, UCL, London, UK
[b] School of Electronic Engineering and Computer Science, QMUL, London, UK
[c] School of Biological and Behavioural Sciences, QMUL, London, UK
[d] University College London Hospitals NHS Foundation Trust, London, UK
[e] UCL School of Pharmacy, UCL, London, UK

ARTICLE INFO

Communicated by S. Liu

Keywords:
Multimodal learning
Missing modalities
Data scarcity
In-context learning

ABSTRACT

Multimodal machine learning with missing modalities is an increasingly relevant challenge arising in various applications such as healthcare. This paper extends the current research into missing modalities to the low-data regime, i.e., a downstream task has both missing modalities and limited sample size issues. This problem setting is particularly challenging and also practical as it is often expensive to get full-modality data and sufficient annotated training samples. We propose to use retrieval-augmented in-context learning to address these two crucial issues by unleashing the potential of a transformer's in-context learning ability. Diverging from existing methods, which primarily belong to the parametric paradigm and often require sufficient training samples, our work exploits the value of the available full-modality data, offering a novel perspective on resolving the challenge. The proposed data-dependent framework exhibits a higher degree of sample efficiency and is empirically demonstrated to enhance the classification model's performance on both full- and missing-modality data in the low-data regime across various multimodal learning tasks. When only 1% of the training data are available, our proposed ICL-CA method outperforms the best baseline by 5.9%, 5.9%, 5.3% and 10.8% on four datasets across various missing states. Notably, our method also reduces the performance gap between full-modality and missing-modality data compared with the baseline. Code is available[1].

1. Introduction

Humankind leverages multimodal data to make intelligent decisions, such as vision, language and sound [1]. Consequently, multimodal machine learning (ML) has emerged as a pivotal learning paradigm in the ML research community, aiming to improve the quality of decision-making by using multimodal data in various fields, e.g., ML-assisted healthcare [2–4] and malicious content detection [5]. However, a major challenge in the application of multimodal ML is the missing-modality issue [6–9], where some data samples do not have complete modalities due to challenges in the data collection process. For instance, in medical applications, some modalities, such as X-ray images [10], are more expensive and/or time-consuming to obtain than others, e.g., Electronic Health Records (EHRs) [11]. Therefore, a multimodality dataset with the missing-modality issue contains samples with complete modalities, i.e., *full-modality data*, and also samples with incomplete modalities, i.e., *missing-modality data*.

Existing research on tackling the missing-modality challenge has two pathways. Before the advent of multimodal transformer [12,13], multimodal learning relies on explicit information fusion with features, the output of modality-dependent backbones. Thus, some research work [14] proposes to learn a parametric model to infer missing modalities. In the era of multimodal transformers, the modality fusion starts from the input layer as a single transformer can handle various input formats, such as vision, language and sound [15]. A recent work [16] proposes the missing aware prompt (MAP) to learn the maximum likelihood estimation for missing modalities at the token level. However, there are two limitations in existing research. Firstly, it is frequently assumed that the sample size during training is adequate so a parametric model can be learned to estimate the missing modalities [14,16], but the sample size is not always sufficient in the real-world [17]. Secondly, there is a notable absence of analysis

* Corresponding author.
E-mail address: zhuo.zhi.21@ucl.ac.uk (Z. Zhi).
[1] GitHubrepository

https://doi.org/10.1016/j.neucom.2025.130502
Received 30 October 2024; Received in revised form 18 March 2025; Accepted 10 May 2025
Available online 29 May 2025

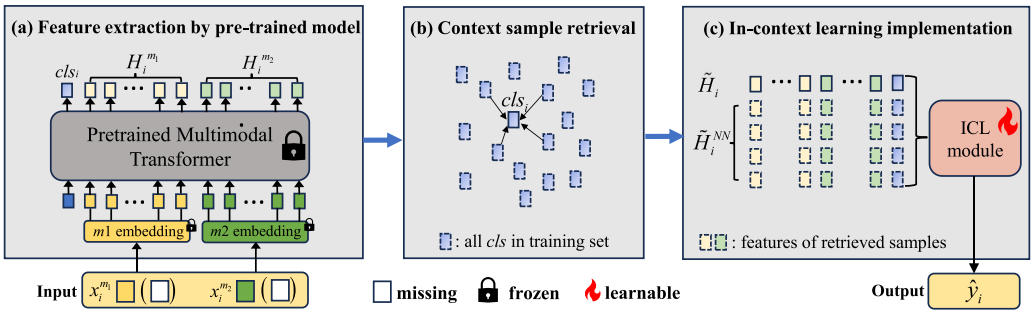


Fig. 1. The overview of the proposed method. (a) Assuming that each sample contains data with 2 modalities $x_i^{m_1}$ and $x_i^{m_2}$, we get the feature $H_i = (H_i^{m_1}, H_i^{m_2}, cls_i)$ of the sample by using a pre-trained multimodal transformer, note that $x_i^{m_1}$ or $x_i^{m_2}$ may be missed. (b) We use the cls token to calculate the cosine similarity between the current sample and all full-modality training samples, and then retrieve the most similar Q samples. (c) We input the pooled feature of the current sample $\tilde{H}_i$ and neighbor samples $\tilde{H}_i^{NN}$ into the ICL module to predict the label $\hat{y}_i$. Note that only the ICL module requires to be trained and the others are frozen. The retrieval-augmented operation is the same for both the training and inference processes. Note that the words 'missing modality' and 'incomplete modality', 'full modality' and 'complete modality' are used interchangeably.

concerning the performance disparity between missing- and full-modality data on multimodal learning.

To further our understanding of the missing-modality challenge in the low-data regime, our paper analyzes the performance of missing- and full-modality data separately in various tasks and training sample sizes. There are two major observations and hypotheses in this paper: (1) The performance of existing methods drops significantly in the low-data regime, and the potential of limited data needs to be more fully exploited. (2) Different tasks depend on full- and missing modality data to different degrees. For low-complexity tasks, the model mainly learns from missing-modality data [18,19], leading to higher performance on missing-modality data compared with full-modality ones. In contrast, the model performs worse on missing-modality data compared with full-modality ones in high-complexity tasks. Therefore, we should not only focus on reconstructing/improving missing-modality data. See Fig. 2 for the empirical evidence. Motivated by our empirical observation, we propose a data-dependent approach based on *retrieval-augmented in-context learning* (ICL) [20,21], to reduce the performance drop of multimodal learning with missing modality in the low-data regime. The proposed method exploits the value of available data and adaptively enhances both missing- and full-modality samples by using the neighboring full-modality samples, as Fig. 1 shows. The core logic of our method involves three key steps: First, we extract multimodal features from input samples using a pre-trained multimodal transformer, such as ViLT [13] accommodating cases where modalities may be missing. Second, we retrieve contextually similar full-modality samples from the training set based on cosine similarity computed using the classification head embeddings. Lastly, we employ an in-context learning module that leverages cross-attention or next-token prediction mechanisms to integrate retrieved contextual features with the current sample's features. This enables the model to implicitly infer missing modality information or refine existing modalities for improved prediction accuracy. Consequently, the ICL module demonstrates improved performance on both missing- and full-modality data across hard and easy tasks, while concurrently diminishing the performance disparity between the two data types. Our experiments validate the effectiveness of the proposed ICL method on various datasets with extensive experiments, see Fig. 2 and Section 4.2. Our main contributions are three-fold:

1. We investigate the data scarcity issue in missing-modality tasks and unveil the drawback of the existing parametric approach in the low-data regime, as its effectiveness often relies on a sufficient sample size. Our empirical study also reveals that the model should adaptively focus on two types of data as dependence on missing-modality data is not necessarily worse than that of full-modality ones.
2. We propose a novel data-dependent in-context learning method to improve the sample efficiency and benefit the learning of both missing- and full-modality data, where the nearest neighbor information of full-modality data is exploited. To the best of our knowledge, our work is among the first to use in-context learning to address the challenge of missing modality in the low-data regime.
3. Our experiment demonstrates the effectiveness of the proposed ICL method on four datasets, including both medical and vision-language multimodal learning tasks. The performance gain on four datasets over the best baseline in the low-data regime are 5.9%, 5.9%, 5.3% and 10.8%.

The paper is organized as follows. Section 2 gives an overview of related research. Section 3 first introduces our empirical observation about the performance gap and learning process difference between missing- and full-modality data, and then elaborates on the proposed method. Section 4 shows the experiment results and the analysis. Finally, Section 5 summarizes this paper, its limitations and future directions.

2. Related work

Missing Modalities in Multimodal Learning. Multimodal models usually assume that the input samples have complete modalities. However, the problem of missing modalities exists in various applications. In ML for healthcare, combining EHR and X-ray as input excels in mortality prediction and phenotype classification task [23], but some patients do not have the time or financial support for X-rays. In ML with vision-language data, a model may not receive image input from the user as a result of network/format issues [16]. The model fails to perform as expected in these situations [9,25]. Consequently, much work has been devoted to improving the robustness of multimodal models under modal absence. [26] optimizes a joint generative-discriminative objective for multimodal data and labels which contain the information required for generating data for missed modality. In [14], multimodal learning with severely missing modalities (SMIL) is designed to reconstruct the missing modalities using modality priors and Bayesian Meta-Learning. [16] introduced two types of missing-aware prompts that can be seamlessly integrated into multimodal transformers. [9] proposes a unified strategy based on multi-task optimization to deal with missing modalities in the transformer-based multimodal model. [2] employs an implicit approach based on Wasserstein distance to achieve the optimal alignment of different modalities which improves the robustness to input noise and missing modalities of the multimodal Transformer. In [27], the proposed modal complementary recovering paradigm strategically integrates both complementary graph-based recovery and topological low-rank adaptation mechanisms to enhance the effectiveness and reliability of incomplete multimodal learning. In much of the existing work, the parametric approach is adopted, which learns a model to handle samples with missing modalities and only uses that model to infer the missed modalities during the

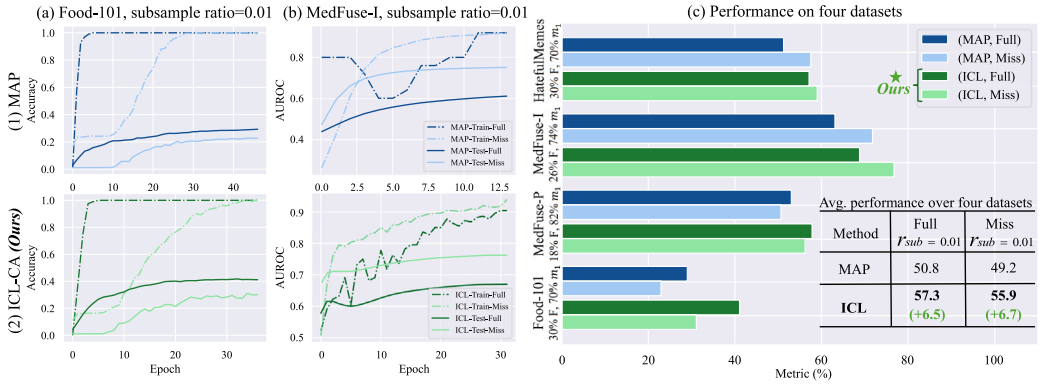


Fig. 2. (a) The learning curve of ICL-CA (ours) and Missing-Aware Prompt (MAP) [16] on the Food-101 dataset in the low data regime. (b) The learning curve of two methods on the MedFuse-I dataset. The subsampling ratio is set to be 0.01. The difference in the learning steps is due to early stopping. F means full-modality, m_1 means text/EHR and m_2 means image/X-ray. During each training process, we calculate the metric of missing- and full-modality samples *separately* and refer to them as '-full' and '-miss'. (c) The performance of the best baseline MAP and our ICL-CA on four multimodality datasets with the missing-modality issue. The y-axis shows the dataset name and missing status. The x-axis is metrics for each dataset, AUROC for MedFuse-I, MedFuse-P and HatefulMemes, and accuracy for Food-101. On each dataset, we compute the metric for test data with full and missing modalities *separately* and show the results in **dark** and light color. The legend means *(Method, Full/Missed-Modality)*. When the task complexity is low, e.g., binary classification tasks like HatefulMemes [22] and MedFuse-I [23], the performance of full-modality data lags behind that of missing-modality data, as fitting the training data does not need full-modality information. When the task complexity is high, e.g., a multi-classification task like Food-101 (101 classes) [24], the full-modality performance surpasses that of the missing-modality, as the task requires all modalities to adequately model the training data. Our ICL is significantly better than MAP on four datasets in four cases as the table below shows. See more details in Section 4.

test stage. Its drawback is that learning such a model requires sufficient training data so the parametric approach cannot perform well when the sample size is low, see Section 4.2. In contrast, our paper uses the semi-parametric approach to deal with the missing-modality issues. In our approach, we enhance the current data by retrieving similar samples during both the training and inference phases. This strategy emphasizes the quality of the samples, thereby reducing reliance on sample size.

Data scarcity in transfer Learning. Multimodal learning based on pre-training and fine-tuning has become popular [28–32]. The performance of the pre-trained model on the target task is highly dependent on the size of the fine-tuning dataset, which is particularly problematic in certain scenarios. For example, scarce positive samples are recorded for some rare diseases [33], or, a small amount of text data is available for some low-resource language tasks [34]. Many existing works focus on facilitating transfer learning under low-data situations. [35] proposes to select features from all layers of the source model to train a classification head for the target domain, which matches performance obtained with fine-tuning on average while reducing training and storage costs. Similarly, [36] introduces the algorithm for the large-scale pre-trained models during low-data fine-tuning, which adaptively selects a more promising subnetwork to perform staging updates based on gradients of back-propagation. From the data perspective, [37] proposes a novel selection strategy to select a subset from pre-training data to help improve the generalization on the target task. Likewise, the prototypical fine-tuning approach is proposed in [38], which automatically learns an inductive bias to improve predictive performance for varying data sizes, especially low-resource settings. In contrast, we explore the application of a data-centric approach within the domain of multimodal learning, specifically addressing scenarios involving missing modalities and data scarcity. Our work demonstrates the effectiveness of the data-centric approach in this novel domain.

In-context learning (ICL). ICL has emerged as a potent transfer learning approach in natural language processing (NLP), where large language models (LLMs) leverage context augmented with a few examples to make predictions, circumventing the need for parameter updates typical in supervised learning [39]. Demonstrating versatility, LLMs apply ICL to perform complex tasks, including mathematical reasoning and commonsense answering [40]. The success of ICL in NLP has recently spurred its adoption in diverse modalities, such as visual [41–44], speech [45,46], and multimodal domains [47–51]. In the work of [52], a vision encoder trained on aligned image-caption data represents images as sequences of continuous embeddings. This approach, using a frozen language model, surprisingly adapts to new tasks through ICL conditioning. Similarly, Flamingo [47], trained on extensive multimodal web corpora, showcases few-shot learning capabilities via ICL. Our paper specifically targets scenarios characterized by missing modalities and limited data, aiming to harness contextual features from full-modality samples. To the best of our knowledge, our work is among the first to use in-context learning to address such a challenge, offering a novel perspective on improving sample efficiency and reducing the performance gap between missing- and full-modality.

3. Proposed method

We first describe the problem definition and then the existing baseline to handle the missing-modality issue. Our empirical observation and the proposed method are elaborated on later.

3.1. Problem setting

We consider the multimodal learning problem with a dataset D containing multimodal input samples, D can be the training/validation/testing dataset. For notation simplicity, we assume there are two modalities in the dataset, i.e., $D = \{x_i^{m_1}, x_i^{m_2}, y_i\}_{i=1}^N$ where the label $y_i \in \{1, \dots, K\}$, but note that our framework can handle any number of modalities in principal. It is assumed that some samples have missing data for a particular modality. For example, some patients do not have the time or financial support for X-rays, and some images from food reviews failed to upload due to network/format issues. More importantly, we assume that the training set size N is limited as a result of the complicated data collection process [11] and expensive human-expert annotations [53]. The prevalence of missing modalities and limited data within our problem context is a common occurrence in critical domains, such as medical data analysis, thus requiring immediate resolution [54,55].

3.2. Empirical observations in the low-data regime

Fig. 2-a1 and b1 show the learning curve of the strong baseline MAP on two datasets with missing modalities. A discernible divergence is

evident in the training curve of full versus missing modalities in the two tasks. For a relatively straightforward task, i.e., MedFuse-I (binary classification task), the training AUROC of full-modality data is lower than that of missing-modality data in many learning steps. In contrast, for the more complex Food-101 dataset (a multi-classification task with 101 classes), the trend is reversed, where the full-modality data have better performance. This observation implies that whether missing-modality data are harder to learn than full-modality data depends on the task complexity. Note that although a similar observation is shown in [19], our work contributes by unveiling this phenomenon in the training of a multimodal transformer model, instead of the joint encoder training in [19]. This insight leads us to conjecture that *only focusing on reconstructing information for the missing modalities is not an optimal solution*, as the missing-modality data are not necessarily more difficult to learn than full-modality data. Consequently, we propose an ICL-based approach, where each sample, regardless of its modality completeness, **adaptively** benefits from its fusion with neighbor full-modality samples. The benefit of ICL is demonstrated in the ICL learning curve of Fig. 2-a2 and b2, where the generalization of both data types is improved.

3.3. Borrowing treasures from your neighbors: A semi-parametric approach

Unlike parametric methods, we adaptively augment full- and missing-modality samples through in-context learning by a limited number of parameters, which fully exploits the value of available data.

The proposed Borrowing Treasures from Your Neighbors method. In-context learning enables LLMs to perform tasks by conditioning an input prompt with exemplar examples without the need for parameter optimization. Drawing inspiration from it, we introduce the method titled *Borrowing Treasures from Your Neighbors*. This approach leverages similar data with full modalities to improve the performance on data containing full and missing modalities, aiming to alleviate the challenges posed by missing modalities and data scarcity. The reason why we only retrieve full-modality training data is that the missing-modality data need a reference of full modalities to implicitly infer the missed modalities, and the full-modality features can be fused with other full-modality ones to improve the generalization. Table 5 shows the strength of only using full-modality neighbors compared with using all training data and only missing-modality data.

As shown in Fig. 1, for each sample $\{x_i^{m_1}, x_i^{m_2}, y_i\}$, the pre-trained multimodal transformer f_{θ_X} infers the feature H_i including features for $x_i^{m_1}$, $x_i^{m_2}$ and a CLS feature cls_i by $H_i = \{H_i^{m_1}, H_i^{m_2}, cls_i\} = f_{\theta_X}(x_i^{m_1}, x_i^{m_2}), H_i^{m_1} \in \mathbb{R}^{L_1 \times d}, H_i^{m_2} \in \mathbb{R}^{L_2 \times d}, cls_i \in \mathbb{R}^{1 \times d}$, where L_1 and L_2 are the number of tokens of embedded $x_i^{m_1}$ and $x_i^{m_2}$, respectively. d is the embedding dimension. Based on the extracted features, the most common approach is to directly train a classifier f_{θ_c} to predict the labels $\hat{y}_i = f_{\theta_c}(H_i^{m_1}, H_i^{m_2}, cls_i)$. When there is missing-modality data, we follow MAP [16] to use default tokens to fill those missing tokens, see Section 4 for details. Then we retrieve the most similar Q samples $H_i^{NN} = \{H_{i,q}^{m_1}, H_{i,q}^{m_2}, cls_{i,q}\}_{q=1}^{Q}$ from the training samples with full modalities, where the features are arranged in descending order according to the similarity. The similarity is determined by cosine similarity and calculated with the cls tokens in our implementation. Finally, we design the ICL module to predict labels from the mean-pooled feature $\tilde{H}_i$ of the current sample and the mean-pooled retrieved *context* $\tilde{H}_i^{NN}$.

3.4. In-context module design

Our proposed method does not update/add any parameters in the pre-train multimodal model. During the training phase, we freeze all the parameters f_{θ_X} of the multimodal transformer (including the input embedding layers). We only update the parameters of the ICL module.

Next, we introduce the details of the ICL module. Transformer-based structures are shown to be capable of in-context learning [39,56]. Inspired by them, we compare two configurations of ICL based on transformer: ICL by cross-attention and ICL by next-token prediction. For ease of explanation, we assume that $N = 2$ and use blocks with different colors to represent tokens for different modalities.

ICL by cross attention (ICL-CA). One approach to perform ICL is to update the current sample's feature using the cross attention with nearest neighbor (NN) samples as keys, as Fig. 3(a) shows. The cross attention function $f_{\theta_{CA}}$ is trained to minimize the classification loss using the classification token, i.e.,

$$\hat{cls}_i = f_{\theta_{CA}}(\tilde{H}_i, \tilde{H}_i^{NN}), \ell_{CA}^{(i)} = \ell_{cls}(\hat{cls}_i, y_i), \tag{1}$$

where ℓ_{cls} means a classification loss such as cross-entropy and $\ell_{CA}^{(i)}$ is the loss value for the ith sample by ICL-CA method. In the cross attention module, the sample interacts with the tokens from similar full-modality samples, and thus implicitly infers missing modalities for missing-modality samples or refines the features for full-modality ones. We give more details of ICL-CA in Section 4.1.

ICL by next-token prediction (ICL-NTP). Another way to apply ICL is to implement the next-token prediction by transformer decoder, which is shown in Fig. 3(b). Write the input of the transformer decoder $[\tilde{H}_i^{NN}; \tilde{H}_i]$ as

$$\begin{aligned}[h_{i,1}^{(1)}, \ldots, h_{i,1}^{(T)}, cls_{i,1}; \ldots; h_{i,Q}^{(1)}, \ldots, h_{i,Q}^{(T)}, cls_{i,Q}; \\ h_{i,Q+1}^{(1)}, \ldots, h_{i,Q+1}^{(T)}, cls_{i,Q+1}],\end{aligned} \tag{2}$$

where $h_{i,q}^{(t)}, q \in \{1, \ldots, Q+1\}, t \in \{1, \ldots, T\}$ is the tth token of the qth neighbor (the $Q+1$th neighbor is the current sample itself). $cls_{i,q}, q \in \{1, \ldots, Q+1\}$ is the cls token of the qth neighbor. The decoder function $f_{\theta(NTP)}$ is trained to predict the next token in an auto-regressive way, i.e.,

$$\hat{h}_{i,q}^{(t)} / \hat{cls}_{i,q}^{(t)} = f_{\theta_{NTP}}(h_{i,1}^{(1)}, \ldots, h_{i,q}^{(t-1)}), \tag{3}$$

$$\ell_{NTP}^{(i)} = \lambda_{NTP} \sum_{q=1}^{Q+1} \sum_{t=1}^{T} (\hat{h}_{i,q}^{(t)} - h_{i,q}^{(t)})^2 + \sum_{q=1}^{Q+1} \ell_{cls}(\hat{cls}_{i,q}, y_{i,q}), \tag{4}$$

where $\lambda_{NTP} = 0.1$ is an adjustable hyperparameter to introduce the loss of feature reconstruction. $y_{i,q}$ is the label of the qth neighbor. $\ell_{NTP}^{(i)}$ is the loss value for the ith sample by ICL-NTP method. While $\hat{cls}_{i,Q+1}$ is used to predict $\hat{y}_i$, we incorporate other tokens (h's) in the loss computation to improve the prediction ability of ICL-NTP. This approach ensures that the outcomes of prior predictions continuously inform the subsequent token prediction, compelling the current sample to assimilate the rich context provided by its neighbors, i.e., each prediction is learned from the accumulation of preceding ones. The detailed settings are described in Section 4.1. Note that the reason why we explore different ICL configurations is to provide a comprehensive understanding of the in-context learning in our problem setting.

4. Experiment

We first introduce the experimental settings and then present the experimental results of our methods and baselines on four datasets, demonstrating the effectiveness of our method in missing modality and low-data tasks.

4.1. Experimental setting

Datasets. We follow existing works using two-modality datasets for a standard comparison [16]. Specifically, we use two real-word medical multimodal datasets containing EHR and X-ray images, i.e., MedFuse-In-hospital mortality (MedFuse-I) [23] and MedFuse-Phenotype (MedFuse-P) [23], and two general vision-language datasets (Hateful Memes [22]) and UPMC Food-101 [24] in our experiment. The details of each dataset are in Appendix A.

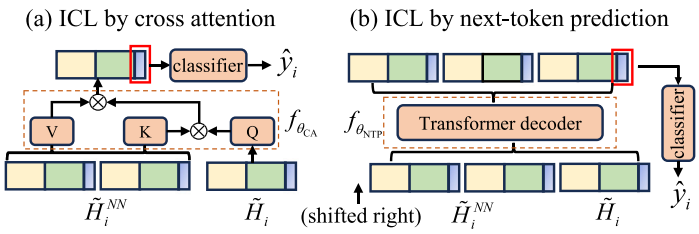


Fig. 3. The illustration of two ICL approaches. (a) ICL by cross attention. (b) ICL by next-token prediction. The yellow and green tokens denote features of two different modalities and the blue token is the *cls* token.

Baseline. We compare our method with the five strong baselines to tackle the missing-modality issue in multimodal transformers. These baselines are divided into two categories according to the range of fine-tuned parameters.

1. Full-parameter fine-tuning. Two methods are involved with full-parameter fine-tuning.
   - Traditional full-parameter fine-tuning (FT-A) [57]. All parameters of the pre-trained model are fine-tuned on the target dataset.
   - Wasserstein modality alignment (WMA) [2]. WMA implicitly employs Wasserstein distance for optimal modality alignment, enhancing the multimodal Transformer's robustness to input noise and missing modalities.
2. Partial-parameter fine-tuning. Three baselines are involved with partial-parameter fine-tuning.
   - Traditional partial-parameter fine-tuning (FT-C) [58]. Only the classifier of the pre-trained model is fine-tuned on the target dataset which is equivalent to removing the ICL module from our proposed method. This practice is widely adopted in transfer learning to balance performance and computational complexity.
   - Missing-aware prompts (MAP) [16]. In this method, empty prompt tokens are initialized to enable cross-layer interactions to learn effective instructions. Only prompt tokens and final classification layer parameters are updated during training.
   - Modal complementary recovering (MCR) [27]. MCR strategically integrates both complementary graph-based recovery and topological low-rank adaptation mechanisms to enhance the effectiveness and reliability of incomplete multimodal learning.

Since our proposed method does not involve fine-tuning the pre-trained model, we mainly compare it with the partial-parameter fine-tuning baselines.

Metrics. We set evaluation metrics based on the specific task type associated with each dataset. For binary classification tasks (MedFuse-I and Hateful Memes) and multi-label classification task (MedFuse-P), we adopt AUROC following [16,23]. For the multi-class classification task (Food-101), accuracy is chosen following [16]. All of these metrics are larger to represent better performance.

Input data processing. Please see the details of input data processing in Appendix A.

Pretrained Multimodal Transformer. We use the pre-trained multimodal transformer, ViLT [13], to extract features following [16]. Please see the details of the pretrained multimodal transformer for different tasks in Appendix A.

ICL module settings. We use a 2-layer transformer and 4 context samples. We choose the loss function of the ICL module according to the task type. Specifically, for binary classification tasks (MedFuse-I andHateful Memes) and multi-label classification task (MedFuse-P), we select binary cross-entropy loss by following [16,23]. For the multi-class classification task (Food-101), we use cross-entropy loss by following [16]. Please see the details of ICL module settings in Appendix A.

Setting of Missing Modalities. For real-world datasets MedFuse-I and MedFuse-P, both of them have missing modalities during data collection: 74% and 82% of patients have missing X-rays, respectively (we keep this missing status instead of artificially creating new missing settings in all experiments to simulate real-world application scenarios). For the other three tasks, we extend the setting in [16] and design six missing statuses to comprehensively evaluate all methods: (1) 30% samples have complete modalities and 70% samples are missing images, (2) 30% samples have complete modalities and 70% samples are missing texts, (3) 30% samples have complete modalities, 35% samples are missing images and 35% samples are missing texts, (4) 50% samples have complete modalities and 50% samples are missing images, (5) 50% samples have complete modalities and 50% samples are missing texts and (6) 50% samples have complete modalities, 25% samples are missing images and 25% samples are missing texts.

Subsampling. We subsample a full dataset to simulate a low-data downstream task. For medical data, we subsample the training dataset to 0.01, 0.02, 0.04, 0.1, 0.2, 0.4. For the other two tasks, we subsample the training dataset to 0.01, 0.02, 0.04, 0.1. Note that subsampling operation is only performed on training set and we use the same test sets for all experiments. To ensure a fair comparison, we maintain the missing status during subsampling. For each downsampling setting, we randomly sample five times to conduct experiments and report the mean value as the final result. We use r_{sub} to refer to the subsampling ratio in the later content.

4.2. Main results

Table 1 presents quantitative results across a range of scenarios, where the target dataset's downsampling ratio r_{sub} is 1%. See Appendix B for the performance of all downsampling ratios. From Table 1, Appendix B, we draw the following observations. (1) Across various datasets and scenarios of missing data, a consistent trend emerges: With sufficient target dataset size (notably for $r_{sub} > 0.1$), WMA and FT-A exhibit superior performance, attributed to the update of all parameters in the target domain. WMA consistently demonstrates superior performance, which we attribute to its use of grid search to determine the optimal degree of modality alignment, effectively mitigating the impact of missing modalities. MAP and MCR follow closely, achieving competitive results by updating fewer parameters. In contrast, when the target data is limited, our proposed ICL method, particularly ICL-CA, demonstrates remarkable efficacy (especially for $r_{sub} \leq 0.1$), surpassing most baseline approaches, even full-parameters fine-tuning results. This trend intensifies as r_{sub} decreases. In addition, we notice that MAP is superior to MCR in almost all experiments, so we mainly compare ICL with MAP later. (2) We separately calculate the performance of the samples with complete modality versus the samples

Table 1
Quantitative results on the MedFuse-I, MedFuse-P, Food-101, and HatefulMemes datasets under various modality-missing scenarios (here we show the result at r_{sub} = 0.01, see Appendix B for all sample sizes.) The bold number indicates the best performance. F means full-modality, m_1 means text/EHR and m_2 means image/X-ray. In this scenario, our proposed ICL-CA method outperforms the best partial-parameter fine-tuning method MAP by 5.9%, 5.9%, 5.3%,10.8% on four datasets (we calculate the improvement of AUROC for Medfuse-I and Medfuse-P, and the average improvement for other two datasets under all missing setting).

Datasets	Missing state	Metric	ICL-CA	ICL-NTP	WMA	FT-A	FT-C	MAP	MCR
MedFuse-I	26% F, 74% m_1	AUROC	**0.750**	0.737	0.722	0.719	0.702	0.691	0.683
MedFuse-P	18% F, 82% m_1	AUROC	**0.577**	0.556	0.533	0.524	0.512	0.518	0.512
HatefulMemes	30% F, 70% m_2	AUROC	**0.576**	0.565	0.542	0.537	0.542	0.528	0.519
	30% F, 70% m_1		**0.577**	0.576	0.552	0.548	0.540	0.531	0.527
	30% F, 35% m_2, 35% m_1		**0.593**	0.583	0.546	0.539	0.532	0.529	0.518
	50% F, 50% m_2		**0.602**	0.591	0.571	0.568	0.579	0.552	0.547
	50% F, 50% m_1		**0.611**	0.609	0.592	0.581	0.574	0.558	0.555
	50% F, 25% m_2, 25% m_1		**0.623**	0.618	0.605	0.595	0.587	0.567	0.554
Food-101	30% F, 70% m_2	Accuracy	0.312	**0.317**	0.261	0.250	0.222	0.222	0.212
	30% F, 70% m_1		**0.342**	0.327	0.279	0.265	0.243	0.247	0.225
	30% F, 35% m_2, 35% m_1		**0.332**	0.321	0.265	0.256	0.229	0.231	0.219
	50% F, 50% m_2		0.351	**0.357**	0.294	0.291	0.273	0.279	0.253
	50% F, 50% m_1		**0.388**	0.369	0.314	0.313	0.283	0.287	0.279
	50% F, 25% m_2, 25% m_1		**0.363**	0.362	0.319	0.308	0.279	0.282	0.260

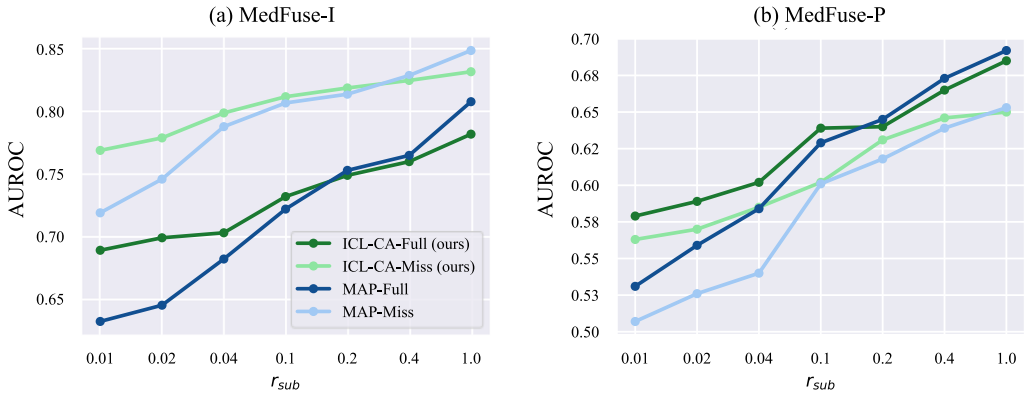


Fig. 4. The performance of MAP and ICL-CA on MedFuse-I and MedFuse-P when using different training set sizes. Our proposed ICL method is highly competitive under low data cases (r_{sub} from 0.01 to 0.1). Crucially, our approach enhances the performance in both full- and missing-modalities, outperforming the best baseline MAP.

with incomplete modality in each experiment, displayed in Figs. 4 and 5 (due to limited space, we only show the results of ICL with the best partial-parameter fine-tuning baseline MAP. For meeting status, we show the results of the first three settings). A notable performance difference is observed between complete and missing modalities across tasks. In simpler binary classification tasks, such as MedFuse-I and HatefulMemes, the performance with full-modality information falls behind that achieved with missing modalities, indicating that complete modality data is not always necessary for fitting the training data. In contrast, for more complex tasks, such as the multi-label classification in MedFuse-P and the multi-class classification in Food-101, the full-modality data exhibit dominant performance. Our ICL method shows remarkable adaptability under these varying conditions. Furthermore, we observe that ICL reduces almost all the performance gap between the two modalities, as detailed in Table 2.

4.3. Ablation study

Performance on task with 3 modalities. We conduct an experiment on the MOSEI dataset [59]. We selected Mean Absolute Error (MAE) as the metric, where smaller values indicate better performance. m_1, m_2, and m_3 refer to the text, vision, and audio modalities, respectively. We set the missing rate for each modality to 70% and $r_{sub} < 0.1$. We compare ICL-CA with the best full-parameter fine-tuning baseline WMA and the best partial-parameter fine-tuning baseline MAP. The results in 3 demonstrate the superiority of our proposed ICL-CA method.

ICL by Masked Feature Modeling (ICL-MF). We explore the efficacy of ICL by employing masked feature modeling with a transformer encoder. In this approach, we randomly mask a certain number of the input tokens with the mask tensor (cls_i is forced to be masked) and calculate the loss separately. For feature tokens H, we calculate the MSE loss between the reconstructed token and the original token. For cls tokens, we train a classifier and compute the loss of the output of the classifier concerning the ground-truth labels. The results of these experiments are in Table 4. We compare the performance of ICL-MF against ICL-CA, ICL-NTP and MAP. This comparison is conducted across four distinct dataset settings: MedFuse-I, MedFuse-P, Food-101 (comprising 30% F and 70% m_1), and HatefulMemes (with the same missing state as Food-101). Additionally, we evaluate them under two subsampling scenarios, specifically at r_{sub} of 0.01 and 0.1.

Table 4, reveals that ICL-MF either underperforms or marginally surpasses ICL-NTP and has a clear gap with ICL-CA across all tested settings. It is speculated that this outcome stems from the intrinsic nature of ICL-MF's use of self-attention, which treats each token uniformly. This approach differs from the mechanism employed in cross attention and next-token prediction, which inherently distinguishes between current and similar samples. However, it is noteworthy that ICL-MF demonstrates a significant performance advantage over the MAP approach.

The Impact of the Number of Neighbors Q. We examine the influence of the number of neighbors on our ICL-CA, as depicted in Fig. 6a.

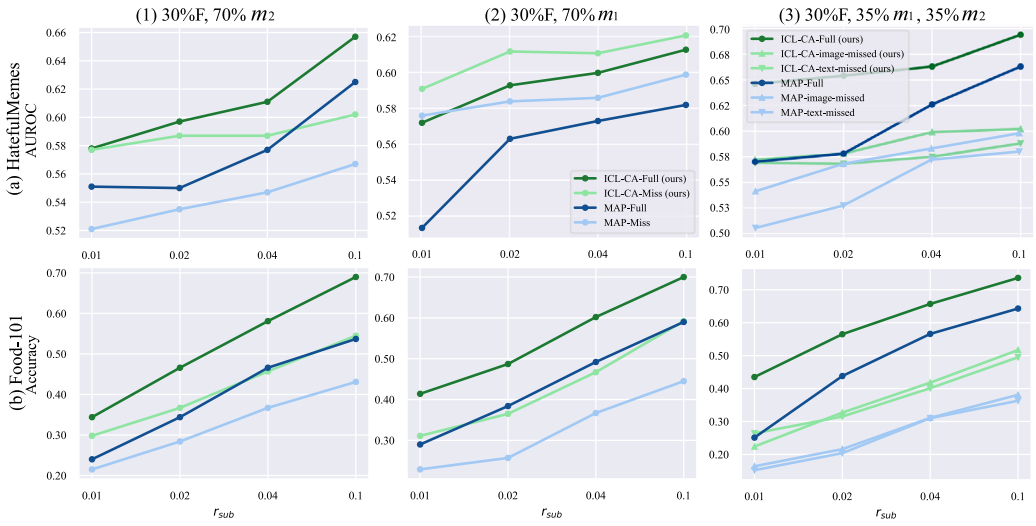


Fig. 5. The performance of MAP and ICL-CA on HatefulMemes and Food-101 when using different training set sizes. The performance of our ICL-CA is much better than that of best baseline MAP in the low-data regime (r_{sub} from 0.01 to 0.1).

Table 2
The relative performance gap between missing-modality and full-modality data on four datasets and all missing states under $r_{sub} = 0.01$. Bold number indicates the best performance. In most of the settings, our proposed ICL-CA shows smaller relative gap compared to the baseline MAP. ICL-CA's averaged relative performance gap (18.6%) is lower than that of MAP (24.4%).

Dataset	Missing state	Metric	ICL-CA			MAP		
			Missing modality	Full modality	Relative gap (%)	Missing modality	Full modality	Relative gap (%)
Medfuse-I	26% F, 74% *m*1	AUROC	0.769	0.689	**11.6**	0.719	0.632	13.8
Medfuse-P	18% F, 82% *m*1	AUROC	0.563	0.579	**2.8**	0.507	0.531	4.7
Hateful Memes	30% F, 70% *m*2	AUROC	0.577	0.578	**0.2**	0.521	0.551	5.8
	30% F, 70% *m*1		0.591	0.572	**3.3**	0.576	0.513	12.3
	30% F, 35% *m*1, 35% *m*2		0.569	0.646	13.5	0.505	0.570	**12.9**
	50% F, 50% *m*2		0.588	0.619	**5.3**	0.520	0.579	11.3
	50% F, 50% *m*1		0.623	0.600	**3.8**	0.573	0.531	7.9
	50% F, 25% *m*1, 25% *m*2		0.615	0.636	**3.4**	0.517	0.584	13.0
Food101	30% F, 70% *m*2	Accuracy	0.298	0.344	15.4	0.215	0.240	**11.6**
	30% F, 70% *m*1		0.311	0.414	33.1	0.229	0.290	**26.6**
	30% F, 35% *m*1, 35% *m*2		0.264	0.435	**64.8**	0.152	0.251	65.1
	50% F, 50% *m*2		0.320	0.393	22.8	0.232	0.282	**21.6**
	350% F, 50% *m*1		0.357	0.442	**23.8**	0.234	0.335	43.2
	50% F, 25% *m*1, 25% *m*2		0.297	0.465	**56.6**	0.188	0.360	91.5
Average					**18.6**			24.4

Table 3
Performance of ICL-CA, WMA and MAP on MOSEI dataset under $r_{sub} < 0.1$. MAE is the metric, where smaller values indicate better performance.

Method	Missing rate	$r_{sub} = 0.01$	$r_{sub} = 0.02$	$r_{sub} = 0.04$
ICL-CA	70% m_1 missed	**1.001**	**0.837**	**0.829**
	70% m_2 missed	**0.985**	**0.834**	**0.827**
	70% m_3 missed	**1.016**	**0.865**	**0.857**
WMA	70% m_1 missed	1.063	0.892	0.851
	70% m_2 missed	0.990	0.876	0.847
	70% m_3 missed	1.096	0.899	0.872
MAP	70% m_1 missed	1.128	1.113	0.998
	70% m_2 missed	1.058	1.003	0.972
	70% m_3 missed	1.224	1.159	1.068

In this analysis, we vary the number of neighbors (1, 2, 4, 8, 16) and observed their effects on two datasets, MedFuse-I and Food-101, under r_{sub} of 0.1 and 0.01. Our findings indicate a marked performance improvement when the number of neighbors is increased from 1 to 4 in all experiments. However, further increases in the number of neighbors do not sustain this upward trend in performance. Therefore, we use $Q = 4$ in our experiment to strike a balance between computational efficiency and efficacy.

The Effect of Pooled Feature Length T. We assess the impact of varying pooled feature lengths (the number of pooled feature tokens in each sample) on our ICL-CA, as illustrated in Fig. 6b. Pooled features of greater length can provide more comprehensive feature information but concurrently increase computational demands. We test pooled feature lengths of 0, 4, 8, 16, and 32 under r_{sub} of 0.01 and 0.1 in the MedFuse-I and Food-101 datasets. A pooled feature length of 0 implies reliance

Table 4
Comparison of ICL-CA, ICL-NTP, ICL-MF and MAP under different datasets. The bold number indicates the best performance.

Datasets	Metric	$r_{sub} = 0.01$				$r_{sub} = 0.1$			
		ICL-CA	ICL-NTP	ICL-MF	MAP	ICL-CA	ICL-NTP	ICL-MF	MAP
MedFuse-I	AUROC	**0.750**	0.737	0.733	0.691	**0.793**	0.789	0.791	0.788
MedFuse-P	AUROC	**0.577**	0.556	0.543	0.518	**0.619**	0.612	0.613	0.614
HatefulMemes 30% F, 70% $m1$	AUROC	**0.577**	0.576	0.561	0.531	**0.617**	0.612	0.608	0.586
Food-101 30% F, 70% $m1$	Accuracy	**0.342**	0.327	0.326	0.247	**0.625**	0.619	0.564	0.489

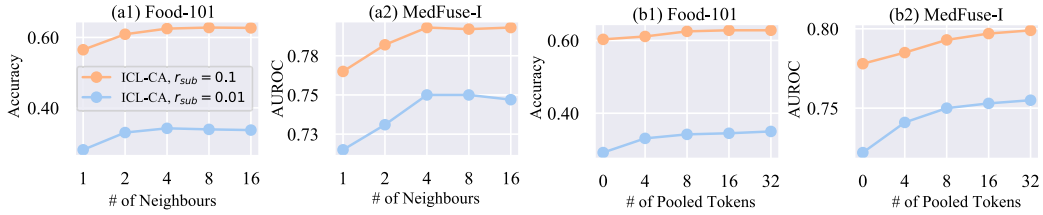


Fig. 6. Comparison of the effect of the number of neighbors and the pooled feature length in the ICL-CA model. (a) Comparison of the effect of the number of neighbors. (b) Comparison of the effect of pooled feature length. We suggest setting the number of neighbors to 4 and the pooled feature length to 8.

Table 5
Performance of ICL-CA by different groups for retrieving neighboring samples under $r_{sub} = 0.01$. NN-all, NN-full and NN-miss refer to using all training data, full-modality data and missing-modality ones respectively, in the retrieval process.

Datasets	Missing state	Metric	NN-all	NN-full	NN-miss
MedFuse-I	26% F, 74% m_1	AUROC	0.732	**0.750**	0.721
MedFuse-P	18% F, 82% m_1	AUROC	0.553	**0.577**	0.542
Hateful-Memes	30% F, 70% m_2	AUROC	0.549	**0.576**	0.523
	30% F, 70% m_1		0.569	**0.577**	0.553
	30% F, 35% m_2, 35% m_1		0.575	**0.593**	0.566
Food-101	30% F, 70% m_2	Accuracy	0.294	**0.312**	0.265
	30% F, 70% m_1		**0.346**	0.342	0.311
	30% F, 35% m_2, 35% m_1		0.266	**0.281**	0.247

solely on the cls token from all samples for ICL. A substantial increase in performance is observed when the pooled feature length is increased from 0 to 8. When the pooled feature length exceeds 8, the gain in performance becomes negligible. Thus, we set the pooled feature length to 8 in this paper.

Groups for retrieving neighboring samples. We compare different groups of samples in training datasets for retrieving neighboring samples, i,e. all the samples, full-modality samples and missing-modality samples. We select the ICL-CA method under $r_{sub} = 0.01$ and three missing settings for this experiment, as shown in Table 5. It shows that employing samples with full modality as the retrieval group yields superior results. This observation indicates that our proposed ICL method can effectively utilize the context provided by the full-modality samples. In addition, using full-modality samples as neighbors enhances computational efficiency due to the reduced sample size of neighbors.

Inference Time. One major concern for the retrieval-based approach is the inference latency. We test the inference time of ICL-CA and MAP on MedFuse-I. We set the batch size to 1 and record the inference time for 100 batches. The average inference time of ICL-CA is 34.41 ms with a std of 4.52 ms. In contrast, MAP has a mean inference time of 40.59 ms and a std of 5.40 ms. The difference in inference time is because MAP has a larger number of tokens (missing-aware prompts) in the transformer.

5. Conclusion

This paper investigates a pivotal challenge in multimodal learning: missing modalities in the low-data regime. Our analysis examines the learning process of both full and missing modalities across tasks of various complexity. Stemming from our findings, we introduce a semi-parametric, retrieval-augmented in-context learning framework to address the challenges. This approach is designed to condition each sample with neighboring full-modality data. The effectiveness of our method is corroborated across diverse datasets, including medical and vision-language prediction tasks. Remarkably, our approach achieves an average performance boost of 5.9%, 5.9%, 5.3% and 10.8% on four datasets over the best baseline in the low-data regime. Furthermore, it effectively narrows the performance disparity caused by modality absence.

CRediT authorship contribution statement

Zhuo Zhi: Writing – review & editing, Writing – original draft, Formal analysis, Data curation, Conceptualization. **Ziquan Liu:** Methodology. **Moe Elbadawi:** Supervision. **Adam Daneshmend:** Supervision. **Mine Orlu:** Supervision. **Abdul Basit:** Supervision. **Andreas Demosthenous:** Supervision. **Miguel Rodrigues:** Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Zhuo zhi reports financial support was provided by Engineering and Physical Sciences Research Council. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC), U.K., under Grant EP/T517793/1.

Appendix A. Experimental settings

Details of Input data processing. For MedFuse-I, we use linear embedding to map EHR to token embeddings and the number of embedded EHR tokens is 48. The number of tokens from X-ray image patches is 144. For MedFuse-P, the token numbers are 96 and 96 for EHR and X-rays. The maximum length of text inputs is 512 for the Food-101 task and 128 for Hateful Memes, and the image processing of the input images is the same as [16].

Details of Pretrained Multimodal Transformer. We use the pre-trained multimodal transformer, ViLT [13], to extract features following [16]. For medical data, we use a pre-trained ViLT model and fine-tune all model parameters on one dataset and then use the fine-tuned model on the other dataset. The reason is that there is a huge gap between the pre-training data, i.e., images and texts, and the downstream data, i.e., EHR and X-ray. Thus, fine-tuning all model parameters helps the model to adapt to the medical data. For the Food-101 task and Hateful Memes, we directly use the ViLT model since there is no such domain gap as in the medical data. For medical datasets, we initialize two kinds of empty tokens and update them in the fine-tuning phase, and then use these tokens on another dataset to represent the missing modality. For Food-101 and HatefulMemes datasets, if the image is missing, we create an image with all pixel values equal to one as dummy input, and if the text is missing, we use an empty string as dummy input by following [16].

Details of ICL module settings. We use a 2-layer transformer and 4 context samples. For computational efficiency, we pool the feature tokens H from ViLT before the input into ICL. The number of pooled tokens is 8. Before the training/testing process, we saved the features inferred by the pretrained multimodal transformer for all full-modality samples in the training set. Then, for each input sample, we first obtain its features by the pretrained multi-modal transformer and only use the cls token to retrieve neighbors through the saved features. Finally, the features of the current sample with its neighbors are input into the ICL module for classification. Note that we only use full-modality training data during the NN search so the computational cost is much less than using all training data.

Details of datasets. The details of each dataset are as follows:

- *MedFuse-In-hospital mortality (MedFuse-I)* [23]. This dataset contains EHR and X-ray data for each patient. The target of this binary classification task is to predict in-hospital mortality after the first 48 h spent in the ICU. The EHR is time-series data with 17 clinical variables, among which five are categorical and 12 are continuous. Each EHR is paired with the last chest X-ray image collected during the ICU stay. The numbers of the samples in the training/val/testing dataset are 18845, 2138 and 5243.
- *MedFuse-Phenotype (MedFuse-P)* [23] This dataset has the same data types as in MedFuse-I. The difference is that this dataset has a larger sample size and the task is multi-label classification to predict whether a set of 25 chronic, mixed, and acute care conditions are assigned to a patient in a given ICU stay. The numbers of the samples in the training/val/testing dataset are 42628, 4802 and 11914.
- *Hateful Memes* [22]. This is a binary classification task. The dataset represents a challenging blend of visual and textual content, specifically designed to tackle the detection of harmful content online. The dataset comprises meme images that are often used in social media contexts, containing layers of nuances in meaning that combine text and imagery. The numbers of the samples in the training/val/testing dataset are 8500, 500 and 1000.
- *UPMC Food-101* [24]. This dataset contains the noisy text–image paired data for 101 kinds of food. The target is to predict the type of food, which is a multi-classification task. The numbers of the samples in the training/val/testing dataset are 61127, 6588 and 25250.

Appendix B. More experimental results

The performance of our method and baselines on *all test samples*. The main paper presents the performance on full-modality and missing-modality test samples *separately* in the figures. Here we give the performance of our method and baselines on *all test samples* as in [16]. Table B.6 presents quantitative results of all test samples across all datasets, methods, and missing states under $r_{sub} \geq 0.1$. Table B.7 presents the results under $r_{sub} < 0.1$.

Table B.6
Quantitative results of the whole test set on the Medfuse-I, Medfuse-P, Food101, and HatefulMemes datasets with different missing rates under various modality-missing scenarios under $r_{sub} \geq 0.1$. Bold number indicates the best performance. With sufficient target dataset size (notably for $r_{sub} > 0.1$), FT-A and WMA exhibits superior performance, attributed to the update of all parameters in the target domain. MAP and MCR follows closely, achieving competitive results by updating fewer parameters. FT-C, on the other hand, performs the worst at all moments, due to the limited number of updated parameters.

r_{sub}	Datasets	Missing state	Metric	ICL-CA	ICL-NTP	WMA	FT-A	FT-C	MAP	MCR
0.1	Medfuse-I	26% F, 74% $m1$	AUROC	**0.793**	0.789	0.792	0.790	0.771	0.788	0.779
	Medfuse-P	18% F, 82% $m1$	AUROC	0.619	0.612	**0.627**	0.621	0.600	0.614	0.603
	Hateful Memes	30% F, 70% $m2$	AUROC	**0.607**	0.598	0.606	0.601	0.577	0.585	0.583
		30% F, 70% $m1$		**0.617**	0.612	0.615	0.609	0.575	0.586	0.579
		30% F, 35% $m1$, 35% $m2$		**0.618**	**0.618**	0.617	0.614	0.579	0.599	0.588
		50% F, 50% $m2$		0.645	0.643	0.642	0.635	0.588	0.607	0.599
		50% F, 50% $m1$		**0.658**	0.654	0.652	0.646	0.603	0.634	0.609
		50% F, 25% $m1$, 25% $m2$		**0.671**	0.666	0.670	0.654	0.620	0.638	0.626
	Food101	30% F, 70% $m2$	Accuracy	**0.595**	0.576	0.570	0.562	0.417	0.463	0.438
		30% F, 70% $m1$		**0.625**	0.619	0.610	0.603	0.450	0.489	0.473
		30% F, 35% $m1$, 35% $m2$		**0.607**	0.602	0.597	0.588	0.434	0.476	0.452
		50% F, 50% $m2$		**0.618**	0.596	0.584	0.580	0.442	0.491	0.455
		50% F, 50% $m1$		**0.629**	0.627	0.608	0.600	0.471	0.533	0.482
		50% F, 25% $m1$, 25% $m2$		**0.621**	0.614	0.600	0.592	0.459	0.520	0.476
0.2	Medfuse-I	26% F, 74% $m1$	AUROC	0.802	0.792	**0.839**	0.832	0.782	0.801	0.788
	Medfuse-P	18% F, 82% $m1$	AUROC	0.634	0.630	**0.708**	0.698	0.616	0.632	0.620
0.4	Medfuse-I	26% F, 74% $m1$	AUROC	0.810	0.806	**0.845**	0.840	0.793	0.815	0.799
	Medfuse-P	18% F, 82% $m1$	AUROC	0.655	0.647	**0.729**	0.720	0.623	0.652	0.625
1.0	Medfuse-I	26% F, 74% $m1$	AUROC	0.820	0.819	**0.855**	0.850	0.804	0.838	0.810
	Medfuse-P	18% F, 82% $m1$	AUROC	0.663	0.660	**0.735**	0.727	0.630	0.668	0.642

Table B.7
Quantitative results of the whole test set on the Medfuse-I, Medfuse-P, Food101, and HatefulMemes datasets with different missing rates under various modality-missing scenarios. Bold number indicates the best performance. When the target data is limited, our proposed ICL method, particularly ICL-CA, demonstrates remarkable efficacy (especially for $r_{sub} < 0.1$), surpassing most baseline approaches. This trend intensifies as r_{sub} decreases.

r_{sub}	Datasets	Missing state	Metric	ICL-CA	ICL-NTP	WMA	FT-A	FT-C	MAP	MCR
0.01	Medfuse-I	26% F, 74% $m1$	AUROC	**0.750**	0.737	0.722	0.719	0.702	0.691	0.683
	Medfuse-P	18% F, 82% $m1$	AUROC	**0.577**	0.556	0.533	0.524	0.512	0.518	0.512
	Hateful Memes	30% F, 70% $m2$	AUROC	**0.576**	0.565	0.542	0.537	0.542	0.528	0.519
		30% F, 70% $m1$		**0.577**	0.576	0.552	0.548	0.540	0.531	0.527
		30% F, 35% $m1$, 35% $m2$,		**0.593**	0.583	0.546	0.539	0.532	0.529	0.518
		50% F, 50% $m2$		**0.602**	0.591	0.571	0.568	0.579	0.552	0.547
		50% F, 50% $m1$		**0.611**	0.609	0.592	0.581	0.574	0.558	0.555
		50% F, 25% $m1$, 25% $m2$		**0.623**	0.618	0.605	0.595	0.587	0.567	0.554
	Food101	30% F, 70% $m2$	Accuracy	0.312	**0.317**	0.261	0.250	0.222	0.222	0.212
		30% F, 70% $m1$		**0.342**	0.327	0.279	0.265	0.243	0.247	0.225
		30% F, 35% $m1$, 35% $m2$,		**0.332**	0.321	0.265	0.256	0.229	0.231	0.219
		50% F, 50% $m2$		0.351	**0.357**	0.294	0.291	0.273	0.279	0.253
		50% F, 50% $m1$		**0.388**	0.369	0.314	0.313	0.283	0.287	0.279
		50% F, 25% $m1$, 25% $m2$		**0.363**	0.362	0.319	0.308	0.279	0.282	0.260
0.02	Medfuse-I	26% F, 74% $m1$	AUROC	0.761	**0.764**	0.758	0.754	0.728	0.722	0.720
	Medfuse-P	18% F, 82% $m1$	AUROC	**0.584**	0.577	0.559	0.553	0.547	0.540	0.529
	Hateful Memes	30% F, 70% $m2$	AUROC	**0.590**	0.581	0.563	0.550	0.545	0.549	0.545
		30% F, 70% $m1$		**0.593**	0.587	0.577	0.570	0.556	0.557	0.548
		30% F, 35% $m1$, 35% $m2$,		0.602	**0.603**	0.579	0.564	0.549	0.548	0.544
		50% F, 50% $m2$		**0.622**	0.613	0.596	0.582	0.568	0.573	0.576
		50% F, 50% $m1$		**0.631**	0.623	0.611	0.605	0.589	0.591	0.583
		50% F, 25% $m1$, 25% $m2$		**0.642**	0.630	0.619	0.611	0.595	0.592	0.587
	Food101	30% F, 70% $m2$	Accuracy	**0.397**	0.373	0.363	0.352	0.287	0.302	0.281
		30% F, 70% $m1$		**0.412**	0.399	0.374	0.368	0.310	0.295	0.286
		30% F, 35% $m1$, 35% $m2$,		**0.406**	0.381	0.372	0.358	0.303	0.299	0.283
		50% F, 50% $m2$		**0.430**	0.416	0.398	0.381	0.331	0.346	0.335
		50% F, 50% $m1$		**0.447**	0.422	0.415	0.399	0.352	0.363	0.356
		50% F, 25% $m1$, 25% $m2$		**0.440**	0.421	0.403	0.390	0.344	0.453	0.347
0.04	Medfuse-I	26% F, 74% $m1$	AUROC	**0.778**	0.777	0.792	0.787	0.752	0.767	0.755
	Medfuse-P	18% F, 82% $m1$	AUROC	**0.597**	0.588	0.585	0.581	0.556	0.562	0.560
	Hateful Memes	30% F, 70% $m2$	AUROC	**0.595**	0.591	0.585	0.583	0.561	0.558	0.543
		30% F, 70% $m1$		**0.600**	0.585	0.580	0.577	0.559	0.567	0.560
		30% F, 35% $m1$, 35% $m2$,		**0.610**	0.607	0.588	0.579	0.562	0.574	0.568
		50% F, 50% $m2$		**0.630**	0.621	0.607	0.595	0.572	0.581	0.579
		50% F, 50% $m1$		**0.643**	0.630	0.622	0.615	0.597	0.599	0.588
		50% F, 25% $m1$, 25% $m2$		**0.651**	0.645	0.629	0.624	0.613	0.619	0.598
	Food101	30% F, 70% $m2$	Accuracy	**0.494**	0.464	0.457	0.448	0.352	0.397	0.382
		30% F, 70% $m1$		**0.508**	0.489	0.470	0.458	0.391	0.405	0.399
		30% F, 35% $m1$, 35% $m2$,		**0.499**	0.470	0.466	0.455	0.383	0.408	0.400
		50% F, 50% $m2$		**0.527**	0.506	0.496	0.490	0.380	0.437	0.411
		50% F, 50% $m1$		**0.538**	0.526	0.509	0.503	0.402	0.458	0.430
		50% F, 25% $m1$, 25% $m2$		**0.535**	0.515	0.500	0.497	0.392	0.444	0.427

Data availability

The code link is given in the paper.

References

[1] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2018) 423–443.

[2] Z. Zhi, Y. Sun, et al., Wasserstein modality alignment makes your multimodal transformer more robust, Trans. Mach. Learn. Res. (2025).

[3] Z. Zhi, M. Elbadawi, A. Daneshmend, M. Orlu, A. Basit, A. Demosthenous, M. Rodrigues, Multimodal diagnosis for pulmonary embolism from EHR data and ct images, in: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC, IEEE, 2022, pp. 2053–2057.

[4] Z. Zhuo, M. Elbadawi, A. Daneshmend, M. Orlu, A. Basit, A. Demosthenous, M. Rodrigues, HgbNet: predicting hemoglobin level/anemia degree from irregular EHR, IEEE Access (2024).

[5] B. Wang, S. Wang, C. Li, R. Guan, X. Li, Harmfully manipulated images matter in multimodal misinformation detection, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 2262–2271.

[6] W. Zhao, K. Yang, P. Ding, C. Na, W. Li, Graph attention contrastive learning with missing modality for multimodal recommendation, Knowl.-Based Syst. (2025) 113035.

[7] Z. Liu, B. Zhou, D. Chu, Y. Sun, L. Meng, Modality translation-based multimodal sentiment analysis under uncertain missing modalities, Inf. Fusion 101 (2024) 101973.

[8] W. Yao, K. Yin, W.K. Cheung, J. Liu, J. Qin, Drfuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 38, 2024, pp. 16416–16424.

[9] M. Ma, J. Ren, L. Zhao, D. Testuggine, X. Peng, Are multimodal transformers robust to missing modality? in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18177–18186.

[10] A.E. Johnson, T.J. Pollard, S.J. Berkowitz, N.R. Greenbaum, M.P. Lungren, C.-y. Deng, R.G. Mark, S. Horng, MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports, Sci. Data 6 (1) (2019) 317.

[11] A.E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T.J. Pollard, S. Hao, B. Moody, B. Gow, et al., MIMIC-IV, a freely accessible electronic health record dataset, Sci. Data 10 (1) (2023) 1.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[13] W. Kim, B. Son, I. Kim, Vilt: Vision-and-language transformer without convolution or region supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 5583–5594.

[14] M. Ma, J. Ren, L. Zhao, S. Tulyakov, C. Wu, X. Peng, Smil: Multimodal learning with severely missing modality, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 2302–2310.

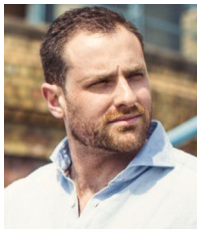
[15] A. Radford, J.W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, in: International Conference on Machine Learning, PMLR, 2023, pp. 28492–28518.

[16] Y.-L. Lee, Y.-H. Tsai, W.-C. Chiu, C.-Y. Lee, Multimodal prompting with missing modalities for visual recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 14943–14952.
[17] Y. Huang, C. Du, Z. Xue, X. Chen, H. Zhao, L. Huang, What makes multi-modal learning better than single (provably), Adv. Neural Inf. Process. Syst. 34 (2021) 10944–10956.
[18] L.A. Vale-Silva, K. Rohr, Long-term cancer survival prediction using multimodal deep learning, Sci. Rep. 11 (1) (2021) 13505.
[19] W. Wang, D. Tran, M. Feiszli, What makes training multi-modal classification networks hard? in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12695–12705.
[20] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G.B. Van Den Driessche, J.-B. Lespiau, B. Damoc, A. Clark, et al., Improving language models by retrieving from trillions of tokens, in: International Conference on Machine Learning, PMLR, 2022, pp. 2206–2240.
[21] O. Ram, Y. Levine, I. Dalmedigos, D. Muhlgay, A. Shashua, K. Leyton-Brown, Y. Shoham, In-context retrieval-augmented language models, 2023, arXiv preprint arXiv:2302.00083.
[22] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The hateful memes challenge: Detecting hate speech in multimodal memes, Adv. Neural Inf. Process. Syst. 33 (2020) 2611–2624.
[23] N. Hayat, K.J. Geras, F.E. Shamout, MedFuse: Multi-modal fusion with clinical time-series data and chest X-ray images, in: Machine Learning for Healthcare Conference, PMLR, 2022, pp. 479–503.
[24] L. Bossard, M. Guillaumin, L. Van Gool, Food-101 – mining discriminative components with random forests, in: European Conference on Computer Vision, 2014.
[25] Z. Zhi, M. Rodrigues, et al., Wasserstein modality alignment makes your multimodal transformer more robust, PMLR (Proceedings of Machine Learning Research), 2024.
[26] Y.-H.H. Tsai, P.P. Liang, A. Zadeh, L.-P. Morency, R. Salakhutdinov, Learning factorized multimodal representations, in: International Conference on Learning Representations, 2018.
[27] M. Ding, H. Lin, J. Zhu, C. Zou, W. Cai, B. Chen, Enhancing incomplete multimodal learning via modal complementary recovering, in: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2025, pp. 1–5.
[28] P. Xu, X. Zhu, D.A. Clifton, Multimodal learning with transformers: A survey, IEEE Trans. Pattern Anal. Mach. Intell. (2023).
[29] J. Lin, H. Yin, W. Ping, P. Molchanov, M. Shoeybi, S. Han, Vila: On pre-training for visual language models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 26689–26699.
[30] B. McKinzie, Z. Gan, J.-P. Fauconnier, S. Dodge, B. Zhang, P. Dufter, D. Shah, X. Du, F. Peng, A. Belyi, et al., MM1: methods, analysis and insights from multimodal LLM pre-training, in: European Conference on Computer Vision, Springer, 2024, pp. 304–323.
[31] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, S.C.H. Hoi, Align before fuse: Vision and language representation learning with momentum distillation, Adv. Neural Inf. Process. Syst. 34 (2021) 9694–9705.
[32] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O.K. Mohammed, S. Singhal, S. Som, et al., Image as a foreign language: Beit pretraining for all vision and vision-language tasks, 2022, arXiv preprint arXiv:2208.10442.
[33] M.A. Mazurowski, P.A. Habas, J.M. Zurada, J.Y. Lo, J.A. Baker, G.D. Tourassi, Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance, Neural Netw. 21 (2–3) (2008) 427–436.
[34] M.A. Hedderich, L. Lange, H. Adel, J. Strötgen, D. Klakow, A survey on recent approaches for natural language processing in low-resource scenarios, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 2545–2568.
[35] U. Evci, V. Dumoulin, H. Larochelle, M.C. Mozer, Head2toe: Utilizing intermediate representations for better transfer learning, in: International Conference on Machine Learning, PMLR, 2022, pp. 6009–6033.
[36] H. Zhang, G. Li, J. Li, Z. Zhang, Y. Zhu, Z. Jin, Fine-tuning pre-trained language models effectively by optimizing subnetworks adaptively, Adv. Neural Inf. Process. Syst. 35 (2022) 21442–21454.
[37] Z. Liu, Y. Xu, Y. Xu, Q. Qian, H. Li, X. Ji, A. Chan, R. Jin, Improved fine-tuning by better leveraging pre-training data, Adv. Neural Inf. Process. Syst. 35 (2022) 32568–32581.
[38] Y. Jin, X. Wang, Y. Hao, Y. Sun, X. Xie, Prototypical fine-tuning: Towards robust performance under varying data sizes, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, 2023, pp. 12968–12976.
[39] Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, Z. Sui, A survey for in-context learning, 2022, arXiv preprint arXiv:2301.00234.
[40] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q.V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Adv. Neural Inf. Process. Syst. 35 (2022) 24824–24837.
[41] F.B. Baldassini, M. Shukor, M. Cord, L. Soulier, B. Piwowarski, What makes multimodal in-context learning work? in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 1539–1550.
[42] X. Wang, W. Wang, Y. Cao, C. Shen, T. Huang, Images speak in images: A generalist painter for in-context visual learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 6830–6839.
[43] X. Wang, X. Zhang, Y. Cao, W. Wang, C. Shen, T. Huang, Seggpt: Segmenting everything in context, 2023, arXiv preprint arXiv:2304.03284.
[44] T. Gupta, A. Kembhavi, Visual programming: Compositional visual reasoning without training, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 14953–14962.
[45] Z. Chen, H. Huang, A. Andrusenko, O. Hrinchuk, K.C. Puvvada, J. Li, S. Ghosh, J. Balam, B. Ginsburg, Salm: Speech-augmented language model with in-context learning for speech recognition and translation, in: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2024, pp. 13521–13525.
[46] S. Wang, C.-H. Yang, J. Wu, C. Zhang, Can whisper perform speech-based in-context learning? in: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2024, pp. 13421–13425.
[47] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al., Flamingo: a visual language model for few-shot learning, Adv. Neural Inf. Process. Syst. 35 (2022) 23716–23736.
[48] S. Huang, L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O.K. Mohammed, Q. Liu, et al., Language is not all you need: Aligning perception with language models, 2023, arXiv preprint arXiv:2302.14045.
[49] C. Zhang, K. Lin, Z. Yang, J. Wang, L. Li, C.-C. Lin, Z. Liu, L. Wang, Mm-narrator: Narrating long-form videos with multimodal in-context learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 13647–13657.
[50] J.Y. Koh, R. Salakhutdinov, D. Fried, Grounding language models to images for multimodal generation, 2023, arXiv preprint arXiv:2301.13823.
[51] B. Huang, C. Mitra, L. Karlinsky, A. Arbelle, T. Darrell, R. Herzig, Multimodal task vectors enable many-shot multimodal in-context learning, Adv. Neural Inf. Process. Syst. 37 (2024) 22124–22153.
[52] M. Tsimpoukelli, J.L. Menick, S. Cabi, S. Eslami, O. Vinyals, F. Hill, Multimodal few-shot learning with frozen language models, Adv. Neural Inf. Process. Syst. 34 (2021) 200–212.
[53] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, B. Ni, MedMNIST v2-a large-scale lightweight benchmark for 2D and 3D biomedical image classification, Sci. Data 10 (1) (2023) 41.
[54] S. Wang, Z. Yan, D. Zhang, H. Wei, Z. Li, R. Li, Prototype knowledge distillation for medical segmentation with missing modality, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2023, pp. 1–5.
[55] Y.-H. Li, Y.-L. Li, M.-Y. Wei, G.-Y. Li, Innovation and challenges of artificial intelligence technology in personalized healthcare, Sci. Rep. 14 (1) (2024) 18994.
[56] R. Zhang, J. Wu, P. Bartlett, In-context learning of a linear transformer block: benefits of the mlp component and one-step gd initialization, Adv. Neural Inf. Process. Syst. 37 (2024) 18310–18361.
[57] K. Lv, Y. Yang, T. Liu, Q. Gao, Q. Guo, X. Qiu, Full parameter fine-tuning for large language models with limited resources, 2023, arXiv preprint arXiv:2306.09782.
[58] Z. Fu, H. Yang, A.M.-C. So, W. Lam, L. Bing, N. Collier, On the effectiveness of parameter-efficient fine-tuning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, 2023, pp. 12799–12807.
[59] A. Bagher Zadeh, P.P. Liang, S. Poria, E. Cambria, L.-P. Morency, Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph, in: I. Gurevych, Y. Miyao (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 2236–2246, http://dx.doi.org/10.18653/v1/P18-1208, URL https://aclanthology.org/P18-1208.

Zhuo Zhi received the B.S degree from Shandong University (SDU), Weihai, China, in 2019, and the M.S degree from the Department of Automatic Test and Control, Harbin Institute of Technology, Harbin, China, in 2021. He is currently a Ph.D candidate at University College London. His research interests focus on machine learning for healthcare and multimodal learning. Zhuo Zhi received the Doctoral Training Partnerships (DTP) award from the Engineering and Physical Sciences Research Council (EPSRC), UK in 2021.

Dr. Ziquan Liu is now a Lecturer (Teaching & Research) at the School of Electronic Engineering and Computer Science, Queen Mary University of London. He obtained his Ph.D. in 2023 at City University of Hong Kong in Video, Image, and Sound Analysis Lab. He obtained the Bachelor of Engineering degree in Information Engineering from Beihang University in 2017. His research interests focus on trustworthy and robust machine learning, uncertainty of foundation models and interpretable machine learning.

Moe Elbadawi received the B.S degree in Pharmacology from University of Bristol in 2010 and the M.S degree in Biomedical Engineering from University of Surrey in 2013. He obtained the Ph.D. degree in Mechanical Engineering from University of Sheffield in 2017. Dr Elbadawi's research centers on the use of digital technologies to advance healthcare, including machine learning, robotics and 3D printing. Dr Moe Elbadawi has been named on the 'Stanford Top 2% Scientist' for two consecutive years (2022, 2023).



Adam Daneshmend has initially trained in pharmacology at the University of Bristol, where he graduated with a first class degree. He has a background in pharmaceuticals, having worked for GlaxoSmithKline in the USA in heart failure research, along with having completed a subsequent degree in economics, finance and management from Bristol. Following this degree, Dr Daneshmend pursued a medical career, graduating from the University of Southampton and entering the specialty of Clinical Pharmacology and Therapeutics. In this time he has worked on several research projects, acted as a sub-investigator for drugs trials during COVID, and continued research into a number of areas of therapeutics. He has worked with the Royal Pharmaceutical Society to provide his medical input on both new and old medications, to improve prescribing within the United Kingdom. His areas of clinical interest include hypertension, diabetes, obesity and prescribing in an acute medical setting.

Mine Orlu is currently Professor of Pharmaceutics at UCL. Following her Pharmacy degree studies at Istanbul University, she continued her postgraduate studies and received her M.Sc. on the subject of colon targeted microspheres in 2003 and her Ph.D. about the interaction of fluorescent labeled nanoparticles with lung cells in 2008 from Istanbul University, Faculty of Pharmacy. During her Ph.D. studies, she held a one year visiting scientist post at King's College London funded by the EU Marie Curie EST programme and received GALENOS Euro Ph.D. in Advanced Drug Delivery. She took up two post-doctoral positions in University of London, The School of Pharmacy — firstly, in Prof Oya Alpar's research group in 2008 and secondly, in Prof Catherine Tuleu's research group from 2009 to 2012. She was appointed as Lecturer in October 2012 and Associate Professor in October 2018 at UCL School of Pharmacy. Prof Orlu's research focuses on the development of advanced drug delivery systems, with a strong emphasis on improving therapeutic outcomes for special patient populations.

Abdul Basit graduated with a First Class Honours degree in Pharmacy in 1993 from the University of Bath. Abdul undertook his pharmacy pre-registration training at Pfizer and became a registered pharmacist in 1994. He joined the School of Pharmacy, University of London (now UCL School of Pharmacy of University College London) in 1994 where he completed his Ph.D. in Pharmaceutics. Since 2010 Abdul holds the position of Professor of Pharmaceutics at the UCL School of Pharmacy and is internationally leading in the field of gastroenterology, oral drug delivery and innovative technologies such as three-dimensional (3D) printing of medicines. Across his career, Abdul has received prestigious awards from the AAPS (Young Investigator Award in Pharmaceutics and Pharmaceutical Technology), Glaxo Smith Kline (Innovative Science Award), AstraZeneca (Pharmaceutical Science Award) and the Academy of Pharmaceutical Sciences (APS Science Award). Abdul was appointed the European Editor of the International Journal of Pharmaceutics. In 2019 and 2020, Abdul was listed amongst the World's Most Highly Influential Researchers by the Web of Science.

Andreas Demosthenous received the B.Eng. degree in electrical and electronic engineering from the University of Leicester, Leicester, U.K., the M.Sc. degree in telecommunications technology from Aston University, Birmingham, U.K., and the Ph.D. degree in electronic and electrical engineering from University College London (UCL), London, U.K., in 1992, 1994, and 1998, respectively. He is currently a Professor with the Department of Electronic and Electrical Engineering, UCL, where he leads the Bioelectronics Group. Dr Demosthenous is a fellow of the Institution of Engineering and Technology (IET), a fellow of the European Alliance for Medical and Biological Engineering Sciences (EAMBES), and a Chartered Engineer (CEng).

Miguel Rodrigues obtained the undergraduate degree in Electrical and Computer Engineering from the Faculty of Engineering of the University of Porto, Portugal and the Ph.D. degree in Electronic and Electrical Engineering from University College London. He is currently a Professor of Information Theory and Processing at University College London. Dr. Rodrigues's research lies in the general areas of information theory, information processing, and machine learning. His most relevant contributions have ranged from the information-theoretic analysis and design of communications systems, information-theoretic security, information-theoretic analysis and design of sensing systems, and the information-theoretic foundations of machine learning.