Borrowing Treasures from Neighbors: In-Context Learning for Multimodal Learning with Missing Modalities and Data Scarcity

Zhuo Zhi [1] **Ziquan Liu** [2] **Moe Elbadawi** [3] **Adam Daneshmend** [4] **Mine Orlu** [5] **Abdul Basit** [5]
Andreas Demosthenous [1] **Miguel Rodrigues** [1]

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 method demonstrates an average improvement of 6.1% over a recent strong baseline across various datasets and missing states. Notably, our method also reduces the performance gap between full-modality and missing-modality data compared with the baseline.

[1]Department of Electronic and Electrical Engineering, University College London, London, UK [2]School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK [3]School of Biological and Behavioural Sciences, Queen Mary University of London, London, UK [4]University College London Hospitals NHS Foundation Trust, London, UK [5]UCL School of Pharmacy, University College London, London, UK. Correspondence to: Zhuo Zhi <zhuo.zhi.21@ucl.ac.uk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

1. Introduction

Humankind leverages multimodal data to make intelligent decisions, such as vision, language and sound (Baltrušaitis et al., 2018). 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 (Hayat et al., 2022) and malicious content detection (Kiela et al., 2020). However, a major challenge in the application of multimodal ML is the missing-modality issue (Suo et al., 2019; Tsai et al., 2018; Suo et al., 2019; Ma et al., 2022), 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 (Johnson et al., 2019), are more expensive and/or time-consuming to obtain than others, e.g., Electronic Health Records (EHRs). (Johnson et al., 2023). 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 (Vaswani et al., 2017; Kim et al., 2021), multimodal learning relies on explicit information fusion with features, the output of modality-dependent backbones. Thus, some research work (Ma et al., 2021) 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 (Radford et al., 2023). A recent work (Lee et al., 2023) proposes 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 (Ma et al., 2021; Lee et al., 2023), but the sample size is not always sufficient in the real-world (Huang et al., 2021). Sec-

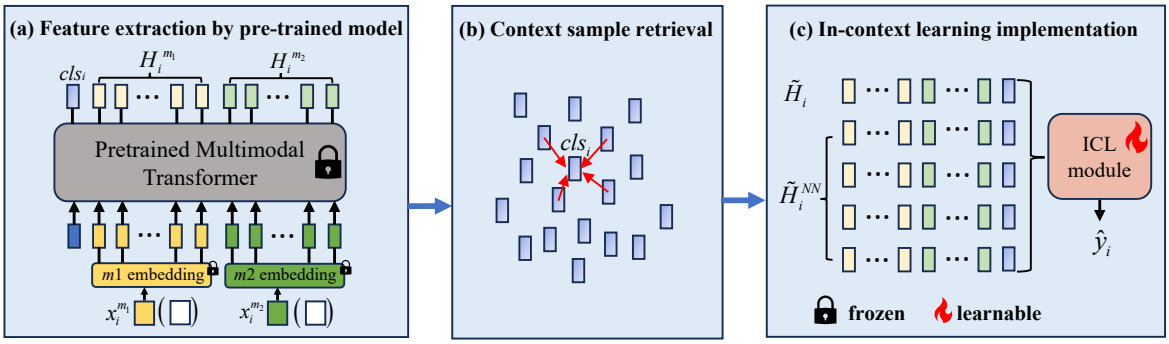

Figure 1. The overview of the proposed method. (**a**) Assuming that each sample contains data with 2 modalities $x_i^{m_1}$ and $x_i^{m_2}$, we get the feature $H_i = ({H_i}^{m_1}, {H_i}^{m_2}, cls_i)$ of the sample by using a pre-trained multimodal transformer, note that $x_i^{m_1}$ or $x_i^{m_2}$ 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.

ondly, there is a notable absence of analysis 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 (Vale-Silva & Rohr, 2021; Wang et al., 2020), 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) (Borgeaud et al., 2022; Ram et al., 2023), 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. Specifically, we train an ICL module on top of the features of a frozen multimodal transformer, such as ViLT (Kim et al., 2021), whose context is full-modality data retrieved from the training set using the cosine-based similarity measure. In cases where modalities are missing, their features are integrated with similar full-modality features, enabling the model to implicitly infer the absent modality for enhanced performance on the target task. For full-modality data, feature refinement is performed using neighbor information to optimize 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 Sec. 3.1. 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 averaged performance gain on four datasets over the baseline MAP in the low-data regime is 6.1%.

The remainder of this paper is organised as follows. Sec. 2 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. Sec. 3 shows the experiment results and the anal-

ysis. Finally, Sec. 4 summarizes this paper, its limitations and future directions. The overview of related research are introduced in Appendix A.

2. 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.

2.1. Problem Setting

We consider the multimodal learning problem with a dataset $\mathcal{D}$ containing multimodal input samples, $\mathcal{D}$ can be the training/validation/testing dataset. For notation simplicity, we assume there are two modalities in the dataset, i.e., $\mathcal{D} = \{x_i^{m_1}, x_i^{m_2}, y_i\}_{i=1}^N$ where the label $y_i \in \{1, \cdots, K\}$, 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 (Johnson et al., 2023) and expensive human-expert annotations (Yang et al., 2023). 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 (Wang et al., 2023b; Zhang et al., 2022a).

Parametric Approach to Mitigating Missing Modalities. As described in Appendix A, parametric methods are employed widely for handling missing modalities in multimodal learning. We introduce this type of approach using a representative work–missing aware prompt (MAP)(Lee et al., 2023). In this method, empty prompt tokens are initialized and concatenated with the input sequences for each layer of the multimodal transformer. In this way, the prompts for the current layer can interact with the prompt tokens inheriting from previous layers, and thus learn more effective instructions for the model prediction. Only the parameters of the prompts and the final classification layer are updated during training.

2.2. Empirical Observations in the Low-Data Regime

Fig. 2-a1 and 2-b1 show the learning curve of 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 (Wang et al., 2020), our work contributes by unveiling this phenomenon in the training of a multimodal transformer model, instead of the joint encoder training in (Wang et al., 2020). 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 2-b2, where the generalization of both data types is improved.

2.3. 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. Tab. 8 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 $f_{\theta_{\text{X}}}$ infers the feature H_i including features for $x_i^{m_1}$, $x_i^{m_2}$ and a CLS feature cls_i by $H_i = \{{H_i}^{m_1}, {H_i}^{m_2}, cls_i\} = f_{\theta_{\text{X}}}(x_i^{m_1}, x_i^{m_2}), {H_i}^{m_1} \in \mathbb{R}^{L_1 \times d}, {H_i}^{m_2} \in \mathbb{R}^{L_2 \times d}, cls_i \in \mathbb{R}^{1 \times d}$, 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 $f_{\theta_{\text{c}}}$ to predict the labels $\hat{y}_i = f_{\theta_{\text{c}}}({H_i}^{m_1}, {H_i}^{m_2}, cls_i)$. When there is missing-modality data, we follow MAP (Lee et al., 2023) to use default tokens to fill those missing tokens, see Sec. 3 for details.

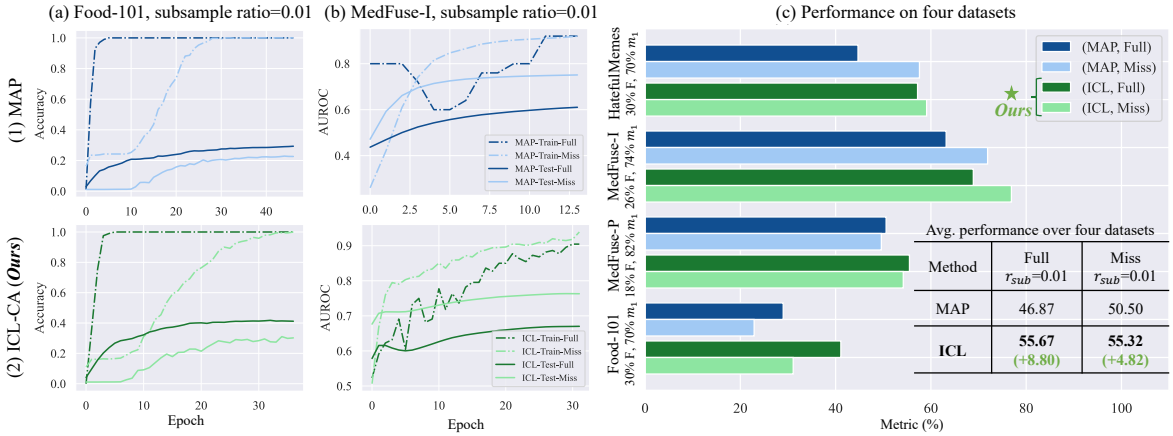

Figure 2. (a) The learning curve of ICL-CA (ours) and Missing-Aware Prompt (MAP) (Lee et al., 2023) on the Food-101 dataset in the low data regime. (b) The learning curve of two methods on the MedFuse-I dataset. The subsampling ratio is set to be 0.01. The difference in the learning steps is due to early stopping. 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 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 (Kiela et al., 2020) and MedFuse-I (Hayat et al., 2022), 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) (Bossard et al., 2014), 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 Sec. 3.

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* $\tilde{H}_i^{NN}$.

2.4. 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 (Brown et al., 2020; Dong et al., 2022). 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.

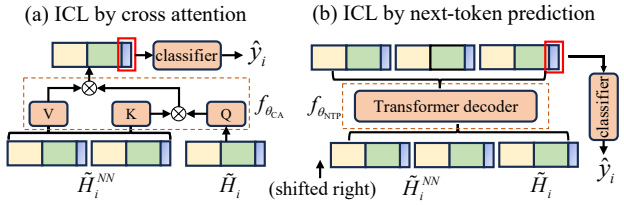

Figure 3. The illustration of two ICL approaches. (a) ICL by cross attention. (b) ICL by next-token prediction. The yellow and green tokens denote features of two different modalities and the blue token is the cls token.

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.,

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

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 B.

ICL by next-token prediction (ICL-NTP). Another way to apply ICL is to implement the next-token prediction by

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 C 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 MAP method by an average of 6.1% over all datasets.

Datasets	Missing state	Metric	ICL-CA	ICL-NTP	FT-A	FT-C	MAP
MedFuse-I	26% F, 74% m_1	AUROC	**0.750**	0.737	0.719	0.702	0.691
		AUPRC	**0.308**	0.286	0.257	0.269	0.285
MedFuse-P	18% F, 82% m_1	AUROC	**0.556**	0.539	0.504	0.490	0.493
		AUPRC	**0.219**	0.204	0.191	0.189	0.190
Food-101	30% F, 70% m_2	Accuracy	0.312	**0.317**	0.250	0.222	0.222
	30% F, 70% m_1		**0.342**	0.327	0.265	0.243	0.247
	30% F, 35% m_2, 35% m_1		0.281	**0.289**	0.228	0.196	0.186
HatefulMemes	30% F, 70% m_2	AUROC	**0.576**	0.565	0.537	0.542	0.528
	30% F, 70% m_1		**0.577**	0.576	0.548	0.540	0.531
	30% F, 35% m_2, 35% m_1		**0.593**	0.583	0.539	0.532	0.529
Average	N.A.	N.A.	**0.451**	0.442	0.404	0.393	0.390

transformer decoder, which is shown in Fig. 3(b). Write the input of the transformer decoder $[\tilde{H}_i^{NN}; \tilde{H}_i]$ as

$$[h_{i,1}^{(1)}, \cdots, h_{i,1}^{(T)}, cls_{i,1}; \cdots; h_{i,Q}^{(1)}, \cdots, h_{i,Q}^{(T)}, cls_{i,Q}; \quad (2)$$
$$h_{i,Q+1}^{(1)}, \cdots, h_{i,Q+1}^{(T)}, cls_{i,Q+1}],$$

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 neighbour 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 $f_{\theta^{(\text{NTP})}}$ is trained to predict the next token in an auto-regressive way, i.e.,

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

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

where $\lambda_{\text{NTP}} = 0.1$ is an adjustable hyperparameter to introduce the loss of feature reconstruction. $y_{i,q}$ is the label of the qth neighbour. $\ell_{\text{NTP}}^{(i)}$ 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 B. 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.

3. 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. Please see details about experiment settings in Appendix B.

3.1. Main Results

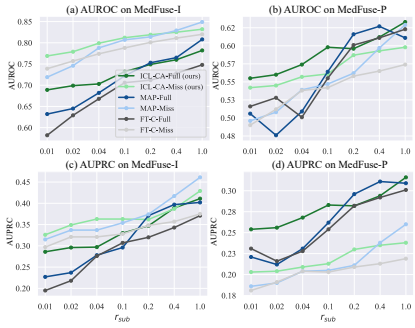

Figure 4. The performance of MAP and ICL-CA on MedFuse-I and MedFuse-P when using different training set sizes. Our proposed ICL method is highly competitive under low data cases (r_{sub} from 0.01 to 0.1). Crucially, our approach enhances the performance in both full- and missing-modalities, outperforming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

Table 1 presents quantitative results across a range of scenarios, where the target dataset's downsampling ratio r_{sub} is 1%. See Appendix C for the performance of all downsampling ratios. From Table 1, Appendix C, Figures 4 and 5, 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 $r_{sub} > 0.1$), FT-A exhibits superior performance, attributed to the update of all parameters in the target domain. MAP follows 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. This trend intensifies as r_{sub} decreases. (2) 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 isn't 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 Tab. 2.

See more details about the ablation study(ICL by Masked Feature Modeling, The Impact of the Number of Neighbors Q, The Effect of Pooled Feature Length T, Groups for retrieving neighboring samples and Inference Time and parameters number) in Appendix D.

Table 2. The relative performance gap between missing-modality and full-modality samples on four datasets. The relative performance gap of all settings is shown in Appendix C, where ICL-CA's averaged relative performance gap (22.1%) is lower than that of MAP (24.1%). For the exception in Food-101, We hypothesize this is because Food-101 is a high-complex multi-classification task that heavily relies on full-modality data, so using full-modality neighbors in our proposed method enhances the performance of full-modality samples and widens the performance gap.

Dataset	HatefulMemes	Food-101	MedFuse-I	MedFuse-P
	30% F, 70% m_1	30% F, 70% m_1		
MAP	28.9%	**26.6%**	13.8%	2.0%
ICL-CA	**3.3%**	33.1%	**11.6%**	**1.8%**

4. 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 6.1% over the baseline in the low-data regime. Furthermore, it effectively narrows the performance disparity caused by modality absence. Our future work will focus on efficient and accurate retrieval methods. We will also extend the method to tasks with more modalities.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41 (2):423–443, 2018.

Bar, A., Gandelsman, Y., Darrell, T., Globerson, A., and Efros, A. Visual prompting via image inpainting. *Advances in Neural Information Processing Systems*, 35: 25005–25017, 2022.

Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., Van Den Driessche, G. B., Lespiau, J.-B., Damoc, B., Clark, A., et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pp. 2206–2240. PMLR, 2022.

Bossard, L., Guillaumin, M., and Van Gool, L. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., and Sui, Z. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

Evci, U., Dumoulin, V., Larochelle, H., and Mozer, M. C. Head2toe: Utilizing intermediate representations for better transfer learning. In *International Conference on Machine Learning*, pp. 6009–6033. PMLR, 2022.

Gupta, T. and Kembhavi, A. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14953–14962, 2023.

Hao, Y., Song, H., Dong, L., Huang, S., Chi, Z., Wang, W., Ma, S., and Wei, F. Language models are general-purpose interfaces. *arXiv preprint arXiv:2206.06336*, 2022.

Hayat, N., Geras, K. J., and Shamout, F. E. Medfuse: Multimodal fusion with clinical time-series data and chest x-ray images. In *Machine Learning for Healthcare Conference*, pp. 479–503. PMLR, 2022.

Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., and Klakow, D. 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, pp. 2545–2568, 2021.

Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O. K., Liu, Q., et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023.

Huang, Y., Du, C., Xue, Z., Chen, X., Zhao, H., and Huang, L. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34:10944–10956, 2021.

Jin, Y., Wang, X., Hao, Y., Sun, Y., and Xie, X. Prototypical fine-tuning: Towards robust performance under varying data sizes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 12968–12976, 2023.

Johnson, A. E., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Mark, R. G., and Horng, S. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

Johnson, A. E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., and Testuggine, D. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33: 2611–2624, 2020.

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

Koh, J. Y., Salakhutdinov, R., and Fried, D. Grounding language models to images for multimodal generation. *arXiv preprint arXiv:2301.13823*, 2023.

Lee, Y.-L., Tsai, Y.-H., Chiu, W.-C., and Lee, C.-Y. Multimodal prompting with missing modalities for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14943–14952, 2023.

Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., and Hoi, S. C. H. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705, 2021.

Liu, Z., Xu, Y., Xu, Y., Qian, Q., Li, H., Ji, X., Chan, A., and Jin, R. Improved fine-tuning by better leveraging pre-training data. *Advances in Neural Information Processing Systems*, 35:32568–32581, 2022.

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

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

Mazurowski, M. A., Habas, P. A., Zurada, J. M., Lo, J. Y., Baker, J. A., and Tourassi, G. D. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural networks*, 21(2-3):427–436, 2008.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

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

Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., and Shoham, Y. In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00083*, 2023.

Suo, Q., Zhong, W., Ma, F., Yuan, Y., Gao, J., and Zhang, A. Metric learning on healthcare data with incomplete modalities. In *IJCAI*, volume 3534, pp. 3540, 2019.

Tsai, Y.-H. H., Liang, P. P., Zadeh, A., Morency, L.-P., and Salakhutdinov, R. Learning factorized multimodal representations. In *International Conference on Learning Representations*, 2018.

Tsimpoukelli, M., Menick, J. L., Cabi, S., Eslami, S., Vinyals, O., and Hill, F. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021.

Vale-Silva, L. A. and Rohr, K. Long-term cancer survival prediction using multimodal deep learning. *Scientific Reports*, 11(1):13505, 2021.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023a.

Wang, S., Yan, Z., Zhang, D., Wei, H., Li, Z., and Li, R. Prototype knowledge distillation for medical segmentation with missing modality. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023b.

Wang, W., Tran, D., and Feiszli, M. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12695–12705, 2020.

Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O. K., Singhal, S., Som, S., et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.

Wang, X., Wang, W., Cao, Y., Shen, C., and Huang, T. 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*, pp. 6830–6839, 2023c.

Wang, X., Zhang, X., Cao, Y., Wang, W., Shen, C., and Huang, T. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023d.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837, 2022.

Xu, P., Zhu, X., and Clifton, D. A. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., and Ni, B. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Zeng, J., Liu, T., and Zhou, J. Tag-assisted multimodal sentiment analysis under uncertain missing modalities. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1545–1554, 2022.

Zhang, C., Chu, X., Ma, L., Zhu, Y., Wang, Y., Wang, J., and Zhao, J. M3care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2418–2428, 2022a.

Zhang, H., Li, G., Li, J., Zhang, Z., Zhu, Y., and Jin, Z. Fine-tuning pre-trained language models effectively by optimizing subnetworks adaptively. *Advances in Neural Information Processing Systems*, 35:21442–21454, 2022b.

Zhang, Z., Zhou, L., Wang, C., Chen, S., Wu, Y., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., et al. Speak foreign languages with your own voice: Cross-lingual neural codec language modeling. *arXiv preprint arXiv:2303.03926*, 2023.

A. Related work

Missing Modality 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 (Hayat et al., 2022), 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 (Lee et al., 2023). The model fails to perform as expected in these situations (Ma et al., 2022). Consequently, much work has been devoted to improving the robustness of multimodal models under modal absence. (Tsai et al., 2018) optimizes a joint generative-discriminative objective for multimodal data and labels which contain the information required for generating data for missed modality. In (Ma et al., 2021), multimodal learning with severely missing modalities (SMIL) is designed to reconstruct the missing modalities using modality priors and Bayesian Meta-Learning. (Lee et al., 2023) introduced two types of missing-aware prompts that can be seamlessly integrated into multimodal transformers. (Ma et al., 2022) proposes a unified strategy based on multi-task optimization to deal with missing modalities in the transformer-based multimodal model. (Zeng et al., 2022) proposes a tag encoding module to assist the transformer's encoder learning with different missing modalities. 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 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 Sec. 3.1. 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 (Xu et al., 2023; Kim et al., 2021; Radford et al., 2021; Li et al., 2021; Wang et al., 2022). 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 (Mazurowski et al., 2008), or, a small amount of text data is available for some low-resource language tasks (Hedderich et al., 2021). Many existing works focus on facilitating transfer learning under low-data situations. (Evci et al., 2022) 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, (Zhang et al., 2022b) 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, (Liu et al., 2022) 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 (Jin et al., 2023), 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 (Dong et al., 2022). Demonstrating versatility, LLMs apply ICL to perform complex tasks, including mathematical reasoning and commonsense answering (Wei et al., 2022). The success of ICL in NLP has recently spurred its adoption in diverse modalities, such as visual (Bar et al., 2022; Wang et al., 2023c;d; Gupta & Kembhavi, 2023), speech (Wang et al., 2023a; Zhang et al., 2023), and multimodal domains (Alayrac et al., 2022; Huang et al., 2023; Hao et al., 2022; Koh et al., 2023; Tsimpoukelli et al., 2021). In the work of (Tsimpoukelli et al., 2021), 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 (Alayrac et al., 2022), 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.

B. Experimental Settings

Datasets. We follow existing works using two-modality datasets for a standard comparison (Lee et al., 2023). Specifically, we use two medical multimodal datasets containing EHR and X-ray images, i.e., MedFuse-In-hospital mortality (MedFuse-I) (Hayat et al., 2022) and MedFuse-Phenotype (MedFuse-P) (Hayat et al., 2022), and two general vision-language datasets (UPMC Food-101 (Bossard et al., 2014) and Hateful Memes (Kiela et al., 2020)) in our experiment. The details of each dataset are as follows:

• *MedFuse-In-hospital mortality (MedFuse-I) (Hayat et al., 2022).* 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 hours 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) (Hayat et al., 2022)* 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.

• *UPMC Food-101 (Bossard et al., 2014).* 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.

• *Hateful Memes (Kiela et al., 2020).* 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.

Baseline. We compare our method with the most recent baseline to tackle the missing-modality issue in multimodal transformers, i.e., missing-aware prompts (MAP) (Lee et al., 2023). We also compare with two methods commonly used in transfer learning: 1) fine-tuning all layers of the pre-trained model on the target dataset (FT-A), and 2) only fine-tuning the classifier of the pre-trained model on the target dataset (FT-C) which is equivalent to removing the ICL module from our proposed method.

Metrics. We set the metrics for each dataset according to the tasks. For MedFuse-I and MedFuse-P, we use the AUROC and AUPRC as evaluation metrics. For Food-101 and Hateful Memes, we use the accuracy and the AUROC respectively.

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 (Lee et al., 2023).

Pretrained Multimodal Transformer. We use the pre-trained multimodal transformer, ViLT (Kim et al., 2021), to extract features following (Lee et al., 2023). 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 (Lee et al., 2023).

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.

Setting of Missing Modalities For MedFuse-I and MedFuse-P, both of them provide the missing setting: 74% and 82% of

patients have missing X-rays, respectively. For the other two tasks, we follow the setting in (Lee et al., 2023) where three kinds of missing states are designed: 1) 30% samples have complete modalities and 70% samples are missing images, 2) 30% samples have complete modalities and 70% samples are missing texts, and 3) 30% samples have complete modalities, 35% samples are missing images and 35% 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. We maintain the original missing-modality rate setting during subsampling. We use r_{sub} to refer to the subsampling ratio in the later content.

C. 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 (Lee et al., 2023). Table 4 presents quantitative results of all test samples across all datasets, methods, and missing states under $r_{sub} \geq 0.1$. Table 5 presents the results under $r_{sub} < 0.1$.

The relative performance gap between missing- and full-modality on all datasets. The main paper presents the relative performance gap, i.e., $(\text{Metric}_{\text{high}} - \text{Metric}_{\text{low}})/\text{Metric}_{\text{low}})$, between missing-modality and full-modality data on four missing states under $r_{sub} = 0.01$. Here we show the details of the performance of missing-modality and full-modality data on all missing states and all datasets in Table 6.

The performance on 3-modalities dataset. We conduct an experiment on the 3-modalities dataset, MOSEI. MOSEI contains more than 23500 sentence utterance videos from more than 1000 online YouTube speakers. The vision, audio and text are involved for sentiment level classification. In this experiment, we set the $r_{sub} < 0.1$ and compare the proposed ICL-CA with FT-A and MAP. We select Mean Absolute Error (MAE) as the metric, where smaller values indicate better perfor025 mance. m1, m2, and m3 refer to the text, vision, and audio modalities, respectively. The results in Table 3 demonstrate the superiority of our proposed ICL-CA method.

Table 3. The performance of the proposed ICL-CA, FT-A and MAP under severe data scarcity and missing modalities on MOSEI dataset.

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**
FT-A	70% m_1 missed	1.092	0.938	0.864
	70% m_2 missed	0.996	0.904	0.860
	70% m_3 missed	1.114	0.993	0.891
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

D. Ablation study

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 7. 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 7, 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

Table 4. 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 exhibits superior performance, attributed to the update of all parameters in the target domain. MAP 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	FT-A	FT-C	MAP
0.1	Medfuse-I	26% F, 74% $m1$	AUROC	**0.793**	0.789	0.790	0.771	0.788
			AUPRC	0.352	0.355	**0.356**	0.321	0.338
	Medfuse-P	18% F, 82% $m1$	AUROC	0.578	0.565	**0.592**	0.556	0.561
			AUPRC	0.234	0.224	**0.246**	0.219	0.224
	Food101	30% F, 70% $m2$	Accuracy	**0.595**	0.576	0.562	0.417	0.463
		30% F, 70% $m1$		**0.625**	0.619	0.603	0.450	0.489
		30% F, 35% $m1$, 35% $m2$		**0.571**	0.566	0.535	0.409	0.453
	Hateful Memes	30% F, 70% $m2$	AUROC	**0.607**	0.598	0.601	0.577	0.585
		30% F, 70% $m1$		**0.617**	0.612	0.609	0.575	0.586
		30% F, 35% $m1$, 35% $m2$		**0.618**	**0.618**	0.614	0.579	0.599
0.2	Medfuse-I	26% F, 74% $m1$	AUROC	0.802	0.792	**0.832**	0.782	0.801
			AUPRC	0.353	0.370	**0.411**	0.339	0.370
	Medfuse-P	18% F, 82% $m1$	AUROC	0.590	0.580	**0.651**	0.576	0.581
			AUPRC	0.243	0.231	**0.290**	0.232	0.237
0.4	Medfuse-I	26% F, 74% $m1$	AUROC	0.810	0.806	**0.840**	0.793	0.815
			AUPRC	0.388	0.399	**0.451**	0.351	0.410
	Medfuse-P	18% F, 82% $m1$	AUROC	0.602	0.593	**0.688**	0.583	0.609
			AUPRC	0.251	0.242	**0.325**	0.237	0.259
1.0	Medfuse-I	26% F, 74% $m1$	AUROC	0.820	0.819	**0.850**	0.804	0.838
			AUPRC	0.420	0.431	**0.490**	0.372	0.444
	Medfuse-P	18% F, 82% $m1$	AUROC	0.611	0.596	**0.704**	0.591	0.630
			AUPRC	0.261	0.244	**0.347**	0.243	0.273

Table 5. 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	FT-A	FT-C	MAP
0.01	Medfuse-I	26% F, 74% $m1$	AUROC	**0.750**	0.737	0.719	0.702	0.691
			AUPRC	**0.308**	0.286	0.257	0.269	0.285
	Medfuse-P	18% F, 82% $m1$	AUROC	**0.556**	0.539	0.504	0.490	0.493
			AUPRC	**0.219**	0.204	0.191	0.189	0.190
	Food101	30% F, 70% $m2$	Accuracy	0.312	**0.317**	0.250	0.222	0.222
		30% F, 70% $m1$		**0.342**	0.327	0.265	0.243	0.247
		30% F, 35% $m1$, 35% $m2$		0.281	**0.289**	0.228	0.196	0.186
	Hateful Memes	30% F, 70% $m2$	AUROC	**0.576**	0.565	0.537	0.542	0.528
		30% F, 70% $m1$		**0.577**	0.576	0.548	0.540	0.531
		30% F, 35% $m1$, 35% $m2$		**0.593**	0.583	0.539	0.532	0.529
0.02	Medfuse-I	26% F, 74% $m1$	AUROC	0.761	**0.764**	0.754	0.728	0.722
			AUPRC	**0.328**	**0.328**	0.299	0.293	0.308
	Medfuse-P	18% F, 82% $m1$	AUROC	**0.559**	0.552	0.530	0.524	0.522
			AUPRC	**0.221**	0.212	0.207	0.198	0.198
	Food101	30% F, 70% $m2$	Accuracy	**0.397**	0.373	0.352	0.287	0.302
		30% F, 70% $m1$		**0.402**	0.389	0.338	0.310	0.295
		30% F, 35% $m1$, 35% $m2$		**0.387**	0.355	0.315	0.262	0.278
	Hateful Memes	30% F, 70% $m2$	AUROC	**0.590**	0.581	0.550	0.545	0.545
		30% F, 70% $m1$		**0.593**	0.587	0.570	0.556	0.557
		30% F, 35% $m1$, 35% $m2$		0.602	**0.603**	0.564	0.549	0.548
0.04	Medfuse-I	26% F, 74% $m1$	AUROC	0.778	0.777	**0.787**	0.752	0.767
			AUPRC	**0.344**	0.336	0.328	0.308	0.319
	Medfuse-P	18% F, 82% $m1$	AUROC	**0.569**	0.557	0.561	0.542	0.545
			AUPRC	**0.228**	0.219	0.213	0.209	0.211
	Food101	30% F, 70% $m2$	Accuracy	**0.494**	0.464	0.448	0.352	0.397
		30% F, 70% $m1$		**0.508**	0.489	0.458	0.391	0.405
		30% F, 35% $m1$, 35% $m2$		**0.484**	0.460	0.422	0.338	0.387
	Hateful Memes	30% F, 70% $m2$	AUROC	**0.595**	0.591	0.583	0.561	0.558
		30% F, 70% $m1$		**0.600**	0.585	0.577	0.559	0.567
		30% F, 35% $m1$, 35% $m2$		0.601	**0.607**	0.579	0.562	0.574

Table 6. 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 (22.1%) is lower than that of MAP (24.1%).

			ICL-CA			MAP		
Dataset	Missing state	Metric	missing modality	full modality	relative gap (%)	missing modality	full modality	relative gap (%)
Medfuse-I	26% F, 74% $m1$	AUROC	0.769	0.689	**11.6**	0.719	0.632	13.8
		AUPRC	0.326	0.286	**14.0**	0.337	0.237	42.2
Medfuse-P	18% F, 82% $m1$	AUROC	0.545	0.555	**1.8**	0.496	0.506	2.0
		AUPRC	0.203	0.254	25.1	0.186	0.221	**18.8**
Food101	30% F, 70% $m2$	Accuracy	0.298	0.344	15.4	0.215	0.240	**11.6**
	30% F, 70% $m1$		0.311	0.414	33.1	0.229	0.290	**26.6**
	30% F, 35% $m1$, 35% $m2$		0.214	0.435	103.3	0.152	0.251	**65.1**
Hateful Memes	30% F, 70% $m2$	AUROC	0.577	0.578	**0.2**	0.521	0.551	5.8
	30% F, 70% $m1$		0.591	0.572	**3.3**	0.576	0.447	28.9
	30% F, 35% $m1$, 35% $m2$		0.569	0.646	**13.5**	0.478	0.601	25.7
Average					**22.1**			24.1

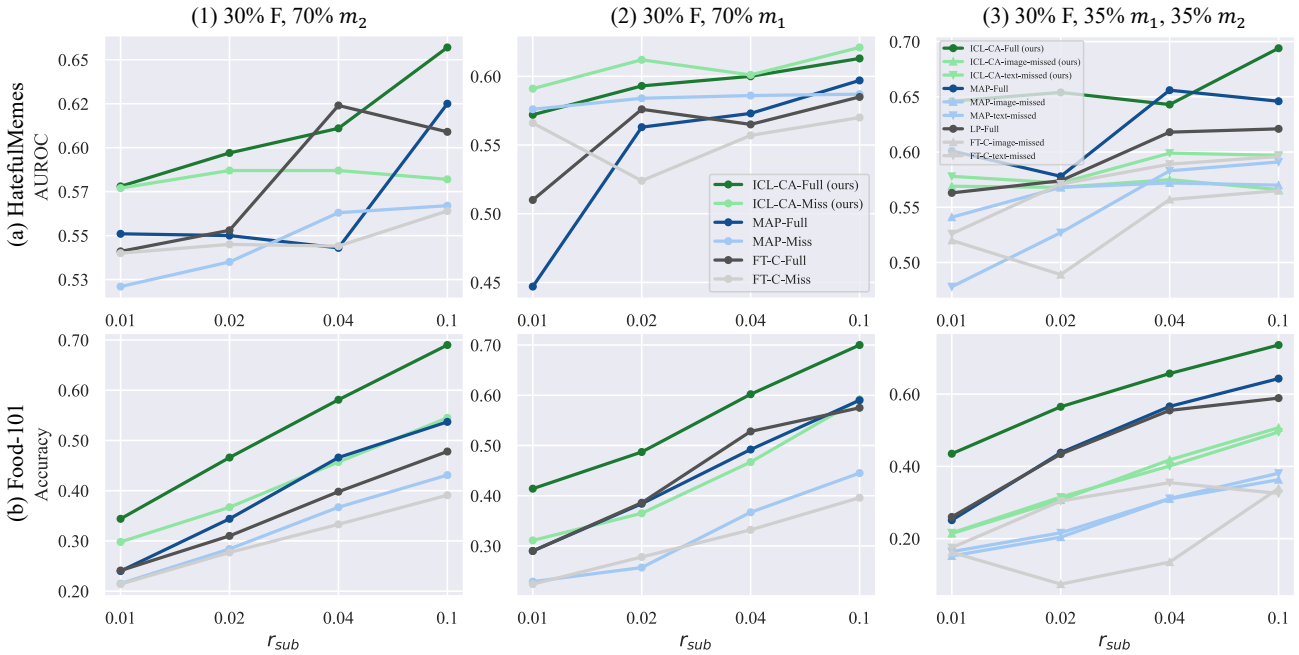

Figure 5. The performance of MAP and ICL-CA on HatefulMemes and Food-101 when using different training set sizes. The performance of our ICL-CA is much better than that of MAP in the low-data regime (r_{sub} from 0.01 to 0.1).

Table 7. 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
	AUPRC	**0.308**	0.286	0.293	0.285	0.352	**0.355**	0.349	0.338
MedFuse-P	AUROC	**0.556**	0.539	0.526	0.493	**0.578**	0.565	0.569	0.561
	AUPRC	**0.219**	0.204	0.199	0.190	**0.234**	0.224	0.228	0.224
Food-101 30% F, 70% $m1$	Accuracy	**0.342**	0.327	0.326	0.247	**0.625**	0.619	0.564	0.489
HatefulMemes 30% F, 70% $m1$	AUROC	**0.577**	0.576	0.561	0.531	**0.617**	0.612	0.608	0.586

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. 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 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

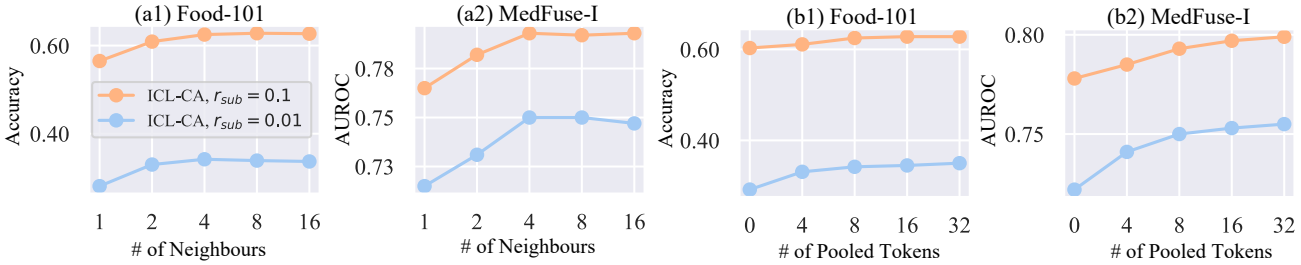

Figure 6. Comparison of the effect of the number of neighbors and the pooled feature length in the ICL-CA model. (a) Comparison of the effect of the number of neighbors. (b) Comparison of the effect of pooled feature length. We suggest setting the number of neighbors to 4 and the pooled feature length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

method under $r_{sub} = 0.01$ for this experiment, as shown in Table 8. 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.

Table 8. 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
		AUPRC	0.269	**0.308**	0.248
MedFuse-P	18% F, 82% m_1	AUROC	0.533	**0.556**	0.520
		AUPRC	0.191	**0.219**	0.184
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
HatefulMemes	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
Average	N.A.	N.A.	0.432	**0.451**	0.414

Inference Time and parameters number. 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 34.41ms with a std of 4.52ms. In contrast, MAP has a mean inference time of 40.59ms and a std of 5.40ms. The difference in inference time is because MAP has a larger number of tokens (missing-aware prompts) in the transformer. We also calculate the number of trainable parameters in Table 9. Although our method has more parameters than MAP, it is less overfitting to the training data in the low-data regime compared with MAP.

Table 9. Number of trainable parameters for different models.

Model	MAP	ICL-CA	FT-A	FT-C
num/million	2.8	5.1	113	0.6