PiCME: Pipeline for Contrastive Modality Evaluation and Encoding in the MIMIC Dataset

Michal Golovanevsky[§], Pranav Mahableshwarkar[§], Ritambhara Singh [‡ †*]

Department of Computer Science, Brown University,
[‡] and Center for Computational Molecular Biology, Brown University
Providence, RI 02912, USA
[§] These authors contributed equally.
[†] E-mail: ritambhara@brown.edu

Carsten Eickhoff*

School of Medicine,
Institute for Bioinformatics and Medical Informatics,
University of Tübingen,
Tübingen, 72074, Germany
E-mail: c.eickhoff@acm.org
**Co-corresponding authors*

Multimodal deep learning can improve clinical prediction by integrating diverse patient data, and contrastive learning aids this by producing reusable representations that transfer across tasks. Yet in clinical settings its use remains largely limited to image-text pairs. We propose the Pipeline for Contrastive Modality Evaluation and Encoding (PiCME), which systematically assesses five clinical data types from MIMIC: discharge summaries, radiology reports, chest X-rays, demographics, and time-series. We pre-train contrastive models on all 26 combinations of two to five modalities, analyze representation alignment, and carry the strongest-aligning combinations forward to downstream evaluation on in-hospital mortality and phenotype prediction. To address performance plateaus as modalities are added, we introduce a Modality-Gated LSTM that weights each modality by its contrastively learned importance. Contrastive models are competitive with supervised baselines for binary mortality prediction at two and three modalities, but are outperformed on multi-label phenotyping and at higher modality counts, with alignment and transfer performance degrading as modalities are added while supervised models plateau. The Modality-Gated LSTM mitigates this decline, improving AUROC from 73.19% to 76.93% and AUPRC from 51.27% to 62.26% in the five-modality setting. We further validate modality importance against attribution scores and assess generalization across demographic subgroups. PiCME is the first to scale contrastive learning across all modality combinations in MIMIC, offering insights into modality selection and training strategy toward more efficient clinical prediction. *

Keywords: Multimodal deep learning; contrastive learning; clinical decision support

*Code and Appendix are available at `https://github.com/rsinghlab/PiCME`

© 2026 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

1. Introduction

1.1. *Background and Significance*

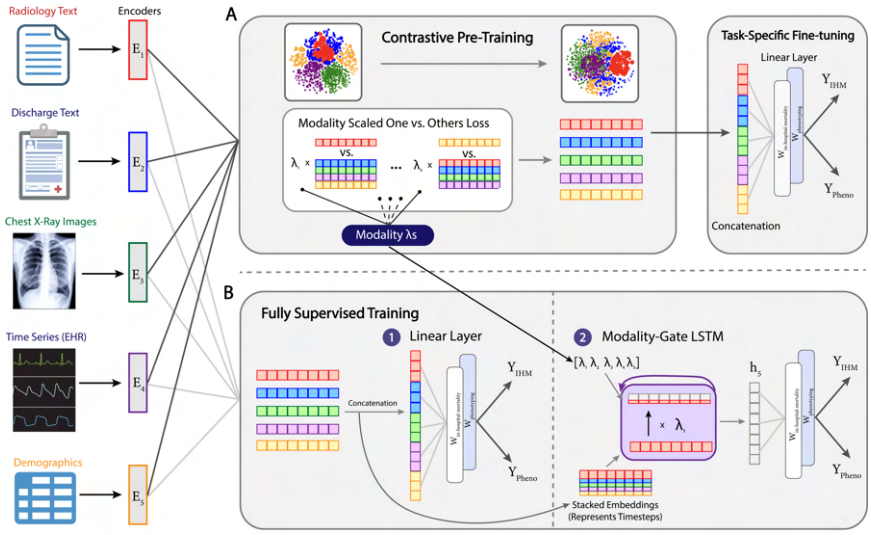


Fig. 1. Overview of the MIMIC Pipeline for Contrastive Modality Evaluation (PiCME). Panel (A): Contrastive learning with One-vs-Others loss, followed by fine-tuning for in-hospital mortality and phenotyping. Panel (B): Fully supervised fine-tuning: (B.1) task-specific modality encoders and (B.2) a modality gate using contrastively learned importance weights to regularize five-modality fusion.

The availability of large-scale, de-identified clinical datasets has fueled data-driven medical research. The MIMIC dataset[1,2] is among the most comprehensive de-identified clinical resources, containing demographics, time-series, notes, and labs, with MIMIC-CXR[3] adding chest X-rays to enable multimodal prediction. Contrastive learning has emerged as an effective framework for multimodal representation learning, producing reusable embeddings that transfer across downstream tasks.[4–6] Building on this success, contrastive learning has gained traction in the medical domain,[7,8] including in MIMIC, where it has been applied to integrate image and text modalities.[9–11]

Most contrastive studies are limited to two modalities, typically image and text, with no systematic evaluation of all combinations. Clinical modalities such as time-series and tabular data remain underexplored, and suboptimal modality choices can yield inconsistent results and higher computational cost.[12]

To address these gaps, we propose the **Pipeline for Contrastive Modality Evaluation and Encoding (PiCME)**, a framework for evaluating modality importance in contrastive learning. We systematically evaluate five clinical data types from MIMIC-IV[2] and MIMIC-CXR:[3] discharge summaries (filtered to admission-time content; Appendix 11), clinical time-series, chest X-rays, radiology reports, and tabular demographics (Figure 1). Our goal is to identify scalable, clinically effective strategies for multimodal integration while providing insight into modality importance. We are the first to train contrastive models on all 26 two- to five-modality combinations. We then select the combinations with the strongest representa-

tion alignment for downstream evaluation on two clinical tasks: in-hospital mortality (binary classification) and phenotype prediction (25-label multi-label classification). Contrastive learning peaks with three modalities—discharge notes, chest X-rays, and demographics—achieving 75.94% AUROC on in-hospital mortality. The supervised baseline also peaks with the same three modalities in phenotyping, reaching 73.84% AUROC, though supervised IHM performance continues to improve with additional modalities. Contrastive transfer performance declines beyond three modalities, likely due to integration noise and feature redundancy, whereas fully supervised models continue to improve modestly but plateau in the four- and five-modality settings.

To overcome this multimodal performance plateau, we introduce the *Modality-Gated Long Short-Term Memory (mLSTM)*, which uses contrastively learned *modality-*λ values to dynamically weight modalities during supervised fine-tuning (Figure 1A). This mechanism enables the model to focus on the most informative modalities. Using all five modalities, the mLSTM achieves the strongest overall performance among evaluated configurations, reaching 76.93% AUROC and 62.26% AUPRC on IHM and 74.36% AUROC and 51.21% AUPRC on phenotyping. Gains over the best three-modality contrastive model are modest in IHM AUROC but substantially larger in AUPRC and in phenotyping.

Beyond predictive performance, we assess the interpretability and robustness of the learned modality weights. Using Integrated Gradients, we validate that the contrastively learned *modality-*λ values align with input attributions during fine-tuning, revealing which modalities drive model predictions. To assess fairness and generalization, we examine performance across demographic subgroups including age, gender, and ethnicity.

Overall, our work provides a comprehensive pipeline for evaluating multimodal integration in MIMIC, offering insights into modality selection, training strategy, and interpretability. This is increasingly important as clinical deep learning systems incorporate diverse data modalities, creating a need for scalable models that generalize across patient populations. Moreover, our results demonstrate that information learned during contrastive pre-training can improve downstream supervised fine-tuning through modality-aware regularization.

2. Related Works

Multimodal learning excels in vision-language tasks[5,6,13] and clinical decision support.[7,14,15] Contrastive learning, particularly CLIP,[16] remains central to multimodal integration, with continued advancements.[4–6] Despite the success of multimodal models, contrastive learning is rarely applied beyond the two-modality setting, primarily integrating vision and language.[7,8,10] This limitation stems from the pairwise design of the InfoNCE loss, which aligns positive pairs while separating negatives. To avoid a quadratically growing number of contrastive loss pairs that fail to capture inter-modality relationships, a loss function that optimizes across all modalities simultaneously is needed. One successful approach is the One-Versus-Others (OvO) loss, which contrasts each modality against the average of the remaining modalities and thereby extends InfoNCE to many modalities in a scalable manner.[17] Related multimodal contrastive work discusses and compares against one-versus-rest style objectives.[18] Our contribution is not the leave-one-out structure itself but the introduction of learnable modality-importance

weights on top of it, which we then reuse for downstream fusion. Given these advantages, OvO loss is well suited for contrastive training on highly multimodal datasets such as MIMIC.

The MIMIC dataset includes MIMIC-IV (demographics, discharge notes, and EHR data) and MIMIC-CXR (chest X-rays and reports). Given its diverse data types, MIMIC-based multimodal studies vary in modality selection, integration methods, and learning paradigms.[19] use supervised learning for chest X-rays and notes, while,[9],[10] and[11] apply contrastive learning to image-text pairs, and[20] contrastively pre-trains on time-series and notes for mortality and phenotyping; all are limited to two modalities.[14] incorporates an LSTM to handle missing modalities, improving multimodal fusion. Expanding beyond two modalities, HAIM[21] integrates demographics, time series, natural language, and medical images, evaluating all combinations of four modalities with Shapley attribution across twelve tasks. HAIM establishes the value of exhaustive modality-combination analysis in a fully supervised setting; PiCME is complementary in that it conducts this analysis for multimodal *contrastive* pre-training, adds a fifth modality, and studies how modality importance learned during pre-training transfers to supervised fusion.

Despite the success of previous MIMIC studies, these works rely on arbitrary modality selection, rarely evaluate contrastive learning beyond two modalities, and seldom examine how modality importance differs between pre-training and fine-tuning. We address these gaps by evaluating all modality configurations and analyzing contributions across both stages.

3. Methods

3.1. *Modality Encoders*

We integrate five modalities to enhance clinical classification: discharge summaries, radiology reports, chest X-rays, clinical time-series, and demographic data. To ensure comparability with prior research, we adopt encoder architectures that have shown strong performance in previous MIMIC frameworks.[9,11,14,21]

Discharge summaries capture the clinical narrative of an admission, while radiology reports contain expert image interpretations. For the in-hospital mortality task, all modalities are restricted to information available within the first 48 hours of the ICU stay: discharge summaries are filtered to remove all hospitalization-course, discharge, and outcome-related sections, retaining only admission-time content (Appendix 11). We encode both using ClinicalBERT,[22] a BERT[23] variant pre-trained on clinical text and widely used in multimodal MIMIC studies.[21,24] To improve efficiency, we fine-tune ClinicalBERT using LoRA, a low-rank adaptation method that limits the number of trainable parameters.[25]

Chest X-ray images are encoded with a ResNet-based model,[26] commonly used in MIMIC-CXR for strong feature extraction.[14] Clinical time-series data, consisting of ICU-recorded vitals and lab results, is modeled using an LSTM,[27] which captures temporal patterns in sequential medical data.[14,21] Demographic data, including age, gender, and ethnicity, is encoded using a multi-layer perceptron (MLP), a standard approach for structured features.[21]

During contrastive pre-training, encoders align multimodal features via self-supervised objectives and are later fine-tuned for downstream tasks. For fully supervised baselines, all encoders are trained from scratch to establish a non-contrastive baseline.

3.2. *Unsupervised Contrastive Pre-Training*

We use two contrastive objectives: pairwise InfoNCE and the multimodal One-Versus-Others (OvO) loss,[17] both self-supervised methods that pull same-patient embeddings together while pushing others apart. For exactly two modalities, we use the standard InfoNCE loss, whereas for three or more modalities, we adopt OvO loss, a non-pairwise method that creates one loss term per modality. In the two-modality case, OvO reduces to InfoNCE since there are only two modalities, leaving no additional modalities to average over when forming the "others" representation. We additionally introduce a learnable weighting scheme in the OvO loss to account for modality importance dynamically.

3.2.1. *Pairwise Contrastive Learning with InfoNCE*

The InfoNCE loss,[28] used in SimCLR,[29] MoCo,[30] and CLIP,[31] pulls together positive pairs (views of the same sample) while pushing apart negatives.

For each sample k in a batch, let x_k^i and x_k^j represent the embeddings from two different modalities, i and j, respectively. The pairwise contrastive prediction loss is defined as:

$$\mathcal{L}_{\text{pair}}^{i,j,k} = -\log \frac{\exp(\text{sim}(x_k^i, x_k^j)/\tau)}{\sum_{m=1}^{N} \exp(\text{sim}(x_k^i, x_m^j)/\tau)} \tag{1}$$

where N is the batch size, τ is a trainable temperature parameter, and $\text{sim}(\cdot,\cdot)$ represents cosine similarity. The final InfoNCE loss for a batch sums over all modality pairs (i, j) across all samples:

$$\mathcal{L}_{\text{InfoNCE}} = \sum_{i,j} \sum_{k=1}^{N} \mathcal{L}_{\text{pair}}^{i,j,k}. \tag{2}$$

As the number of modalities increases, each loss term captures only pairwise relationships rather than interactions among multiple modalities. To address this limitation, we adopt the One-Versus-Others loss, which extends contrastive learning to settings with more modalities.

3.2.2. *Multimodal Contrastive Learning with One-Versus-Others (OvO) Loss*

For three or more modalities, we move beyond the pairwise paradigm and instead use the One-Versus-Others (OvO) loss, also known as One-Versus-Rest or Leave-One-Out Loss.[17,18] This approach has been applied in attention mechanisms[12] and multi-view representation learning.

In contrast to InfoNCE, which enforces similarity between modality pairs, OvO encourages each modality to align with the representation of all remaining modalities. For each sample k and modality i, we define the "other" embedding by averaging over all remaining modalities:

$$\bar{x}_k^{\neg i} = \frac{1}{K-1} \sum_{\substack{j=1 \\ j \neq i}}^{K} x_k^j. \tag{3}$$

The OvO contrastive loss is then defined as:

$$\mathcal{L}_{\text{OvO}}^{i,k} = -\log \frac{\exp(\text{sim}(x_k^i, \bar{x}_k^{\neg i})/\tau)}{\sum_{m=1}^{N} \exp(\text{sim}(x_k^i, \bar{x}_m^{\neg i})/\tau)}. \tag{4}$$

To allow the model to learn the relative importance of each modality dynamically, we extend the OvO loss by introducing modality-specific learnable mixture weights λ_i for each modality i. Rather than treating all modalities equally, the model learns to assign higher importance to informative modalities while down-weighting less relevant ones. The weighted OvO loss is defined as:

$$\mathcal{L}^{i,k}_{\text{weighted-OvO}} = -\lambda_i \log \frac{\exp(\text{sim}(x^i_k, \bar{x}^{\neg i}_k)/\tau)}{\sum_{m=1}^{N} \exp(\text{sim}(x^i_k, \bar{x}^{\neg i}_m)/\tau)}, \tag{5}$$

where each λ_i is a trainable parameter optimized jointly with the contrastive objective. After each step, the λ values are normalized using the softmax function such that $\sum_i \lambda_i = 1$. The final weighted loss is:

$$\mathcal{L}_{\text{weighted-OvO}} = \sum_{i=1}^{K} \sum_{k=1}^{N} \mathcal{L}^{i,k}_{\text{weighted-OvO}}. \tag{6}$$

These weights act as a regularizer, letting the model adjust each modality's contribution to improve robustness. A side ablation isolating the contrastive objective indicates that both the leave-one-out structure and the learnable weights contribute to five-modality transfer performance, with the largest gains on phenotyping (Appendix 12).

By combining standard contrastive learning for two modalities with a modality-weighted OvO approach for three or more modalities, our method captures both pairwise interactions and global multimodal dependencies, ensuring robust representation learning across diverse clinical modalities.

3.3. *Supervised Fine-Tuning*

3.3.1. *Predictive Tasks and Training Objectives*

We evaluate our approach on two clinical tasks: **in-hospital mortality** (IHM) prediction and **phenotyping**. IHM is a binary classification task that predicts whether a patient will survive their hospital stay using data from the first 48 hours of an ICU admission, following.[14] Phenotyping is a multi-label classification task that predicts whether a patient has any of 25 chronic, mixed, or acute care conditions during an ICU stay. For both tasks, we report the Area Under the Receiver Operating Characteristic (AUROC) and Area Under the Precision-Recall (AUPRC) curves, following.[14,21,24] Details on hyperparameter tuning and data splits are provided in Appendix Sections 7 and 8.

3.3.2. *Contrastive Pre-Training for Downstream Tasks*

For contrastive learning, each modality is first encoded into an embedding optimized using the contrastive objective. After pre-training, the frozen encoders are applied to the held-out fine-tuning patients, whose embeddings are used for downstream classification (Figure 1A).

To integrate multiple modalities, we concatenate the contrastively trained embeddings into a unified representation, following common fusion strategies such as HAIM[21] and MedFuse.[14] The resulting representation is

$$x \in \mathbb{R}^{nm} \tag{7}$$

where n is the embedding size and m is the number of modalities. This fused vector is passed to a multi-layer perceptron (MLP) for task-specific prediction.

3.3.3. *Fully Supervised Baseline*

As a comparison, we implement fully supervised baselines for each modality combination. These models omit contrastive pre-training and instead train all modality encoders from scratch. The architecture mirrors the contrastive setup, using the same modality encoders (Section 3.1) and a shared classification head operating on concatenated embeddings (Figure 1B.1).

3.3.4. *Modality-Gated LSTM*

We introduce the Modality-Gated LSTM (mLSTM), an extension of the Long Short-Term Memory (LSTM) network in which the *cell state* is modulated by modality-specific importance weights λ_i learned during contrastive pre-training. This allows the network to scale each modality's contribution according to its learned importance.

Unlike a conventional LSTM, which recurs over time, the mLSTM recurs over modalities. Its input is a fixed-order sequence of projected modality embeddings, one per modality, so that recurrence step i corresponds to modality i under the fixed ordering $(\mathcal{T}_{\mathcal{R}}, \mathcal{T}_{\mathcal{D}}, \mathcal{I}, \mathcal{TS}, \mathcal{D})$; combinations using fewer than five modalities preserve the relative order of the modalities present. The mLSTM retains the standard LSTM gating structure, with an input gate I_i, forget gate F_i, candidate memory $\widetilde{C}_i$, and output gate O_i computed at each step. We additionally apply the scalar λ_i associated with modality i to the candidate memory contribution at that step, where $\sum_{i=1}^{|\Lambda|} \lambda_i = 1$ and $|\Lambda|$ is the number of modalities. With $C_0 = H_0 = \mathbf{0}$:

$$\begin{aligned} \vec{\lambda}^{(i)} &= \langle \lambda_i, \ldots, \lambda_i \rangle, \quad |\vec{\lambda}^{(i)}| = h, \\ C_i &= F_i \odot C_{i-1} + (I_i \odot \widetilde{C}_i) \odot \vec{\lambda}^{(i)}, \\ H_i &= O_i \odot \tanh(C_i). \end{aligned}$$

The final hidden state $H_{|\Lambda|} \in \mathbb{R}^h$, produced after all modalities have been consumed, is passed to a standard MLP for task-specific prediction. The λ_i values are taken from the preceding contrastive pre-training stage and held frozen during supervised training; they may alternatively be specified manually. All remaining mLSTM and MLP parameters, along with the modality encoders, are trained from scratch under task supervision. The modality gates are therefore the only component inherited from contrastive pre-training, which is why the mLSTM is reported alongside the fully supervised models in Table 1 rather than as a contrastive-transfer result.

4. Results

We first analyze how modality combinations align in a shared representation space, observing diminishing returns as modalities are added. We then evaluate downstream performance: contrastive transfer peaks at three modalities and declines, while supervised baselines plateau. The Modality-Gated LSTM addresses this plateau using contrastively learned importance.

4.1. *More Modalities, Less Alignment: Challenges in Task-Agnostic Contrastive Learning*

We use t-SNE to project modality embeddings into two dimensions, as it captures local structure better than PCA or UMAP for visualizing alignment. Figure 3 shows how embeddings from two to five modalities evolve over training epochs. Each point represents an embedding produced by the modality encoders described in Section 3.1. Alignment is strongest in the two- and three-modality settings, whereas four- and five-modality combinations remain more dispersed. Radiology reports, in particular, remain distinct throughout training, suggesting that their domain-specific language is more difficult to align with structured or temporal modalities such as demographics and time-series.

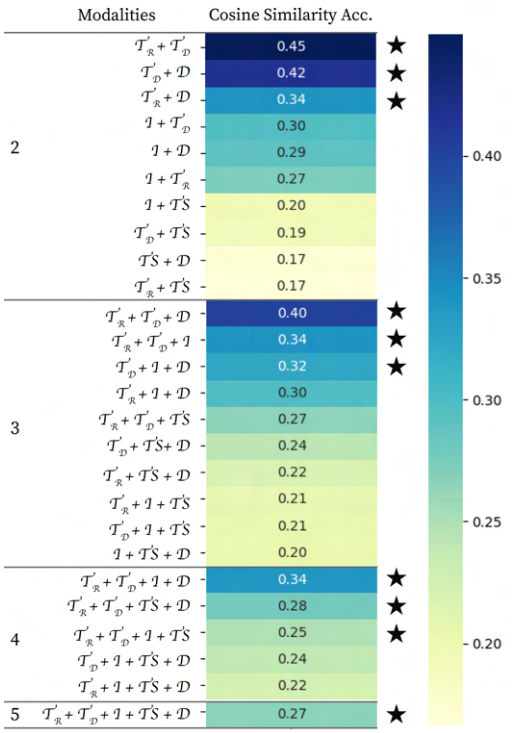


Fig. 2. Top-5 cosine similarity accuracy. 5 Modalities: Radiology Text ($\mathcal{T}_{\mathcal{R}}$), Discharge Text ($\mathcal{T}_{\mathcal{D}}$), Images ($\mathcal{I}$), Demographics ($\mathcal{D}$), and Time-Series ($\mathcal{TS}$).

To complement this qualitative analysis, we compute a top-5 alignment score based on cosine similarity in the shared embedding space. For each embedding, we retrieve nearest neighbors from the full pool of embeddings across all modalities, excluding the query embedding itself. A sample is counted as aligned if at least one of its five nearest neighbors belongs to the same patient. Same-modality neighbors are permitted but not required, so this metric measures patient-level neighborhood consistency in the joint embedding space rather than strict cross-modal retrieval.

Figure 2 shows that top-5 similarity decreases as additional modalities are introduced. The highest accuracy (45%) is achieved using radiology and discharge text. Among three-modality combinations, the best-performing set reaches 40% accuracy with radiology text, discharge notes, and demographics. We hypothesize that the two text modalities benefit from being encoded with ClinicalBERT, providing a shared latent space that facilitates alignment. Demographics also adapts well as its low-dimensional structured representation (e.g., age, sex, ethnicity, and marital status) can be readily mapped into the shared embedding space.

By contrast, time-series data consistently reduce alignment quality. Their high variance and temporal dynamics make them more difficult to reconcile with static or text-based modalities (Appendix 6). When included in two- or three-modality settings, time-series embeddings lower cosine similarity scores, highlighting the difficulty of learning coherent joint representations.

The highest-aligning modality combinations (starred) are selected for downstream fine-tuning, alongside $\mathcal{T}_{\mathcal{R}} + \mathcal{I} + \mathcal{D}$ as a discharge-free control.

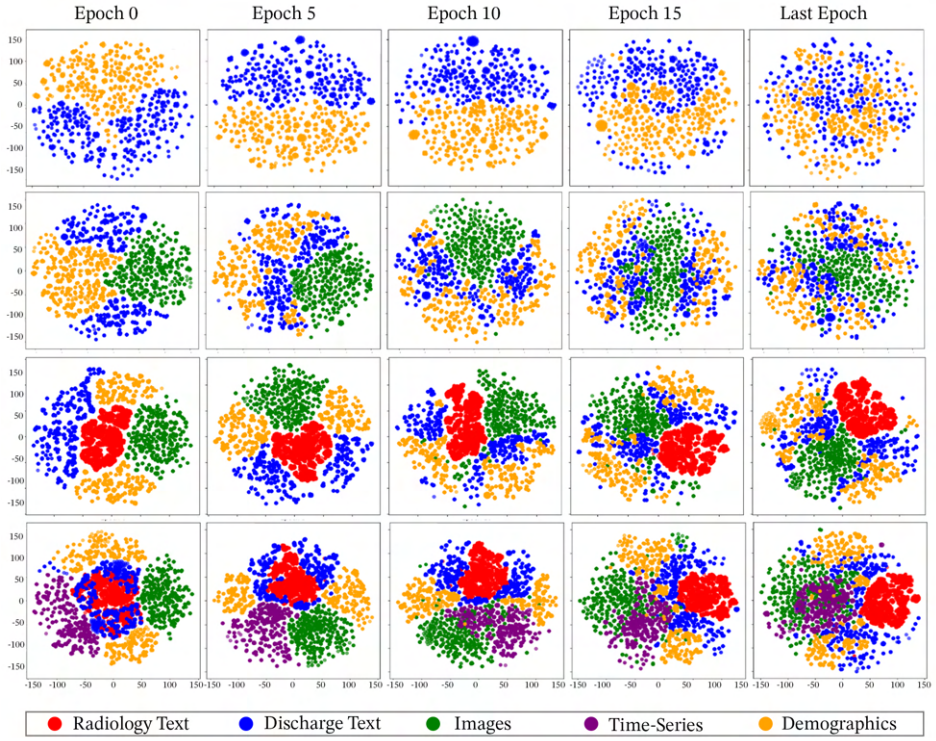


Fig. 3. T-SNE visualization of contrastive learning through epochs.

4.2. *Effectiveness of Contrastive Pre-training and Modality-Gated LSTM on Task-Specific Fine-Tuning*

Table 1 reports AUROC and AUPRC for three settings: contrastive pre-training with frozen encoders, fully supervised training from scratch, and the Modality-Gated LSTM.

Contrastively pre-trained embeddings often match or outperform fully supervised models on IHM in two- and three-modality settings, while supervised training is generally stronger for phenotyping. The strongest contrastive performance is achieved using discharge text, images, and demographics ($\mathcal{T}_\mathcal{D} + \mathcal{I} + \mathcal{D}$), reaching 75.94% AUROC and 58.32% AUPRC for IHM, outperforming all two- and three-modality supervised counterparts. This demonstrates the effectiveness of contrastive learning when integrating a small number of semantically aligned modalities. The advantage is most pronounced for IHM, whereas phenotyping, a 25-label task, is more challenging for contrastive representations. To verify that this result is not driven by discharge-summary content, we also evaluate the discharge-free combination $\mathcal{T}_\mathcal{R} + \mathcal{I} + \mathcal{D}$, which reaches 71.38% AUROC and 50.16% AUPRC on IHM. Early clinical narrative therefore contributes additional signal, but the core finding holds without it (Appendix 11).

As modalities are added, contrastive IHM AUROC drops from 75.94% at three to 61.44% at five, with phenotyping showing a similar decline. These diminishing returns likely reflect redundancy and noise, consistent with the weaker alignment in Section 4.1 (Figure 3).

Fully supervised models outperform contrastive models in the four- and five-modality settings, improving from 65.50% AUROC with three modalities to 73.02% with four and

Table 1. Fine-tuning Results for In-Hospital Mortality and Phenotyping Prediction. We use five modalities: Radiology Text ($\mathcal{T}_\mathcal{R}$), Discharge Text ($\mathcal{T}_\mathcal{D}$), Images ($\mathcal{I}$), Demographics ($\mathcal{D}$), and Time-Series ($\mathcal{TS}$). The final values are computed using the average of 10 random initializations to ensure robustness.

Modalities	Approach	In-Hospital Mortality		Phenotyping	
		AUROC	AUPRC	AUROC	AUPRC
$\mathcal{T}_\mathcal{R} + \mathcal{T}_\mathcal{D}$	Contrastive	57.52 ± 1.71	27.25 ± 1.75	68.90 ± 0.08	43.25 ± 0.45
$\mathcal{T}_\mathcal{R} + \mathcal{T}_\mathcal{D}$	Supervised Baseline	58.70 ± 2.30	31.30 ± 2.98	65.09 ± 0.77	38.05 ± 0.78
$\mathcal{T}_\mathcal{D} + \mathcal{D}$	**Contrastive**	**64.29** ± 0.91	**37.56** ± 0.58	62.81 ± 0.17	36.08 ± 0.24
$\mathcal{T}_\mathcal{D} + \mathcal{D}$	**Supervised Baseline**	61.02 ± 2.92	28.80 ± 2.09	**73.76** ± 0.88	**48.20** ± 1.06
$\mathcal{T}_\mathcal{R} + \mathcal{D}$	Contrastive	61.34 ± 0.25	27.93 ± 0.48	64.10 ± 0.09	36.20 ± 0.25
$\mathcal{T}_\mathcal{R} + \mathcal{D}$	Supervised Baseline	63.50 ± 1.58	30.50 ± 1.25	72.87 ± 0.94	45.47 ± 0.98
$\mathcal{T}_\mathcal{R} + \mathcal{T}_\mathcal{D} + \mathcal{D}$	Contrastive	61.23 ± 2.43	28.20 ± 1.81	68.32 ± 0.11	42.61 ± 0.22
$\mathcal{T}_\mathcal{R} + \mathcal{T}_\mathcal{D} + \mathcal{D}$	Supervised Baseline	62.82 ± 2.80	29.65 ± 2.15	73.40 ± 1.00	46.13 ± 1.30
$\mathcal{T}_\mathcal{R} + \mathcal{T}_\mathcal{D} + \mathcal{I}$	Contrastive	63.97 ± 5.50	32.88 ± 5.20	62.24 ± 0.50	35.11 ± 0.53
$\mathcal{T}_\mathcal{R} + \mathcal{T}_\mathcal{D} + \mathcal{I}$	Supervised Baseline	66.77 ± 3.05	35.59 ± 3.33	72.36 ± 0.50	46.38 ± 0.41
$\mathcal{T}_\mathcal{R} + \mathcal{I} + \mathcal{D}$	Contrastive	71.38 ± 0.18	50.16 ± 0.67	59.27 ± 1.04	34.56 ± 0.57
$\mathcal{T}_\mathcal{R} + \mathcal{I} + \mathcal{D}$	Supervised Baseline	64.96 ± 3.67	31.92 ± 0.17	70.20 ± 0.97	43.01 ± 0.79
$\mathcal{T}_\mathcal{D} + \mathcal{I} + \mathcal{D}$	**Contrastive**	**75.94** ± 0.50	**58.32** ± 2.16	62.93 ± 0.14	36.15 ± 0.29
$\mathcal{T}_\mathcal{D} + \mathcal{I} + \mathcal{D}$	**Supervised Baseline**	65.50 ± 1.94	39.14 ± 3.93	**73.84** ± 0.26	**49.34** ± 0.57
$\mathcal{T}_\mathcal{R} + \mathcal{T}_\mathcal{D} + \mathcal{I} + \mathcal{D}$	Contrastive	62.98 ± 2.52	29.69 ± 2.13	67.21 ± 0.23	40.08 ± 0.40
$\mathcal{T}_\mathcal{R} + \mathcal{T}_\mathcal{D} + \mathcal{I} + \mathcal{D}$	Supervised Baseline	66.92 ± 3.04	38.24 ± 5.67	73.13 ± 1.19	48.17 ± 0.42
$\mathcal{T}_\mathcal{R} + \mathcal{T}_\mathcal{D} + \mathcal{I} + \mathcal{D}$	**Modality-Gated LSTM**	62.82 ± 7.47	32.81 ± 4.36	**73.32** ± 0.52	50.51 ± 1.16
$\mathcal{T}_\mathcal{R} + \mathcal{T}_\mathcal{D} + \mathcal{TS} + \mathcal{D}$	Contrastive	64.45 ± 0.52	34.36 ± 0.45	68.69 ± 0.10	42.76 ± 0.24
$\mathcal{T}_\mathcal{R} + \mathcal{T}_\mathcal{D} + \mathcal{TS} + \mathcal{D}$	Supervised Baseline	73.02 ± 2.35	50.49 ± 3.82	72.45 ± 0.98	45.32 ± 0.38
$\mathcal{T}_\mathcal{R} + \mathcal{T}_\mathcal{D} + \mathcal{TS} + \mathcal{D}$	Modality-Gated LSTM	73.70 ± 3.13	50.22 ± 2.58	72.81 ± 0.34	49.72 ± 0.48
$\mathcal{T}_\mathcal{R} + \mathcal{T}_\mathcal{D} + \mathcal{I} + \mathcal{TS}$	Contrastive	63.18 ± 0.94	31.70 ± 0.63	68.50 ± 0.27	41.92 ± 0.44
$\mathcal{T}_\mathcal{R} + \mathcal{T}_\mathcal{D} + \mathcal{I} + \mathcal{TS}$	Supervised Baseline	71.02 ± 3.62	39.41 ± 3.97	72.91 ± 0.44	46.98 ± 0.44
$\mathcal{T}_\mathcal{R} + \mathcal{T}_\mathcal{D} + \mathcal{I} + \mathcal{TS}$	**Modality-Gated LSTM**	**73.73** ± 2.38	**54.32** ± 3.96	73.21± 0.61	**50.52** ± 0.11
$\mathcal{T}_\mathcal{R} + \mathcal{T}_\mathcal{D} + \mathcal{I} + \mathcal{TS} + \mathcal{D}$	Contrastive	61.44 ± 3.42	30.88 ± 2.37	68.38 ± 0.28	42.65 ± 0.46
$\mathcal{T}_\mathcal{R} + \mathcal{T}_\mathcal{D} + \mathcal{I} + \mathcal{TS} + \mathcal{D}$	Supervised Baseline	73.19 ± 2.62	51.27 ± 6.65	73.43 ± 0.52	47.44 ± 0.51
$\mathcal{T}_\mathcal{R} + \mathcal{T}_\mathcal{D} + \mathcal{I} + \mathcal{TS} + \mathcal{D}$	**Modality-Gated LSTM**	**76.93** ± 2.82	**62.26** ± 2.79	**74.36** ± 0.71	**51.21** ± 0.77

73.19% with five on IHM. However, gains saturate between four and five modalities, and the five-modality supervised result remains below the best three-modality contrastive result of 75.94%. Supervised training therefore extracts additional signal from extra modalities, but simple concatenation does not fully exploit it.

To overcome these limitations, we introduce the Modality-Gated LSTM, which uses contrastively learned modality importance scores (Table 2) to regulate each modality's contribution during integration. It achieves the best overall performance, reaching 76.93% AUROC and 62.26% AUPRC for IHM, and 74.36% AUROC and 51.21% AUPRC for phenotyping. Note that replacing the modality gates with a vanilla LSTM over the same ordered embedding sequence reduces IHM performance to 72.13% AUROC and 45.35% AUPRC, indicating that the gains stem from the contrastively derived modality weights rather than recurrent fusion alone (Appendix 14). These results highlight the importance of structured modality integration for complex multimodal inputs, while also motivating future investigation into alternative modality integration stratgies (Appendix 6).

For comparison, we retrain MedFuse[14] on our filtered cohort, where it reaches 73.91% AU-

ROC and 43.36% AUPRC on IHM, which the five-modality mLSTM exceeds (Appendix 13).

These findings offer practical guidance: for binary tasks with two or three modalities, contrastive pre-training is often sufficient and yields reusable embeddings, while for higher modality counts and multi-label tasks, the Modality-Gated LSTM mitigates redundancy via contrastively learned importance.

4.3. *Analyzing Model Performance Across Demographics and Modalities*

4.3.1. *Demographic Performance Across Models*

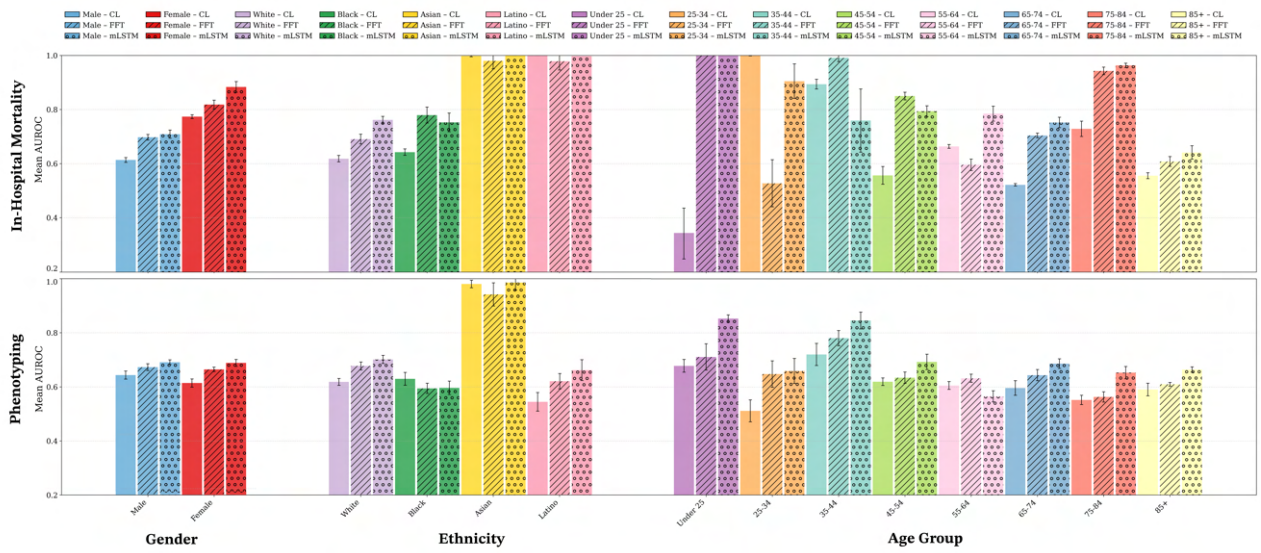


Fig. 4. AUROC scores of the five-modality models across gender, ethnicity, and age group for In-Hospital Mortality (top) and Phenotyping (bottom). Models compared include Contrastive Learning Pretraining (CL), Fully Fine-Tuned (FFT), and Modality-Gated LSTM (mLSTM).

Figure 4 reports AUROC for five-modality models on IHM and phenotyping across age, gender, and ethnicity, comparing contrastive pretraining (CL), the fully fine-tuned baseline (FFT), and the mLSTM. Subgroup sizes are in Appendix Table A1.

Across nearly all subgroups and both tasks, mLSTM achieves the highest AUROC, indicating that dynamic modality weighting improves both accuracy and generalization. Models perform better on females in IHM despite their lower representation (866 vs. 1164 males), whereas this disparity is less pronounced in phenotyping. mLSTM consistently performs best across both genders.

Performance also varies across ethnic groups. CL shows gains for underrepresented groups in phenotyping, particularly Asian and Black patients. In IHM, however, mLSTM performs more consistently in smaller subgroups, including Asian and Latino patients (combined $n = 51$). Patients younger than 25 exhibit lower and more variable performance, likely due to limited sample size and greater life-stage variability. CL performs well in middle-aged groups (e.g., 55–64), while mLSTM remains stable across age brackets.

Overall, mLSTM is the most robust method across demographic subgroups, combining strong performance with consistent generalization. Although CL performs well for some populations, its greater variability in underrepresented groups highlights the value of adaptive modality integration.

4.3.2. *Validating Modality Lambdas with Integrated Gradients*

Table 2. Contrastively learned modality-λ (five-modality setting) and Integrated Gradient (IG) scores for each modality across the two predictive tasks.

Modality	***modality*-λ**	**In-Hospital Mortality**	**Phenotyping**
Discharge Text ($\mathcal{T}_\mathcal{D}$)	0.297	$\text{IG}_{\mathcal{T}_\mathcal{D}} = 0.264$	$\text{IG}_{\mathcal{T}_\mathcal{D}} = 0.336$
Radiology Text ($\mathcal{T}_\mathcal{R}$)	0.245	$\text{IG}_{\mathcal{T}_\mathcal{R}} = 0.249$	$\text{IG}_{\mathcal{T}_\mathcal{R}} = 0.122$
Demographics ($\mathcal{D}$)	0.187	$\text{IG}_\mathcal{D} = 0.237$	$\text{IG}_\mathcal{D} = 0.250$
Images ($\mathcal{I}$)	0.172	$\text{IG}_\mathcal{I} = 0.184$	$\text{IG}_\mathcal{I} = 0.207$
Time-Series ($\mathcal{TS}$)	0.100	$\text{IG}_{\mathcal{TS}} = 0.066$	$\text{IG}_{\mathcal{TS}} = 0.086$

To validate the *modality*-λ values, we compare them against Integrated Gradients (IG), which estimates input importance relative to a zero baseline.[32] Unlike the contrastive weights, IG reflects supervised task predictions, offering a complementary check computed on the independently-trained supervised classifier rather than on the mLSTM itself. Using concatenated embeddings as classifier input (Figure 1A.2), we compute aggregated modality attributions for both tasks.

As shown in Table 2, discharge text receives the highest attribution scores across both tasks. This is consistent with Table 1, where models incorporating discharge text achieve the strongest performance across both contrastive and fully supervised settings. Conversely, time-series embeddings receive substantially lower attribution scores, matching the weaker multimodal alignment observed in Figure 3. Overall, the attribution scores align with the contrastively learned modality-λ values in Table 2.

5. Conclusions

We introduce the Pipeline for Contrastive Modality Evaluation and Encoding (PiCME), the first framework to evaluate contrastive pre-training across all 26 combinations of two to five modalities in the MIMIC dataset. Our analysis identifies both the strengths and limitations of contrastive pre-training for multimodal clinical prediction. Contrastive learning performs well with a small number of modalities, producing reusable representations that transfer across tasks. Contrastive performance peaks with three modalities, discharge summaries, chest X-rays, and demographics, and declines as additional modalities are introduced, while supervised fusion plateaus. To address this limitation, we propose the Modality-Gated LSTM, which dynamically weights modality contributions using contrastively learned importance scores. This achieves the best five-modality performance while improving generalization, and we validate the learned importance with attribution methods. Overall, PiCME provides a scalable pipeline and practical guidance for modality selection and training strategy.

References

1. A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi and R. G. Mark, Mimic-iii, a freely accessible critical care database, *Scientific data* **3**, 1 (2016).
2. 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, *Scientific data* **10**, p. 1 (2023).
3. A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark and S. Horng, Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports, *Scientific data* **6**, p. 317 (2019).
4. X. Zhai, B. Mustafa, A. Kolesnikov and L. Beyer, Sigmoid loss for language image pre-training, in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023.
5. J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong and S. C. H. Hoi, Align before fuse: Vision and language representation learning with momentum distillation, *Advances in neural information processing systems* **34**, 9694 (2021).
6. J. Li, D. Li, C. Xiong and S. Hoi, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in *International conference on machine learning*, 2022.
7. Z. Wang, Z. Wu, D. Agarwal and J. Sun, Medclip: Contrastive learning from unpaired medical images and text, *arXiv preprint* (2022).
8. Z. Huang, F. Bianchi, M. Yuksekgonul, T. J. Montine and J. Zou, A visual–language foundation model for pathology image analysis using medical twitter, *Nature medicine* **29**, 2307 (2023).
9. S.-C. Huang, L. Shen, M. P. Lungren and S. Yeung, Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition, in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.
10. Y. Zhang, H. Jiang, Y. Miura, C. D. Manning and C. P. Langlotz, Contrastive learning of medical visual representations from paired images and text, in *Machine learning for healthcare conference*, 2022.
11. F. Wang, Y. Zhou, S. Wang, V. Vardhanabhuti and L. Yu, Multi-granularity cross-modal alignment for generalized medical visual representation learning, *Advances in Neural Information Processing Systems* **35**, 33536 (2022).
12. M. Golovanevsky, E. Schiller, A. Nair, E. Han, R. Singh and C. Eickhoff, One-versus-others attention: Scalable multimodal integration for biomedical data, in *Biocomputing 2025: Proceedings of the Pacific Symposium*, 2024.
13. H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen and Y. J. Lee, Llava-next: Improved reasoning, ocr, and world knowledge, *arXiv preprint* (2024).
14. N. Hayat, K. J. Geras and F. E. Shamout, Medfuse: Multi-modal fusion with clinical time-series data and chest x-ray images, in *Machine Learning for Healthcare Conference*, 2022.
15. M. Golovanevsky, C. Eickhoff and R. Singh, Multimodal attention-based deep learning for alzheimer's disease diagnosis, *Journal of the American Medical Informatics Association* **29**, 2014 (2022).
16. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger and I. Sutskever, Learning transferable visual models from natural language supervision, *arXiv preprint* (2021).
17. R. Thapa, B. He, M. R. Kjaer, H. Moore, G. Ganjoo, E. Mignot and J. Zou, Sleepfm: Multimodal representation learning for sleep across brain activity, ecg and respiratory signals, *arXiv preprint arXiv:2405.17766* (2024).
18. Z. Piran, M. Klein, J. Thornton and M. Cuturi, Contrasting multiple representations with the

multi-marginal matching gap, *arXiv preprint arXiv:2405.19532* (2024).
19. A. Agostini, D. Chopard, Y. Meng, N. Fortin, B. Shahbaba, S. Mandt, T. M. Sutter and J. E. Vogt, Weakly-supervised multimodal learning on mimic-cxr, *arXiv preprint arXiv:2411.10356* (2024).
20. R. King, T. Yang and B. J. Mortazavi, Multimodal pretraining of medical time series and notes, in *Machine Learning for Health (ML4H)*, 2023.
21. L. R. Soenksen, Y. Ma, C. Zeng, L. Boussioux, K. Villalobos Carballo, L. Na, H. M. Wiberg, M. L. Li, I. Fuentes and D. Bertsimas, Integrated multimodal artificial intelligence framework for healthcare applications, *NPJ digital medicine* **5**, p. 149 (2022).
22. E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann and M. McDermott, Publicly available clinical bert embeddings, *arXiv preprint arXiv:1904.03323* (2019).
23. J. Devlin, Bert: Pre-training of deep bidirectional transformers for language understanding/arxiv preprint, *arXiv preprint arXiv:1810.04805* (2018).
24. J. Liu, D. Capurro, A. Nguyen and K. Verspoor, Attention-based multimodal fusion with contrast for robust clinical prediction in the face of missing modalities, *Journal of Biomedical Informatics* **145**, p. 104466 (2023).
25. E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, Lora: Low-rank adaptation of large language models., *ICLR* **1**, p. 3 (2022).
26. K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.
27. S. Hochreiter and J. Schmidhuber, Long short-term memory, *Neural computation* **9**, 1735 (1997).
28. A. v. d. Oord, Y. Li and O. Vinyals, Representation learning with contrastive predictive coding, *arXiv preprint arXiv:1807.03748* (2018).
29. T. Chen, S. Kornblith, M. Norouzi and G. Hinton, A simple framework for contrastive learning of visual representations, in *International conference on machine learning*, 2020.
30. K. He, H. Fan, Y. Wu, S. Xie and R. Girshick, Momentum contrast for unsupervised visual representation learning, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.
31. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, Learning transferable visual models from natural language supervision, in *International conference on machine learning*, 2021.
32. M. Sundararajan, A. Taly and Q. Yan, Axiomatic attribution for deep networks, in *International conference on machine learning*, 2017.