2025 S.T. Yau High School Science Award

Research Report

The Team

Name of student team member: Santosh Patapati

School: Panther Creek High School

City, Province/State, Country: Frisco, Texas, USA

Name of supervising teacher: Gerry Skiles

Job Title: Computer Science Teacher

School/Institution: Panther Creek High School

City, Province/State, Country: Frisco, Texas, USA

Title of Research Report

CLARGA: Adaptive Residual Graph Attention for Contrastive Multimodal Representation Learning

Date

8/23/2025

Table of Contents

CLARGA: Adaptive Residual Graph Attention for Contrastive Multimodal Representation Learning

Santosh Patapati
Panther Creek High School
1875 PGA Pkwy, Frisco, TX 75034
santosh.patapati.311@k12.friscoisd.org

Abstract

Multimodal representation learning is becoming increasingly important due to the growing availability of diverse multimodal data across various domains. Particularly, the ability to adapt to arbitrary numbers or types of modalities is useful for improving flexibility. We propose CLARGA, a general-purpose multimodal representation learning architecture that builds a learned attention-weighted graph over modality features and uses Graph Attention Networks to fuse them. CLARGA is trained end-to-end with combined supervised and contrastive loss, which aligns modalities while maintaining each modality's own strength. We demonstrate CLARGA's effectiveness in diverse multimodal representation learning tasks across 7 datasets spanning finance, human-computer interaction, general multimedia classification, and complex affective computing. It consistently outperforms baselines, ablations, and recent state-of-the-art approaches. Particularly, we demonstrate the highest known performance on the DAIC-WoZ dataset for multimodal depression identification. Our results show that CLARGA is an accurate and robust general-purpose fusion framework suitable for a wide range of complex multimodal learning tasks.

Keywords—representation learning; graph construction; modality fusion; graph attention networks (GAT); contrastive multimodal learning; cross-modal alignment

1. Introduction

Multimodal data—images, audio, text, time-series, and more—are now commonplace across a variety of domains [9]. Effectively representing and fusing information that arrives in different sensory or semantic formats has therefore become a core challenge of machine learning, with a wide range of applications including chemistry [42], physics [32], and healthcare [34]. Naïve strategies that simply concatenate feature vectors ("early fusion") or average modality-level decisions ("late fusion") often struggle with heterogeneity, quadratic growth in parameters, and sensitive behavior when some inputs are missing [8, 9, 70]. Likewise, end-to-end attention-based fusion models typically require paired data for every modality at training time [30, 74]. Such models may scale quadratically with the number of modalities, making them costly or infeasible for real-world applications [22, 55].

We introduce Contrastive Learning with Adaptive Residual Graph Attention (CLARGA), a general-purpose fusion architecture that accepts arbitrary numbers and types of modalities. It models modality embeddings as nodes in a fully-connected, per-sample graph. A lightweight multi-head Graph Attention Network (GAT) [79] computes learned edge weights, enabling each modality to draw information from each other with sub-quadratic complexity. Residual connections and Layer Normalization stabilize deeper message passing and mitigate oversmoothing. When training, CLARGA employs a hybrid objective that couples supervised task loss with the InfoNCE contrastive term, maximizing a lower bound on mutual information between each unimodal embedding and the fused representation, thereby aligning modalities even when some are noisy or messy.

Our contributions are as follows:

1. We introduce CLARGA, a graph-based plug-and-play architecture for multimodal representation learning supporting arbitrary numbers and types of modalities.
2. We couple this adaptive graph fusion with a lightweight contrastive alignment objective that draws each unimodal embedding toward the fused representation.
3. To handle missing inputs, we employ a learnable mask embedding that substitutes absent modalities and integrates with the attention mechanism for robust, variable-set fusion.
4. We construct novel proofs that build on existing theory. Specifically, we provide: (i) universality of the fusion block as an approximator of any continuous permutation-invariant function; (ii) a Lipschitz

bound quantifying the impact of a missing modality on the fused vector; and (iii) a Rademacher-complexity excess-risk bound for the joint supervised–contrastive loss.

5. Across seven public benchmarks, CLARGA consistently outperforms strong baselines, state-of-the-art fusion methods, and ablations, achieving the best-known results on the DAIC-WoZ dataset.

2. Related Works

2.1. Early Fusion Architectures

Early work used feature-level fusion, where raw features from all modalities are concatenated and fed into a single model, and decision-level fusion, where each modality is processed by a separate model and their final predictions are aggregated [61, 66]. Despite their simplicity, both approaches struggle with heterogeneity, quadratic growth of parameters, and missing inputs.

2.2. Attention-based & Transformer-style Fusion

Recent systems are beginning to use learned attention mechanisms that let the model weight modalities on a per-sample basis. Cross-modal attention [85] has become popular for fusing information across modalities. MulT [74] injects attention across every stream so that one modality can guide another without explicit temporal alignment. (Nagrani et al., 2021) [55] builds on this with the Multimodal bottleneck Transformer to funnel all cross-modal interactions through a few shared "bottleneck" tokens per layer, compressing attention and cutting its quadratic cost while creating rich cross-modal exchange.

2.3. Graph-based Fusion

Graph-based architectures represent heterogeneous inputs as nodes and edges in a graph, enabling flexible relational modeling. Graph Attention Networks [79] extend this by learning attention weights on edge. Graph-based fusion, a newer approach, begins by building a graph in which nodes encode information about each modality (e.g., entire audio or video modalities), or even finer-grained elements (e.g., objects or words) [20, 41]. Nodes are connected by edges which encode the relationships between the modalities.

2.4. Learning Shared Representations

Beyond task-specific fusion, models can learn modality-agnostic embeddings that can be reused across tasks. Deep CCA [3, 11] trains per-modality encoders to maximize correlation of paired samples. Other variants combine this with task-specific losses (e.g., ViLBERT [45]) to preserve modality-specific details while aligning shared structures.

2.5. Contrastive Alignment

Contrastive learning has delivered gains in zero-shot transfer. CLIP [63] aligns image and text encoders by pulling matches pairs together in embedding space and pushing mismatches apart. For example, DGI [80] uses a contrastive InfoMax objective [28] to maximize mutual information between local node-patch embeddings and global graph summary for unsupervised representation learning.

2.6. Handling Missing Modalities

Real-world deployments can suffer from missing modalities, which model architectures have to account for. Many models that address this issue employ imputation [62, 73, 89], robust training [47, 86], or shared-specific factorization [75, 84]. Across these methods, learned mask tokens and contrastive alignment losses have emerged as lightweight yet effective tools. The former gives models an explicit symbol for missing data [49, 56, 64], while the latter keeps partial-input embeddings close to the manifold learned from complete data [21, 44, 88]. Both have been demonstrated to be effective independently.

3. Methodology

3.1. Problem Setup

3.1.1. Data and Notation

We consider a supervised dataset

$$\mathcal{D} = \left\{ (x_1^{(n)}, \ldots, x_M^{(n)},\ y^{(n)}) \right\}_{n=1}^{N},$$

where each example may contain up to M heterogeneous modalities. For the m-th modality, we denote the input domain by $\mathcal{X}_m$. A modality-specific encoder

$$f_m : \mathcal{X}_m \longrightarrow \mathbb{R}^d$$

maps raw input to a d-dimensional latent representation

$$h_m = f_m(x_m).$$

Missing modalities are common in real data. Whenever modality m is absent we substitute h_m with a learned mask embedding $h_{\text{mask}} \in \mathbb{R}^d$. A diagonal binary mask

$$M = \operatorname{diag}(m_{11}, \ldots, m_{MM}),\ m_{ii} = 1 \text{ iff modality } i \text{ is missing},$$

is stored so that the graph-attention block can conveniently ignore self-edges of masked nodes (details in §3).

Each sample is accompanied by a task label y (class label, regression target, or multi-label vector). We train the model with a supervised loss $\mathcal{L}_{\text{sup}}$ combined with a contrastive alignment term.

3.1.2. Goal

Our aim is to learn the parameter set

$$\theta = \Big\{ f_1, \ldots, f_M,\ W_q, W_k, W_g,\ q_F,\ g \Big\},$$

where: (1) W_q, W_k project node features to query / key space for the multi-head graph-attention layers; (2) W_g contains the weights of the residual GAT message-passing blocks; (3) q_F is a learnable query vector that aggregates node embeddings into the global fusion vector z_{fusion}; (4) g is a shallow prediction head.

The parameters are optimized for predictive accuracy and modality-fusion alignment, so that: (1) The fused prediction $\hat{y} = g(z_{\text{fusion}})$ minimizes the task loss on the training set; (2) Each modality embedding h_m shares high mutual information with z_{fusion}, enforced through a batch-wise InfoNCE objective.

3.2. Proposed Approach

In this section, we detail the full CLARGA pipeline in general use cases.

3.2.1. Modality Encoders

For each of the M input modalities (e.g. image, audio, text, tabular, time series), we employ a dedicated encoder that projects raw data into a shared d-dimensional feature space. In our experiments, we fix $d = 3$ [1]

If a modality is missing for a given sample, we substitute its embedding by a learnable mask vector $h_{\text{mask}} \in \mathbb{R}^{256}$. A binary mask tensor accompanies the embeddings so later layers can ignore unusable nodes.

3.2.2. Adaptive Graph Attention

The M modality vectors $\{h_i\}_{i=1}^{M}$ form the nodes of a sample-specific, fully connected directed graph. To quantify how much information each modality should gather from every other, we employ a multi-head graph-attention mechanism.

First, each node is projected into a shared query-key space:

$$q_i = W_q h_i, \qquad k_j = W_k h_j, \qquad W_q, W_k \in \mathbb{R}^{128\times 256}.$$

For every attention head $h \in \{1, \ldots, H\}$ (with $H = 4$ in our default configuration) we compute the raw compatibility score

$$e_{ij}^{(h)} \;=\; \mathrm{LeakyReLU}\big(q_i^{(h)\top} k_j^{(h)}\big).$$

To prevent a node from redundantly attending to itself and to exclude missing modalities, diagonal scores and rows corresponding to masked inputs are set to $-\infty$. The softmax operation along each destination node then yields normalized attention coefficients.

$$\alpha_{ij}^{(h)} \;=\; \frac{\exp\big(e_{ij}^{(h)}\big)}{\sum_{k\neq i} \exp\big(e_{ik}^{(h)}\big)}.$$

The resulting $\alpha_{ij}^{(h)}$ values determine, on a per-sample basis, the strength with which modality j influences modality i during subsequent message-passing layers.

3.2.3. Residual Graph-Attention Layers

To propagate information across modalities we apply a stack of L residual graph-attention layers. In our experiments, we fix $L = 3$. At layer $\ell \in \{0, 1\}$ each node i aggregates messages from its neighbours via the head-averaged attention coefficients introduced in §3.2.2.

$$m_i^{(\ell)} \;=\; \frac{1}{H} \sum_{h=1}^{H} \sum_{j=1}^{M} \alpha_{ij}^{(h)}\, h_j^{(\ell)}.$$

The aggregated message is concatenated with the node's current state and linearly transformed,

$$\tilde{h}_i^{(\ell)} \;=\; W_g\big[h_i^{(\ell)} \| m_i^{(\ell)}\big], \qquad W_g \in \mathbb{R}^{256\times 512},$$

after which a residual connection and Layer Normalization produce the updated embedding,

$$h_i^{(\ell+1)} \;=\; \mathrm{LayerNorm}\big(h_i^{(\ell)} + \mathrm{Dropout}(\tilde{h}_i^{(\ell)})\big).$$

Empirically, three such layers provide a sufficient field without too much computation, although we have not yet tested deeper stacks.

3.2.4. Fusion Read-Out

Once message passing is complete, the model must collapse the M context-enriched node embeddings into a single multimodal representation. We introduce a learnable query vector $q_F \in \mathbb{R}^{128}$ and compute scalar relevance scores.

$$s_i \;=\; q_F^{\top} W_k h_i^{(L)},$$

which are converted to attention weights $\beta_i = \mathrm{softmax}(s_1, \ldots, s_M)$. The final fused vector is the corresponding weighted sum:

$$z_{\text{fusion}} \;=\; \sum_{i=1}^{M} \beta_i\, h_i^{(L)}.$$

A dropout layer with probability 0.1 is applied to z_{fusion} before it enters the task-specific prediction head.

[1] The encoders employed in experimentation include ResNet, 1D-CNNs, BERT-based encoder models [18], and more. The encoders can be either trainable or frozen (see Section 4).

3.2.5. Optimization Objective

CLARGA is trained with a dual-term loss that couples supervised learning with a modality-fusion alignment loss.

Firstly, we incorporate task loss. For a labeled example (x, y) the supervised term is

$$\mathcal{L}_{\text{sup}} = \begin{cases} \text{CE}\big(g(z_{\text{fusion}}),\, y\big), & \text{classification,} \\ \text{MSE}\big(g(z_{\text{fusion}}),\, y\big), & \text{regression,} \end{cases}$$

where g denotes the shallow prediction head.

Secondly, to ensure that every modality remains well aligned with the fused representation, we adopt the InfoNCE contrastive loss [78] with batch negatives. In our experiments, we fix $\tau = 0.07$.

3.3. Theoretical Analysis

We are able to establish three guarantees for parts of CLARGA. First, the fusion block is a universal approximator for continuous permutation-invariant functions on heterogenous modality sets (§3.3.1). Second, the architecture is Lipschitz-robust to single-modality dropout, so the fused representation (and hence the prediction) degrades in proportion to the missing input's norm (§3.3.2). Third, the hybrid supervised-contrastive objective inherits a data-dependent $\text{O}(\sqrt{\log \mathcal{N}(\varepsilon)/n})$ excess-risk bound recently proved for supervised contrastive learning under non-IID tuple sampling (§3.3.3). We also provide commentary on existing proofs and literature regarding InfoNCE alignment loss (§3.3.4) and on how residual-LayerNorm GAT layers prevent oversmoothing (§3.3.5).

3.3.1. Universality of the CLARGA Fusion Block

Proposition 1. *Let*

$$f : (\mathbb{R}^d)^M \to \mathbb{R}^p$$

be any continuous permutation invariant function. For every compact $\mathcal{K}$ *and every* $\varepsilon! >!0$*, there exists a choice of weights* θ *in a three-layer, multi-head CLARGA fusion block such that*

$$\sup_{x \in \mathcal{K}} \big\| \text{CLARGA}_\theta(x) - f(x) \big\| < \varepsilon.$$

Proof sketch.

1. Deep-Sets form. (Zaheer et al., 2017) shows any continuous, permutation-invariant f can be decomposed as $\rho\big(\sum_i \phi(x_i)\big)$ [87].
2. Attention subsumes summation. A multi-head GAT with shared query and key projections computes $\sum_i \alpha_i \phi(x_i)$. Setting all logits equal forces $\alpha_i = 1/M$, recovering the Deep-Sets sum. Learnable logits therefore strictly enlarge the function class.
3. Continuity and invariance. Because the softmax is continous and symmetric, the map $(x_1, \ldots, x_M) \mapsto \sum_i \alpha_i \phi(x_i)$ stays inside the invariant function space $\mathcal{S}$.
4. Density preservation. Composing with a universal MLP ρ maintains density in $\mathcal{S}$ [16, 31].

Therefore the CLARGA fusion block is dense in $\mathcal{S}$, extending invariant-network universality results for attention-based fusion [51]. The full proof and explanation with architectural information is written in Appendix A. □

3.3.2. Lipschitz Robustness to Missing Modalities

Proposition 2 (Lipschitz robustness)**.** *Assume each encoder* f_m *is* L*-Lipschitz and that the fusion weights satisfy* $\sum_i \beta_i = 1$ *with* $\beta_i \geq 0$*. Replacing a single modality* k *by the mask embedding* h_{mask} *yields*

$$\big\| z_{\text{fusion}}^{\text{full}} - z_{\text{fusion}}^{\text{masked}} \big\| \;\leq\; L\, \beta_k \, \|x_k\|.$$

If the task head g *is further constrained to be* K*-Lipschitz (e.g. via spectral normalisation), the prediction perturbation obeys*

$$\big\| g(z^{\text{full}}) - g(z^{\text{masked}}) \big\| \leq K L\, \beta_k \, \|x_k\|.$$

Full proof and explanation in Appendix B.

Proof sketch. The encoder perturbation obeys $|h_k - h_{\text{mask}}| \leq L|x_k|$. All other encoders remain unchanged. Graph-attention layers are 1-Lipschitz when attention coefficients are treated as fixed in the forward pass [5]. Because each subsequent residual GAT layer and LayerNorm is non-expansive, the perturbation norm after L layers is still bounded by $L|x_k|$. Finally, the fusion step is a convex combination with coefficient β_k, scaling the deviation by at most β_k.The optional classifier contributes a multiplicative K factor, completing the bound. Full proof appears in Appendix B. □

3.3.3. Generalization Bound for Supervised–Contrastive Objective

Proposition 3 (Rademacher complexity bound)**.** *Let* $\mathcal{H}$ *be the class of CLARGA networks whose parameter matrices have Frobenius norm bounded by* B *and whose activation functions are* 1*-Lipschitz. Let* $\widehat{h}$ *minimise the empirical hybrid loss* $\mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda_c \mathcal{L}_{\text{NCE}}$ *over* n *i.i.d. examples. Then, for any* $0 < \delta < 1$*, with probability at least* $1 - \delta$*,*

$$\mathcal{E}(\widehat{h}) - \mathcal{E}^\star \;\leq\; \tilde{O}\Big(\sqrt{\frac{B^2\, d_{\text{eff}} \,+\, \log(1/\delta)}{n}}\Big),$$

where d_{eff} *is the effective rank of the network's Jacobian and* $\tilde{O}$ *hides poly-log factors in the batch size used for negatives.*

Proof sketch.

1. Hybrid loss Lipschitzness. Both $\mathcal{L}_{\text{sup}}$ (cross-entropy with bounded logits) and $\mathcal{L}_{\text{NCE}}$ (softmax with temperature) are 1-Lipschitz in the network outputs given spectral-normalised weights.
2. Rademacher complexity. The empirical Rademacher complexity of $\mathcal{H}$ is bounded by $\tilde{O}!\left(B\sqrt{d_{\text{eff}}/n}\right)$ following (Bartlett and Mendelson, 2002) [10].
3. McDiarmid concentration. Lipschitz continuity of the hybrid loss ensures a standard concentration inequality, giving the stated high-probability bound.

A full proposition, group of lemmas, and formal proof are written in Appendix C.

□

3.3.4. Mutual-Information View

InfoNCE is a variational lower bound on mutual information between two random variables and becomes tight when the critic is optimal [78]. In CLARGA the critic is the cosine-similarity in the fused space (§3.2.6), so maximizing $\mathcal{L}_{\text{NCE}}$ encourages each modality embedding hm to retain information predictive of $z$fusion. While stronger f-divergence bounds exist [46], we found no empirical benefit over $\tau = 0.07$ when testing on AV-MNIST [81] and ENRICO [17].

A complete proposition, formal proof, and additional commentary are written in Appendix D.

3.3.5. Depth, Residual Connections, and Oversmoothing

Finally, residual connections and LayerNorm provably mitigate oversmoothing in linearised GNNs [67]. Since each CLARGA layer matches the standard residual-LayerNorm, the lower-bound of (Scholkemper et al., 2025) ensures that node representations cannot collapse entirely. This justifies our choice of $L = 3$ (§3.2.4).

A complete proposition, formal proof, and additional commentary are written in Appendix E.

4. Experimental Setup

4.1. Datasets

Table 1. Information on the datasets utilized in experimentation

Dataset	Prediction Task	Modalities	Count
AV-MNIST	Digit	Image, Audio	56.0k
MM-IMDb	Movie Genre	Image, Text	25.9k
STOCKS-F&B	Stock Returns	Timeseries	75.5k
STOCKS-HEALTH	Stock Returns	Timeseries	75.5k
STOCKS-TECH	Stock Returns	Timeseries	75.5k
ENRICO	User Interface	Image, Set	1,460
DAIC-WoZ	Depression	Video, Audio, Text	189

We evaluate CLARGA across a diverse range of datasets to evaluate its generalization and robustness across a variety of applications (see Table 1).

4.1.1. AV-MNIST

The Audio Visual-MNIST (AV-MNIST) [81] dataset contains spoken audio and image pairs for digit classification tasks. It is a synthetic benchmark where each sample pairs highly noisy MNIST images [38] and TIDIGITS audio [40], making it far more difficult than the original MNIST dataset.

For all architectures evaluated on AV-MNIST, we employ a trainable encoder that has not received any pretraining to conduct the initial processing of each modality. The images are encoded using a 4-layer CNN and the spectrograms a 2-layer CNN. They are both finally projected by fully connected layers for processing by the evaluated models. Here, we opt to use smaller, trainable models rather than pre-trained models to better isolate the performance of the CLARGA and other architectures themselves.

4.1.2. MM-IMDb

From the Multimodal-IMDB (MM-IMDb) [4] dataset, we extract poster images and plot summaries for every movie provided in the dataset. Images and summaries are encoded by a VGG16 [69] and Google word2vec [52] model before being passed into the evaluated architecture for movie genre classification from 23 options.

4.1.3. STOCKS

The STOCKS datasets, introduced in (Liang et al., 2021) [43], are collections of stock market time series data across three categories. Namely: (1) STOCKS-F&B, which has 14 input and 4 output stocks in the GICS Restaurants or Packaged Food & Meats category [53], (2) STOCKS-HEALTH, which consists of 56 input and 7 output stocks in the Health Care category, and (3) STOCKS-TECH, which has 94 input and 6 output stocks categorized by GICS as Information Technology or Communication Services.

Every input stock (consisting of 500 trading days) is treated as a separate time-series mode, with the goal of predicting returns over the next day. To adapt the task to the baseline and state-of-the-art models, we discretize the continuous return variable R into three non-overlapping categories: (1) Low, where $0 \leq R < 0.1$, (2) $Medium$, where $0.1 \leq R < 0.5$, and (3) $High$, where $R \geq 0.5$. Mean Absolute Error (MAE) is calculated by mapping the three classes to numbers ($Low \rightarrow 0$, $Medium \rightarrow 1$, $High \rightarrow 2$) and then deriving MAE as usual. Each modality is encoded by the same CNN-BiLSTM network, which consists of 3 CNNs, 1 BiLSTM [15], and one fully-connected layer acting as projection.

4.1.4. ENRICO

ENRICO [39] is a higher-quality subset of the ENRICO dataset [17] consisting of Android app screens categorized by their design topics. We extract UI screenshots and view hierarchy from the dataset. The view hierarchy is treated as a set as it contains an unordered collection of UI elements that each contain metadata and their spatial and structured layout [43].

A frozen pre-trained ResNet-18 [24] model with its head replaced by a projection layer is used for encoding. We employ a frozen pre-trained model as an encoder here due to the relatively small size of ENRICO and the level of complexity the task already contains.

4.1.5. DAIC-WoZ

The DAIC-WoZ dataset [23] consists of data from 189 psychotherapy clinical interview recordings. Every recording is accompanied by a Patient Health Question-8 (PHQ-8) [35] score, a common inventory used in psychiatry [6, 68], which is used to classify the associated participant as either depressed or not depressed.

DAIC-WoZ faces a major data scarcity issue (see Table 1). To mitigate this, we train the proposed approach on 8-second segments and augment the training split using the techniques proposed in (Patapati, 2024) [59]. To further mitigate this issue, we run 10-fold cross-validation and average the results across all runs [72].

We apply four encoders across three modalities when training and testing CLARGA on the DAIC-WoZ. Facial Action Units (FAUs) [2, 19] are pre-computed and used as input for a BiLSTM model to encode the video modality. Text transcripts are processed by MentalRoBERTa [33] and followed by a fully-connected layer. Audio is converted into Mel Frequency Cepstral Coefficients (MFCCs) [1, 71] that are processed by a BiLSTM model. Audio is also processed by wav2vec and a subsequent fully-connected layer. Everything except for the pre-trained backbone is unfrozen and trainable. These are all then processed by CLARGA for final classification.

Due to the level of compute necessary for running machine learning on the DAIC-WoZ, we train and evaluate only CLARGA and compare it against models specifically designed for the dataset to analyze its ability to perform in highly specific downstream tasks.

4.2. Baseline and State-of-the-Art Models

We train and evaluate 4 models on the benchmark datasets to compare against CLARGA. Two of these models are custom baselines used to establish simple, modality-agnostic fusion strategies for benchmarking, enabling us to quantify the gains from CLARGA's graph-attention fusion over naïve concatenation and averaging approaches. The other two, Multimodal Lego (MM-Lego) [26] and FuseMix [82], are recent state-of-the-art architectures for multimodal processing and representation learning.

4.2.1. Early-Fusion Baseline

This baseline first encodes each modality independently using a modality-specific encoder. The resulting feature vectors are concatenated into a single representation, which is then passed through an MLP for classification.

4.2.2. Late-Fusion Baseline

The averaged late-fusion baseline encodes each modality independently using a modality-specific encoder and classifier head. The model then averages the predicted class probabilities from all modalities to produce the final prediction. This approach treats each modality's prediction equally and does not learn modality-specific fusion weights.

4.2.3. Multimodal Lego

MM-Lego wraps each pretrained encoder in a "LegoBlock" that projects its output into a uniform, frequency-domain latent space to avoid signal interference across modalities. These blocks are stacked into a fusion model that passes the shared latent states through each modality's cross-attention update in turn. We fine-tune the combined network of MM-Lego, using the process named LegoFuse [26], for as many epochs as the other architectures are fine-tuned or trained.

4.2.4. FuseMix

FuseMix freezes unimodal encoders and then pre-computes their latent outputs for each modality. It then generates new paired embeddings by linearly interpolating corresponding latent vectors across modalities to keep them semantically aligned. Finally, two multi-layer perceptrons project these into a common space and align them with a contrastive loss.

As FuseMix requires unimodal encoders to be frozen, we pre-train the backbones on AV-MNIST and STOCKS datasets for 5 epochs before integrating them into FuseMix.

4.3. Ablation Study

To isolate the effects of different components within CLARGA on overall performance, we conduct an ablation study across four ablations.

4.3.1. Uniform Attention

To isolate the benefit of learning per-sample attention, we replace the adaptive weights $\alpha_{(ij)}$ with uniform weights, setting

$$\alpha_{ij} = \frac{1}{M-1} \quad \text{for all } j \neq i.$$

This static graph forces each modality to contribute equally. By comparing the uniform-α variant to the full CLARGA with learned attention, we can see the extent to which adaptive weighting improves fusion.

4.3.2. No Residual Connections

To isolate the effect of the skip connection, we remove the residual term in each GNN layer so that

$$h_i^{(\ell+1)} = \sigma\big(W_g\, m_i^{(\ell)}\big)$$

with no added $h_i^{(\ell)}$. This forces each layer to rely solely on aggregated neighbor messages and tests how much the residual helps prevent oversmoothing.

4.3.3. No Contrastive Alignment

We set the InfoNCE weight to zero ($\lambda_{(c)} = 0$), so the model is trained purely with the supervised loss. This removes the optimization for cross-modal alignment and tests how much the contrastive term regularizes learning. By comparing this variant to the full CLARGA, we measure how much the InfoNCE objective improves generalization.

4.3.4. Early-Fusion (Mean)

We test a simple early-fusion approach by averaging all initial nodes before any graph processing:

$$z_{\text{fusion}}^{(0)} = \frac{1}{M} \sum_n h_n^{(0)}.$$

We then treat $z_{\text{fusion}}^{(0)}$ as a single-node graph with no edges and pass it directly through the decoder. This isolates the benefit of the graph-based message-passing and attention fusion.

4.4. Modality Dropping Robustness Test

To assess robustness to missing information, we evaluate CLARGA, every ablation, and MM-Lego under three scenarios on AV-MNIST: (1) all modalities (audio and image) present, (2) image modality dropped at test time, (3) audio modalitiy dropped at test time. For each dropped-modality scenario, the corresponding input is replaced with a null value or masked out. We report classification accuracy for each setting. We are unable to perform this experiment for FuseMix as it is unable to handle missing modalities.

5. Results and Discussion

We discuss the four complementary studies (as detailed previously in §4): (1) cross-dataset benchmarking (Table 2), (2) robustness analysis on AV-MNIST (Table 3), and (3) domain-specific evaluation on DAIC-WoZ (Table 4).

5.1. Cross-dataset Benchmarking

Across seven benchmarks, CLARGA attains the highest or joint-highest accuracy on every task while maintaining very low GFLOPs when compared to the state-of-the-art models. On MM-IMDb, CLARGA improves move-genre prediction accuracy to 69%, a 3 percent gain over MM-Lego and a 5 percent gain over FuseMix [24, 82]. For UI-topic classification (ENRICO), CLARGA reaches 83% accuracy, which is a 3% margin over the next best model and 20% above the simple early-fusion baseline. This demonstrates the effectiveness of CLARGA on smaller datasets.

On the AV-MNIST benchmark, CLARGA performs equal to MM-Lego and the Uniform Attention ablation with an accuracy of 77%[3]. Interestingly, we observe that performance of state-of-the-art models, ablations, and CLARGA are all similar when tested on AV-MNIST. We believe this is due to the relatively straight-forward and simple nature of the dataset, meaning that much of the performance is based on the capabilities of the encoders themselves. Despite this, CLARGA performs equal to the best and we observe that it possesses the greatest robustness to missing modalities on AV-MNIST, as discussed in §5.2.

CLARGA leads over all other models by 1%-4% on STOCKS datasets. We observe that CLARGA scales well to higher numbers of modalities[4] very effectively, as its gap in performance compared to other approaches increases as STOCKS datasets incorporate more modalities (100 modalities on STOCKS-TECH versus 18 on STOCKS-F&B).

Additionally, based on results across all the datasets, we can identify that contrastive alignment gives the highest benefit when modalities are semantically distant. On MM-IMDb (image + text) and ENRICO (image + set-like data) the No-Contrastive variant fall behind CLARGA by 6% and 14%, respectively. This is consistent with recent papers which demonstrate that InfoNCE narrows statistical distance between heterogeneous embeddings, enabling simpler classifiers to use joint cues linking different modes [90].

Put together, the results on these datasets demonstrate CLARGA's ability to adapt to a wide range of modalities (image, audio, text, time-series) across a different domains (pattern recognition, finance, human-computer interaction) very effectively. This is made more impressive when considering the computational complexity of CLARGA, which is relatively low compared to the state-of-the-art approaches and made insignificant when training larger pipelines.

5.2. Robustness on AV-MNIST

Table 3 shows that CLARGA's adaptive design holds up well when one channel is absent. Removing the image modality, which appears to be the more informative modality [76], reduces CLARGA's accuracy from 77% to 48%. While this drop in accuracy is substantial, it is smaller than the declines in the ablations' performance. Early-Fusion (Mean) drops in performance by 21% and even the Uniform-Attention variant sees a drop of 17%. Remove residual links or the contrastive term reduces perfor-

[2] We exclude papers which evaluate on sub-segments of the DAIC-WoZ, as this has been shown to unfairly inflate accuracy metrics [12, 59]

[3] Although the performance observed here is very poor compared to what is expected on MNIST [14, 37], we would like to emphasize the increased difficulty of AV-MNIST due to the noise, information reduction, and artifacts introduced (as discussed in §4.1.1) [60, 76]

[4] For the STOCKS datasets we follow (Liang et al., 2021) and the subsequent MultiBench analysis, which treat the return series of each listed company as a separate time-series modality [43] (see §3.1.1 and §3.2)

Table 2. Performance of Models Across Datasets and GFLOPs Calculated on AV-MNIST (focusing only on method-specific components)

Model	**Accuracy (%)**			**F&B**		**HEALTH**		**TECH**		**GFLOPs**
	MM-IMDb	ENRICO	AV-MNIST	Acc. (%)	MAE	Acc. (%)	MAE	Acc. (%)	MAE	
Early-Fusion	61	63	66	51	0.53	58	0.50	64	0.45	1.214
Late-Fusion	56	65	69	53	0.51	59	0.49	63	0.46	1.214
Multimodal Lego	66	76	**77**	57	0.47	63	0.44	66	0.41	0.036
FuseMix	64	80	75	54	0.50	60	0.47	62	0.43	–
Uniform Attention	65	80	**77**	59	**0.45**	66	0.43	66	0.42	0.02
No Residual Connection	65	77	75	55	0.48	64	0.47	65	0.45	0.02
No Contrastive Alignment	63	69	74	49	0.58	58	0.50	62	0.49	0.02
Early-Fusion (Mean)	63	70	73	58	0.49	61	0.49	63	0.48	0.02
CLARGA	**69**	**83**	**77**	**60**	**0.45**	**68**	**0.41**	**70**	**0.40**	0.02

Table 3. Impact of dropping modalities on AV-MNIST dataset

Model	**Acc. (%)**	**Kept Modality**	
		Image (%)	Audio (%)
FuseMix	75	–	–
Multimodal Lego	**77**	57 (-20 pp)	43 (-34 pp)
Uniform Attention	**77**	60 (-17 pp)	45 (-32 pp)
No Res. Connection	75	57 (-18 pp)	42 (-33 pp)
No Contrastive Alignment	74	54 (-20 pp)	38 (-36 pp)
Early-Fusion (Mean)	73	52 (-21 pp)	35 (-38 pp)
CLARGA	**77**	**62 (-15 pp)**	**48 (-29 pp)**

Table 4. Performance against state-of-the-art approaches on the DAIC-WoZ dataset.[2] The first section consists of models evaluated on the AVEC 2016 benchmark [77], while the second consists of K-fold [72] or Leave-One-Subject-Out (LOSO) trained models [36].

Model	**Accuracy (%)**	**Approach**
(Ma et al., 2016) [48]	72.0	CNN-LSTM [29, 38]
(Vázquez-Romero et al., 2020) [83]	72.0	Ensemble CNN
(Muzammel et al., 2021) [54]	77.2	LSTM + MLP [29]
(Patapati, 2024) [59]	85.1	BiLSTM + GPT-4 [15, 57]
CLARGA	**91.4**	Adaptive Residual GAT
(Othmani et al., 2022) [58]	87.4	VGGish + 1D-CNN [27]
(Patapati, 2024) (LOSO) [59]	91.0	BiLSTM + GPT-4 [15, 57]
(Muzammel et al., 2021) (LOSO) [54]	95.5	LSTM + MLP [29]
CLARGA (K-Fold)	**95.7**	Adaptive Residual GAT

mance further. These gains in robustness show that, by using learned edge weights and residual message passing, CLARGA enables the remaining audio information to compensate far more effectively for the missining vision information than static or mean-pooling-based approaches. This demonstrates that adaptive weighting and contrastive alignment are the main factors behind this robustness. The pattern is even clearer when the image modality is dropped (audio only). CLARGA has a 29% loss in accuracy, but still outperforms every other model (see Table 3).

MM-Lego [26] applies ”LegoBlocks”, wrapping each pretrained encoder in a small adapter that projects its output into a common latent space and then updates it through cross-attention with other modalities. When we remove the audio modality, MM-lego faces a 20% loss in accuracy. When the image input is removed, its accuracy falls by 34%. These results place MM-lEgo between full CLARGA and the simpler ablations in terms of robustness. In other words, model-merging adapters give some level of robustness. However, without CLARGA's per-node, sample-specific attention and contrastive alignment, they cannot match its ability to recover when a modality vanishes.

5.3. Complex Downstream Task: DAIC-WoZ

Table 4 summarizes results on multi-modal depression. This is a highly difficult and very specific downstream task, where models must pick up on and learn extremely subtle cues correlating to mental health [7, 25]. The complexity of the DAIC-WoZ is shown by the fact that even for models constructed specifically for the dataset, the vast majority do not surpass 80% accuracy [50, 54, 65, 77], and fewer 90%.

CLARGA attains 91.4% accuracy on the AVEC-2016 challenge benchmark, surpassing all baselines proposed in the original challenge and recent state-of-the-art models. Under 10-fold cross-validation, CLARGA reaches 95.7% accuracy, surpassing models evaluated using LOSO.[5]

We believe that the performance improvement stems from two main factors. Firstly, graph attention balances the uneven predictive strength [54] of the different features and modalities present in the DAIC-WoZ. Secondly, contrastive alignment mitigates the well-known scarcity and imabalance of psychiatric data [13, 50, 54, 59]. This level of performance shows CLARGA's ability to adaptively fuse very different modalities and pick up on faint cross-modal

[5]The only reason we do not scale to LOSO testing is due to its high compute cost. This biases the comparison in favor of LOSO-based models due to their better access to training data, putting CLARGA at a disadvantage, as demonstrated in previous DAIC-WoZ-based papers [54, 59]

cues. Such cues are only revealed given the context of mutliple separate modalities. This also demonstrates the framework's high-quality, general-purpose design.

6. Conclusion

CLARGA delivers a versatile and efficient framework that consistently excels across diverse multimodal representation learning challenges. By combining adaptive graph attention, residual message passing, and contrastive alignment, it not only achieves state-of-the-art accuracy across several benchmarks but also adapts well to missing inputs. Our theoretical guarantees and low computatonal cost further demonstrate its real-world practicality. These findings show that CLARGA offers a robust and practical approach for general-purpose multimodal representation learning.

References

[1] Zrar Kh. Abdul and Abdulbasit K. Al-Talabani. Mel frequency cepstral coefficient and its applications: A review. *IEEE Access*, 10:122136–122158, 2022. 6

[2] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. Openface: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science, 2016. 6

[3] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1247–1255, Atlanta, Georgia, USA, 2013. PMLR. 2

[4] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A. González. Gated multimodal units for information fusion. In *ICLR Workshop Track*, 2017. Workshop track, International Conference on Learning Representations. 5

[5] Raghu Arghal, Eric Lei, and Shirin Saeedi Bidokhti. Robust graph neural networks via probabilistic lipschitz constraints. *arXiv preprint arXiv:2112.07575*, 2021. 4

[6] Jorge Arias de la Torre, Gemma Vilagut, Amy Ronaldson, Jose M. Valderas, Ioannis Bakolis, Alex Dregan, Antonio J. Molina, Fernando Navarro-Mateu, Katherine Pérez, Xavier Bartoll-Roca, Matilde Elices, Víctor Pérez-Sola, Antoni Serrano-Blanco, Vicente Martín, and Jordi Alonso. Reliability and cross-country equivalence of the 8-item version of the patient health questionnaire (phq-8) for the assessment of depression: results from 27 countries in europe. *Lancet Regional Health–Europe*, 31:100659, 2023. 6

[7] Umut Arıöz, Matej Križan, Polona Selič, Livia Kramer, Judith Eckle-Kohler, Luca Cirillo, and Patrizia Cerutti. Scoping review on the multimodal classification of depression and experimental study on existing multimodal models. *Diagnostics*, 12(11):2683, 2022. 8

[8] Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16(6):345–379, 2010. 1

[9] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2018. 1

[10] Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002. 5

[11] Adrian Benton, Huda Khayrallah, Biman Gujral, Dee Ann Reisinger, Sheng Zhang, and Raman Arora. Deep generalized canonical correlation analysis. *arXiv preprint arXiv:1702.02519*, 2017. 2

[12] Sergio Burdisso, Esaú Villatoro-Tello, Srikanth Madikeri, and Petr Motlicek. Node-weighted graph convolutional network for depression detection in transcribed clinical interviews. In *Interspeech 2023*, pages 3617–3621, 2023. 7

[13] Caio B Casella, Antonis A Kousoulis, Brandon A Kohrt, Jason Bantjes, Christian Kieling, Pim Cuijpers, Sarah Kline, Konstantinos Kotsis, Guilherme V Polanczyk, Dan J Stein, Peter Szatmari, Kathleen R Merikangas, Zeina Mneimneh, and Giovanni A Salum. Data gaps in prevalence rates of mental health conditions around the world: a retrospective analysis of nationally representative data. *Lancet Global Health*, 13(5):e879–e887, 2025. 8

[14] Dan C. Cireşan, Ueli Meier, Jonathan Masci, Luca M. Gambardella, and Jürgen Schmidhuber. High-performance neural networks for visual object classification. *arXiv preprint arXiv:1102.0183*, 2011. 7

[15] Zhiyong Cui, Ruimin Ke, Ziyuan Pu, and Yinhai Wang. Deep bidirectional and unidirectional lstm recurrent neural network for network-wide traffic speed prediction, 2019. 5, 8

[16] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989. 4

[17] Biplab Deka, Zifeng Huang, Chad Franzen, Joshua Hibschman, Daniel Afergan, Yang Li, Jeffrey Nichols, and Ranjitha Kumar. Rico: A mobile app dataset for building data-driven design applications. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, page 845–854, New York, NY, USA, 2017. Association for Computing Machinery. 5

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. 3

[19] Paul Ekman and Wallace V. Friesen. Facial action coding system (facs) [database record]. APA PsycTests, 1978. 6

[20] Yasha Ektefaie, George Dasoulas, Ayush Noori, Maha Farhat, and Marinka Zitnik. Multimodal learning with graphs, 2023. 2

[21] Amir Ghorbani et al. Attention-based multimodal fusion with contrast for robust clinical prediction. *Journal of Biomedical Informatics*, 150:104569, 2023. 2

[22] Michal Golovanevsky, Eva Schiller, Akira Nair, Eric Han, Ritambhara Singh, and Carsten Eickhoff. One-versus-others attention: Scalable multimodal integration for biomedical data. *arXiv preprint arXiv:2307.05435*, 2023. 1

[23] Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128, Reykjavik, Iceland, 2014. European Language Resources Association (ELRA). 6

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6, 7

[25] Lang He, Mingyue Niu, Prayag Tiwari, Pekka Marttinen, Rui Su, Jiewei Jiang, Chenguang Guo, Hongyu Wang, Songtao Ding, Zhongmin Wang, Wei Dang, and Xiaoying Pan.

Deep learning for depression recognition with audiovisual cues: A review. *arXiv preprint arXiv:2106.00610*, 2021. 8

[26] Konstantin Hemker, Nikola Simidjievski, and Mateja Jamnik. Multimodal lego: Model merging and fine-tuning across topologies and modalities in biomedicine, 2025. 6, 8

[27] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. Cnn architectures for large-scale audio classification, 2017. 8

[28] R. Devon Hjelm, Alexander Fedorov, Kaveh Lavoie-Marchildon, Karan Grewenig, Philip Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. Originally released as arXiv:1808.06670. 2

[29] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997. 8

[30] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Kazuhiro Sumi, John R. Hershey, and Tim K. Marks. Attention-based multimodal fusion for video description. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[31] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. 4

[32] E. A. Huerta, Gabrielle Allen, Igor Andreoni, Leo … Singer, et al. Enabling real-time multi-messenger astrophysics discoveries with deep learning. *Nature Reviews Physics*, 1:600–608, 2019. 1

[33] Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. MentalBERT: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190, Marseille, France, 2022. European Language Resources Association. 6

[34] Adrienne Kline, Hanyin Wang, Yikuan Li, Saya Dennis, Meghan Hutch, Zhenxing Xu, Fei Wang, Feixiong Cheng, and Yuan Luo. Multimodal machine learning in precision health: A scoping review. *npj Digital Medicine*, 5(1):171, 2022. 1

[35] Kurt Kroenke, Tara W. Strine, Robert L. Spitzer, Janet B. W. Williams, Joyce T. Berry, and Ali H. Mokdad. 6

[36] Sajeev Kunjan, T. S. Grummett, K. J. Pope, D. M. W. Powers, S. P. Fitzgibbon, T. Bastiampillai, M. Battersby, and T. W. Lewis. The necessity of leave one subject out (loso) cross validation for eeg disease diagnosis. In *Brain Informatics: 14th International Conference, BI 2021, Virtual Event, September 17–19, 2021, Proceedings*, page 558–567, Berlin, Heidelberg, 2021. Springer-Verlag. 8

[37] Yann LeCun, Bernhard Boser, John S. Denker, Donnie Henderson, and Richard E. Howard. Backpropagation applied to handwritten zip code recognition. Neural Computation, December 1989. Error rate 0.8% on MNIST. 7

[38] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 5, 8

[39] Luis A. Leiva, Asutosh Hota, and Antti Oulasvirta. Enrico: A dataset for topic modeling of mobile ui designs. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, New York, NY, USA, 2021. Association for Computing Machinery. 5

[40] R. Gary Leonard and George R. Doddington. TIDIGITS audio database (ldc93s10). Linguistic Data Consortium, Philadelphia, PA, 1993. LDC Catalog No. LDC93S10; `https://catalog.ldc.upenn.edu/LDC93S10`. 5

[41] Jiang Li, Xiaoping Wang, Guoqing Lv, and Zhigang Zeng. Graphmft: A graph network based multimodal fusion technique for emotion recognition in conversation. *arXiv preprint arXiv:2208.00339*, 2022. 2

[42] Junxian Li, Di Zhang, Xunzhi Wang, Zeying Hao, Jingdi Lei, Qian Tan, Cai Zhou, Wei Liu, Yaotian Yang, Xinrui Xiong, Weiyun Wang, Zhe Chen, Wanli … Ouyang, Yuqiang Li, and Dongzhan Zhou. Chemvlm: Exploring the power of multimodal large language models in chemistry area. *arXiv preprint arXiv:2408.07246*, 2024. 1

[43] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie (Yufan) Chen, Peter Wu, Michelle A. Lee, Yuke Zhu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Multibench: Multiscale benchmarks for multimodal representation learning. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. 5, 6, 7

[44] Ronghao Lin and Haifeng Hu. Missmodal: Increasing robustness to missing modality in multimodal sentiment analysis. *Transactions of the Association for Computational Linguistics*, 11:321–338, 2023. 2

[45] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019. 2

[46] Yiwei Lu, Guojun Zhang, Sun Sun, Hongyu Guo, and Yaoliang Yu. f-micl: Understanding and generalizing infonce-based contrastive learning, 2024. 5

[47] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? *arXiv preprint arXiv:2204.05454*, 2022. 2

[48] Xingchen Ma, Hongyu Yang, Qiang Chen, Di Huang, and Yunhong Wang. Depaudionet: An efficient deep model for audio based depression classification. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, page 35–42, New York, NY, USA, 2016. Association for Computing Machinery. 8

[49] Kaustubh Maheshwari et al. Missing modality robustness in semi-supervised multi-modal semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1234–1243, 2024. 2

[50] Suresh Mamidisetti, Abhishek Singh Dhadwal, Praveen Kumar, and Sainaba Parammal. Multimodal depression detection using audio, visual and textual cues: A survey. *NeuroQuantology*, 20(4):325–336, 2022. 8

[51] Haggai Maron, Ethan Fetaya, Nimrod Segol, and Yaron Lipman. On the universality of invariant networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4363–4371. PMLR, 2019. 4

[52] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013. 5

[53] MSCI Inc. and S&P Dow Jones Indices. *Global Industry Classification Standard (GICS)*. MSCI Inc., New York, NY, august 2024 edition, 2024. First published January 7, 2020; updated August 2024. 5

[54] Muhammad Muzammel, Hanan Salam, and Alice Othmani. End-to-end multimodal clinical depression recognition using deep neural networks: A comparative analysis. *Computer Methods and Programs in Biomedicine*, 211:106433, 2021. 8

[55] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *Advances in Neural Information Processing Systems 34*, 2021. 1, 2

[56] Niki Nezakati, Md Kaykobad Reza, Ameya Patil, Mashhour Solh, and M. Salman Asif. Mmp: Towards robust multimodal learning with masked modality projection, 2024. 2

[57] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, et al. Gpt-4 technical report, 2024. 8

[58] Alice Othmani, Assaad-Oussama Zeghina, and Muhammad Muzammel. A model of normality inspired deep learning framework for depression relapse prediction using audiovisual data. *Computer Methods and Programs in Biomedicine*, 226:107132, 2022. 8

[59] Santosh V. Patapati. Integrating large language models into a tri-modal architecture for automated depression classification on the daic-woz, 2024. 6, 7, 8

[60] Fernando Perez-Rua, Pierre Vielzeuf, Sébastien Pateux, Patrick Pérez, Cristóbal Garcia, Matthieu Cord, and Lluis G. Pérez. Mfas: Multimodal fusion architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 6926–6935, 2019. Section3.1 on AV-MNIST energy reduction. 7

[61] Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, and Amir Hussain. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, 174:50–59, 2016. 2

[62] Pranav Poudel, Aavash Chhetri, Prashnna K. Gyawali, Georgios Leontidis, and Binod Bhattarai. Multimodal federated learning with missing modalities through feature imputation network. In *Medical Image Understanding and Analysis: 29th Annual Conference (MIUA)*, pages 289–299. Springer, 2025. 2

[63] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 2

[64] Madina Ramazanova et al. Exploring missing modality in multimodal egocentric datasets. In *CVPR Workshop on Multimodal Understanding of Live Assistants (MULA)*, 2025. 2

[65] Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja Pantic. Avec 2017—real-life depression and affect recognition workshop and challenge. In *Proceedings of the 7th Audio/Visual Emotion Challenge and Workshop*. ACM, 2017. 8

[66] Alina Roitberg, Kunyu Peng, Zdravko Marinov, Constantin Seibold, David Schneider, and Rainer Stiefelhagen. A comparative analysis of decision-level fusion for multimodal driver behaviour understanding. *arXiv preprint arXiv:2204.04734*, 2022. 2

[67] Michael Scholkemper, Xinyi Wu, Ali Jadbabaie, and Michael T. Schaub. Residual connections and normalization can provably prevent oversmoothing in gnns, 2025. 5

[68] Cheolmin Shin, Seung-Hoon Lee, Kyu-Man Han, Ho-Kyoung Yoon, and Changsu Han. Comparison of the usefulness of the PHQ-8 and PHQ-9 for screening for major depressive disorder: Analysis of psychiatric outpatient data. *Psychiatry Investigation*, 16(4):300–305, 2019. 6

[69] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. 5

[70] Cees G. M. Snoek, Marcel Worring, and Arnold W. M. Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, pages 399–402. ACM, 2005. 1

[71] S. Stevens, S. J. Volkmann, and B. Newman, E.˙ A scale for the measurement of the psychological magnitude of pitch. *Journal of the Acoustical Society of America*, 8(3):185–190, 1937. 6

[72] Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–132, 1974. 6, 8

[73] Minh Nguyen Tran, Sharath Pankanti, Tirthankar Bhattacharya, and Eric Achtert. Missing modalities imputation via cascaded residual autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 913–921, 2017. 2

[74] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy, 2019. Association for Computational Linguistics. 1, 2

[75] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*, 2018. 2

[76] Yuan Tseng, Layne Berry, Yi-Ting Chen, I-Hsiang Chiu, Hsuan-Hao Lin, Max Liu, Puyuan Peng, Yi-Jen Shih,

Hung-Yu Wang, Haibin Wu, Po-Yao Huang, Chun-Mao Lai, Shang-Wen Li, David Harwath, Yu Tsao, Shinji Watanabe, Abdelrahman Mohamed, Chi-Luen Feng, and Hung-yi Lee. Av-superb: A multi-task evaluation benchmark for audio-visual representation models. *arXiv preprint arXiv:2309.10787*, 2023. Appendix A.1.2 describes AV-MNIST's PCA and noise setup. 7

[77] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, page 3–10, New York, NY, USA, 2016. Association for Computing Machinery. 8

[78] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. 4, 5

[79] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018. 1, 2

[80] Petar Veličković, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, and R. Devon Hjelm. Deep graph infomax. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. 2

[81] Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. Centralnet: a multilayer approach for multimodal fusion. In *Computer Vision – ECCV 2018 Workshops*, pages 575–589. Springer, 2019. 5

[82] Noël Vouitsis, Zhaoyan Liu, Satya Krishna Gorti, Valentin Villecroze, Jesse C. Cresswell, Guangwei Yu, Gabriel Loaiza-Ganem, and Maksims Volkovs. Data-efficient multimodal fusion on a single gpu. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024. Open Access version provided by the Computer Vision Foundation. 6, 7

[83] Adrián Vázquez-Romero and Ascensión Gallardo-Antolín. Automatic detection of depression in speech using ensemble convolutional neural networks. *Entropy*, 22(6):688, 2020. 8

[84] Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15875–15884, 2023. 2

[85] Dayu Wei, Fuchun Li, Rui Zhao, Xubo Han, and Yun Wu. Multi-modality cross-attention network for image and sentence matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2127–2136, 2020. 2

[86] Sangmin Woo, Sumin Lee, Yeonju Park, Muhammad Adi Nugroho, and Changick Kim. Towards good practices for missing modality robust action recognition. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*. AAAI Press, 2023. 2

[87] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets. In *Advances in Neural Information Processing Systems*, pages 3391–3401, 2017. 4

[88] He Zhang, Wei Liu, Rong Chen, et al. Com: Contrastive masked-attention model for incomplete multimodal data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[89] Yue Zhang, Chengtao Peng, Qiuli Wang, Dan Song, Kaiyan Li, and S. Kevin Zhou. Unified multi-modal image synthesis for missing modality imputation. *arXiv preprint arXiv:2304.05340*, 2023. 2

[90] Zhengyang Zhou, Yunrui Li, Pengyu Hong, and Hao Xu. Multimodal fusion with relational learning for molecular property prediction. *Communications Chemistry*, 8(1):200, 2025. 7

Appendix

A Universality of the CLARGA Fusion Block

Proposition 1 (Restatement of Proposition 1 in Main text)**.** *Let* $f : (\mathbb{R}^d)^M \to \mathbb{R}^p$ *be any* continuous *and* permutation-invariant *function; i.e.* $f(x_{\pi(1)}, \dots, x_{\pi(M)}) = f(x_1, \dots, x_M)$ *for every permutation* $\pi \in S_M$*. For every compact set* $\mathcal{K} \subset (\mathbb{R}^d)^M$ *and every* $\varepsilon > 0$*, there exists a parameter configuration* $\theta^\star$ *of a three-layer, multi-head CLARGA fusion block such that*

$$\sup_{x \in \mathcal{K}} \big\| \mathrm{CLARGA}_{\theta^\star}(x) - f(x) \big\| \ < \ \varepsilon. \tag{A.1}$$

Proof. The argument proceeds in three stages.

1. DeepSets normal form. (Zaheer et al., 2017) proved that the set of functions of the form $\rho\big(\sum_{i=1}^{M} \phi(x_i)\big)$ with ρ, ϕ continuous is *dense* in the space $\mathcal{S}$ of continuous permutation-invariant maps on compact domains. Hence it suffices to approximate $h(x_1, \dots, x_M) = \rho\big(\sum_i \phi(x_i)\big)$ to arbitrary accuracy.

2. Summation via attention. Consider a single-head graph-attention layer with *shared* linear query/key projections $W_q, W_k \in \mathbb{R}^{m \times d}$ and value map ϕ. Its output (before message passing) is

$$\sum_{i=1}^{M} \alpha_i \, \phi(x_i), \quad \alpha_i = \frac{\exp\big(\langle W_q x_i, W_k x_i \rangle\big)}{\sum_j \exp\big(\langle W_q x_j, W_k x_j \rangle\big)}. \tag{A.2}$$

Choose $W_q = W_k = \mathbf{0}$ so that every inner product is zero; then $\alpha_i = \frac{1}{M}$ and the layer implements the *mean* $\frac{1}{M} \sum_i \phi(x_i)$. Multiplying the value matrix by M rescales the mean to a *sum*. Thus the family of attention layers *contains* the DeepSets aggregator. Because the softmax weights α_i are themselves learnable via W_q, W_k, the attention mechanism strictly *enlarges* the representable function class.

3. Universal approximation. Fix $\varepsilon > 0$ and compact $\mathcal{K}$. By the universal-approximation theorem for two-layer ReLU (or LeakyReLU) MLPs, there exist finite-width MLPs $\phi_\varepsilon : \mathbb{R}^d \to \mathbb{R}^r$ and $\rho_\varepsilon : \mathbb{R}^r \to \mathbb{R}^p$ such that

$$\sup_{x \in \mathcal{K}} \Big\| \rho_\varepsilon\Big(\sum_{i=1}^{M} \phi_\varepsilon(x_i)\Big) - f(x) \Big\| < \frac{\varepsilon}{2}. \tag{A.3}$$

We embed these into CLARGA's three-layer fusion block as follows:

1. **Attention-sum (Layer 1):** Use a single attention head whose value map is the MLP ϕ_ε. Set $W_q = W_k = 0$ so that $\alpha_i = 1/M$. Multiply the head's output by M to obtain exactly $\sum_{i=1}^{M} \phi_\varepsilon(x_i)$. Dropout is disabled at approximation time; LayerNorm is absorbed into the subsequent linear transform.

2. **Nonlinear projection (Layer 2):** Apply a LeakyReLU activation to the summed vector, then a learned linear map U projecting into $\mathbb{R}^r$.

3. **Final read-out (Layer 3):** Apply a LeakyReLU activation followed by a linear map V so that $x \mapsto V\big(\text{LeakyReLU}(U(x))\big) \equiv \rho_\varepsilon(x)$. The residual skip and LayerNorm can be absorbed into V.

By construction, this block computes exactly $\rho_\varepsilon\big(\sum_i \phi_\varepsilon(x_i)\big)$ on $\mathcal{K}$. Hence

$$\sup_{x\in\mathcal{K}}\big\|\text{CLARGA}_{\theta^\star}(x) - f(x)\big\| \le \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon, \tag{A.4}$$

completing the proof. □

Commentary

Proposition 1 extends the classical DeepSets universality result to CLARGA's *attention-based* fusion: if the logits are forced equal, the layer reproduces the set sum; with learnable logits it can *adapt* the weights α_i to each sample, strictly increasing expressive power while retaining permutation invariance. Because both softmax and MLPs are continuous, the resulting function class remains dense in the invariant space $\mathcal{S}$. Hence, no matter how complex the true multimodal fusion rule is (provided it is continuous and order-agnostic), a suitably wide three-layer CLARGA block can approximate it arbitrarily well on any bounded domain.

B Lipschitz Robustness to Missing Modalities

Proposition 2 (Restatement of Proposition 2 in Main Text)**.** *Assume*

1. *each encoder* $f_m : \mathcal{X}_m \to \mathbb{R}^d$ *is L-Lipschitz, i.e.* $\|f_m(x) - f_m(x')\| \le L\|x - x'\|$ *for all* $x, x' \in \mathcal{X}_m$*;*
2. *every linear weight matrix that appears inside a graph-attention or projection layer is spectrally normalised so its operator norm is at most* 1*;*[1]
3. *the fusion coefficients satisfy* $\beta_i \ge 0$ *and* $\sum_{i=1}^M \beta_i = 1$*;*
4. *the prediction head* $g : \mathbb{R}^d \to \mathbb{R}^p$ *is* K*-Lipschitz (e.g. enforced by spectral normalisation).*

Let $z_{\text{fusion}}^{\text{full}}$ *denote the fused representation obtained from the complete modality set* $\{x_i\}_{i=1}^M$ *and* $z_{\text{fusion}}^{\text{masked}}$ *the fused representation when modality* k *is* missing *and replaced by the learned mask embedding* h_{mask}*. Then*

$$\big\|z_{\text{fusion}}^{\text{full}} - z_{\text{fusion}}^{\text{masked}}\big\| \le L\,\beta_k\,\|x_k\|, \tag{B.1}$$

$$\big\|g(z_{\text{fusion}}^{\text{full}}) - g(z_{\text{fusion}}^{\text{masked}})\big\| \le K\,L\,\beta_k\,\|x_k\|. \tag{B.2}$$

Proof. Let $x_k^0 \in \mathcal{X}_k$ be a fixed reference input (e.g. the zero vector) and define the mask embedding so that

$$h_{\text{mask}} = f_k(x_k^0). \tag{B.3}$$

Set

$$\delta_0 := f_k(x_k) \;-\; f_k(x_k^0). \tag{B.4}$$

By encoder Lipschitzness,

$$\|\delta_0\| = \|f_k(x_k) - f_k(x_k^0)\| \le L\,\|x_k - x_k^0\| \;\le\; L\,\|x_k\|. \tag{B.5}$$

[1] LayerNorm, residual addition, ReLU, and dropout are (weakly) non-expansive, hence 1-Lipschitz. Softmax is 1-Lipschitz on probability-simplex-valued logits under the ℓ_1 norm; we treat the attention coefficients as *fixed* during the forward pass because the perturbation concerns only the input features.

Stability of one graph-attention layer. Fix any one GAT layer (with H heads). Under our spectral-norm assumptions each of the following maps is 1-Lipschitz:

- the query/key projections $h \mapsto W_q h$ and $h \mapsto W_k h$,
- the softmax-over-dot-products $\ell \mapsto \alpha(\ell)$ on any compact logit domain,
- the message-aggregation $h \mapsto \alpha\, h$,
- the linear update and residual+LayerNorm $h \mapsto h + W_g[h \| m]$.

Hence the combined attention-plus-update block is 1-Lipschitz, and

$$\|\delta_\ell\| \;\le\; \|\delta_{\ell-1}\|, \quad \ell = 1, \ldots, L. \tag{B.6}$$

By induction, $\|\delta_L\| \le \|\delta_0\| \le L\|x_k\|$.

Fusion step. The fused vector is a convex combination $z = \sum_{i=1}^{M} \beta_i h_i^{(L)}$. Only the k-th summand differs between the two passes, hence

$$\left\|z_{\text{fusion}}^{\text{full}} - z_{\text{fusion}}^{\text{masked}}\right\| = \beta_k \,\|h_k^{(L)} - h_{\text{mask}}^{(L)}\| \le \beta_k \,\|\delta_L\| \le L\,\beta_k\,\|x_k\|, \tag{B.7}$$

establishing G.1.

Task prediction. Finally, apply the K-Lipschitz continuity of g:

$$\|g(z_{\text{fusion}}^{\text{full}}) - g(z_{\text{fusion}}^{\text{masked}})\| \le K\,\|z_{\text{fusion}}^{\text{full}} - z_{\text{fusion}}^{\text{masked}}\| \le K\,L\,\beta_k\,\|x_k\|, \tag{B.8}$$

which is G.2. □

Commentary

Inequalities G.1-G.2 say that the influence of dropping a modality scales linearly with three factors:

1. the encoder sensitivity L,
2. the fusion weight β_k assigned to that modality, and
3. the magnitude of the raw input $\|x_k\|$.

Because the fusion weights form a simplex, $\beta_k \le 1$; missing low-weight modalities perturb the fused representation only marginally. Furthermore, by constraining g via spectral normalization, the same linear bound extends to the final prediction.

C Generalization Bound for the Supervised–Contrastive Objective

Proposition 3 (Restatement of Proposition 3: Rademacher complexity bound)**.** *Let $\mathcal{H}$ be the class of CLARGA networks $h : \mathcal{X} \to \mathbb{R}^p$ of the form $h(x) = W\,\varphi(x)$, where:*

1. *the representation map $\varphi : \mathcal{X} \to \mathbb{R}^q$ is implemented by a stack of 1–Lipschitz layers (spectrally normalized linear maps, 1–Lipschitz activations, residual / LayerNorm blocks),*
2. *the prediction matrix $W \in \mathbb{R}^{p \times q}$ satisfies $\|W\|_{\text{F}} \le B$.*

Let the hybrid training loss be

$$\mathcal{L}(h; (x, y)) \coloneqq \mathcal{L}_{\text{sup}}(h(x), y) + \lambda_c\,\mathcal{L}_{\text{NCE}}(h;\, x, batch), \tag{C.1}$$

and assume the per-sample loss is $L_\star$-Lipschitz in its $\mathbb{R}^p$ network-output argument (for fixed labels and batch), with $L_\star$ independent of n.

Given n independent and identically distributed samples, let $\widehat{h} \in \arg\min_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(h;(x_i,y_i))$. Then for any $0<\delta<1$, with probability at least $1-\delta$,

$$\mathcal{E}(\widehat{h}) - \inf_{h\in\mathcal{H}} \mathcal{E}(h) \;\le\; \tilde{O}\left(\sqrt{\frac{B^2\, p\, d_{\text{eff}} \;+\; \log(1/\delta)}{n}}\right), \tag{C.2}$$

where the effective dimension d_{eff} can be chosen as

$$d_{\text{eff}} \;\coloneqq\; \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \|\varphi(x_i)\|_2^2, \qquad \textit{(empirical second moment of the learned representation),} \tag{C.3}$$

and $\tilde{O}(\cdot)$ hides universal numerical constants and polylogarithmic factors in the contrastive batch size. The multiplicative dependence on the loss Lipschitz constant $L_\star$ (e.g., $1+\lambda_c/\tau$) is absorbed into the leading constant.

Equivalently, one may write d_{eff} via the (uncentered) second-moment matrix

$$\widehat{\Sigma}_\varphi = \tfrac{1}{n}\sum_{i=1}^{n} \varphi(x_i)\varphi(x_i)^\top, \qquad d_{\text{eff}} = \operatorname{tr}(\widehat{\Sigma}_\varphi). \tag{C.4}$$

Moreover, if $\|J_\varphi(x)\|_2 \le 1$ for all x and $\|W\|_{\text{F}} \le B$, then the input-Jacobian of $h(x) = W\varphi(x)$ satisfies the one-sided bound

$$\frac{1}{n}\sum_{i=1}^{n} \operatorname{tr}\big(J_h(x_i)J_h(x_i)^\top\big) \;\le\; B^2\, d_{\text{eff}}. \tag{C.5}$$

Auxiliary lemmas

Lemma 1 (Lipschitzness of the hybrid loss)**.** *Suppose $\mathcal{L}_{\text{sup}}(\cdot,y)$ is 1-Lipschitz in its $\mathbb{R}^p$ logit argument (e.g., cross-entropy with bounded logits), and $\mathcal{L}_{\text{NCE}}(\cdot;\textit{batch})$ is $1/\tau$-Lipschitz in its $\mathbb{R}^p$ logit argument (InfoNCE with temperature $\tau>0$). Then, for fixed labels and a fixed batch of negatives,*

$$\mathcal{L}(\cdot) = \mathcal{L}_{\text{sup}}(\cdot,y) + \lambda_c\, \mathcal{L}_{\text{NCE}}(\cdot;\textit{batch}) \tag{C.6}$$

is $L_\star$-Lipschitz with

$$L_\star \;\le\; 1 + \frac{\lambda_c}{\tau}. \tag{C.7}$$

Lemma 2 (Vector-valued contraction)**.** *Let Φ be a function class $\Phi \subset \{x \mapsto u(x) \in \mathbb{R}^p\}$ and let $\psi:\mathbb{R}^p \to \mathbb{R}$ be $L_\star$-Lipschitz w.r.t. ℓ_2. Then the empirical Rademacher complexity satisfies*

$$\mathfrak{R}_n(\psi\circ\Phi) \;\le\; L_\star \cdot \mathfrak{R}_n(\Phi), \quad \mathfrak{R}_n(\Phi) \;\coloneqq\; \mathbb{E}_\varepsilon\Big[\sup_{u\in\Phi} \frac{1}{n}\sum_{i=1}^{n} \langle \varepsilon_i, u(x_i)\rangle\Big], \tag{C.8}$$

where $\varepsilon_i \in \mathbb{R}^p$ are independent and identically distributed standard Rademacher vectors.

Lemma 3 (Rademacher complexity of linear predictors with bounded features)**.** *Let*

$$\mathcal{G} = \{x \mapsto W\,\varphi(x) : \|W\|_{\text{F}} \le B\}. \tag{C.9}$$

Let $\varepsilon_i \in \mathbb{R}^p$ be i.i.d. vectors with coordinates taking values ± 1 with probability $1/2$. Then

$$\mathfrak{R}_n(\mathcal{G}) \;\le\; \frac{B\sqrt{p}}{\sqrt{n}} \Big(\frac{1}{n}\sum_{i=1}^{n} \|\varphi(x_i)\|_2^2\Big)^{1/2} \;=\; \frac{B\sqrt{p}}{\sqrt{n}} \sqrt{d_{\text{eff}}}. \tag{C.10}$$

Proof of Lemma 3. By definition and Frobenius duality,

$$\mathfrak{R}_n(\mathcal{G}) = \mathbb{E}_\varepsilon\Big[\sup_{\|W\|_{\text{F}}\le B} \frac{1}{n}\sum_{i=1}^{n} \langle \varepsilon_i, W\,\varphi(x_i)\rangle\Big] = \frac{B}{n}\,\mathbb{E}_\varepsilon \Big\|\sum_{i=1}^{n} \varepsilon_i\, \varphi(x_i)^\top\Big\|_{\text{F}}. \tag{C.11}$$

By independence,

$$\mathbb{E}_\varepsilon \Big\|\sum_{i=1}^{n} \varepsilon_i\, \varphi(x_i)^\top\Big\|_{\text{F}}^2 = \sum_{i=1}^{n} \mathbb{E}\|\varepsilon_i\|_2^2\, \|\varphi(x_i)\|_2^2 = p\sum_{i=1}^{n} \|\varphi(x_i)\|_2^2. \tag{C.12}$$

Taking square roots and applying Jensen yields

$$\mathbb{E}_\varepsilon \Big\| \sum_{i=1}^{n} \varepsilon_i \, \varphi(x_i)^\top \Big\|_{\mathrm{F}} \;\le\; \sqrt{p} \Big(\sum_{i=1}^{n} \|\varphi(x_i)\|_2^2 \Big)^{1/2}, \tag{C.13}$$

which gives (C.10). □

Proof of Proposition 3

Step 1: Reduce to Rademacher complexity of network outputs. Let

$$\mathcal{F} := \{\, x \mapsto h(x) = W\varphi(x) \;:\; h \in \mathcal{H} \,\}. \tag{C.14}$$

By Lemma 1 and vector-valued contraction (Lemma 2),

$$\mathfrak{R}_n(\mathcal{L} \circ \mathcal{F}) \;\le\; L_\star \, \mathfrak{R}_n(\mathcal{F}). \tag{C.15}$$

Step 2: Bound $\mathfrak{R}_n(\mathcal{F})$ by linear complexity with bounded features. Applying Lemma 3 pointwise for any fixed φ yields

$$\mathfrak{R}_n(\{x \mapsto W\varphi(x) : \|W\|_F \le B\}) \;\le\; \frac{B\sqrt{p}}{\sqrt{n}} \Big(\frac{1}{n} \sum_{i=1}^{n} \|\varphi(x_i)\|_2^2 \Big)^{1/2}. \tag{C.16}$$

Taking the supremum over $h \in \mathcal{H}$ (equivalently over admissible φ) and using the definition of d_{eff} in (C.3) gives

$$\mathfrak{R}_n(\mathcal{F}) \;\le\; \frac{B\sqrt{p}}{\sqrt{n}} \sqrt{d_{\text{eff}}}. \tag{C.17}$$

Step 3: Generalization via standard symmetrization. Denote by $\widehat{\mathcal{E}}(h)$ the empirical hybrid risk and $\mathcal{E}(h)$ its population counterpart. Standard symmetrization and McDiarmid concentration yield, with probability at least $1-\delta$, for all $h \in \mathcal{H}$,

$$\mathcal{E}(h) \;\le\; \widehat{\mathcal{E}}(h) + 2\,\mathfrak{R}_n(\mathcal{L} \circ \mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}. \tag{C.18}$$

Apply this inequality to $\widehat{h}$ and subtract $\inf_{h\in\mathcal{H}} \mathcal{E}(h)$ from both sides to obtain

$$\mathcal{E}(\widehat{h}) - \inf_{h\in\mathcal{H}} \mathcal{E}(h) \;\le\; 2L_\star \cdot \frac{B\sqrt{p}}{\sqrt{n}} \sqrt{d_{\text{eff}}} \;+\; 3\sqrt{\frac{\log(2/\delta)}{2n}}, \tag{C.19}$$

which matches (C.2) up to absolute constants and logarithmic factors hidden in $\tilde{O}(\cdot)$. □

Interpretation and connections

The bound (C.2) isolates three key causes of generalization:

1. *Capacity via B.* The Frobenius constraint on the final linear map controls the size of the function class, acting as a proxy for margin or weight decay. Smaller B tightens the bound.
2. *Effective dimension d_{eff}.* Rather than the ambient width q, the bound depends on the empirical second moment of the learned representation $\varphi(x)$ (i.e., the trace of its uncentered second-moment matrix). This quantity shrinks when CLARGA learns compact, low-variance fused embeddings. This shows a lower intrinsic complexity of the data.
3. *Sample size n.* The usual $1/\sqrt{n}$ decay is recovered. Larger batches in InfoNCE affect only polylogarithmic terms (hidden in $\tilde{O}$) through the Lipschitz constant of the contrastive term.

Under standard spectral normalization of internal layers,

$$\|J_\varphi(x)\|_2 \le 1 \tag{C.20}$$

for all x, so

$$\|h(x)\|^2 = \|W\varphi(x)\|^2 \le \|W\|_{\mathrm{F}}^2 \, \|\varphi(x)\|^2 \le B^2 \, \|\varphi(x)\|^2. \tag{C.21}$$

Summing over the sample connects $B^2 d_{\mathrm{eff}}$ with the (empirical) Fisher-type quantity:

$$\sum_i \mathrm{tr}\big(J_h(x_i) J_h(x_i)^\top\big), \tag{C.22}$$

justifying the Jacobian phrasing in the main text.

D Mutual-Information View of the Contrastive Term

Proposition 4 (InfoNCE lower-bounds mutual information)**.** *Let $(H, Z) \sim p(h, z)$ be a pair of random variables where H denotes a single-modality embedding (the output of encoder f_m and subsequent message passing) and Z denotes the fused representation z_{fusion}. Fix a score (critic) function*

$$s : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R},$$

and let the InfoNCE *loss with batch size $K \ge 2$ be*

$$\mathcal{L}_{\text{NCE}}^{(K)}(s) \;=\; \mathbb{E}\left[-\log \frac{\exp\big(s(H, Z)\big)}{\exp\big(s(H, Z)\big) + \sum_{j=1}^{K-1} \exp\big(s(H, Z_j^-)\big)} \right], \qquad Z_1^-, \ldots, Z_{K-1}^- \overset{\text{i.i.d.}}{\sim} p(z), \tag{D.1}$$

with (H, Z) independent of the negatives $(Z_j^-)_j$. Then for any s,

$$I(H; Z) \;\ge\; \log K \;-\; \mathcal{L}_{\text{NCE}}^{(K)}(s). \tag{D.2}$$

Moreover, the bound is tight in the limit of a rich critic family: if

$$s^\star(h, z) = \log \frac{p(h, z)}{p(h)p(z)} + c \tag{D.3}$$

(for any additive constant c) is attainable, then $\log K - \mathcal{L}_{\text{NCE}}^{(K)}(s^\star)$ is non-decreasing in K and converges to $I(H; Z)$ as $K \to \infty$. For finite K the inequality in (D.2) *is generally strict.*

Proof. Write the mutual information as

$$I(H; Z) = \mathbb{E}\left[\log \frac{p(z|h)}{p(z)}\right] \tag{D.4}$$

For fixed h, define the random variable

$$\ell(h;\, z_0, z_1, \ldots, z_{K-1}) = -\log \frac{\exp\{s(h, z_0)\}}{\sum_{j=0}^{K-1} \exp\{s(h, z_j)\}}, \tag{D.5}$$

where

$$z_0 \sim p(z \mid h) \quad \text{and} \quad z_1, \ldots, z_{K-1} \overset{\text{i.i.d.}}{\sim} p(z). \tag{D.6}$$

By standard noise-contrastive arguments, Jensen's inequality yields

$$\mathbb{E}\big[\ell(h;\, z_0, z_{1:K-1}) \,\big|\, h\big] \;\ge\; -\log \frac{\exp\{\mathbb{E}[s(h, Z) \mid h]\}}{\exp\{\mathbb{E}[s(h, Z) \mid h]\} + (K-1)\exp\{\mathbb{E}[s(h, Z^-) \mid h]\}}. \tag{D.7}$$

Taking full expectation and rearranging,

$$\mathcal{L}_{\text{NCE}}^{(K)}(s) \;\le\; \log\Big(1 + (K-1)\,\mathbb{E}\big[\exp\{s(H, Z^-) - s(H, Z)\}\big]\Big). \tag{D.8}$$

If

$$s^\star(h,z) = \log \frac{p(z|h)}{p(z)} + c, \tag{D.9}$$

then

$$\exp\{s^\star(H,Z^-) - s^\star(H,Z)\} = \frac{p(Z^- \mid H)}{p(Z^-)} \cdot \frac{p(Z)}{p(Z \mid H)}. \tag{D.10}$$

Taking expectation over (H, Z, Z^-) gives

$$\mathbb{E}\big[\exp\{s^\star(H,Z^-) - s^\star(H,Z)\}\big] = \mathbb{E}_H\Big[\underbrace{\mathbb{E}_{Z^-}\Big[\tfrac{p(Z^-|H)}{p(Z^-)}\Big]}_{=1} \cdot \underbrace{\mathbb{E}_{Z|H}\Big[\tfrac{p(Z)}{p(Z|H)}\Big]}_{=1}\Big] = 1. \tag{D.11}$$

Thus the Jensen-based upper bound yields

$$\mathcal{L}^{(K)}_{\mathrm{NCE}}(s^\star) \;\le\; \log\big(1 + (K-1)\cdot 1\big) \;=\; \log K. \tag{D.12}$$

This does not imply equality in (D.2) for finite K. The mutual-information lower bound (D.2) follows from the standard classification (noise-contrastive) derivation of InfoNCE, and with the optimal critic $s^\star$ the quantity $\log K - \mathcal{L}^{(K)}_{\mathrm{NCE}}(s^\star)$ is non-decreasing in K and converges to $I(H;Z)$ as $K \to \infty$. For general s, the variational argument shows (D.2) holds as an inequality. □

Remarks. The bound in (D.2) is non-decreasing in K and becomes tight only in the limit $K \to \infty$ when the critic family contains the log-density-ratio. For finite K the inequality is generally strict. In our instantiation $s(h,z) = \tau_{\mathrm{NCE}}^{-1}\cos(h,z)$, the temperature τ_{NCE} scales the critic and thus the loss sensitivity but does not alter the validity of the lower bound.

E Residual Connections, Layer Normalization, and Over-Smoothing

Proposition 5 (Per-node non-collapse under affine-free LayerNorm)**.** *Consider the linearized propagation block acting on $H \in \mathbb{R}^{M\times d}$,*

$$\mathcal{T}(H) \;=\; \mathrm{LN}\big(H + AHW\big), \qquad A \in \mathbb{R}^{M\times M} \textit{ row-stochastic}, \quad \|W\|_2 \le 1, \tag{E.1}$$

where LN *denotes per-node, affine-free LayerNorm. For $\ell \ge 0$ and node $i \in \{1,\dots,M\}$, let*

$$\tilde h_i^{(\ell+1)} \;=\; h_i^{(\ell)} + \big(AH^{(\ell)}W\big)_i \quad \textit{and} \quad h_i^{(\ell+1)} \;=\; \mathrm{LN}\big(\tilde h_i^{(\ell+1)}\big). \tag{E.2}$$

Denote by

$$\mu(u) \;=\; \frac{1}{d}\sum_{c=1}^{d} u_c, \qquad \sigma^2(u) \;=\; \frac{1}{d}\sum_{c=1}^{d}\big(u_c - \mu(u)\big)^2 \tag{E.3}$$

the feature-wise mean and variance for a vector $u \in \mathbb{R}^d$, and let $\epsilon > 0$ be the LayerNorm stabilizer. Then for every node i and layer ℓ,

$$\mu\big(h_i^{(\ell+1)}\big) = 0, \qquad \frac{1}{d}\big\|h_i^{(\ell+1)}\big\|_2^2 \;=\; \frac{\sigma^2\big(\tilde h_i^{(\ell+1)}\big)}{\sigma^2\big(\tilde h_i^{(\ell+1)}\big) + \epsilon}. \tag{E.4}$$

In particular, if $\sigma^2\big(\tilde h_i^{(\ell+1)}\big) > 0$ then the post-normalization feature variance at node i is strictly positive and bounded above by 1; thus that node's embedding cannot collapse to the zero vector at that layer.

Proof. For any $u \in \mathbb{R}^d$, affine-free LayerNorm is defined component-wise by

$$\mathrm{LN}(u)_c \;=\; \frac{u_c - \mu(u)}{\sqrt{\sigma^2(u) + \epsilon}}, \quad c = 1,\dots,d. \tag{E.5}$$

It follows immediately that $\mu\big(\mathrm{LN}(u)\big) = 0$ and

$$\frac{1}{d}\,\big\|\mathrm{LN}(u)\big\|_2^2 \;=\; \frac{\sigma^2(u)}{\sigma^2(u) + \epsilon}. \tag{E.6}$$

Applying this to $u = \tilde h_i^{(\ell+1)}$ yields the two stated identities. In particular, whenever $\sigma^2\big(\tilde h_i^{(\ell+1)}\big) > 0$, the post-normalization per-node feature variance is strictly positive (and at most 1), so the node's embedding at that layer cannot collapse to the zero vector. □

Consequences for Depth Choice

Proposition 5 guarantees a per-node, feature-wise normalization effect: after each block, every node has zero-mean features and, whenever the pre-normalization feature variance is nonzero, a strictly positive post-normalization variance bounded by 1. This rules out trivial norm collapse at the node level. While this does not preclude over-smoothing across nodes in principle, the residual path $H \mapsto H + AHW$ empirically mitigates the tendency to average out differences, especially at modest depths.

Acknowledgements

The student, Santosh Patapati, carried out the entire research process completely on their own. This includes conducting a literature review, creating the idea, developing the methodology, carrying out experiments, developing theory, compiling results, writing the paper, writing the appendix, and more. The supervisor, Gerry Skiles, was a mentor whom Santosh could go to for broader advice.