SUPPLEMENTARY:
"Multimodal Graph Representation Learning over Arbitrary Sets of Modalities"

A Details of Models Used for Comparison in Experiments

A.1 Early Fusion Baseline

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

A.2 Late Fusion Baseline

The averaged late fusion baseline encodes each modality independently using a dedicated 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 fusion weights specific to each modality.

A.3 Multimodal Lego

MM-Lego wraps each pretrained encoder in a "LegoBlock" that projects its output into a uniform 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 **?**, for as many epochs as the other architectures are fine tuned or trained.

A.4 FuseMix

FuseMix freezes unimodal encoders and then precomputes 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 multilayer perceptrons project these into a common space and align them with a contrastive loss.

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

B 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. $f(x_{\pi(1)}, \ldots, x_{\pi(M)}) = f(x_1, \ldots, x_M)$ 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*

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

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 $h(x_1, \ldots, x_M) = \rho\big(\sum_i \phi(x_i)\big)$ 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

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

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

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

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 $x \mapsto V\big(\mathrm{LeakyReLU}(U(x))\big) \equiv \rho_\varepsilon(x)$. The residual skip and LayerNorm can be absorbed into V.

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

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

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.

C 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*

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

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

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

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

Set

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

By encoder Lipschitzness,

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

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 and update block is 1-Lipschitz, and

$$\|\delta_\ell\| \;\le\; \|\delta_{\ell-1}\|, \quad \ell = 1, \dots, D. \tag{C.6}$$

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

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

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

establishing G.1.

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

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

which is G.2. □

[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.

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, $\beta_k \leq 1$; 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.

D 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 $\|W\|_{\mathrm{F}} \leq B$.*

Let the hybrid training loss be

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

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 $1-\delta$,

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

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

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

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

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

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

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{tr}\big(J_h(x_i)J_h(x_i)^\top\big) \;\leq\; B^2\,d_{\mathrm{eff}}. \tag{D.5}$$

Auxiliary lemmas

Lemma 1 (Lipschitzness of the hybrid loss)**.** *Suppose $\mathcal{L}_{\mathrm{sup}}(\cdot,y)$ is 1-Lipschitz in its $\mathbb{R}^p$ logit argument (e.g., cross-entropy with bounded logits), and $\mathcal{L}_{\mathrm{NCE}}(\cdot;\textit{batch})$ 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,

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

is $L_\star$-Lipschitz with

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

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*

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

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*

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

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

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

Proof of Lemma 3. By definition and Frobenius duality,

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

By independence,

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

Taking square roots and applying Jensen yields

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

which gives (D.10). □

Proof of Proposition 3

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

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

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

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

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

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

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

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

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 $h \in \mathcal{H}$,

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

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

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

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

Interpretation and connections

The bound (D.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,

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

for all x, so

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

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

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

justifying the Jacobian phrasing in the main text.

E 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*

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

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

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

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

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

(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 (E.2) *is generally strict.*

Proof. Write the mutual information as

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

For fixed h, define the random variable

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

where

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

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

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

Taking full expectation and rearranging,

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

If

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

then

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

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

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

Thus the Jensen-based upper bound yields

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

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

Remarks. The bound in (E.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 $s(h,z) = \tau_{\text{NCE}}^{-1} \cos(h,z)$, the temperature τ_{NCE} scales the critic and thus the loss sensitivity but does not alter the validity of the lower bound.

F Residual Connections, Layer Normalization, and Over-Smoothing

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

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

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

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

Denote by

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

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,

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

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

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

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

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

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.