LEGRAMJEPA: NEGATIVE-FREE MULTIMODAL GRAPH FOUNDATION MODELS IN LATENT SPACE WITH VOLUMETRIC ALIGNMENT AND SIGREG

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised learning on text-attributed graphs enables graph foundation models that transfer to out-of-distribution graphs without labels. The standard approach aligns a graph encoder and a text encoder with a contrastive objective, pulling positive graph-text pairs together and pushing mismatched pairs apart. This depends on large numbers of negative pairs, and extending it beyond two modalities requires a pairwise alignment term for each pair of modalities, so the number of terms grows quadratically. We present LeGramJEPA, a latent-space approach to multimodal graph self-supervised learning that uses no negatives and whose alignment cost grows linearly in the number of modalities. LeGramJEPA replaces the contrastive term with a volumetric alignment objective and constrains each encoder's output distribution toward an isotropic Gaussian with SIGReg to prevent embedding collapse. The volume term couples all modalities in a single quantity per anchor, reducing the number of alignment terms from quadratic to linear in modalities compared to cross-modal methods. Pretraining with LeGramJEPA is competitive with contrastive baselines on zero-shot node classification and improves over them on link prediction. Under a three-modality (graph, text, image) training and transfer scheme, LeGramJEPA consistently outperforms its contrastive counterpart across modality pairs.

1 INTRODUCTION

Text-Attributed Graphs (TAGs) are an important data format for machine learning, as their applications span domains whose data comprises both text and relational structure: e.g., recommendation systems, social networks, e-commerce platforms, and citation networks (Hu et al., 2020a; Yan et al., 2023). TAGs also offer rich text attributes in addition to the graph structure. Because node features live in a shared textual space, a model pretrained on one TAG can in principle be applied to another whose label space it has never seen, making TAG foundation models a promising approach to graph foundation models (GFM) (Zhu et al., 2025b; Fang et al., 2025).

Prior TAG methods either inject relational structure into a language model, reconstruct masked inputs (Jin et al., 2023a; He et al., 2025; Zhao et al., 2024; Yang et al., 2021; Jin et al., 2023b), or use text as a universal feature space to unify tasks across domains (Liu et al., 2024; Li et al., 2024; Wang et al., 2024; He et al., 2024; Zhao et al., 2023; Chen et al., 2024; Tang et al., 2024). The first body of work is generative and evaluated in-domain, reporting no cross-graph transfer; the second is supervised or multi-task, so its transfer is bounded by how well the source and target label spaces align. The strongest zero-shot transfer reported between domains in TAGs instead comes from a third area of work, which abandons labels and input-space reconstruction for alignment in latent space, adopting the CLIP formulation (Radford et al., 2021) in a graph–text setting (Zhu et al., 2025b; Fang et al., 2025; Zhang et al., 2025b). As with contrastive learning, the negatives are what prevent a degenerate solution in which the encoder maps every input to a single point, so performance depends on the number and quality of negative pairs (Robinson et al., 2021).

In this work we remove that dependency. Rather than pushing apart negatives, we align positive pairs by prediction in latent space and prevent degenerate solutions with an explicit constraint on the embeddings, so training needs only positives. This approach has been developed in single-modality

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

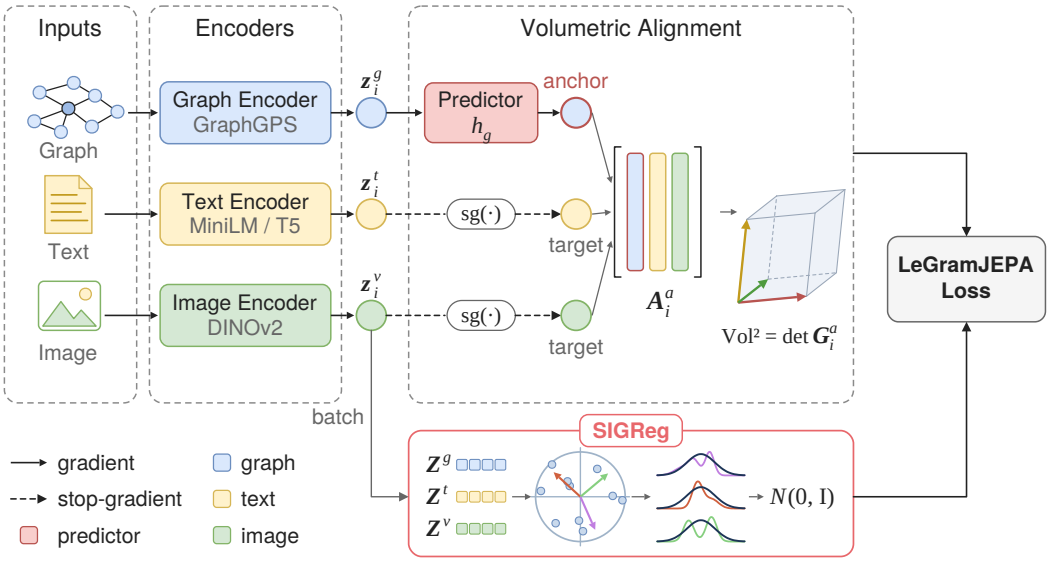


Figure 1: Overview of LeGramJEPA, a negative-free, multimodal self-supervised graph learning approach for TAGs. Each modality is encoded into $\boldsymbol{z}_i^a$. For each anchor modality a, the predictor output $h_a(\boldsymbol{z}_i^a)$ and the stop-gradient targets $\mathrm{sg}(\boldsymbol{z}_i^b)_{b\neq a}$ form the columns of $\boldsymbol{A}_i^a$. Their squared volume $\det \boldsymbol{G}_i^a$ gives one alignment term per anchor, so there are k terms instead of the $k(k-1)$ required by pairwise alignment. SIGReg pushes each modality's batch $\boldsymbol{Z}^a$ toward $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ along random 1-D projections, which prevents collapse without negatives.

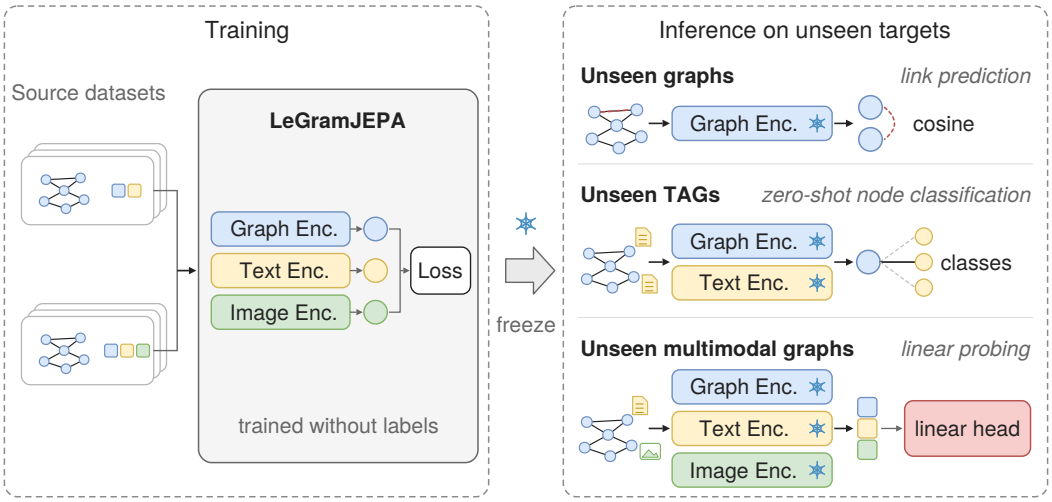


Figure 2: Training and deployment of LeGramJEPA. *Left:* a single model is pretrained without labels on several graph–text and graph–text–image source datasets. *Right:* the frozen encoders transfer to unseen graphs (link prediction), unseen TAGs (zero-shot node classification against class descriptions) and unseen multimodal graphs (linear probing).

settings for graphs (Thakoor et al., 2022; Bini & Marchand-Maillet, 2025; Skenderi et al., 2025) and images (Assran et al., 2023), where collapse is prevented either by architectural asymmetry or by an explicit statistical constraint on the embedding distribution (Balestriero & LeCun, 2025). Extensions to the multimodal vision–language setting include asymmetric work such as (Chen et al., 2025) and information-maximization methods (Kuhn et al., 2026). It has not, however, been brought to the graph–text setting, despite TAGs being an accessible multimodal setting.

Cross-modal alignment is a popular method of alignment in multimodal work typically written as a term over a pair of modalities, predicting one modality's embedding from the other (Radford et al., 2021; Chen et al., 2025; Kuhn et al., 2026). Extending this to $k > 2$ modalities by summing over modality pairs makes the number of alignment terms grow quadratically in k. As such, Cicchetti et al. (2025) replaces this pairwise construction with a single quantity: the volume spanned by k modality embeddings obtained from a Gram determinant. It has so far been shown only as a replacement for a contrastive objective; we extend it to the information maximization setting.

We therefore propose LeGramJEPA, to our knowledge, the first negative-free multimodal self-supervised graph learning approach with volumetric alignment (Figure 1).

1.1 Positioning.

Graph-text self-supervised learning is exhausted by two objective families: contrastive alignment (Zhu et al., 2025b; Fang et al., 2025; Zhang et al., 2025b) and relation-aware reconstruction (Jin et al., 2023a; He et al., 2025; Zhao et al., 2024). To our knowledge, no published latent-space graph multimodal method is negative-free, and none constrains the embedding distribution rather than relying on the uniformity implicit in InfoNCE (Wang & Isola, 2020). LeGramJEPA addresses this gap. It aligns modalities through a volumetric term, replaces the implicit uniformity term with a per-modality distributional constraint, and so gives a negative-free multimodal graph objective whose alignment term is linear rather than quadratic in the number of modalities.

We make the following contributions:

- We introduce a novel learning objective based on volumetric alignment coupling all modalities in a single determinant term per anchor and reducing the alignment term to linear in the number of modalities. We instantiate this as LeGramJEPA, a negative-free multimodal graph foundation model that pairs volume alignment with SIGReg.
- We show that LeGramJEPA transfer learning matches the contrastive state-of-the-art on zero-shot node classification and improves on it for link prediction.
- We show that pretraining with LeGramJEPA under a three-modality (graph, text, image) transfer scheme consistently outperforms its contrastive counterpart across modality pairs.

2 Related Work

2.1 Structure-aware language model pretraining.

Earlier work in the line of TAGs treats structures as a supervision signal for the text encoder, discarding the graph at inference. SPECTER (Cohan et al., 2020) trains document embeddings against a citation-derived triplet loss. Similarly, Patton (Jin et al., 2023a) gives a self-supervised objective a network context by combining masked language modelling with masked node prediction. UniGraph (He et al., 2025) induces a network context by masked language modelling with only samples of text from neighbouring nodes while including a GNN head whose nodes are initialized with CLS tokens. P2TAG (Zhao et al., 2024) combines masked language modelling with graph pretraining and adapts downstream through prompt tuning. GraphFormers (Yang et al., 2021) and EdgeFormers (Jin et al., 2023b) instead nest message passing inside transformer blocks, so structure enters the attention computation. Most of these methods, with the exception of UniGraph, are trained and evaluated under the same domain of graphs; thus, they report no cross-graph transfer. Unlike these methods, which learn by reconstructing masked tokens, our work compares objectives that align graph and text embeddings in latent space.

2.2 Unified text-space graph foundation models.

A parallel line of work uses natural language as a universal feature space to unify graphs and tasks of varying domains. OFA (Liu et al., 2024) describes nodes and edges in text, encodes them with a shared language model, standardizes node, link and graph tasks through nodes-of-interest, and appends prompt substructures for in-context inference. ZeroG (Li et al., 2024) follows a similar procedure but improves cross-dataset zero-shot transfer through improving feature misalignment

and negative transfer. GFT (Wang et al., 2024) improves transfer through identifying computation trees as the transferable unit, and quantizing them into a discrete vocabulary. Most of this family is supervised or multi-task rather than self-supervised.

2.3 GRAPH-TEXT ALIGNMENT.

A large body of work on multimodal learning utilizes a two-tower approach with a graph encoder and a text encoder and aligns them with an InfoNCE objective, inheriting the CLIP formulation (Radford et al., 2021). GraphCLIP (Zhu et al., 2025b) follows this, aligning graph and text embeddings via a contrastive loss; pulling positive pairs together while disaligning negative pairs. This method scales this to cross-domain transfer by utilizing LLM-generated summaries of sampled subgraphs to supply the text side of each pair. UniGLM (Fang et al., 2025) pursues a similar objective; however, it uses adaptive positive sampling and a lazy embedding-update mechanism against a frozen language encoder. TAGA (Zhang et al., 2025b) constructs two complementary views: node texts serialized into a document following graph topology, and the text corpus reorganized as a graph, aligning them via a contrastive loss. Recent work targets some failure modes of these methods, namely degenerate solutions, by either using weighted pairs according to estimated quality (Liu et al., 2026) or by arguing that a nonzero graph-text representation gap is beneficial, removing the problem of complete embedding collapse (Zhang et al., 2025a). However, every method in this family is contrastive, relying heavily on quality negative pairs.

2.4 NEGATIVE-FREE AND INFORMATION-MAXIMIZATION SSL

Self-supervised methods that avoid negatives avoid degenerate solutions in one of two ways. The first uses architectural asymmetry: a predictor on the gradient branch and a stop-gradient on the target, optionally with a momentum encoder (Grill et al., 2020; Chen & He, 2021; Assran et al., 2023). Such mechanisms have only been partially characterized (Tian et al., 2021) and adapted to graphs through bootstrapping, with Laplacian augmentations, and multi-views (Thakoor et al., 2022; Bini & Marchand-Maillet, 2025; Skenderi et al., 2025). The second replaces that asymmetry with an explicit statistical constraint on the embedding distribution (Bardes et al., 2022; Zbontar et al., 2021; Balestriero & LeCun, 2025; Zimmermann et al., 2025). These methods have been similarly extended to multiple modalities with architectural asymmetry (Chen et al., 2025; Lei et al., 2025) and by constraints on the embedding distribution Kuhn et al. (2026).

3 PRELIMINARIES

This work focuses on Text-Attributed Graphs, for which the nodes have text as their features. Formally, we define this type of data and the zero-shot learning on it.

Definition 3.1 (Text-Attributed Graph)**.** A Text-Attributed Graph (TAG) $\mathcal{G}$ is a graph such that $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \{T_n\}_{n=1}^{N}\}$, where $\mathcal{V}$ represents the set of $N = |\mathcal{V}|$ nodes $v_i \in \mathcal{V}$, and $\mathcal{E}$ represents the set of edges $(v_i, v_j) \in \mathcal{E}$. The raw text is represented by $T_n \in \mathcal{T}^L$ for node $n \in [1, 2, \ldots, N]$, where $\mathcal{T}$ is the token space and L is the maximum sequence length. This can be extended to a feature matrix $\boldsymbol{X} \in \mathbb{R}^{N \times d}$ where the raw text is processed into high-dimensional d text embeddings.

Definition 3.2 (Zero-Shot Learning)**.** Let $\mathcal{G}^s$ be source data used for pretraining and $\mathcal{G}^t$ a target TAG disjoint from $\mathcal{G}^s$, with label space $\mathcal{Y}^t$. A model $f_{\theta^\star}$ pretrained on $\mathcal{G}^s$ is evaluated zero-shot on $\mathcal{G}^t$ if its predictions follow from only $\theta^\star$. That is, there is no learning on $\mathcal{G}^t$.

Definition 3.3 (Node Classification)**.** Given a target TAG $\mathcal{G}^t$ with label sentences $\{S_c\}_{c \in \mathcal{Y}^t}$, the predicted class of node v_n is the label whose sentence is most similar to its representation,

$$\hat{y}_n \;=\; \arg\max_{c \,\in\, \mathcal{Y}^t} \; \mathrm{sim}(\boldsymbol{z}_n,\; \boldsymbol{z}_{S_c})\,. \tag{1}$$

We write $\mathrm{sim}(\cdot, \cdot)$ for cosine similarity.

Definition 3.4 (Link Prediction)**.** Given a target TAG $\mathcal{G}^t$, a candidate pair $(u, v) \in \mathcal{V}^t \times \mathcal{V}^t$ is scored by the similarity of the two node representations,

$$s(u, v) \;=\; \mathrm{sim}(\boldsymbol{z}_u,\; \boldsymbol{z}_v)\,, \tag{2}$$

Performance is measured by the ranking that s induces over held-out edges against sampled non-edges.

4 METHODOLOGY

Let us define our data batch as $D = \{(G_i, T_i)\}_{i=1}^{B}$, with graphs $G_i \in \mathcal{G}$ and their attributed text $T_i \in \mathcal{T}$; every node of an input subgraph carries text, or a subgraph inherits text from its nodes. Here we have graph and text as our two modalities and, in Section 4.2, allow $k \geq 2$ modalities indexed by $a \in \{1, \ldots, k\}$. Each modality has an encoder mapping its input to a space $\mathbb{R}^d$. For the graph–text case we write $f_\theta : \mathcal{G} \to \mathbb{R}^d$ for the graph encoder and $g_\phi : \mathcal{T} \to \mathbb{R}^d$ for the text encoder, and denote the embedding of modality a for sample i by $\boldsymbol{z}_i^a \in \mathbb{R}^d$. We write $\boldsymbol{Z}^a = \{\boldsymbol{z}_i^a\}_{i=1}^B$ for the batch of modality-a embeddings and $\mathrm{sg}(\cdot)$ for the stop-gradient operator.

The goal is to build a multi-modal latent-space architecture which does not rely on negatives. LeJEPA (Balestriero & LeCun, 2025) learns without negatives and without architectural asymmetry by pairing a predictive term, which aligns views of a sample together, and SIGReg (Appendix C.3). SIGReg constrains each view's batch of embeddings toward an isotropic Gaussian and prevents collapse.

However, in a multimodal setting, complete removal of this asymmetry has been shown to hurt zero-shot and linear probing performance (Kuhn et al., 2026). As such, we follow suit with LeVLJEPA (Kuhn et al., 2026) which uses asymmetry defined in SimSiam (Chen & He, 2021): a predictor and a stop-gradient. The prediction flows from an anchor modality through a predictor, and the remaining modalities it aligns to are detached via the stop-gradient. Here, modalities are used in place of views. That is, the modalities of a sample are the positives to be aligned, and SIGReg regularizes the distribution of each modality's embeddings across the batch.

4.1 CROSS-MODAL PREDICTION WITH STOP-GRADIENT

For two modalities, LeVLJEPA attaches a two-layer MLP predictor to each output embedding. Under a graph–text modality, this would be defined as: $h_g : \mathbb{R}^d \to \mathbb{R}^d$, which intuitively can be thought of as predicting the text embedding from the graph embedding, and $h_t : \mathbb{R}^d \to \mathbb{R}^d$ the reverse. Each target is detached, so gradients flow only through the predictor and its own encoder. The loss is the MSE between the predictor output and its detached target, plus a SIGReg term per modality:

$$\mathcal{L}_{\text{cross}} = \frac{1}{B} \sum_{i=1}^{B} \Big[\big\| h_g(\boldsymbol{z}_i^g) - \mathrm{sg}(\boldsymbol{z}_i^t) \big\|_2^2 + \big\| h_t(\boldsymbol{z}_i^t) - \mathrm{sg}(\boldsymbol{z}_i^g) \big\|_2^2 \Big]. \tag{3}$$

This cross term is a sum over pairs of modalities. One predictor output is aligned to one detached target. Generalizing it to k modalities by aligning every anchor to every other modality yields $k(k-1)$ terms. That is, quadratic in the number of modalities. For three modalities, this is six terms; we give the explicit expansion of this in Appendix E.

Cicchetti et al. (2025) address a similar quadratic problem for a contrastive objective by replacing pairwise alignment with a single geometric quantity. They use the volume of the k-parallelotope spanned by the modality embeddings. We extend this methodology to the information-maximization setting, while preserving the predictor/stop-gradient asymmetry required in a multi-modal setting.

4.2 LEGRAMJEPA

For sample i and each anchor modality a, we form the $d \times k$ matrix whose columns are the k modality directions, with the anchor column carrying the anchor predictor output $h_a(\boldsymbol{z}_i^a)$ and the remaining $k-1$ columns carrying the detached targets $\mathrm{sg}(\boldsymbol{z}_i^b)$, $b \neq a$:

$$\boldsymbol{A}_i^a = \big[\, h_a(\boldsymbol{z}_i^a) \;\big|\; \mathrm{sg}(\boldsymbol{z}_i^b)_{\,b \neq a} \,\big] \in \mathbb{R}^{d \times k}. \tag{4}$$

The single anchor column and $k-1$ detached columns are the volumetric analogue of the pairwise predictor and stop-gradient terms of equation 10. Gradients enter only through the anchor's predictor and encoder. Its Gram matrix and squared spanned volume are:

$$\boldsymbol{G}_i^a = (\boldsymbol{A}_i^a)^\top \boldsymbol{A}_i^a \in \mathbb{R}^{k \times k}, \qquad \mathrm{Vol}(\boldsymbol{A}_i^a)^2 = \det \boldsymbol{G}_i^a. \tag{5}$$

The objective sums one volume term per anchor and one SIGReg term per modality:

$$\mathcal{L}_{\text{LeGramJEPA}} = \frac{1-\lambda_{\text{tot}}}{kB} \sum_{a=1}^{k} \sum_{i=1}^{B} \det \boldsymbol{G}_i^a + \sum_{a=1}^{k} \lambda_a \, \text{SIGReg}(\boldsymbol{Z}^a), \qquad \lambda_{\text{tot}} = \sum_{a=1}^{k} \lambda_a \leq 1. \quad (6)$$

Minimizing this loss drives each sample's modalities toward alignment through the volume term, while SIGReg keeps each modality's batch of embeddings toward an isotropic Gaussian and so prevents collapse. We minimize the squared volume $\det \boldsymbol{G}_i^a$ rather than the volume $\sqrt{\det \boldsymbol{G}_i^a}$. Since the square root has an unbounded gradient as the modalities approach alignment ($\det \boldsymbol{G}_i^a \to 0$), whereas the squared volume stays well conditioned there.

The number of alignment terms is k rather than $k(k-1)$, linear in the number of modalities. The overall illustration of this methodology is given in Figure 1.

4.3 Encoder Architecture

To fairly evaluate our methodology against the CLIP framework in graph learning, we use the exact encoder architecture described in Zhu et al. (2025b). In particular, we use a graph transformer, GraphGPS (Rampásek et al., 2022), and a distilled sentence-level text encoder, MiniLM (Wang et al., 2020b; Reimers & Gurevych, 2019b).

Additionally, each encoder output is passed through a one-layer MLP projection, which must be batch-normalized, as work has shown that layer normalization prevents the SIGReg anti-collapse term from being optimized sufficiently (Kuhn et al., 2026; Maes et al., 2026). The graph and sentence transformer include layer normalization in their final layer; as such, this projection is sometimes required for better performance. We treat the inclusion of this projection as a hyperparameter.

5 Experiments

We pre-train LeGramJEPA on source datasets OOD from the targets, evaluating in two regimes. The first is zero-shot transfer on text-attributed graphs under the two-modality (graph, text) protocol defined in Zhu et al. (2025b), covering node classification (Section 5.2) and link prediction (Section 5.3) on seven target graphs. The second is a three-modality (graph, text, image) transfer setting on the MM-Graph benchmark (Zhu et al., 2025a) (Section 5.4). Figure 2 illustrates our training and evaluation scheme.

Datasets. For the two-modality transfer experiments, we follow Zhu et al. (2025b) exactly, pretraining on their source datasets: obgn-Arxiv (Wang et al., 2020a), ArXiv_2023 (He et al., 2023), PubMed (Yang et al., 2016), obgn-Products (Hu et al., 2020b), and Reddit (Huang et al., 2024b). To test transfer results, the target datasets are Cora (Sen et al., 2008), CiteSeer (Giles et al., 1998), WikiCS (Mernyei & Cangea, 2020), Instagram (Huang et al., 2024a), Ele-Photo, Ele-Comp, and Books-History (Yan et al., 2023). To fairly evaluate our method on this benchmark, we similarly use the LLM-summarized subgraph captions provided by Zhu et al. (2025b). For the three-modality experiments, we pretrain on Amazon-Sports, Amazon-Cloth, and Goodreads-LP, and evaluate on the held-out datasets Ele-Fashion and Goodreads-NC (Zhu et al., 2025a); we use DINOv2 features for the image modality, and T5 features for the text modality.

Baselines. We reproduce the grouping of (Zhu et al., 2025b): text-only encoders (BERT (Devlin et al., 2019), SBERT (Reimers & Gurevych, 2019a), DeBERTa (He et al., 2021), E5 (Wang et al., 2022), Qwen2 (Yang et al., 2024), LLaMA3.1 (Team, 2024)). Label-supervised graph-LLM methods: GraphGPT (Tang et al., 2024), LLaGA (Chen et al., 2024), OFA (Liu et al., 2024), ZeroG (Li et al., 2024). Label-free graph and text SSL: DGI (Veličković et al., 2019), GRACE (Zhu et al., 2020), BGRL (Thakoor et al., 2022), GraphMAE (Hou et al., 2022), G2P2 (Wen & Fang, 2023), and the graph-summary contrastive method GraphCLIP. Baseline numbers are as reported by Zhu et al. (2025b). Detailed explanations of baselines are provided in Appendix B.

Architecture control. To attribute performance differences to the loss function rather than to capacity or architecture, LeGramJEPA and GraphCLIP runs use the same encoders defined in Section 4.3. All three operate at the same maximum parameter budget of 150 million.

Table 1: Zero-shot node classification accuracy (%) across target datasets. Source pretraining data are described in 5. Boldface indicates the best performance per dataset. Baseline results are as reported by Zhu et al. (2025b).

Method	Params	Cora	CiteSeer	WikiCS	Instagram	Ele-Photo	Ele-Comp.	Books-Hist.
BERT	110M	19.56±0.98	33.26±2.35	29.37±0.00	57.02±0.57	21.80±0.14	13.88±0.29	9.95±0.42
SBERT	66M	54.35±1.26	50.47±0.90	48.16±0.00	48.34±1.23	35.96±0.44	41.82±0.22	30.45±0.19
DeBERTa	184M	16.42±1.26	16.42±1.26	15.29±0.00	39.81±0.58	12.38±0.26	10.62±0.15	8.70±0.26
E5	110M	44.65±0.82	42.57±0.54	31.49±0.00	61.28±0.97	35.14±0.28	16.54±0.14	12.92±0.48
Qwen2	7B	61.44±1.29	53.57±0.86	58.72±0.25	39.13±0.78	45.55±0.12	59.18±0.20	23.79±0.34
Qwen2	72B	62.18±0.98	60.97±0.87	60.91±0.08	47.70±0.31	52.41±0.39	60.88±0.30	53.56±0.64
LLaMA3.1	8B	57.75±1.21	53.54±1.71	58.32±0.21	39.37±1.14	34.38±0.25	46.98±0.21	22.28±0.18
LLaMA3.1	70B	65.72±1.24	62.79±1.24	62.82±0.04	43.68±0.52	51.26±0.53	61.62±0.42	53.33±0.55
GraphGPT	7B	23.25±1.45	18.04±1.45	6.30±0.26	45.12±1.16	7.62±0.22	29.71±0.83	15.92±0.14
LLaGA	7B	21.44±0.65	16.07±1.15	2.65±0.72	41.12±0.94	6.50±0.53	23.10±0.33	11.17±0.58
OFA	30M	37.25±1.38	29.64±0.19	45.52±1.06	32.71±0.16	33.03±0.64	22.09±0.39	16.87±0.93
ZeroG	66M	62.32±1.91	52.55±1.23	54.93±0.06	48.97±0.78	45.12±0.65	56.20±0.35	40.74±0.65
DGI	128M	24.03±1.40	18.71±1.22	18.86±0.25	61.42±1.12	13.96±0.17	27.12±0.03	15.77±0.02
GRACE	128M	13.69±1.27	22.88±1.49	16.07±0.32	62.23±0.93	10.16±0.13	10.94±0.12	32.39±0.11
BGRL	128M	20.80±1.06	26.50±1.22	18.35±0.22	61.45±0.82	5.21±0.22	24.12±0.22	16.28±0.35
GraphMAE	128M	23.25±1.07	20.75±0.88	12.14±0.20	62.39±0.84	12.53±0.08	8.36±0.06	21.76±0.17
G2P2	63M	41.51±0.78	51.02±0.62	31.92±0.15	52.87±0.78	22.21±0.12	32.52±0.13	26.18±0.25
GraphCLIP	150M	**67.31**±1.76	63.13±1.13	**70.19**±0.10	64.05±0.34	53.40±0.64	**62.04**±0.21	**53.88**±0.35
w/o Freeze	150M	61.29±1.37	61.66±1.01	58.92±0.10	60.15±0.75	45.90±0.65	53.24±0.26	29.40±0.31
LeGramJEPA	150M	65.68±1.36	**67.40**±2.46	62.32±0.92	**65.04**±0.89	**54.86**±0.64	61.14±0.39	48.47±0.78

Evaluation. Node classification holds out a random 20% of nodes for testing and reports accuracy over five seeds. MRR and AUC are used as the metrics in link prediction. MRR samples 100 negatives. More details can be found in Appendix D.1. Linear probing evaluation is detailed in D.2.

5.1 Transfer Learning

Sections 5.2 and 5.3 evaluate the two-modality model under zero-shot transfer (Definition 3.2). Node predictions are the nearest label sentence in embedding space (Definition A.1). Edge scores are the cosine similarity of the two node representations (Definition A.2). Section 5.4 then evaluates the three-modality model under linear probing on held-out graphs.

5.2 Zero-Shot Node Classification

Table 1 reports zero-shot node classification accuracy. LeGramJEPA is competitive with the contrastive state of the art without using negatives: it improves over GraphCLIP on CiteSeer, Instagram, and Ele-Photo, while GraphCLIP retains an advantage on Cora, WikiCS, Ele-Comp, and Books History. GraphCLIP's margins are consistent with its contrastive objective directly optimizing the score used at evaluation (the similarity of matched against unmatched graph-text pairs), an observation similarly noted by Kuhn et al. (2026). LeGramJEPA trains only on positive pairs yet remains within a few points on most targets. As such, this is evidence that non-trivial alignment for this evaluation emerges without negatives, in scenarios where negatives have an advantage.

5.3 Zero-Shot Link Prediction

The clearest gains appear on link prediction. Under AUC, LeGramJEPA outcompetes all the baselines of Zhu et al. (2025b), which are only reported on Cora, WikiCS and Books-History (Table 2). We extend this to the MRR metric (Appendix D.1) but include only GraphCLIP as our baseline. Since it previously outcompeted all other baselines, and we are interested in comparing with the contrastive counterpart. Under MRR (Figure 3), LeGramJEPA outperforms GraphCLIP on four of the seven targets, with the largest margins on the two e-commerce graphs, ties on Instagram, and trails on Cora and CiteSeer.

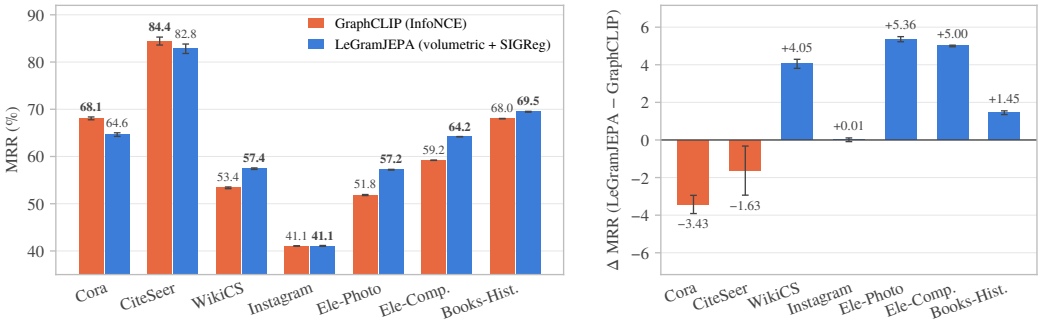


Figure 3: Zero-shot link prediction (MRR, %). *Left:* absolute scores. *Right:* difference (LeGramJEPA − GraphCLIP), with the combined standard deviation as error bars. Positive values favour LeGramJEPA.

Table 2: Zero-shot link prediction across target datasets (AUC, %). Boldface indicates the best value per column. Baseline results are as reported by Zhu et al. (2025b).

Method	Cora	WikiCS	Books-Hist.
SBERT	77.76±0.54	81.55±0.09	87.02±0.07
E5	69.57±0.60	80.69±0.06	69.93±0.07
GRACE	67.19±0.56	69.33±0.06	76.90±0.05
GraphMAE	70.17±0.77	73.39±0.04	83.26±0.07
OFA	75.36±0.89	89.03±0.09	86.47±0.10
ZeroG	81.20±0.82	88.82±0.07	92.85±0.12
GraphCLIP	83.15±0.76	92.67±0.10	96.04±0.05
LeGramJEPA	**97.77**±0.32	**93.90**±0.03	**98.57**±0.02

During link prediction, the candidate edge is scored by comparing two node embeddings (Appendix D.1), so the text encoder does not play a role at test time, and the result depends only on the geometry of the node embedding space. Under GraphCLIP, the embedding space of the node would be shaped indirectly by positive and negative text pairs. SIGReg directly pushes the graph embeddings towards an isotropic Gaussian. Empirically, this provides better geometry for link prediction.

5.4 Three-Modality Transfer Learning

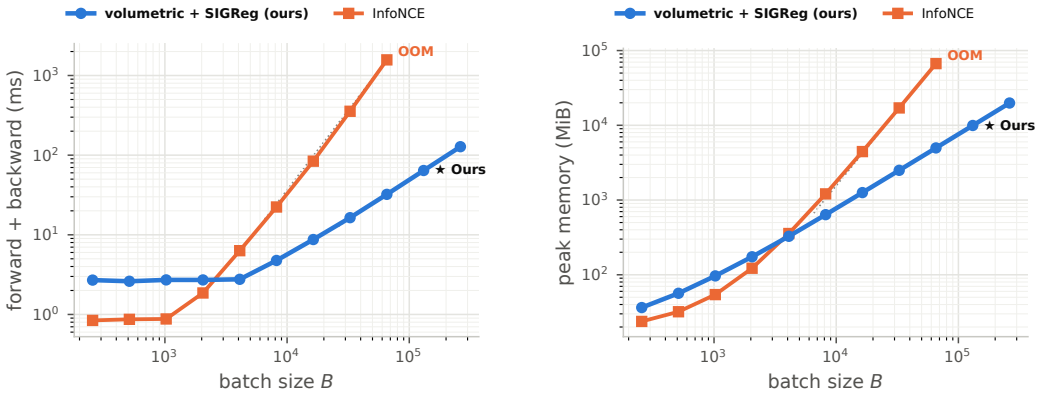
Table 3 evaluates linear-probe node classification on target graphs after pretraining on our source datasets (Section 5). LeGramJEPA improves over GraphCLIP on both target datasets and both reported modality pairs, with the largest margin on Goodreads-NC. LeGramJEPA easily produces a three-modality model, and although adding the image modality does not improve over the graph–text pair, the three-modality model still outperforms both GraphCLIP variants. The lower block of the table includes supervised methods, so it is expected to perform better. It is therefore encouraging that our pre-training, evaluated with a linear probe, with no non-linearity, comes close: within two points of the strongest supervised model on Ele-Fashion, where it also beats GCN, and within four on Goodreads-NC, where it beats the feature-only MLP. Explanations of the baselines are provided in Appendix B.

5.5 Efficiency Study

In Figure 4 we investigate the time and space complexity of alignment objectives by comparing how the cost of volumetric and InfoNCE grow with the batch size B. We time each alignment term on synthetic embeddings (d=768, k=2) with no encoder, so the measurement is only of the loss. Fitting $y \propto B^p$ over $B \geq 2048$ gives $p = 1.03$ for the volumetric objective and $p = 1.95$ for InfoNCE (Figure 4a). This matches our derivation of linear and quadratic time complexity in Propositions 2

Table 3: Linear-probe node classification accuracy (%) on target graphs after pretraining on Amazon-Sports, Amazon-Cloth and Goodreads-LP. G, T and I denote the graph, text (T5) and image (DINOv2) modalities. Per column, bold marks the best self-supervised result and underlining the best supervised one. In supervised methods, node features are initialized as the concatenation [T ‖ I] of the text and image embeddings. We quote these rows from Zhu et al. (2025a).

	Method	Modality	Ele-Fashion	Goodreads-NC
Self-supervised	GraphCLIP	G-T	81.6 ±0.06	75.6 ±0.03
	GraphCLIP	G-I	82.1 ±0.07	72.3 ±0.05
	LeGramJEPA (ours)	G-T	**83.7** ±0.02	**79.1** ±0.04
	LeGramJEPA (ours)	G-I	83.2 ±0.04	76.9 ±0.03
	LeGramJEPA (ours)	G-T-I	83.6 ±0.03	77.6 ±0.04
Supervised	GCN	G + [T ‖ I]	79.4 ±0.04	81.7 ±0.03
	SAGE	G + [T ‖ I]	<u>85.3</u> ±0.09	<u>83.0</u> ±0.08
	MLP	[T ‖ I]	84.9 ±0.01	68.8 ±0.03



(a) Fitted $B^{1.03}$ vs $B^{1.95}$ over $B \geq 2048$.

(b) InfoNCE retains a $B \times B$ logit matrix for the backward pass.

Figure 4: Cost of the alignment and InfoNCE term. Synthetic embeddings (d=768, k=2) are used; encoders are excluded. Intersection is at $B^\star \approx 2554$.

and 1, respectively. The two methods intersect at $B^\star \approx 2554$, and below it the volumetric term is the more expensive of the two. Under memory, InfoNCE retains the $B \times B$ similarity matrix until the backward pass (Lemma 1), so its memory grows with B^2. In some cases, InfoNCE runs out of device memory where the volumetric alignment still runs. Appendix G.1 gives the formal arguments, and Appendix H the benchmark setup.

6 Conclusion

In this work, we proposed LeGramJEPA, a negative-free approach to multimodal graph self-supervised learning in latent space. A single volumetric term per anchor aligns all modalities of a sample, so the number of alignment terms grows with k rather than $k(k-1)$, and SIGReg keeps each modality's embeddings from collapsing without the need for negatives. LeGramJEPA is trained only on positive pairs and is competitive with GraphCLIP on zero-shot node classification and improves over it on link prediction. Under three modalities it outperforms its contrastive counterpart on both modality pairs. Additionally, the proposed volumetric alignment method costs linear time and memory on the batch size, compared to quadratic for InfoNCE. This study demonstrates a negative-free latent-space approach for GFM which can be used on real-world graphs with several modalities.

AI USE STATEMENT

We used generative AI tools to improve the clarity and presentation of the writing.

ETHICS STATEMENT

We are not aware of any ethical concerns raised by this work.

REPRODUCIBILITY STATEMENT

We released an anonymized GitHub repository with the code and instructions needed to reproduce all experiments here. Proofs of our claims are provided in Appendix G.1, and the full evaluation protocol is described in Appendix D.1.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael G. Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 15619–15629. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01499. URL https://doi.org/10.1109/CVPR52729.2023.01499.

Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics. *CoRR*, abs/2511.08544, 2025. doi: 10.48550/ARXIV.2511.08544. URL https://doi.org/10.48550/arXiv.2511.08544.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=xm6YD62D1Ub.

Lorenzo Bini and Stéphane Marchand-Maillet. Self-supervised graph learning via spectral bootstrapping and laplacian-based augmentations. *CoRR*, abs/2506.20362, 2025. doi: 10.48550/ARXIV.2506.20362. URL https://doi.org/10.48550/arXiv.2506.20362.

Delong Chen, Mustafa Shukor, Théo Moutakanni, Willy Chung, Jade Yu, Tejaswi Kasarla, Allen Bolourchi, Yann LeCun, and Pascale Fung. VL-JEPA: joint embedding predictive architecture for vision-language. *CoRR*, abs/2512.10942, 2025. doi: 10.48550/ARXIV.2512.10942. URL https://doi.org/10.48550/arXiv.2512.10942.

Runjin Chen, Tong Zhao, Ajay Kumar Jaiswal, Neil Shah, and Zhangyang Wang. Llaga: Large language and graph assistant. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7809–7823. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/chen24bh.html.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 15750–15758. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.01549. URL https://openaccess.thecvf.com/content/CVPR2021/html/Chen_Exploring_Simple_Siamese_Representation_Learning_CVPR_2021_paper.html.

Giordano Cicchetti, Eleonora Grassucci, Luigi Sigillo, and Danilo Comminiello. Gramian multimodal representation learning and alignment. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=ftGnpZrW7P.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. SPECTER: document-level representation learning using citation-informed transformers. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 2270–2282. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.ACL-MAIN.207. URL https://doi.org/10.18653/v1/2020.acl-main.207.

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 (NAACL)*, 2019.

Yi Fang, Dongzhe Fan, Sirui Ding, Ninghao Liu, and Qiaoyu Tan. Uniglm: Training one unified language model for text-attributed graphs embedding. In Wolfgang Nejdl, Sören Auer, Meeyoung Cha, Marie-Francine Moens, and Marc Najork (eds.), *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining, WSDM 2025, Hannover, Germany, March 10-14, 2025*, pp. 973–981. ACM, 2025. doi: 10.1145/3701551.3703586. URL https://doi.org/10.1145/3701551.3703586.

C. Lee Giles, Kurt D. Bollacker, and Steve Lawrence. CiteSeer: An automatic citation indexing system. In *Proceedings of the Third ACM Conference on Digital Libraries, DL '98, Pittsburgh, PA, USA, June 23-26, 1998*, pp. 89–98. ACM, 1998. doi: 10.1145/276675.276685.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent - A new approach to self-supervised learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations (ICLR)*, 2021.

Xiaoxin He, Xavier Bresson, Thomas Laurent, and Bryan Hooi. Explanations as features: Llm-based features for text-attributed graphs. *CoRR*, abs/2305.19523, 2023. doi: 10.48550/ARXIV.2305.19523. URL https://doi.org/10.48550/arXiv.2305.19523.

Xiaoxin He, Xavier Bresson, Thomas Laurent, Adam Perold, Yann LeCun, and Bryan Hooi. Harnessing explanations: Llm-to-lm interpreter for enhanced text-attributed graph representation learning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=RXFVcynVe1.

Yufei He, Yuan Sui, Xiaoxin He, and Bryan Hooi. Unigraph: Learning a unified cross-domain foundation model for text-attributed graphs. In Yizhou Sun, Flavio Chierichetti, Hady W. Lauw, Claudia Perlich, Wee Hyong Tok, and Andrew Tomkins (eds.), *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, V.1, KDD 2025, Toronto, ON, Canada, August 3-7, 2025*, pp. 448–459. ACM, 2025. doi: 10.1145/3690624.3709277. URL https://doi.org/10.1145/3690624.3709277.

Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. GraphMAE: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2022.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan,

and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020a. URL https://proceedings.neurips.cc/paper/2020/hash/fb60d411a5c5b72b2e7d3527cfc84fd0-Abstract.html.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020b. URL https://proceedings.neurips.cc/paper/2020/hash/fb60d411a5c5b72b2e7d3527cfc84fd0-Abstract.html.

Xuanwen Huang, Kaiqiao Han, Yang Yang, Dezheng Bao, Quanjin Tao, Ziwei Chai, and Qi Zhu. Can GNN be good adapter for LLMs? In *Proceedings of the ACM Web Conference 2024, WWW 2024*, pp. 893–904. ACM, 2024a. doi: 10.1145/3589334.3645627.

Xuanwen Huang, Kaiqiao Han, Yang Yang, Dezheng Bao, Quanjin Tao, Ziwei Chai, and Qi Zhu. Can GNN be good adapter for llms? In Tat-Seng Chua, Chong-Wah Ngo, Ravi Kumar, Hady W. Lauw, and Roy Ka-Wei Lee (eds.), *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pp. 893–904. ACM, 2024b. doi: 10.1145/3589334.3645627. URL https://doi.org/10.1145/3589334.3645627.

Bowen Jin, Wentao Zhang, Yu Zhang, Yu Meng, Xinyang Zhang, Qi Zhu, and Jiawei Han. Patton: Language model pretraining on text-rich networks. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 7005–7020. Association for Computational Linguistics, 2023a. doi: 10.18653/V1/2023.ACL-LONG.387. URL https://doi.org/10.18653/v1/2023.acl-long.387.

Bowen Jin, Yu Zhang, Yu Meng, and Jiawei Han. Edgeformers: Graph-empowered transformers for representation learning on textual-edge networks. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023b. URL https://openreview.net/forum?id=2YQrqe4RNv.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Lukas Kuhn, Giuseppe Serra, Randall Balestriero, and Florian Buettner. Levljepa: End-to-end vision-language pretraining without negatives. *CoRR*, abs/2607.00784, 2026. doi: 10.48550/ARXIV.2607.00784. URL https://doi.org/10.48550/arXiv.2607.00784.

Hongyang Lei, Xiaolong Cheng, Qi Qin, Dan Wang, Huazhen Huang, Qingqing Gu, Yetao Wu, and Luo Ji. M3-JEPA: multimodal alignment via multi-gate moe based on the joint-embedding predictive architecture. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/lei25b.html.

Yuhan Li, Peisong Wang, Zhixun Li, Jeffrey Xu Yu, and Jia Li. Zerog: Investigating cross-dataset zero-shot transferability in graphs. In Ricardo Baeza-Yates and Francesco Bonchi (eds.), *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pp. 1725–1735. ACM, 2024. doi: 10.1145/3637528.3671982. URL https://doi.org/10.1145/3637528.3671982.

Hao Liu, Jiarui Feng, Lecheng Kong, Ningyue Liang, Dacheng Tao, Yixin Chen, and Muhan Zhang. One for all: Towards training one graph model for all classification tasks. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=4IT2pgc9v6.

Yuhang Liu, Minglai Shao, Zengyi Wo, Yunlong Chu, Bing Hao, Shengzhong Liu, Ruijie Wang, and Jianxin Li. Learning noise-resilient and transferable graph-text alignment via dynamic quality assessment. In Alistair Moffat, Falk Scholer, Hannah Bast, Marc Najork, and Min Zhang (eds.), *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2026, MelbourneVICAustralia, July 20-24, 2026*, pp. 1209–1220. ACM, 2026. doi: 10.1145/3805712.3809579. URL https://doi.org/10.1145/3805712.3809579.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. Leworldmodel: Stable end-to-end joint-embedding predictive architecture from pixels, 2026. URL https://arxiv.org/abs/2603.19312.

Péter Mernyei and Cătălina Cangea. Wiki-CS: A Wikipedia-based benchmark for graph neural networks. *CoRR*, abs/2007.02901, 2020. URL https://arxiv.org/abs/2007.02901.

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. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Ladislav Rampásek, Michael Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/5d4834a159f1547b267a05a4e2b7cf5e-Abstract-Conference.html.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, 2019a.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019b. URL https://arxiv.org/abs/1908.10084.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=CR1XOQ0UTh-.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29(3):93–106, 2008. doi: 10.1609/aimag.v29i3.2157.

Geri Skenderi, Hang Li, Jiliang Tang, and Marco Cristani. Graph-level representation learning with joint-embedding predictive architectures. *Trans. Mach. Learn. Res.*, 2025, 2025. URL https://openreview.net/forum?id=v47f4DwYZb.

Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin Su, Suqi Cheng, Dawei Yin, and Chao Huang. Graphgpt: Graph instruction tuning for large language models. In Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang (eds.), *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pp. 491–500. ACM, 2024. doi: 10.1145/3626772.3657775. URL https://doi.org/10.1145/3626772.3657775.

Llama Team. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. doi: 10.48550/ARXIV.2407.21783. URL https://doi.org/10.48550/arXiv.2407.21783.

Shantanu Thakoor, Corentin Tallec, Mohammad Gheshlaghi Azar, Mehdi Azabou, Eva L. Dyer, Rémi Munos, Petar Velickovic, and Michal Valko. Large-scale representation learning on graphs via bootstrapping. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=0UXT6PpRpW.

Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10268–10278. PMLR, 2021. URL http://proceedings.mlr.press/v139/tian21a.html.

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

Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. Microsoft academic graph: When experts are not enough. *Quant. Sci. Stud.*, 1(1):396–413, 2020a. doi: 10.1162/QSS_A_00021. URL https://doi.org/10.1162/qss_a_00021.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9929–9939. PMLR, 2020. URL http://proceedings.mlr.press/v119/wang20k.html.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020b. URL https://proceedings.neurips.cc/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Zehong Wang, Zheyuan Zhang, Nitesh V. Chawla, Chuxu Zhang, and Yanfang Ye. GFT: graph foundation model with transferable tree vocabulary. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/c23ccf9eedf87e4380e92b75b24955bb-Abstract-Conference.html.

Zhihao Wen and Yuan Fang. Augmenting low-resource text classification with graph-grounded pre-training and prompting. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2023.

Hao Yan, Chaozhuo Li, Ruosong Long, Chao Yan, Jianan Zhao, Wenwen Zhuang, Jun Yin, Peiyan Zhang, Weihao Han, Hao Sun, Weiwei Deng, Qi Zhang, Lichao Sun, Xing Xie, and Senzhang Wang. A comprehensive study on text-attributed graphs: Benchmarking and rethinking. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/37d00f567a18b478065f1a91b95622a0-Abstract-Datasets_and_Benchmarks.html.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang,

Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report. *CoRR*, abs/2407.10671, 2024. doi: 10.48550/ARXIV.2407.10671. URL https://doi.org/10.48550/arXiv.2407.10671.

Junhan Yang, Zheng Liu, Shitao Xiao, Chaozhuo Li, Defu Lian, Sanjay Agrawal, Amit Singh, Guangzhong Sun, and Xing Xie. Graphformers: Gnn-nested transformers for representation learning on textual graph. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 28798–28810, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f18a6d1cde4b205199de8729a6637b42-Abstract.html.

Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. Revisiting semi-supervised learning with graph embeddings. In Maria-Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 40–48. JMLR.org, 2016. URL http://proceedings.mlr.press/v48/yanga16.html.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320. PMLR, 2021. URL http://proceedings.mlr.press/v139/zbontar21a.html.

Heng Zhang, Tianyi Zhang, Yuling Shi, Xiaodong Gu, Yaomin Shen, Zijian Zhang, Yilei Yuan, Hao Zhang, and Jin Huang. Can representation gaps be the key to enhancing robustness in graph-text alignment? *CoRR*, abs/2510.12087, 2025a. doi: 10.48550/ARXIV.2510.12087. URL https://doi.org/10.48550/arXiv.2510.12087.

Zheng Zhang, Yuntong Hu, Bo Pan, Chen Ling, and Liang Zhao. TAGA: text-attributed graph self-supervised learning by synergizing graph and text mutual transformations. In Meeyoung Cha, Chanyoung Park, Noseong Park, Carl Yang, Senjuti Basu Roy, Jessie Li, Jaap Kamps, Kijung Shin, Bryan Hooi, and Lifang He (eds.), *Proceedings of the 34th ACM International Conference on Information and Knowledge Management, CIKM 2025, Seoul, Republic of Korea, November 10-14, 2025*, pp. 4263–4272. ACM, 2025b. doi: 10.1145/3746252.3761357. URL https://doi.org/10.1145/3746252.3761357.

Huanjing Zhao, Beining Yang, Yukuo Cen, Junyu Ren, Chenhui Zhang, Yuxiao Dong, Evgeny Kharlamov, Shu Zhao, and Jie Tang. Pre-training and prompting for few-shot node classification on text-attributed graphs. In Ricardo Baeza-Yates and Francesco Bonchi (eds.), *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pp. 4467–4478. ACM, 2024. doi: 10.1145/3637528.3671952. URL https://doi.org/10.1145/3637528.3671952.

Jianan Zhao, Meng Qu, Chaozhuo Li, Hao Yan, Qian Liu, Rui Li, Xing Xie, and Jian Tang. Learning on large-scale text-attributed graphs via variational inference. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=q0nmYciuuZN.

Jing Zhu, Yuhang Zhou, Shengyi Qian, Zhongmou He, Tong Zhao, Neil Shah, and Danai Koutra. Mosaic of modalities: A comprehensive benchmark for multimodal graph learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 14215–14224. Computer Vision Foundation / IEEE, 2025a. doi: 10.1109/CVPR52734.2025.01326. URL https://openaccess.thecvf.com/

content/CVPR2025/html/Zhu_Mosaic_of_Modalities_A_Comprehensive_Benchmark_for_Multimodal_Graph_Learning_CVPR_2025_paper.html.

Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Deep graph contrastive representation learning. *arXiv preprint arXiv:2006.04131*, 2020.

Yun Zhu, Haizhou Shi, Xiaotang Wang, Yongchao Liu, Yaoke Wang, Boci Peng, Chuntao Hong, and Siliang Tang. Graphclip: Enhancing transferability in graph foundation models for text-attributed graphs. In *Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025- 2 May 2025*, pp. 2183–2197. ACM, 2025b. doi: 10.1145/3696410.3714801. URL https://doi.org/10.1145/3696410.3714801.

Eric Zimmermann, Harley Wiltzer, Justin Szeto, David Alvarez-Melis, and Lester Mackey. Kerjepa: Kernel discrepancies for euclidean self-supervised learning. *CoRR*, abs/2512.19605, 2025. doi: 10.48550/ARXIV.2512.19605. URL https://doi.org/10.48550/arXiv.2512.19605.

A Further Preliminaries

Definition A.1 (Zero-Shot Node Classification)**.** Given a target TAG $\mathcal{G}^t$ with label sentences $\{S_c\}_{c\in\mathcal{Y}^t}$, the predicted class of node v_n is the label whose sentence is most similar to its representation,

$$\hat{y}_n \;=\; \arg\max_{c\,\in\,\mathcal{Y}^t} \; \mathrm{sim}(\boldsymbol{z}_n,\; \boldsymbol{z}_{S_c})\,. \tag{7}$$

We write $\mathrm{sim}(\cdot,\cdot)$ for cosine similarity.

Definition A.2 (Zero-Shot Link Prediction)**.** Given a target TAG $\mathcal{G}^t$, a candidate pair $(u, v) \in \mathcal{V}^t \times \mathcal{V}^t$ is scored by the similarity of the two node representations,

$$s(u, v) \;=\; \mathrm{sim}(\boldsymbol{z}_u,\; \boldsymbol{z}_v)\,, \tag{8}$$

Performance is measured by the ranking that s induces over held-out edges against sampled non-edges.

B Baselines

Text-only encoders (Table 1). These are applied directly, without training on the source data. Each node's text and each label's text are embedded, and the prediction is the label with the highest cosine similarity. The graph is not used.

- **BERT** (Devlin et al., 2019): Transformer encoder pretrained with masked language modelling.
- **SBERT** (Reimers & Gurevych, 2019a): BERT fine-tuned with a siamese objective to produce sentence embeddings.
- **DeBERTa** (He et al., 2021): BERT variant with attention that keeps content and position separate.
- **E5** (Wang et al., 2022): Text embedding model trained contrastively on large sets of weakly paired text.
- **Qwen2** (Yang et al., 2024): Decoder-only LLM, used here as a text encoder.
- **LLaMA3.1** (Team, 2024): Decoder-only LLM, used here as a text encoder.

Label-supervised graph-LLM methods (Table 1). These are trained on the source datasets with their labels, then applied to the targets.

- **GraphGPT** (Tang et al., 2024): Aligns graph tokens with an LLM through graph–text matching, then instruction-tunes on labelled tasks.
- **LLaGA** (Chen et al., 2024): Turns each node's neighbourhood into a sequence of embeddings projected into the LLM's token space, then instruction-tunes on labelled tasks.

- **OFA** (Liu et al., 2024): Describes nodes and classes in text and adds class nodes to the graph, so all classification tasks share one format. Trained with labels.
- **ZeroG** (Li et al., 2024): Fine-tunes a language model (LoRA) on labelled source data so that node text and class-description text align, with prompt nodes added to the graph.

Label-free self-supervised methods (Table 1). These are pretrained on the source datasets without labels. Zero-shot predictions use cosine similarity between node embeddings and label embeddings.

- **DGI** (Veličković et al., 2019): Maximises agreement between node embeddings and a summary of the whole graph, using a corrupted graph as negatives.
- **GRACE** (Zhu et al., 2020): Contrasts each node's embeddings across two augmented views, with all other nodes as negatives.
- **BGRL** (Thakoor et al., 2022): Negative-free. An online encoder predicts a target encoder's embedding of another view, and the target is an exponential moving average of the online encoder.
- **GraphMAE** (Hou et al., 2022): Masks node features and reconstructs them with a GNN encoder–decoder, using a scaled cosine error.
- **G2P2** (Wen & Fang, 2023): Contrastive graph–text pretraining that pairs a node with its own text, its neighbours' text and a summary text. Classifies with label-text prompts.

Graph–summary contrastive.

- **GraphCLIP** (Zhu et al., 2025b): Two-tower contrastive alignment between subgraph embeddings and LLM-generated subgraph summaries, with the text encoder frozen. Table 1 uses its reported numbers. In Table 4 we train it from scratch on our multimodal pretraining pool (G-T and G-I) and evaluate it with the same frozen-encoder linear probe as ours.

Supervised multimodal baselines (Table 4). These are trained with labels on each target, with a 3-layer MLP classification head; numbers are as reported by Zhu et al. (2025a).

- **GCN** (Kipf & Welling, 2017): Each node is initialised with the concatenation of its T5 text embedding and its DINOv2 image embedding.
- **SAGE** (Hamilton et al., 2017): Each node is initialised with the concatenation of its T5 text embedding and its DINOv2 image embedding.
- **MLP**: Uses the concatenated node features alone, ignoring the graph structure.

C Info-Max Background

C.1 LeJEPA Loss Function

Formally, given B samples with V views each, $\boldsymbol{z}_{n,v} = f_\theta(\boldsymbol{x}_{n,v}) \in \mathbb{R}^d$, it pairs a predictive term $\mathcal{L}_{\text{pred}}$, which drives the V views of a sample together, with SIGReg (Appendix C.3), which constrains each view's batch of embeddings toward an isotropic Gaussian and so prevents collapse:

$$\mathcal{L}_{\text{LeJEPA}}\big(\{\boldsymbol{x}_{n,v}\}_{n,v=1}^{B,V}\big) = \frac{\lambda}{V}\sum_{v=1}^{V}\text{SIGReg}\big(\{\boldsymbol{z}_{n,v}\}_{n=1}^{B}\big) + \frac{1-\lambda}{B}\sum_{n=1}^{B}\mathcal{L}_{\text{pred}}\big(\{\boldsymbol{z}_{n,v}\}_{v=1}^{V}\big),\ \lambda \in [0,1] \tag{9}$$

C.2 LeVLJEPA Loss Function

$$\mathcal{L}_{\text{cross}} = \frac{1}{B}\sum_{i=1}^{B}\Big[\big\|h_g(\boldsymbol{z}_i^g) - \text{sg}(\boldsymbol{z}_i^t)\big\|_2^2 + \big\|h_t(\boldsymbol{z}_i^t) - \text{sg}(\boldsymbol{z}_i^g)\big\|_2^2\Big]. \tag{10}$$

The full two-modality objective adds a per-modality distributional constraint:

$$(1-\lambda_g-\lambda_t)\,\mathcal{L}_{\text{cross}} + \lambda_g\,\text{SIGReg}(\boldsymbol{Z}^g) + \lambda_t\,\text{SIGReg}(\boldsymbol{Z}^t), \qquad \lambda_g, \lambda_t \geq 0,\ \lambda_g + \lambda_t \leq 1. \tag{11}$$

C.3 SIGREG

SIGReg constrains a set of embeddings to follow an isotropic Gaussian in $\mathbb{R}^d$.

Reduction to one dimension. The Cramér–Wold theorem states that a distribution on $\mathbb{R}^d$ is determined by its one-dimensional projections: if $\langle \boldsymbol{\omega}, \boldsymbol{z} \rangle$ and $\langle \boldsymbol{\omega}, \boldsymbol{z}' \rangle$ agree in distribution for every $\boldsymbol{\omega} \in \mathbb{S}^{d-1}$, then $\boldsymbol{z}$ and $\boldsymbol{z}'$ agree in distribution. Every one-dimensional projection of an isotropic Gaussian is standard Gaussian, so the d-dimensional constraint reduces to a family of univariate tests.

SIGReg draws M unit-norm directions $\boldsymbol{\omega}_m$ and projects the embeddings $\boldsymbol{Z}$ onto each,

$$\boldsymbol{h}^{(m)} = \boldsymbol{Z}\boldsymbol{\omega}_m \in \mathbb{R}^n. \tag{12}$$

Univariate test. For the univariate test, SIGReg Balestriero & LeCun (2025) uses the Epps–Pulley statistic. Given projected, standardized samples $h_1, \ldots, h_n$ with empirical characteristic function $\varphi_n(t; \boldsymbol{h}) = \frac{1}{n}\sum_{i=1}^{n} e^{\imath t h_i}$, the statistic integrates the squared discrepancy against the standard Gaussian characteristic function under a Gaussian weight $w(t)$,

$$T_n \;=\; \int_{\mathbb{R}} \left| \varphi_n(t; \boldsymbol{h}^{(m)}) - e^{-t^2/2} \right|^2 w(t)\, \mathrm{d}t, \tag{13}$$

which is evaluated by fixed-grid quadrature.

Regularizer. With M directions $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_M$ drawn uniformly on $\mathbb{S}^{d-1}$ and resampled at every step, the regularizer for a view matrix $\boldsymbol{Z}^v \in \mathbb{R}^{n \times d}$ is the mean statistic over directions,

$$\mathrm{SIGReg}(\boldsymbol{Z}^v) \;=\; \frac{1}{M}\sum_{m=1}^{M} T_n(\boldsymbol{Z}^v \boldsymbol{\omega}_m)\,. \tag{14}$$

It is zero when the projections are standard Gaussian along every sampled direction and positive otherwise.

D EVALUATION

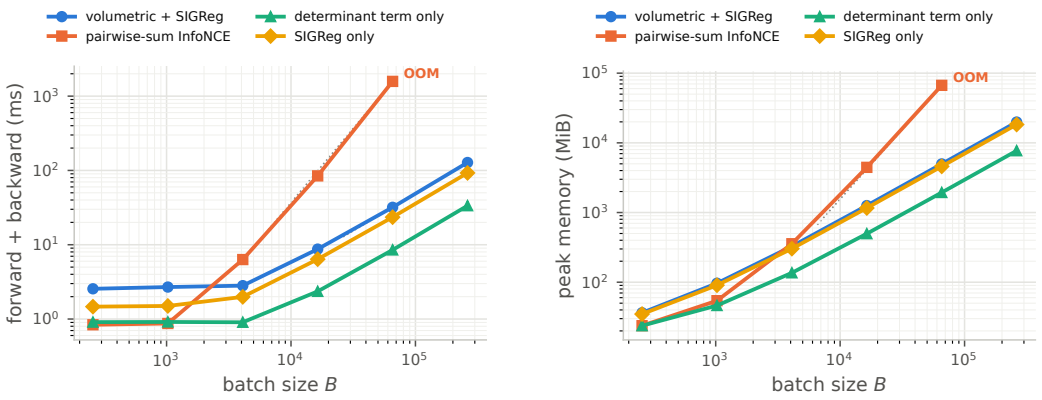
D.1 CLASSIFICATION

All evaluation uses the pretrained encoder under the training schema defined in (Definition 3.2). Reported figures are the mean and standard deviation over five seeds.

For node classification, we hold out a random 20% of nodes for testing and report accuracy over five seeds. We expose three scoring directions: $\cos(z_g, h_t(z_t))$, $\cos(h_g(z_g), z_t)$, $\cos(z_g, z_t)$, where z_g and z_t represent the embedding outputs of the graph and text encoder, respectively, and $h_g(\cdot)$, $h_t(\cdot)$ represent the predictors on top of the encoders. These predictors, being a part of our objective, give us some flexibility in which similarity measure works best per dataset, as per Definition A.1. As such, we report the best-performing similarity measure for our model.

For link prediction (Definition A.2) we report two metrics, both scoring a candidate edge (u, v).

- *AUC* follows the protocol of Zhu et al. (2025b). Each held-out positive is matched against one uniformly drawn non-edge, and we report the probability that a true edge scores above a random one, i.e the area under the ROC curve over the matched positive and negative sets.
- *MRR* is a stricter, and more modern retrieval-style measure for link prediction. For each positive (u, v) we score the true target v against R negative targets (u, v'), where each v' is sampled uniformly at random and rejected if (u, v') is a real edge. The true target v is ranked among the $1 + R$ candidates and the reciprocal of its rank is averaged over all positives. We report with 100 negatives.



(a) Fitted $B^{1.03}$ vs $B^{1.95}$ over $B \geq 2048$.

(b) InfoNCE retains a $B \times B$ logit matrix for the backward pass.

Figure 5: Cost of the alignment term on synthetic embeddings (d=768, k=2), encoders are excluded. Intersection is at $B^\star \approx 2554$. For batch sizes before this point the volumetric term is the more expensive of the two.

D.2 Linear Probing

For linear probing on the multimodal targets (Section 5.4), the pretrained encoder is frozen. A single linear layer is then trained on these embeddings with cross-entropy, using the train, validation and test splits provided with each dataset. We report test accuracy at the probe epoch with the highest validation accuracy.

E Pairwise Cross-Modal Alignment Scales Quadratically

Extending the cross-modal term equation 10 to k modalities aligns every anchor to every other modality. With a predictor h_a per modality,

$$\mathcal{L}^{(k)}_{\text{cross}} = \frac{1}{B} \sum_{i=1}^{B} \sum_{a=1}^{k} \sum_{\substack{b=1 \\ b \neq a}}^{k} \left\| h_a(\boldsymbol{z}_i^a) - \text{sg}(\boldsymbol{z}_i^b) \right\|_2^2, \tag{15}$$

a sum over the $k(k-1)$ ordered modality pairs. For $k = 3$ modalities, graph g, text t, image m, this expands to six terms, two per anchor:

$$\begin{aligned}
\text{anchor } g: &\quad \left\| h_g(\boldsymbol{z}^g) - \text{sg}(\boldsymbol{z}^t) \right\|_2^2 + \left\| h_g(\boldsymbol{z}^g) - \text{sg}(\boldsymbol{z}^m) \right\|_2^2, \\
\text{anchor } t: &\quad \left\| h_t(\boldsymbol{z}^t) - \text{sg}(\boldsymbol{z}^g) \right\|_2^2 + \left\| h_t(\boldsymbol{z}^t) - \text{sg}(\boldsymbol{z}^m) \right\|_2^2, \\
\text{anchor } m: &\quad \left\| h_m(\boldsymbol{z}^m) - \text{sg}(\boldsymbol{z}^g) \right\|_2^2 + \left\| h_m(\boldsymbol{z}^m) - \text{sg}(\boldsymbol{z}^t) \right\|_2^2,
\end{aligned}$$

The count grows as $k(k-1) = \Theta(k^2)$. LeGramJEPA replaces the two terms under each anchor with the single volume $\det \boldsymbol{G}_i^a$ of equation 4, giving k terms in place of $k(k-1)$: three volume terms rather than six for $k = 3$.

F Extended Efficiency Study

Figure 4 shows that the volumetric alignment's cost, empirically, SIGReg contributes a larger part and the determinant the smaller; both are linear in B which matches our discussions in Appendix G.1.

G Space and Time Complexity

G.1 Cost of the Alignment Term

This appendix makes precise the claim that the contrastive alignment term is quadratic in the batch size while the volumetric term is linear. The statements concern the alignment term alone: the encoders are identical across objectives by the architecture control of Section 4.3.

We show this for $k = 2$, matching the graph–text setting of our experiments, and drop k from the notation. B is the batch size, d the shared embedding width, and $\boldsymbol{Z}^a \in \mathbb{R}^{B\times d}$ the batch of modality-a embeddings with rows $\boldsymbol{z}_i^a$.

G.1.1 Contrastive alignment

Lemma 1 (Score matrix)**.** For row normalized $\hat{\boldsymbol{Z}}^1, \hat{\boldsymbol{Z}}^2 \in \mathbb{R}^{B\times d}$, constructing matrix $\boldsymbol{S} = \hat{\boldsymbol{Z}}^1(\hat{\boldsymbol{Z}}^2)^\top$ costs $\Theta(B^2d)$ time and retains $\Theta(B^2)$ memory.

Proposition 1 (Contrastive alignment is quadratic in B)**.** Let us write the score matrix $\boldsymbol{S} = \hat{\boldsymbol{Z}}^1(\hat{\boldsymbol{Z}}^2)^\top$, so that $\boldsymbol{S}_{ij}$ is the similarity between the graph embedding of sample i and the text embedding of sample j. The entries on the diagonal represent positive similarities $\boldsymbol{S}_{ii}$, and the $B-1$ entries $\boldsymbol{S}_{ij}$ with $j \neq i$ are the in-batch negatives.

$$\mathcal{L}_{\mathrm{NCE}} = \frac{1}{2B}\sum_{i=1}^{B}\left[-\log\frac{\exp(\boldsymbol{S}_{ii}/\tau)}{\sum_{j=1}^{B}\exp(\boldsymbol{S}_{ij}/\tau)} - \log\frac{\exp(\boldsymbol{S}_{ii}/\tau)}{\sum_{j=1}^{B}\exp(\boldsymbol{S}_{ji}/\tau)}\right] \tag{16}$$

The symmetric InfoNCE objective costs $\Theta(B^2d)$ operations and retains $\Theta(B^2)$ memory.

Proof. Row-normalizing the two embedding matrices costs $\Theta(Bd)$ operations and retains $\Theta(Bd)$ memory. By Lemma 1 the score matrix contributes $\Theta(B^2d)$ operations and $\Theta(B^2)$ memory. The softmax and cross-entropy add a further $\Theta(B^2)$, $\Theta(B)$ operations respectively.

$$\text{operations:}\quad \underbrace{\Theta(B^2d)}_{\text{scores}} + \underbrace{\Theta(B^2)}_{\text{softmax}} + \underbrace{\Theta(B)}_{\text{CE}} + \underbrace{\Theta(Bd)}_{\text{normalization}} = \Theta(B^2d), \tag{17}$$

$$\text{memory:}\quad \underbrace{\Theta(B^2)}_{\boldsymbol{S},\,\boldsymbol{P},\,\text{grad buffer}} + \underbrace{\Theta(Bd)}_{\hat{\boldsymbol{Z}}^1,\,\hat{\boldsymbol{Z}}^2} = \Theta(B^2), \tag{18}$$

Equation 18 holds whenever $B \gtrsim d$ as $O(Bd) = O(B^2)$.

□

G.1.2 Volumetric alignment

Proposition 2 (Volumetric alignment is linear in B)**.** With $\boldsymbol{A}_i^a = [\, h_a(\boldsymbol{z}_i^a) \mid \mathrm{sg}(\boldsymbol{z}_i^b)_{b\neq a} \,] \in \mathbb{R}^{d\times 2}$ and $\boldsymbol{G}_i^a = (\boldsymbol{A}_i^a)^\top \boldsymbol{A}_i^a$ the Gram matrix.

$$\mathcal{L}_{\mathrm{vol}} = \frac{1}{2B}\sum_{a=1}^{2}\sum_{i=1}^{B}\det \boldsymbol{G}_i^a, \tag{19}$$

The volumetric alignment term costs $\Theta(Bd)$ operations and retains $\Theta(Bd)$ memory.

Proof. Fix an anchor a and a sample i. Normalising the two columns of $\boldsymbol{A}_i^a$, each of length d, costs $\Theta(d)$ operations. The Gram matrix $\boldsymbol{G}_i^a$ has four entries, each an inner product of length d, costing $\Theta(d)$, in general costing $\Theta(dk^2)$. Its determinant costs $\Theta(k^3)$ by LU factorisation, which is $\Theta(1)$ at $k = 2$ and is dominated by the $\Theta(dk^2)$ whenever $d \geq k$. This is what we are concerned with since $d >= 768$ in our experiments. The per-anchor per-sample cost is therefore $\Theta(d)$ operations, and each anchor retains its stacked $\boldsymbol{A}^a \in \mathbb{R}^{B\times d\times 2}$.

For the backward pass, $\det \boldsymbol{G}_i^a$ is a scalar function of sample i alone and only the anchor column carries gradient, so $\partial\mathcal{L}_{\mathrm{vol}}/\partial h_a(\boldsymbol{z}_i^a)$ is a single d-vector per sample: the gradient buffer is $\Theta(Bd)$

per anchor. Summing over B samples and 2 anchors,

$$\text{operations:} \quad \underbrace{\Theta(Bd)}_{\text{normalization}} + \underbrace{\Theta(Bd)}_{\text{Gram entries}} + \underbrace{\Theta(B)}_{\text{determinants}} = \Theta(Bd), \tag{20}$$

$$\text{memory:} \quad \underbrace{\Theta(Bd)}_{\boldsymbol{A}^1,\,\boldsymbol{A}^2} + \underbrace{\Theta(B)}_{\boldsymbol{G}_i^1,\,\boldsymbol{G}_i^2} + \underbrace{\Theta(Bd)}_{\text{grad. buffer}} = \Theta(Bd), \tag{21}$$

□

The remaining term of the objective, the sliced Epps–Pulley statistic of SIGReg, is likewise linear in the batch size: Balestriero & LeCun (2025, §4.2.3) report $\mathcal{O}(N)$ time and memory in the minibatch size N for their implementation, the number of slices and quadrature points entering only as constants of the configuration. The volumetric objective is therefore linear in B in both time and memory, against $\Theta(B^2 d)$ and $\Theta(B^2)$ for InfoNCE.

H ALIGNMENT-TERM BENCHMARK

Metric The benchmark times the forward and backward pass of each alignment objective and records its peak memory, with no encoder in the computation. The embeddings are random tensors of shape $B \times d$, one per modality, so the backward pass computes gradients with respect to them and stops.

Objectives. We use $k = 2$ modalities and width $d = 768$. The volumetric alignment is our training implementation: the determinant term of Proposition 2, evaluated per sample, plus SIGReg on each modality with $M = 128$ slices and $T = 17$ quadrature points. InfoNCE is the symmetric objective of equation 16, with the temperature initialised at 0.07. Both objectives receive the same embeddings. We also time the determinant term and SIGReg separately.

Results. The fitted time exponents are 1.03 ± 0.02 for the volumetric objective and 1.95 ± 0.05 for InfoNCE. The two costs cross at $B^\star \approx 2554$. InfoNCE runs out of memory at $B = 262{,}144$, where the volumetric alignment still maintains memory. Separating the volumetric objective, SIGReg is the larger of its two terms and the determinant term the smaller, and each is linear in B on its own.

Scope. These numbers describe the alignment term on one device. They do not give the cost of a full training step, which includes the encoder, and the value of $B^\star$ depends on the hardware and implementation. The exponents do not: they follow from the operation counts in Appendix G.1.