A Graph Foundation Model with Cross-Modal Alignment and Modality-Aware Expert Fusion for Multi-Modal Graphs

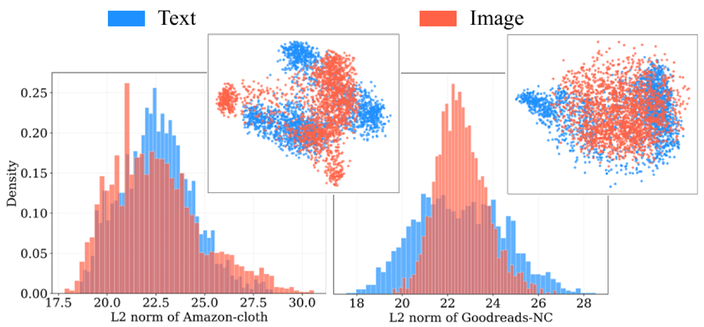
Dongxiao He Ankang Yang Jitao Zhao Di Jin

School of Computer Science and Technology, Tianjin University, Tianjin, China.

ICML International Conference On Machine Learning

Motivation

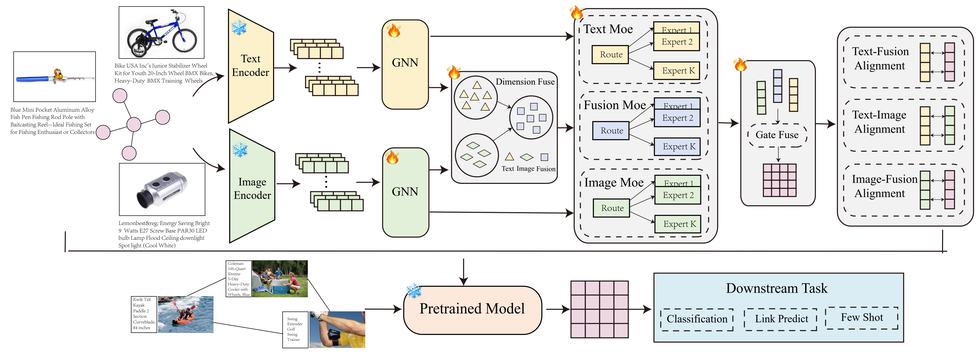
Text and image embeddings remain distributionally different across multi-modal graph datasets, even after CLIP encoding.



Naive early fusion:
Text features + Image features → Average / Concatenate → Shared GNN

Limitations:
Dilute modality-specific information
One modality may dominate graph propagation
Cross-modal alignment remains insufficient

Methodology



CLIP-based modality encoding:

$$\mathbf{x}_i^{(t)} = f_{\text{CLIP}}^{(t)}(\mathbf{s}_i^{(t)}), \qquad \mathbf{x}_i^{(v)} = f_{\text{CLIP}}^{(v)}(\mathbf{s}_i^{(v)}),$$

Modality-specific graph propagation:

$$\mathbf{z}_i^{(t)} = f_{\text{GNN}}^{(t)}\left(\mathcal{G},\ \{\mathbf{x}_j^{(t)}\}_{j\in\mathcal{V}}\right), \quad \mathbf{z}_i^{(v)} = f_{\text{GNN}}^{(v)}\left(\mathcal{G},\ \{\mathbf{x}_j^{(v)}\}_{j\in\mathcal{V}}\right),$$

Modality-Aware MoE

Intra-group routing: selects top-k experts within each modality group

$$\mathbf{e}_i^{(s)} = \sum_{j\in\mathcal{S}_i^{(s)}} \alpha_{i,j}^{(s)}\ \phi_j^{(s)}(\mathbf{z}_i^{(s)}).$$

Inter-modal routing: balances text, image, and fused expert

$$\mathbf{z}_i^{(f)} = \pi_i^{(t)}\mathbf{e}_i^{(t)} + \pi_i^{(v)}\mathbf{e}_i^{(v)} + \pi_i^{(g)}\mathbf{e}_i^{(g)}.$$

Cross-Modal Contrastive Loss

$$\mathcal{L}_{a,b} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(\mathbf{z}_i^{(a)},\mathbf{z}_i^{(b)}))}{\sum_{j=1}^{N}\exp(\text{sim}(\mathbf{z}_i^{(a)},\mathbf{z}_j^{(b)}))}, \quad \mathcal{L} = w_{tv}\mathcal{L}_{t,v} + w_{vf}\mathcal{L}_{v,f} + w_{tf}\mathcal{L}_{t,f},$$

Experiments

Self-Supervised Pre-training Evaluation

	In-distribution			In-domain Generalization			Out-of-domain Generalization			
	Products	Goodreads-LP	Amazon-cloth	Amazon-sports	Goodreads-NC	Ele-fashion	Arxiv	Wiki-CS	FB15K237	WN18RR
Use CLIP to encode raw multimodal data as input features.										
BGRL	72.57±0.05	6.49±0.02	17.35±0.03	23.66±0.05	69.28±0.10	74.92±0.03	65.07±0.04	71.35±0.02	90.12±0.15	75.02±0.14
DGI	71.91±0.06	7.23±0.07	18.39±0.04	24.07±0.11	70.39±0.08	83.72±0.02	61.07±0.02	69.79±0.16	91.43±0.17	73.26±0.11
GraphMAE2	70.46±0.15	7.77±0.12	16.25±0.11	27.44±0.14	72.49±0.05	80.23±0.04	64.15±0.18	74.82±0.19	89.96±0.14	73.97±0.14
GRACE	73.44±0.08	4.77±0.02	18.15±0.05	25.59±0.05	67.47±0.12	79.70±0.02	58.11±0.04	67.28±0.02	90.12±0.15	73.02±0.19
GCOPE	76.87±0.14	8.91±0.11	17.49±0.12	24.24±0.13	75.43±0.15	77.61±0.14	63.92±0.09	75.94±0.18	91.03±0.16	75.61±0.13
SAMGPT	78.53±0.11	9.13±0.06	19.14±0.06	28.24±0.13	79.02±0.09	83.82±0.07	69.79±0.17	74.63±0.18	91.76±0.10	74.96±0.15
FUG	74.34±0.17	7.94±0.10	20.49±0.11	25.33±0.12	74.62±0.14	81.29±0.01	68.11±0.05	73.31±0.12	91.49±0.12	74.35±0.10
Use raw multimodal data as input features.										
CLIP-text	64.83±0.09	3.71±0.09	13.77±0.09	25.13±0.06	69.45±0.07	83.53±0.02	62.75±0.03	64.32±0.19	89.72±0.21	74.86±0.13
CLIP-image	–	2.77±0.07	13.09±0.07	16.33±0.02	64.29±0.12	80.26±0.05	–	–	–	–
UniGraph2	80.82±0.08	8.76±0.06	24.25±0.12	29.52±0.14	79.78±0.07	84.83±0.11	70.54±0.04	77.07±0.06	92.30±0.03	75.20±0.09
CAME	**81.84**±0.17	**15.42**±0.08	**28.88**±0.12	**37.68**±0.19	**83.21**±0.11	**85.13**±0.03	**71.66**±0.05	**78.57**±0.13	**94.59**±0.07	**81.97**±0.11

Fewshot Evaluation

	In-distribution			Out-of-domain Generalization					
	Goodreads-NC-5-way			Arxiv-5-way			Wiki-CS-5-way		
	1-shot	3-shot	5-shot	1-shot	3-shot	5-shot	1-shot	3-shot	5-shot
Use CLIP to encode raw multimodal data as input features.									
BGRL	28.73±9.16	39.32±11.45	44.19±11.88	38.93±11.02	54.79±12.26	59.57±12.28	33.98±10.40	44.75±11.22	49.86±11.98
DGI	26.34±8.65	33.67±10.28	38.00±10.63	40.49±11.26	55.23±11.91	61.93±11.44	31.55±9.23	45.04±10.60	51.65±10.96
GraphMAE2	25.21±7.94	32.11±9.68	37.22±10.64	37.48±10.67	56.67±12.14	64.31±11.74	31.92±12.72	43.39±11.71	50.78±11.13
GRACE	38.95±11.22	48.71±11.73	52.48±11.66	41.02±11.14	54.69±12.56	61.17±12.17	36.14±12.43	48.38±11.50	53.79±10.39
GCOPE	37.60±11.35	48.88±10.17	54.34±9.91	47.50±12.89	56.10±12.35	59.91±12.46	39.04±12.96	53.26±11.04	59.53±10.77
SAMGPT	29.95±10.09	34.96±10.36	37.65±10.63	48.37±12.43	59.57±11.34	64.26±11.26	39.52±11.73	49.38±11.28	54.02±10.78
FUG	42.76±12.61	50.49±12.86	53.09±12.68	47.20±13.06	56.77±12.88	60.27±12.46	32.32±10.15	46.83±11.16	49.40±11.48
Use raw multimodal data as input features.									
CLIP-text	34.94±10.75	45.73±11.22	49.26±12.19	42.30±11.25	58.72±12.12	65.03±11.27	32.81±9.70	46.95±10.87	54.31±10.63
CLIP-image	27.33±8.65	35.55±10.29	39.95±11.32	–	–	–	–	–	–
UniGraph2	39.66±11.33	53.06±11.77	58.00±11.65	43.11±11.90	60.74±11.96	66.96±11.13	35.56±10.04	50.95±11.04	58.12±10.54
CAME	**48.03**±12.54	**60.26**±12.30	**60.47**±11.01	**49.38**±13.18	**63.78**±12.17	**67.10**±11.67	**47.65**±12.86	**55.30**±11.92	**61.89**±12.95

Ablation on key components

Setting	Ele-fashion	Goodreads-NC	Amazon-cloth	Amazon-sports
CAME	**85.13**±0.03	**83.21**±0.11	**28.88**±0.12	**37.68**±0.19
w/o $L_{t,v}$	84.67±0.04	82.46±0.09	20.64±0.15	34.30±0.12
w/o $L_{t,f}$	85.04±0.03	83.13±0.07	27.27±0.10	37.24±0.16
w/o $L_{v,f}$	82.44±0.07	82.58±0.11	24.62±0.14	36.26±0.10
only $L_{t,v}$	83.95±0.05	82.20±0.09	21.66±0.17	32.99±0.13
only $L_{t,f}$	80.74±0.11	80.70±0.12	19.19±0.11	34.82±0.18
only $L_{v,f}$	84.35±0.04	81.48±0.08	22.03±0.10	35.12±0.14
w/o Moe	84.39±0.03	82.55±0.06	26.37±0.12	36.36±0.10
w/o dim-gate	84.75±0.06	82.26±0.07	27.90±0.13	37.41±0.09
w/o m-GNN	84.52±0.04	81.92±0.06	21.19±0.13	26.83±0.11