SAMGPT: Text-free Graph Foundation Model for Multi-domain Pre-training and Cross-domain Adaptation

Xingtong Yu, Zechuan Gong, Chang Zhou, Yuan Fang, Hui Zhang

Motivation

Problem

- How to build foundation models has emerged as an important question, paving a plausible path toward artificial general intelligence.

Challenges

- Different domains exhibit various structural characteristics

C1: How do we harmonize structural variance across multiple domains during pre-training?

C2: How do we adapt multi-domain prior structural knowledge to cross-domain downstream tasks?

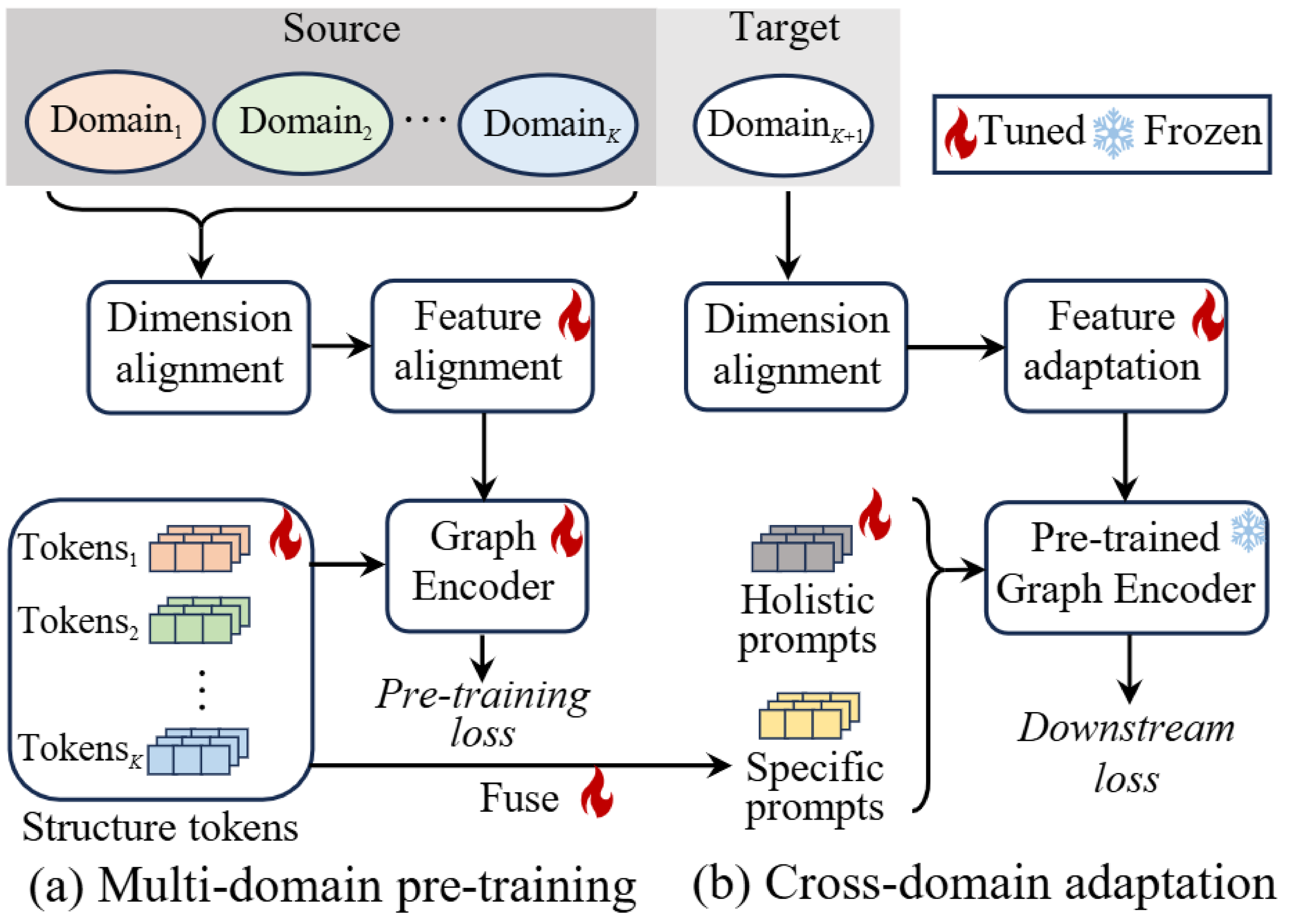

Figure 1: Motivation of SAMGPT.

Proposed Method: SAMGPT

Multi-domain pre-training

Dimension alignment

$$\tilde{\mathbf{X}}_i = \mathrm{DAL}_{\mathcal{S}_i}(\mathbf{X}_i),$$

Feature alignment

$$\mathbf{H}^{\mathrm{FAL}} = \mathrm{GE}(\mathrm{FAL}(\mathcal{G}_S, \tilde{\mathcal{X}}_S; \Psi); \Theta),$$

$$\tilde{\mathcal{X}}_S = \{\tilde{\mathbf{X}}_i : G_i \in \mathcal{G}_S\}$$

Structure alignment

Pretext tokens

$$\mathcal{T}_{S_i} = \{\mathbf{t}^l_{S_i} : l \in \{1, \dots, L\}\}$$

Add token to each layer of graph encoder, guiding structure-based aggregation

$$\mathbf{h}^l_v = \mathrm{Aggr}(\mathbf{h}^{l-1}_v, \{\mathbf{t}^l_{S_i} \odot \mathbf{h}^{l-1}_u : u \in \mathcal{N}_v\}; \theta^l),$$
$$\forall v \in V_i,$$

Structure-aligned embedding matrix

$$\mathbf{H}^{\mathrm{SAL}} = \mathrm{Stack}(\mathbf{H}^{\mathrm{SAL}}_1, \dots, \mathbf{H}^{\mathrm{SAL}}_K)$$

Overall embedding

$$\mathbf{H}^{\mathrm{AL}} = \mathbf{H}^{\mathrm{FAL}} + \alpha \mathbf{H}^{\mathrm{SAL}}$$

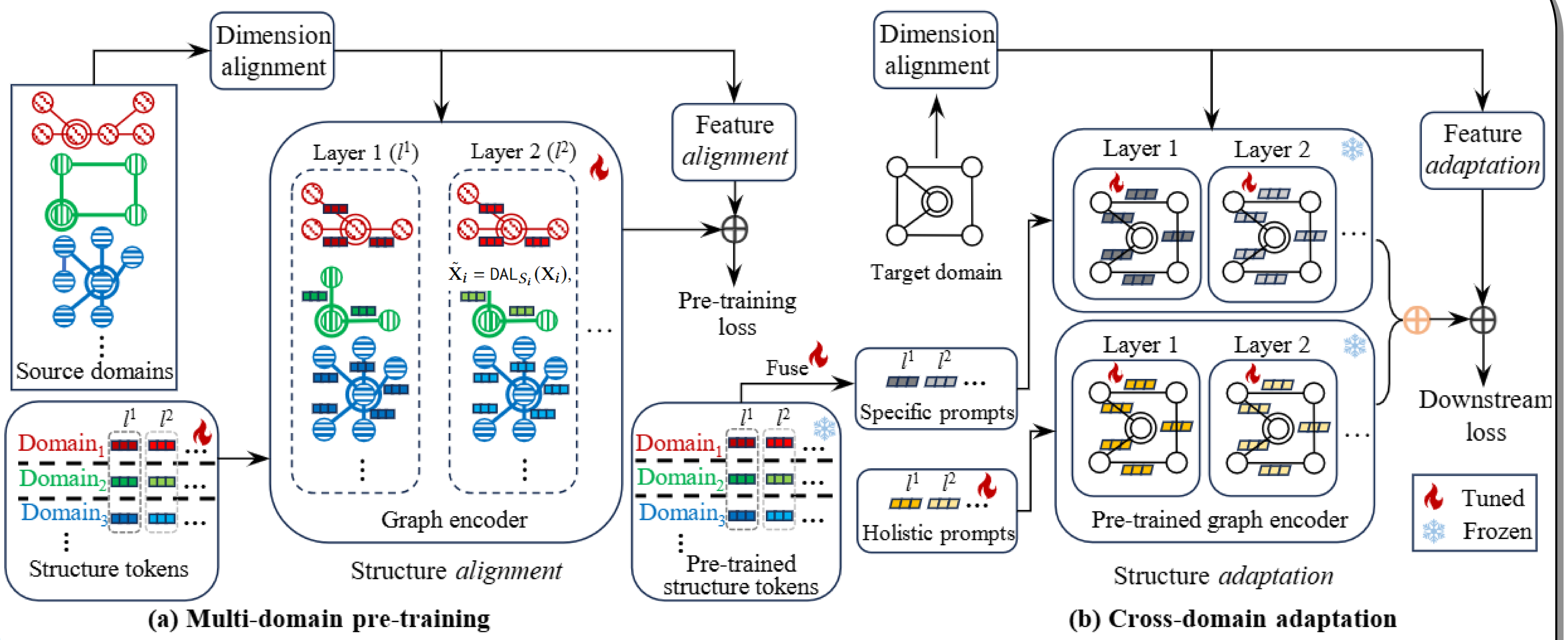

Figure 2: Overall framework of SAMGPT.

Prompt tuning

Feature adaptation

$$\mathbf{H}^{\mathrm{FAD}} = \mathrm{GE}(\mathrm{FAD}(G, \tilde{\mathbf{X}}; \Gamma); \Theta_{\mathrm{pre}})$$

Holistic prompts

$$\mathcal{P}_{\mathrm{hol}} = \{\mathbf{p}^1_{\mathrm{hol}}, \dots, \mathbf{p}^L_{\mathrm{hol}}\}$$

Specific prompts

$$\mathcal{P}_{\mathrm{spe}} = \{\mathbf{p}^1_{\mathrm{spe}}, \dots, \mathbf{p}^L_{\mathrm{spe}}\}$$

$$\mathbf{p}^l_{\mathrm{spe}} = \sum_{i=1}^{K} \lambda^l_i \mathbf{t}^l_{S_i}$$

Prompt modification

$$\mathbf{h}^l_v = \mathrm{Aggr}(\mathbf{h}^{l-1}_v, \{\mathbf{p}^l_{\mathrm{hol}} \odot \mathbf{h}^{l-1}_u : u \in \mathcal{N}_v\}; \theta^l_{\mathrm{pre}}), \quad \forall v \in V$$

Fuse embeddings

$$\mathbf{H}^{\mathrm{SAD}} = \mathbf{H}^{\mathrm{hol}} + \beta \mathbf{H}^{\mathrm{spe}}$$

Overall embedding

$$\mathbf{H}^{\mathrm{AD}} = \mathbf{H}^{\mathrm{FAD}} + \alpha \mathbf{H}^{\mathrm{SAD}}$$

Experiment

Table 2: Accuracy (%) of one-shot *node classification* with standard deviations. Each column represents a target domain, using other columns as source domains. The best method in each column is bolded, and the runner-up is underlined.

Method \ Target domain	Cora	Citeseer	Pubmed	Photo	Computers	Facebook	LastFM
GCN	29.53 ± 7.56	26.29 ± 6.50	23.32 ± 11.56	26.96 ± 12.94	24.40 ± 5.62	20.45 ± 5.62	9.21 ± 3.11
GAT	24.27 ± 9.26	21.56 ± 8.09	22.28 ± 9.76	17.85 ± 10.22	23.03 ± 12.12	29.27 ± 6.47	9.01 ± 2.61
DGI	33.40 ± 10.48	25.80 ± 8.27	47.22 ± 9.50	30.89 ± 10.54	25.75 ± 12.45	34.36 ± 9.57	14.14 ± 6.31
GraphCL	27.72 ± 9.37	35.02 ± 8.46	48.89 ± 9.03	34.78 ± 11.56	23.79 ± 12.28	34.85 ± 7.07	18.93 ± 7.32
GPPT	27.18 ± 4.88	25.90 ± 4.68	39.82 ± 8.79	31.58 ± 10.27	19.94 ± 9.61	34.73 ± 3.99	20.98 ± 3.98
GraphPrompt	28.26 ± 12.68	32.51 ± 8.73	47.47 ± 9.15	48.11 ± 9.89	42.82 ± 11.67	40.44 ± 9.68	19.84 ± 7.23
GPF	32.17 ± 6.56	36.79 ± 7.70	41.28 ± 8.14	47.47 ± 8.19	35.75 ± 7.12	40.45 ± 6.34	27.26 ± 5.50
Hassani	33.35 ± 6.93	33.66 ± 7.24	39.87 ± 8.16	48.48 ± 7.07	39.99 ± 7.91	37.70 ± 5.79	27.16 ± 4.94
GCOPE	35.62 ± 11.93	**38.33** ± 9.28	45.38 ± 9.87	52.87 ± 9.19	45.65 ± 10.69	40.63 ± 8.50	28.84 ± 7.59
SAMGPT	**47.80** ± 11.88	36.38 ± 9.10	**50.25** ± 10.43	**58.71** ± 8.69	**48.22** ± 8.17	**42.70** ± 8.73	**33.36** ± 8.11

Table 3: Accuracy (%) of one-shot *graph classification* with standard deviations. Each column represents a target domain, using other columns as source domains. The best method in each column is bolded, and the runner-up is underlined.

Method \ Target domain	Cora	Citeseer	Pubmed	Photo	Computers	Facebook	LastFM
GCN	30.64 ± 10.31	26.90 ± 7.15	38.84 ± 11.82	15.60 ± 8.77	21.94 ± 14.51	31.33 ± 9.47	28.83 ± 9.60
GAT	27.80 ± 7.85	27.50 ± 7.13	21.66 ± 8.70	15.74 ± 7.62	16.02 ± 13.46	21.20 ± 7.31	27.80 ± 7.85
InfoGraph	34.98 ± 10.15	35.87 ± 9.84	48.67 ± 12.29	25.70 ± 11.73	19.02 ± 14.09	31.26 ± 9.65	23.29 ± 7.99
GraphCL	42.70 ± 10.64	36.66 ± 8.67	47.53 ± 11.52	33.07 ± 12.31	16.02 ± 13.47	21.99 ± 13.00	21.30 ± 10.45
GraphPrompt	37.38 ± 14.03	36.66 ± 9.19	**49.55** ± 10.25	50.79 ± 12.31	43.09 ± 11.45	41.71 ± 10.61	32.62 ± 8.54
GPF	39.62 ± 8.52	36.73 ± 7.66	45.08 ± 10.36	47.57 ± 10.16	35.70 ± 8.71	34.84 ± 5.14	34.31 ± 7.05
Hassani	36.86 ± 10.74	35.78 ± 8.80	43.97 ± 13.27	41.55 ± 13.08	29.49 ± 13.86	35.57 ± 9.00	25.39 ± 8.14
GCOPE	38.85 ± 10.99	**39.93** ± 9.82	47.05 ± 11.74	53.93 ± 9.74	45.60 ± 10.96	40.26 ± 9.53	34.68 ± 7.70
SAMGPT	**55.35** ± 13.62	38.75 ± 9.40	48.69 ± 10.16	**58.75** ± 11.67	**48.72** ± 11.18	**43.71** ± 9.54	**48.28** ± 9.72

Table 4: Data ablation study with an increasing number of source domains, while fixing *Cora* as the target domain.

Method	Number of source domains 1	2	3	4
GraphPrompt	35.53±12.06	37.13±11.79	36.90±11.23	38.54±11.84
GCOPE	39.47±12.14	36.63± 9.46	35.28±11.99	38.61±12.74
SAMGPT	**40.43**±11.00	**41.97**±11.01	**42.30**±11.56	**45.95**±12.96

Table 5: Model ablation study on key components of SAMGPT.

Methods	Structure tokens	Holistic prompts	Specific prompts	Target domain for node classification Cora	Photo	Facebook	Target domain for graph classification Cora	Photo	Facebook
Variant 1	×	×	×	36.36 ± 12.71	49.10 ± 9.94	35.36 ± 9.06	45.44 ± 13.47	52.45 ± 12.37	38.74 ± 10.26
Variant 2	×	×	✓	40.62 ± 11.79	56.23 ± 9.04	39.80 ± 10.39	45.63 ± 13.52	57.78 ± 11.64	42.22 ± 10.95
Variant 3	✓	×	×	44.26 ± 10.92	56.61 ± 10.14	41.11 ± 8.34	52.88 ± 12.25	58.14 ± 12.01	43.12 ± 9.76
Variant 4	✓	✓	×	46.10 ± 12.02	57.76 ± 10.00	40.46 ± 8.89	54.52 ± 14.32	58.12 ± 12.30	43.15 ± 10.12
SAMGPT	✓	✓	✓	**47.80** ± 11.88	**58.71** ± 8.69	**42.70** ± 8.73	**55.35** ± 13.62	**58.75** ± 11.67	**43.71** ± 9.54