# Quantizing Text-attributed Graphs for Semantic-Structural Integration


Jianyuan Bo[1], Hao Wu[2], Yuan Fang[1]

[1]Singapore Management University, Singapore

[2]Beijing Normal University

{jybo.2020, yfang}@smu.edu.sg, wuhao@bnu.edu.cn

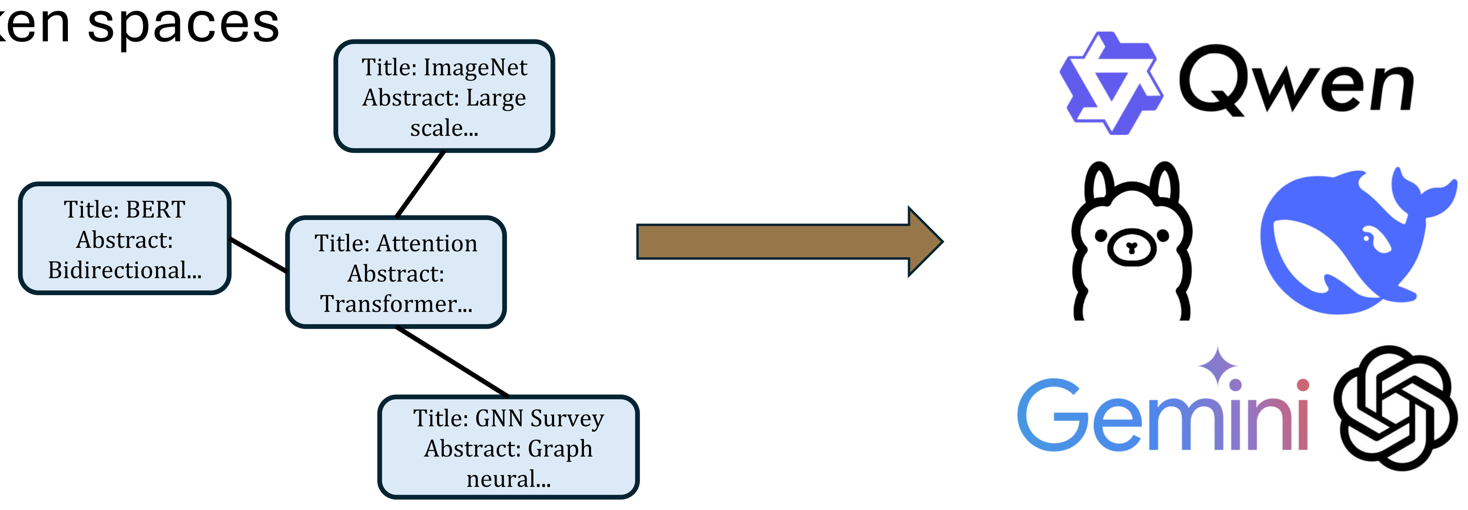


August 3–7, 2025 KDD2025


## Introduction

### 🎯 Text-Attributed Graphs (TAGs) + Large Language Models

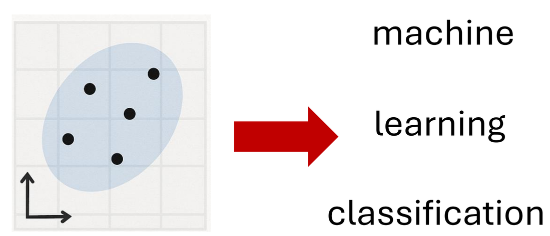
- TAGs contain rich semantic + structural information
- LLMs excel at semantic understanding but struggle with graph structures
- Need to bridge continuous graph embeddings ↔ discrete LLM token spaces




## Motivation

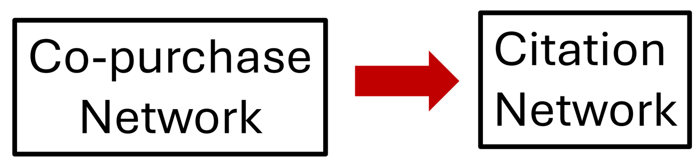
### Problem 1: Semantic-Structural Integration

- Graph embeddings are continuous vectors
- LLMs work with discrete tokens




### Problem 2: Transfer Learning Requirements

- Need labeled source data
- Limits adaptability across domains



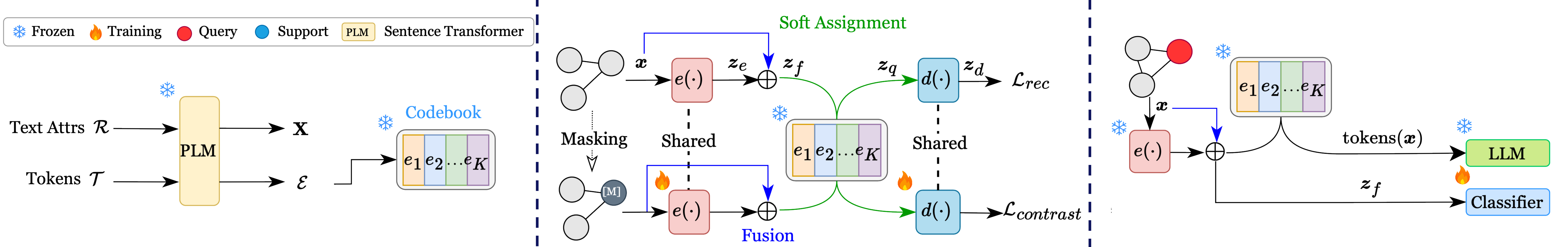

### 🔥 Current Limitations

❌ Expensive alignment mechanisms between GNN and LLM

❌ Manual graph verbalization loses structural details

❌ Require labeled data for cross-dataset transfer

## Proposed Method: STAG




(a) Codebook Construction (b) Pre-training (c) Inference

## Method

### STAG: Soft Tokenization for Text-Attributed Graphs

### 🔑 Three Key Innovations

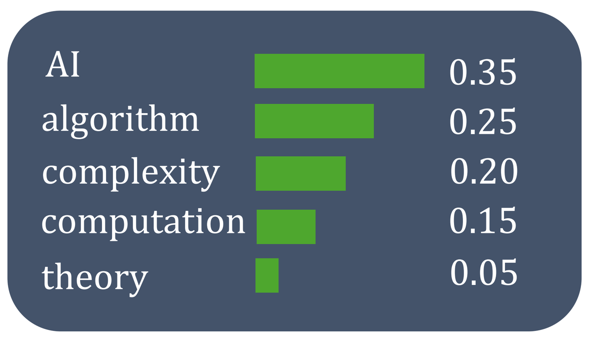
#### 1 Semantic-Structural Fusion

- Fuse GNN embeddings + original text features

$$z_f = \phi \cdot \frac{\mathbf{W}_f z_e}{\|\mathbf{W}_f z_e\|_2} + \psi \cdot \frac{x}{\|x\|_2}$$

#### 2 Soft Assignment Strategy

- Map nodes to token distributions (not single tokens)
- KL divergence guides quantization to semantic tokens
- Prevents overfitting, improves transferability




#### 3 Dual-Branch Training

- Reconstruction branch: preserves semantics
- Contrastive branch: captures neighborhood structure
- No labeled data required!

### Inference with LLM

**System Prompt**: You are a node classifier. Given a list of tokens representing a node's features, predict its class from the following options: [Research Paper, Dataset, Software].

**Few-shot examples**: Node tokens: [research, methodology, experiment] Class: Research Paper

Node tokens: [benchmark, statistics, collection] Class: Dataset

Node tokens: [implementation, code, library] Class: Software

**Test Node**: Node tokens: [algorithm, computation, optimization] Predict the class:

## Experiments

### Few-shot Node Classification

| Pre-train data | Method | LLM | Target data | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Cora | Cora Full | CiteSeer | PubMed | WikiCS | ogbn-arxiv | ogbn-products |
| Same as target | GCN | ✗ | 76.10±4.26 | 82.81±7.40 | 59.95±6.92 | 66.35±6.71 | 70.55±8.26 | 76.61±7.72 | 80.08±7.41 |
| | GAT | ✗ | 79.60±5.00 | 84.72±7.83 | 60.85±6.78 | 67.40±7.18 | 77.95±7.81 | 80.80±7.99 | 81.94±7.07 |
| No pre-train | Raw Text | ✓ | 63.40±9.07 | 71.66±7.84 | 62.10±5.35 | 85.00±5.48 | 77.15±6.92 | 54.74±9.21 | 87.58±5.48 |
| | Raw Feat + Quantization | ✓ | 54.85±6.28 | 73.74±8.10 | 56.40±5.48 | 48.30±7.78 | 73.40±8.19 | 63.75±9.98 | 65.84±9.96 |
| | Raw Feat + Linear Probing | ✗ | 70.25±7.22 | 81.29±7.47 | 63.00±6.72 | 68.30±6.28 | 78.05±7.47 | 83.05±7.40 | 77.53±7.16 |
| Cora Full | DGI | ✗ | 77.05±5.12 | 83.32±8.12 | 63.85±5.39 | 68.20±7.57 | 78.65±6.90 | 81.30±8.51 | 79.90±7.20 |
| | GraphMAE2 | ✗ | 77.70±6.92 | 84.74±7.42 | 65.25±5.84 | 66.35±6.09 | 80.95±4.96 | 80.04±8.15 | 73.93±7.57 |
| | GPPT | ✗ | 27.16±7.61 | 67.90±12.72 | 28.66±7.60 | 21.53±10.91 | 29.00±8.08 | 36.92±10.32 | 24.32±5.13 |
| | G2P2 | ✗ | 74.90±7.47 | 81.10±7.44 | 59.65±9.68 | 67.85±8.02 | 69.90±10.52 | 68.75±10.14 | 70.97±10.03 |
| | Prodigy | ✗ | 39.50±6.75 | 60.80±6.38 | 42.90±5.02 | 43.68±6.91 | 43.25±6.91 | 47.85±6.89 | 30.70±5.94 |
| | OFA | ✗ | 45.95±4.52 | 56.95±5.31 | 36.80±5.50 | 49.40±4.75 | 46.45±4.67 | 50.80±4.73 | 33.60±4.26 |
| | STAG | ✓ | 67.60±6.72 | 80.95±8.02 | 62.45±7.02 | 54.50±7.83 | 79.20±8.41 | 71.56±10.32 | 69.34±9.93 |
| | + Linear Probing | ✗ | 78.50±5.62 | 86.04±6.70 | **66.70**±5.36 | **69.00**±6.31 | **84.05**±5.78 | 82.99±8.10 | 79.62±7.12 |
| | + Prompt Tuning | ✓ | 73.30±4.77 | 85.20±7.59 | 65.40±5.98 | 66.20±5.70 | 79.45±7.53 | 79.18±8.28 | 73.94±9.67 |
| | + Prompt Tuning* | ✗ | **78.65**±5.93 | **86.66**±7.67 | 65.80±7.03 | 68.25±6.80 | 83.55±5.94 | **83.57**±8.30 | **80.48**±6.86 |

### Zero-shot Node Classification

| Pre-train data | Method | LLM | Target data | | | |
|---|---|---|---|---|---|---|
| | | | Cora | Cora Full | WikiCS | ogbn-arxiv |
| No pre-train | Raw Feat + Q | ✓ | 47.10±5.98 | 60.33±10.88 | 70.40±8.88 | 25.48±5.54 |
| | Raw Feat + $C$ | ✗ | 62.20±8.45 | 77.23±8.96 | 73.85±8.02 | 72.85±10.43 |
| Cora Full | G2P2 | ✗ | 60.45±7.58 | 64.29±11.56 | 50.25±8.43 | 19.66±6.38 |
| | OFA | ✗ | 20.30±2.93 | 23.85±3.58 | 21.45±3.99 | 17.60±3.74 |
| | STAG | ✓ | 48.05±6.15 | 62.63±11.70 | **76.25**±8.48 | 26.01±7.52 |
| | STAG + $C$ | ✗ | **66.55**±7.48 | **82.90**±9.52 | 75.15±7.81 | **74.23**±9.35 |

## Task Generalization

### Zero-shot Binary Link Prediction

| Method | Cora | ogbn-products |
|---|---|---|
| LLaGA | 87.35 | 92.99 |
| STAG | 63.00 | 92.65 |
| STAG (non LLM) | **93.20** | **96.85** |

### N-way 5-shot Edge Classification

| Method | WN18RR | FB15K237 |
|---|---|---|
| OFA | 34.35 | 19.55 |
| STAG | 41.75 | 56.60 |
| STAG + Linear Probing | **58.30** | **74.80** |