TOWARDS EFFECTIVE FEDERATED MULTIMODAL GRAPH LEARNING VIA NAVIGATING MULTIFACETED HETEROGENEITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal-attributed graphs (MAGs), where nodes carry heterogeneous semantic content across multiple modalities while edges encode relational dependencies, have been widely adopted across diverse domains. Federated multimodal graph learning (FMGL) extends federated graph learning (FGL) to MAGs, enabling collaborative optimization across decentralized MAGs without exposing raw data. However, naively applying existing FGL methods to FMGL is insufficient, as they fail to navigate the *multifaceted heterogeneity* inherent in decentralized MAGs, including task heterogeneity across diverse client objectives, modality heterogeneity from discrepant modality quality and semantic domains, and topology heterogeneity arising from divergent topological patterns with low cross-modality correlation. To address these challenges, we propose **Fed**erated multimodal graph learning with **T**opology-aware **C**ross-modal **R**outing (**FedTCR**), the first systematic algorithm designed for FMGL. To handle task heterogeneity, FedTCR employs a two-stage paradigm that comprises federated task-agnostic pre-training followed by isolated task-oriented fine-tuning. To jointly address modality and topology heterogeneity, FedTCR introduces a topology-aware cross-modal routing mechanism. Concretely, each client distills modality-specific knowledge into compact prototypes via topology-aware importance-weighted aggregation informed by graph structure; the server then evaluates cross-client cross-modal relationships among these structure-informed prototypes and routes informative ones as contrastive references, driving a tri-level cross-modal contrastive learning scheme that jointly aligns cross-client modalities while preserving discrimination. Experiments across 7 domains demonstrate that FedTCR outperforms state-of-the-art baselines on both graph-centric and modality-centric tasks.

1 INTRODUCTION

Multimodal-attributed graphs (MAGs) represent entities as nodes enriched with heterogeneous semantic content across multiple modalities, while edges encode relational dependencies among these entities. This expressive data structure has been widely adopted across diverse domains Yanglet et al. (2025); Li & Nabavi (2024). Motivated by such scenarios, multimodal graph learning (MGL) has garnered increasing attention, both enhancing graph-centric tasks through rich multimodal attributes Zheng et al. (2025) and improving modality-specific tasks via topological structures Mai et al. (2023).

Despite these advances, existing MGL approaches predominantly assume centralized data access, where all MAG data reside within a single learning system. In practice, however, multiple organizations (e.g., Instagram, X) independently maintain private MAGs that encode complementary knowledge, while privacy regulations and commercial constraints prohibit direct data aggregation, preventing isolated MGL models from leveraging this cross-platform collective intelligence.

To this end, recent work Li et al. (2026) proposed federated multimodal graph learning (FMGL), extending federated graph learning (FGL) Li et al. (2024b) to MAGs. FGL enables collaborative optimization through iterative client-server communication, transmitting model parameters or gradients without exposing raw graph data. While conceptually promising, this paradigm lacks systematic analysis of key challenges and targeted solutions. We argue that naively extending traditional FGL

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.

methods to FMGL is insufficient, as they fail to address the *multifaceted heterogeneity* inherent in decentralized MAGs:

Task Heterogeneity. The rich multimodal semantics of MAGs naturally lead to diverse application requirements across clients. Beyond conventional graph-level tasks (e.g., node classification), clients may pursue modality-centric objectives such as cross-modal retrieval or multimodal generation. However, mainstream FGL algorithms Zhang et al. (2021); Baek et al. (2023); Huang et al. (2024) are designed for a uniform task type and exclusively focus on graph-centric objectives Li et al. (2024b); Zhu et al. (2025). Their architectures and optimization procedures are tightly coupled to a single task type, preventing collaboration when clients pursue different objectives.

Modality Heterogeneity. Distinct MAG sources and collection preferences across clients yield significant discrepancies in both modality quality and semantic domains. Clients with noisy or domain-biased modalities may propagate misleading knowledge during cross-client aggregation, degrading others' high-quality representations. Existing FGL methods address only unimodal attribute heterogeneity, lacking per-modality reliability assessment and cross-client cross-modal alignment capabilities.

Topology Heterogeneity. Graph topology provides crucial structural priors for modeling entity relationships. In FMGL, however, client MAGs often exhibit diverse topological patterns due to varying user behaviors or data collection processes. Existing FGL methods evaluate topological quality via homophily-based metrics on local subgraphs Zhu et al. (2024); Li et al. (2023); Huang et al. (2024), implicitly assuming consistent homophily patterns across clients. This assumption is fragile in MAGs, where node pairs strongly homophilic under one modality may display weak homophily or even heterophily under another Zheng et al. (2025), making traditional FGL methods unable to reliably distill and share topology priors across clients.

Building upon these insights, we propose **Fed**erated multimodal graph learning with **T**opology-aware **C**ross-modal **R**outing (**FedTCR**), the first systematic algorithm addressing multifaceted heterogeneity in FMGL. To tackle ***task heterogeneity***, FedTCR adopts a two-stage paradigm: federated task-agnostic pre-training followed by isolated task-oriented fine-tuning, where clients collaboratively learn a shared multimodal graph encoder without assuming unified downstream objectives and then independently adapt it to their specific applications. To jointly address ***modality and topology heterogeneity***, FedTCR introduces a topology-aware cross-modal routing mechanism that synergistically integrates two key designs: (1) *topology-aware prototype construction*, where each client estimates node importance from graph topology and performs importance-weighted aggregation to distill modality-specific knowledge into compact prototypes; (2) *cross-modal prototype routing*, where the server evaluates cross-client cross-modal relationships among these structure-informed prototypes and routes informative ones as positive or negative references for each client. The routed references drive a tri-level cross-modal contrastive learning scheme (operating at node, neighbor, and client levels) that jointly aligns cross-client modalities while preserving discrimination.

Contributions. (1) **Valuable Insights.** We identify the *multifaceted heterogeneity* in FMGL, revealing fundamental challenges overlooked by existing FGL methods. (2) **Novel Method.** We propose FedTCR, the first optimization algorithm explicitly designed for FMGL. FedTCR adopts a two-stage paradigm with task-agnostic federated pre-training to handle task heterogeneity, and introduces a topology-aware cross-modal routing mechanism to jointly address modality and topology heterogeneity. (3) **State-of-the-art Performance.** Experiments across 7 domains demonstrate that FedTCR achieves consistent and substantial improvements over state-of-the-art baselines on both graph-centric and modality-centric tasks.

2 Related Work

Multimodal Graph Learning. MGL enriches graph learning with heterogeneous semantic content from multiple modalities. A line of work constructs modality-specific relational structures and fuses them for recommendation, e.g., MMGCN Wei et al. (2019), MGAT Tao et al. (2020), and LGMRec Guo et al. (2024). Another line exploits the interplay between multimodal semantics and topology for graph-oriented objectives Guo et al. (2025); Zheng et al. (2025); Mai et al. (2023).

More recent studies inject graph context into generative models, conditioning LLMs on multimodal neighborhoods for graph-to-text generation Yoon et al. (2023) and steering diffusion models with graph-guided signals for graph-to-image synthesis Jin et al. (2024). Complementary benchmarks such as OpenMAG Wan et al. (2026) systematize MAG datasets, encoders, and task protocols. Despite this progress, existing MGL studies predominantly assume centralized data access and cannot accommodate privately held MAGs distributed across organizations.

Federated Graph Learning. FGL extends federated learning to graph-structured data, enabling collaborative training without exposing raw graphs Fu et al. (2022). Existing studies mainly fall into three categories: graph-level FGL, which mitigates cross-dataset heterogeneity via gradient-aware client clustering (GCFL+ Xie et al. (2021)); subgraph-level FGL, which addresses subgraph heterogeneity and missing cross-client links (Fed-PUB Baek et al. (2023), FedGTA Li et al. (2023), FedSage+ Zhang et al. (2021), AdaFGL Li et al. (2024a), FGSSL Huang et al. (2024), FGGP Wan et al. (2024), FedTAD Zhu et al. (2024), FedDEP Zhang et al. (2024b), FedGCN Yao et al. (2023), FedGL Chen et al. (2021)); and node-level FGL over ego-networks (FedEgo Zhang et al. (2023)), standardized by benchmarks such as FedGraphNN He et al. (2021), FederatedScope-GNN Wang et al. (2022), and OpenFGL Li et al. (2024b). However, these methods are designed for unimodal graphs and a uniform task type, and thus cannot navigate the multifaceted heterogeneity inherent in federated multimodal graph learning, which is the focus of this work.

3 Preliminaries

Multimodal-attributed Graphs. We consider a FMGL system consisting of K clients coordinated by a central server. Each client k privately maintains a MAG $\mathcal{G}^k = (\mathcal{V}^k, \mathcal{E}^k, \{\mathbf{X}^{k,(m)}\}_{m\in\mathcal{M}})$, where $\mathcal{V}^k$ and $\mathcal{E}^k$ denote the node and edge sets, respectively. $\mathcal{M}$ denotes the set of available modalities. For each node $v_i \in \mathcal{V}^k$ under modality m, a modality-specific feature vector $\mathbf{x}_i^{k,(m)} \in \mathbb{R}^{d_m}$ is obtained using a pre-trained modality encoder (e.g., Sentence-BERT Reimers & Gurevych (2019) or T5 Raffel et al. (2020) for text, ViT Wu et al. (2020a) or DINOv2 Oquab et al. (2023) for images). The node features of modality m are organized as a matrix $\mathbf{X}^{k,(m)} \in \mathbb{R}^{|\mathcal{V}^k|\times d_m}$. All modalities share a common relational structure represented by the adjacency matrix $\mathbf{A}^k \in \mathbb{R}^{|\mathcal{V}^k|\times|\mathcal{V}^k|}$, with degree matrix $\mathbf{D}^k$ defined as $\mathbf{D}^k_{ii} = \sum_j \mathbf{A}^k_{ij}$ and $\mathbf{D}^k_{ij} = 0$ for $i \neq j$. Following standard graph signal processing, the symmetrically normalized adjacency matrix is defined as $\tilde{\mathbf{A}}^k = (\mathbf{D}^k)^{-1/2}\mathbf{A}^k(\mathbf{D}^k)^{-1/2}$, and the corresponding graph Laplacian is $\mathbf{L}^k = \mathbf{I} - \tilde{\mathbf{A}}^k$.

Federated Training Protocol. Due to privacy constraints, clients cannot share raw graph data; instead, they collaborate through iterative server-client communication. FedTCR adopts a two-stage paradigm: *federated task-agnostic pre-training* followed by *isolated task-oriented fine-tuning*. During pre-training, at each communication round the server samples a subset $\mathcal{S}$ of clients; each selected client $k \in \mathcal{S}$ downloads the global parameters $\hat{\mathbf{W}}$ and any server-side messages (e.g., routed prototypes) and performs local optimization on its private MAG $\mathcal{G}^k$; the server then aggregates the updated parameters via node-count-weighted averaging, $\hat{\mathbf{W}} = \sum_{k\in\mathcal{S}} \frac{|\mathcal{V}^k|}{N}\mathbf{W}^k$ with $N = \sum_{k\in\mathcal{S}}|\mathcal{V}^k|$, and performs additional procedures (e.g., cross-modal prototype routing). After pre-training converges, each client independently fine-tunes the encoder with a task-specific head without further communication.

Downstream Tasks. We consider two graph-centric tasks (node classification and link prediction) and two modality-centric tasks (modality retrieval and modality generation covering G2Text and G2Image), with formal definitions deferred to Appendix B. In this work, we focus on learning a task-agnostic multimodal graph encoder through federated pre-training, which can be subsequently adapted to these diverse downstream applications on individual clients.

4 Methodology

We present FedTCR (Fig. 1), an optimization algorithm explicitly designed for FMGL to address the multifaceted heterogeneity in Sec. 1. FedTCR adopts a two-stage paradigm to handle task heterogeneity: federated task-agnostic pre-training (Sec. 4.1) followed by isolated task-oriented

fine-tuning (Sec. 4.2), and jointly addresses modality and topology heterogeneity during pre-training through a topology-aware cross-modal routing mechanism.

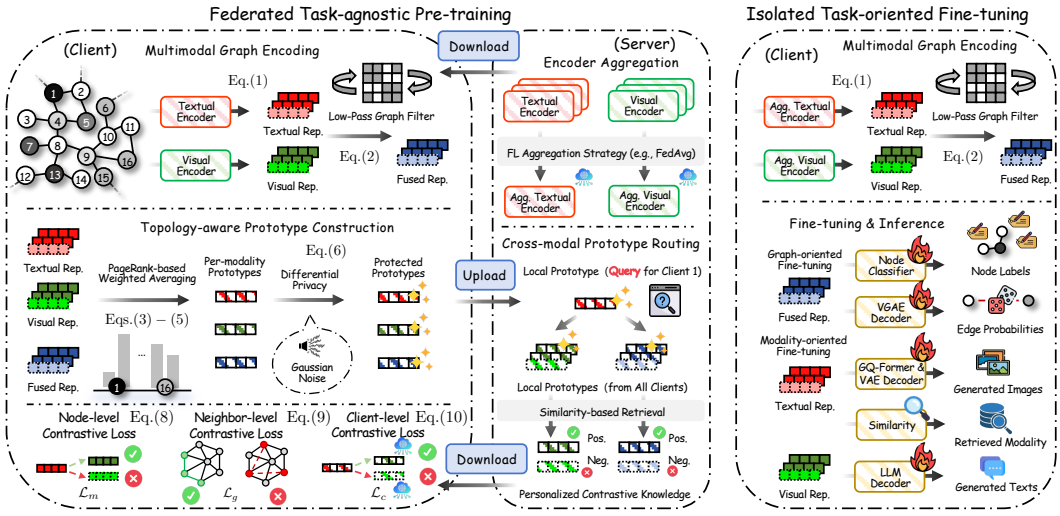


Figure 1: Overview of FedTCR, which addresses multifaceted heterogeneity in FMGL through federated task-agnostic pre-training with topology-aware cross-modal routing, followed by isolated task-oriented fine-tuning.

4.1 Federated Task-agnostic Pre-training

To address task heterogeneity, FedTCR decouples task-agnostic representation learning from task-specific adaptation. During pre-training, clients collaboratively learn a shared multimodal graph encoder without assuming unified downstream objectives, thereby capturing transferable multimodal and structural knowledge across decentralized MAGs. At each communication round, pre-training proceeds in two stages: (1) Each client performs local multimodal graph encoding, constructs topology-aware modality prototypes, and optimizes multimodal representations; (2) The server aggregates encoder parameters via weighted averaging and performs cross-modal prototype routing to deliver personalized contrastive knowledge to each client for the next round of optimization.

Multimodal Graph Encoding. For each client k with MAG $\mathcal{G}^k$, we first project each modality's raw features into a shared d-dimensional latent space via modality-specific linear transformations $\mathbf{\Phi}^{(m)} \in \mathbb{R}^{d_m \times d}$:

$$\mathbf{z}_i^{k,(m)} = \frac{(\mathbf{X}^{k,(m)}\mathbf{\Phi}^{(m)})_i}{\|(\mathbf{X}^{k,(m)}\mathbf{\Phi}^{(m)})_i\|_2}, \quad \forall i \in \mathcal{V}^k,\ m \in \mathcal{M}, \tag{1}$$

where $(\cdot)_i$ denotes the i-th row, and $\mathbf{Z}^{k,(m)} \in \mathbb{R}^{|\mathcal{V}^k| \times d}$ stacks these rows to stabilize training. The fused representation is obtained by averaging across modalities: $\bar{\mathbf{Z}}^k = \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \mathbf{Z}^{k,(m)}$. We then apply low-pass graph filtering over the local topology to produce the multimodal graph representation:

$$\mathbf{H}^k = \frac{1}{\alpha+1}\sum_{l=0}^{L}\left(\frac{\alpha}{\alpha+1}\tilde{\mathbf{A}}^k\right)^l \bar{\mathbf{Z}}^k, \tag{2}$$

where $\alpha > 0$ controls the smoothing strength, $\tilde{\mathbf{A}}^k$ is the symmetrically normalized adjacency matrix, and L is the propagation depth. As we show in Theorem 1, many widely used GNNs (e.g., GCN Kipf & Welling (2017), SGC Wu et al. (2019)) can be viewed as special cases of this graph-filtering formulation.

Topology-aware Prototype Construction. In FMGL, we distill each client's modality-specific knowledge into compact prototypes for cross-client alignment. Naive mean aggregation treats all nodes equally and ignores the local MAG topology; inspired by Blondel et al. (2008); Lovász (1993), we instead assign each node a topology-aware weight measuring its representativeness for the graph's community. Without loss of generality, we use PageRank algorithm Page et al. (1999) to compute

these weights, so that importance propagates over multiple hops rather than depending only on 1-hop degree. Let $\alpha \in (0,1)$ be the damping factor; the PageRank vector $\boldsymbol{\pi}^k \in \mathbb{R}^{|\mathcal{V}^k|}$ is the stationary distribution

$$\boldsymbol{\pi}^k = \frac{1-\alpha}{|\mathcal{V}^k|}\mathbf{1} + \alpha\left((\mathbf{D}^k)^{-1}\mathbf{A}^k\right)^\top \boldsymbol{\pi}^k, \tag{3}$$

where $\mathbf{A}^k$ is the adjacency matrix and $\mathbf{D}^k$ is the diagonal degree matrix; the transpose gives the transition probabilities from neighbors into each node. We compute $\boldsymbol{\pi}^k$ by power iteration, then set $\omega_i^k = \pi_i^k$ after convergence, formulated as:

$$\pi_i^{k,(t+1)} = \frac{1-\alpha}{|\mathcal{V}^k|} + \alpha \sum_{j\in\mathcal{V}^k} \frac{\mathbf{A}^k_{ji}}{d_j^k}\pi_j^{k,(t)}, \tag{4}$$

where $d_j^k = \sum_l \mathbf{A}^k_{jl}$ is the degree of node v_j. $\boldsymbol{\pi}^k$ is normalized so that $\sum_{i\in\mathcal{V}^k}\omega_i^k = 1$ with $\omega_i^k = \pi_i^k$.

Using these normalized weights, we define modality-specific prototypes by importance-weighted aggregation:

$$\mathbf{P}^{k,(m)} = \sum_{i\in\mathcal{V}^k} \omega_i^k \cdot \mathbf{z}_i^{k,(m)}, \quad \forall m \in \bar{\mathcal{M}}, \tag{5}$$

where $\bar{\mathcal{M}} = \mathcal{M}\cup\{|\mathcal{M}|+1\}$ adds the fused modality; $\mathbf{P}^{k,(m)}$ for $m \in \bar{\mathcal{M}}$ is computed from $\mathbf{Z}^{k,(m)}$, and $\mathbf{Z}^{k,(|\mathcal{M}|+1)} = \mathbf{H}^k$ (i.e., Eq. (2)). Each prototype thus distills the local MAG's modality-specific semantics and structure into a single vector. Together with the updated encoder parameters, they are uploaded to the server for aggregation and used in the subsequent routing. To protect privacy, each client further applies (ε,δ)-differential privacy (DP) Dwork (2008); Abadi et al. (2016) before uploading: each prototype $\mathbf{P}^{k,(m)}$ is L2-clipped with a bound $C>0$ and then perturbed with Gaussian noise calibrated to the target (ε,δ):

$$\begin{aligned}\tilde{\mathbf{P}}^{k,(m)} &= \mathbf{P}^{k,(m)} \cdot \min\Big(1, \frac{C}{\|\mathbf{P}^{k,(m)}\|_2}\Big),\\ \hat{\mathbf{P}}^{k,(m)} &= \tilde{\mathbf{P}}^{k,(m)} + \mathcal{N}(0,\sigma^2 C^2 \mathbf{I}_d),\end{aligned} \tag{6}$$

where σ is the noise scale. The server performs cross-modal prototype routing (Eq. (7)) on the noisy prototypes, and ε and C jointly control the privacy-utility tradeoff.

Cross-modal Prototype Routing. After receiving all clients' topology-aware prototypes and encoder parameters, the server (1) aggregates encoder parameters $\{\boldsymbol{\Phi}^{(m)}\}_{m\in\mathcal{M}}$ via a mainstream FL algorithm (e.g., FedAvg McMahan et al. (2017)), and (2) performs cross-modal prototype routing to produce and deliver personalized contrastive knowledge to each client, which guides the next round of federated optimization.

For each client a, modality $i \in \bar{\mathcal{M}}$, and each other modality $j \neq i$, the server selects from $\{\mathbf{P}^{b,(j)}\}_{b\neq a}$ the most similar prototype as the positive and the most dissimilar as the negative, which can be formulated as follows:

$$\begin{aligned}\mathbf{P}_{\text{pos}}^{a,(i,j)} &= \hat{\mathbf{P}}^{b^*,(j)}, \quad b^* = \arg\max_{b\neq a} \cos(\hat{\mathbf{P}}^{a,(i)}, \hat{\mathbf{P}}^{b,(j)}),\\ \mathbf{P}_{\text{neg}}^{a,(i,j)} &= \hat{\mathbf{P}}^{b',(j)}, \quad b' = \arg\min_{b\neq a} \cos(\hat{\mathbf{P}}^{a,(i)}, \hat{\mathbf{P}}^{b,(j)}).\end{aligned} \tag{7}$$

Here $\cos(\cdot,\cdot)$ denotes cosine similarity. The server sends the routed pairs $\{(\mathbf{P}_{\text{pos}}^{a,(i,j)}, \mathbf{P}_{\text{neg}}^{a,(i,j)})\}_{i,j}$ to each client a for client-level contrastive learning (Eq. (10)). Enforcing this contrast yields two effects: (1) per-modality cross-client alignment encourages complementary knowledge transfer and mitigates propagation of noisy or domain-biased modalities, addressing modality heterogeneity; (2) aligning topology-aware prototypes across modality boundaries harmonizes diverse topological patterns, addressing the topology heterogeneity inherent in decentralized MAGs.

Local Optimization Objectives. Each client optimizes a composite objective comprising two local multimodal graph learning losses and one cross-client alignment loss, following the cross-contrastive paradigm. Letting $\psi(\mathbf{u},\mathbf{v}) = \exp(\cos(\mathbf{u},\mathbf{v})/\tau)$ with temperature τ, and denoting $\mathbf{z}_i^{k,(|\mathcal{M}|+1)} = \mathbf{h}_i^k$

as the fused graph-filtered representation treated as an additional modality, we write $\psi_{ij}^{(m_1,m_2)}$ as shorthand for $\psi\big(\mathbf{z}_i^{k,(m_1)}, \mathbf{z}_j^{k,(m_2)}\big)$:

(1) ***Node-level Contrastive Loss*** maximizes the inter-modality agreement of the same node while contrasting it against imposter nodes from the other modality:

$$\mathcal{L}_m = \sum_{\substack{m_1 \in \bar{\mathcal{M}} \\ m_2 \in \bar{\mathcal{M}} \setminus \{m_1\}}} \sum_{i \in \mathcal{V}^k} -\log \frac{\psi_{ii}^{(m_1,m_2)}}{\sum_{j \in \mathcal{V}^k} \psi_{ij}^{(m_1,m_2)}}. \tag{8}$$

(2) ***Neighbor-level Contrastive Loss*** captures the semantic relations underlying the graph topology. For each node $v_i \in \mathcal{V}^k$, we simulate random walks on client k's graph to collect a set of topological neighbors $\mathcal{P}_i^k$ as positives, and randomly sample $\mathcal{N}_i^k \subset \mathcal{V}^k \setminus \mathcal{P}_i^k$ as negatives:

$$\mathcal{L}_g = \sum_{\substack{m_1 \in \bar{\mathcal{M}} \\ m_2 \in \bar{\mathcal{M}} \setminus \{m_1\}}} \sum_{i \in \mathcal{V}^k} -\log \frac{\sum_{j \in \mathcal{P}_i^k} \psi_{ij}^{(m_1,m_2)}}{\sum_{j \in \mathcal{P}_i^k \cup \mathcal{N}_i^k} \psi_{ij}^{(m_1,m_2)}}. \tag{9}$$

(3) **Client-level Contrastive Loss** leverages the personalized contrastive knowledge from the server (Eq. (7)) to perform prototype-level contrastive learning across clients and modalities. Writing $\psi_{\text{pos}}^{(i,j)} = \psi\big(\mathbf{P}^{k,(i)}, \mathbf{P}_{\text{pos}}^{k,(i,j)}\big)$ and $\psi_{\text{neg}}^{(i,j)} = \psi\big(\mathbf{P}^{k,(i)}, \mathbf{P}_{\text{neg}}^{k,(i,j)}\big)$:

$$\mathcal{L}_c = \sum_{\substack{i \in \bar{\mathcal{M}} \\ j \in \bar{\mathcal{M}} \setminus \{i\}}} -\log \frac{\psi_{\text{pos}}^{(i,j)}}{\psi_{\text{pos}}^{(i,j)} + \psi_{\text{neg}}^{(i,j)}}. \tag{10}$$

Notably, Eq. (10) is activated from the second round after the server collects all client prototypes.

Finally, the overall client-side optimization objective during pre-training is formulated as follows:

$$\mathcal{L} = \lambda_m \mathcal{L}_m + \lambda_g \mathcal{L}_g + \lambda_c \mathcal{L}_c, \tag{11}$$

where λ_m, λ_g, and λ_c are balancing coefficients.

4.2 Isolated Task-oriented Fine-tuning

Once federated pre-training converges, each client retains the aggregated encoder and independently fine-tunes on its local MAG without further communication. Client k attaches a task-specific head g^k and optimizes it jointly with the encoder via a task-specific loss $\mathcal{L}_{\text{task}}$: for *graph-centric tasks* (e.g., node classification, link prediction), the head takes the graph-filtered representation $\mathbf{H}^k$ (Eq. (2)); for *modality-centric tasks* (e.g., modality retrieval), it takes the modality-specific embeddings $\mathbf{Z}^{k,(m)}$ (Eq. (1)) to retain modality-level characteristics. This isolation is deliberate: it avoids conflicting gradients from heterogeneous downstream objectives while preserving the transferable knowledge from pre-training. The complete procedure is given in Appendix J.

5 Experiments

We evaluate FedTCR along six research questions, with full reproducibility details deferred to Appendix D and Appendix B: **Q1**: Can FedTCR outperform existing methods on graph-centric and modality-centric tasks under homogeneous-task clients (Sec. 5.2)? **Q2**: Can it enable effective collaboration across task-heterogeneous clients (Sec. 5.3)? **Q3**: What does each module of FedTCR contribute (Sec. 5.4)? **Q4**: How robust is it under data sparsity (Sec. 5.5)? **Q5**: How sensitive is it to key hyperparameters (Sec. 5.6)? **Q6**: How does it perform under differentially private prototype perturbation (Sec. 5.7)?

5.1 Experimental Setup

Datasets and Simulation Strategy. We evaluate FedTCR on 8 publicly available MAG datasets spanning 7 domains: Movies Ni et al. (2019), Grocery and Toys Ni et al. (2019); Hou et al. (2024), RedditS Desai et al. (2021), BiliDance Zhang et al. (2024a), ele-fashion Ni et al. (2019); Hou et al.

(2024), Flickr30k Plummer et al. (2015), and SemArt Garcia & Vogiatzis (2018). Following subgraph FL benchmarks Li et al. (2024b), each MAG is partitioned into client subgraphs with the Louvain algorithm Blondel et al. (2008): 5 clients per MAG in the *homogeneous-task* configuration, and 4 clients per MAG in the *heterogeneous-task* one (one MAG per task, 20 clients across the 5 tasks in total). Dataset statistics and descriptions are in Appendix D.

Baselines. We compare against four method families: (1) *Isolated Training* (MMGCN Wei et al. (2019), MGAT Tao et al. (2020), LGMRec Guo et al. (2024)); (2) *Classical FL* (FedAvg McMahan et al. (2017), SCAFFOLD Karimireddy et al. (2020), FedDC Gao et al. (2022)) with an MMGCN backbone; (3) *Multimodal FL* (PEPSY Nguyen et al. (2025), FedMVP Che et al. (2024), FedMAC Nguyen et al. (2024)); and (4) *Unimodal FGL* (FedSage+ Zhang et al. (2021), FedGTA Li et al. (2023), FedSPA Tan et al. (2025b), FedIIH Yu et al. (2025), FedSSP Tan et al. (2024), S2FGL Tan et al. (2025a), FedLap Aliakbari et al. (2025)). Unimodal FGL methods are graph-centric only and thus inapplicable to modality-centric tasks; label-dependent methods (e.g., FedGTA, FedIIH) cannot perform link prediction. These cases are marked as "N/A". Baseline details are in Appendix E.

Evaluation Metrics. We adopt task-specific metrics: ACC for node classification, AUC for link prediction, R@1 for modality retrieval, and R-L (ROUGE-L) for G2Text with C-S (CLIP-Score) for G2Image.

5.2 Homogeneous Tasks (Answer for **Q1**)

To answer **Q1**, we compare FedTCR against all four baseline families on the 8 MAGs under the homogeneous-task setting (5 clients per MAG, all sharing the MAG's designated task); Table 1 reports the full per-dataset comparison.

Graph-centric Tasks. FedTCR consistently achieves the best performance across all graph-centric benchmarks (Table 1), improving over the second-best method by +1.50% ACC on Movies and +0.10% on Grocery for node classification, and surpassing the runner-up by +4.45% AUC on RedditS and +2.31% on BiliDance for link prediction. Classical FL methods with the MMGCN backbone outperform isolated training on the link prediction benchmarks, confirming the benefit of federated collaboration on structure-dependent tasks, but they cannot exploit modality- and topology-specific knowledge across clients as FedTCR does.

Modality-centric Tasks. FedTCR also demonstrates clear superiority on modality-centric tasks: +7.75% R@1 on Toys and +1.72% on ele-fashion for modality retrieval, and +6.58% R-L on Flickr30k (G2Text) and +3.64% C-S on SemArt (G2Image) for modality generation, over the best competing method. This consistent superiority shows that topology-aware cross-modal routing effectively transfers complementary multimodal knowledge across clients, while task-agnostic pre-training learns representations that generalize across downstream applications.

5.3 Heterogeneous Tasks (Answer for **Q2**)

To answer **Q2**, we evaluate FedTCR under a task-heterogeneous setting comprising 5 downstream tasks, each instantiated on a different MAG partitioned into 4 clients: node classification on Movies, link prediction on RedditS, modality retrieval on Toys, G2Text on Flickr30k, and G2Image on SemArt. Existing FL and FGL methods cannot join such a federation at all, as their architectures couple the federated parameters and the training objective to a single task type; their only viable collaboration is same-task federation, which is already covered by Table 1 where all of them trail FedTCR. We therefore conduct a controlled self-comparison that varies only the collaboration scope over identical per-task partitions: *Isolated* (no federation), FedTCR (homo.) (4 same-task clients), and FedTCR (hete.) (all 20 clients). As shown in Fig. 2, FedTCR (homo.) already improves over Isolated training by +7.06% on average, and FedTCR (hete.) further improves over FedTCR (homo.) by +2.44% on average. The gains are the largest on the modality-generation tasks (+3.87% on average), which depend most heavily on the complementary cross-modal knowledge transferred by the topology-aware routing, and remain positive on graph-centric tasks (+1.00%). Since the arms differ only in the collaboration scope, these results establish that cross-task collaboration carries genuine net benefit rather than interference, enabling cross-task collective intelligence that is architecturally out of reach for existing methods.

Table 1: Performance comparison under the homogeneous-task setting (5 clients). 'N/A' indicates the algorithm is not applicable to the task. Best results are in bold and the second-best are underlined.

	Node Classification (ACC)		Link Prediction (AUC)		Modality Retrieval (R@1)		Modality Generation (R-L/C-S)	
Methods	Movies	Grocery	RedditS	BiliDance	Toys	ele-fashion	Flickr30k	SemArt
MMGCN	53.47±0.32	80.39±0.51	89.52±0.23	76.06±0.42	49.29±0.46	81.74±0.28	24.56±0.23	66.12±0.28
MGAT	52.12±0.34	80.41±0.49	90.23±0.25	76.21±0.44	48.89±0.43	82.12±0.31	24.34±0.21	65.78±0.30
LGMRec	53.34±0.29	80.56±0.53	90.89±0.22	76.68±0.38	49.78±0.45	82.45±0.26	25.12±0.24	66.89±0.26
Isolated	53.07±0.36	80.93±0.57	88.34±0.26	62.89±0.46	49.44±0.52	81.37±0.31	23.95±0.26	65.34±0.31
FedAvg	49.07±0.43	78.18±0.68	92.32±0.31	72.30±0.56	45.31±0.62	68.96±0.37	24.02±0.33	65.56±0.37
SCAFFOLD	49.56±0.45	78.66±0.63	92.78±0.28	72.78±0.52	45.89±0.58	69.67±0.39	23.89±0.29	65.89±0.34
FedDC	49.87±0.40	78.43±0.66	93.01±0.30	73.01±0.54	46.23±0.64	69.89±0.35	24.45±0.31	66.12±0.38
PEPSY	48.67±0.47	78.13±0.73	93.45±0.33	73.56±0.60	47.21±0.66	70.12±0.40	24.78±0.35	66.45±0.40
FedMVP	48.36±0.50	77.98±0.71	93.21±0.36	74.18±0.63	44.73±0.69	69.52±0.44	25.01±0.32	66.78±0.43
FedMAC	49.01±0.44	78.35±0.76	93.67±0.31	74.02±0.57	46.89±0.63	70.44±0.41	25.23±0.34	67.01±0.38
FedSage+	50.23±0.36	79.21±0.56	92.89±0.26	72.67±0.46	N/A	N/A	N/A	N/A
FedGTA	50.89±0.32	79.56±0.51	N/A	N/A	N/A	N/A	N/A	N/A
FedSPA	48.89±0.64	79.34±1.01	93.12±0.46	73.12±0.82	N/A	N/A	N/A	N/A
FedIIH	51.03±0.34	79.62±0.53	N/A	N/A	N/A	N/A	N/A	N/A
FedSSP	48.35±0.43	78.26±0.67	93.52±0.31	73.45±0.55	43.11±0.61	70.31±0.36	N/A	N/A
S2FGL	48.56±0.89	79.12±1.38	92.97±0.63	72.89±1.13	N/A	N/A	N/A	N/A
FedLap	50.77±0.30	79.43±0.47	93.23±0.22	73.23±0.39	N/A	N/A	N/A	N/A
FedTCR (Ours)	**54.27±0.25**	**81.01±0.40**	**97.84±0.18**	**78.45±0.32**	**53.64±0.36**	**83.87±0.22**	**26.89±0.18**	**69.45±0.22**

5.4 Ablation Study (Answer for **Q3**)

To answer **Q3**, we ablate five components of FedTCR: (1) *w/o TW*: replaces PageRank importance weighting with uniform averaging in prototype construction (Eq. (5)); (2) *w/o CR*: removes cross-modal prototype routing and the associated client-level loss $\mathcal{L}_c$ (Eq. (10)); (3) *w/o* $\mathcal{L}_m$ and (4) *w/o* $\mathcal{L}_g$: remove the node-level (Eq. (8)) and neighbor-level (Eq. (9)) contrastive losses, respectively; and (5) *w/o TS*: removes the pre-training stage and directly performs end-to-end federated task-specific training. All variants are trained under the identical protocol and hyperparameters as the full model, so any performance gap is attributable solely to the removed component.

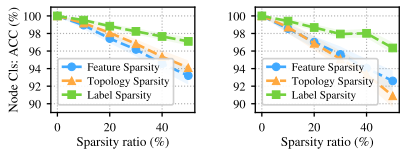
Table 2: Ablation study (5 clients). Each row removes one component.

	Node Cls. (ACC)	Link Pred. (AUC)	Retrieval (R@1)	G2Text (R-L)	G2Image (C-S)
Variants	Movies	RedditS	Toys	Flickr30k	SemArt
w/o TW	53.69±0.30	96.89±0.21	52.45±0.42	26.23±0.22	68.56±0.25
w/o CR	53.03±0.33	95.67±0.24	51.12±0.47	25.56±0.26	68.04±0.28
w/o $\mathcal{L}_m$	53.92±0.28	97.12±0.20	52.91±0.40	26.45±0.19	68.78±0.24
w/o $\mathcal{L}_g$	53.58±0.31	96.45±0.22	52.01±0.44	26.01±0.23	68.34±0.26
w/o TS	53.36±0.34	96.12±0.25	51.67±0.48	25.78±0.25	67.98±0.29
FedTCR (Full)	**54.27±0.25**	**97.84±0.18**	**53.64±0.36**	**26.89±0.18**	**69.45±0.22**

As shown in Table 2, removing any component consistently degrades performance, demonstrating their complementary contributions. *w/o CR* incurs the largest drop (e.g., -2.28% ACC on Movies and -4.70% R@1 on Toys), highlighting the importance of cross-modal prototype routing for exchanging task-relevant knowledge beyond parameter aggregation. The degradation of *w/o TS* further verifies that task-agnostic pre-training is crucial for learning transferable representations before task-specific adaptation. In addition, *w/o TW* consistently underperforms the full model, validating the benefit of topology-aware prototype construction, while the larger impact of *w/o* $\mathcal{L}_g$ compared with *w/o* $\mathcal{L}_m$ suggests that neighborhood-level supervision provides stronger guidance than node-level alignment.

5.5 Robustness Analysis (Answer for **Q4**)

To answer **Q4**, we evaluate FedTCR under three data sparsity scenarios controlled by ratio α: *feature sparsity* (feature masking), *topology sparsity* (edge removal), and *label sparsity* (label removal), in both task settings. As shown in Fig. 3, FedTCR remains robust across all sparsity types and task settings: at $\alpha = 50\%$ in the homogeneous setting it preserves 93.2%, 94.1%, and 97.1% of its original ACC under feature, topology, and label sparsity, respectively. The degradation pattern is setting-dependent: feature sparsity hurts most under homogeneous clients, where corrupted modality representations propagate directly into prototype construction, whereas topology sparsity dominates in the heterogeneous-task federation, where divergent objectives make topology-aware routing the main channel of cross-client support; label sparsity stays the mildest in both settings, since task-agnostic pre-training requires no downstream supervision.



(a) Homogeneous (5 clients) (b) Heterogeneous (20 clients)

Figure 3: Retained ACC on Movies under three sparsity types at varying ratios, relative to the unperturbed setting. Bands: one std over five seeds.

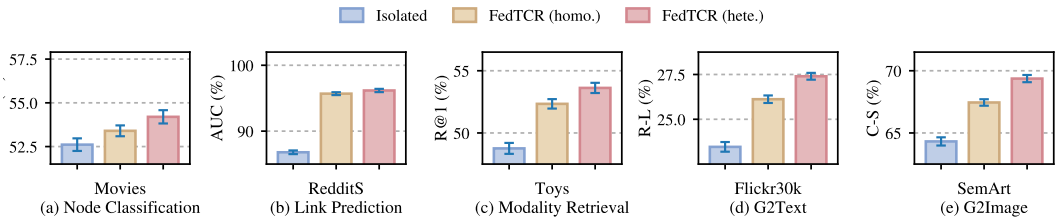


Figure 2: Performance under the task-heterogeneous setting (4 clients for each dataset, i.e., one MAG per task). FedTCR (homo.) federates only the 4 same-task clients, while FedTCR (hete.) pre-trains all 20 clients across the 5 tasks.

5.6 Hyperparameter Sensitivity (Answer for **Q5**)

To answer **Q5**, we sweep the loss weights λ_m, λ_g, λ_c (summing to 1 by default), the temperature τ, and the propagation depth L. As shown in Table 3, FedTCR performs best at the default configuration and remains stable around it; the model is primarily sensitive to λ_c and τ: an overly small λ_c weakens the cross-client alignment signal, while an overly large τ makes the contrastive distribution overly uniform. In contrast, varying λ_m, λ_g, or L changes the accuracy only marginally, indicating that the defaults already lie in a flat optimal region and that FedTCR requires little hyperparameter tuning in practice.

Table 3: Hyperparameter sensitivity on Movies. Each row sweeps one hyperparameter; the others are fixed at their defaults.

Hyperparameter		Node Classification Accuracy (%)				
λ_m	$\{0.1, 0.2, \mathbf{0.3}, 0.4, 0.5\}$	53.96	54.12	**54.27**	54.18	54.02
λ_g	$\{0.05, 0.1, \mathbf{0.2}, 0.3, 0.5\}$	53.88	54.09	**54.27**	54.16	53.97
λ_c	$\{0.1, 0.3, \mathbf{0.5}, 0.7, 1.0\}$	52.41	53.72	**54.27**	54.05	53.83
τ	$\{0.1, 0.3, \mathbf{0.5}, 0.7, 1.0\}$	54.03	54.19	**54.27**	53.86	52.58
L	$\{1, 2, \mathbf{3}, 4, 5\}$	54.01	54.17	**54.27**	54.21	54.08

5.7 Privacy-Utility Tradeoff (Answer for **Q6**)

To answer **Q6**, we evaluate FedTCR with (ε, δ)-differentially private prototype uploading (Eq. (6)), fixing $\delta = 10^{-5}$ and $C = 1.0$ while varying ε. As shown in Table 4, performance is nearly lossless at $\varepsilon \geq 8$ (over 99.4% retained), since prototypes are importance-weighted aggregates over many nodes, which limits per-client sensitivity, and the server-side aggregation further averages the perturbations out across clients. Even at $\varepsilon = 1$, FedTCR retains over 97.4% of its non-private performance, staying clearly ahead of the strongest non-private baseline on RedditS and Toys and within 1.1% of MMGCN on Movies.

Table 4: Performance under (ε, δ)-DP prototype perturbation ($\delta = 10^{-5}$, $C = 1.0$); mean±std over five seeds.

Privacy budget ε	Movies (ACC)	RedditS (AUC)	Toys (R@1)
∞ (w/o DP)	**54.27±0.25**	**97.84±0.18**	**53.64±0.36**
16	54.14±0.28	97.76±0.20	53.56±0.39
8	53.98±0.30	97.61±0.22	53.42±0.42
4	53.73±0.34	97.34±0.26	53.14±0.46
2	53.31±0.37	96.93±0.29	52.78±0.50
1	52.89±0.43	96.34±0.35	52.25±0.55

6 Conclusion

We identify the *multifaceted heterogeneity* inherent in federated multimodal graph learning (FMGL), spanning task, modality, and topology heterogeneity across decentralized MAGs. We propose FedTCR, the first systematic algorithm for FMGL, which employs a two-stage paradigm (federated task-agnostic pre-training followed by isolated task-oriented fine-tuning) and a topology-aware cross-modal routing mechanism that distills modality-specific knowledge into structure-informed prototypes for cross-client contrastive alignment. Experiments across 8 datasets show that FedTCR consistently achieves state-of-the-art performance on graph-centric and modality-centric tasks, laying a foundation for privacy-preserving multimodal graph learning.

AI Usage Disclosure

Generative AI tools were used during the preparation of this manuscript for limited purposes, including language refinement, literature retrieval and discovery, and research ideation support. Specifically, AI tools assisted with improving writing clarity, organizing related work, exploring potential research directions, and discussing possible methodological designs and experimental considerations.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Javad Aliakbari, Johan Östman, Ashkan Panahi, et al. Subgraph federated learning via spectral methods. *arXiv preprint arXiv:2510.25657*, 2025.

Jinheon Baek, Wonyong Jeong, Jiongdao Jin, Jaehong Yoon, and Sung Ju Hwang. Personalized subgraph federated learning. In *International conference on machine learning*, pp. 1396–1415. PMLR, 2023.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008 (10):P10008, 2008.

Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*, 2013.

Liwei Che, Jiaqi Wang, Xinyue Liu, and Fenglong Ma. Leveraging foundation models for multimodal federated learning with incomplete modality. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 401–417. Springer, 2024.

Chuan Chen, Weibo Hu, Ziyue Xu, and Zibin Zheng. Fedgl: Federated graph learning framework with global self-supervision. *arXiv preprint arXiv:2105.03170*, 2021.

Karan Desai, Gaurav Kaul, Zubin Trivadi Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people, for the people. In *Advances in Neural Information Processing Systems, NeurIPS, Datasets and Benchmarks Track, NeurIPS DB Track*, 2021.

Cynthia Dwork. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pp. 1–19. Springer, 2008.

Xingbo Fu, Binchi Zhang, Yushun Dong, Chen Chen, and Jundong Li. Federated graph machine learning: A survey of concepts, techniques, and applications. *ACM SIGKDD Explorations Newsletter*, 24(2):32–47, 2022.

Liang Gao, Huazhu Fu, Li Li, Yingwen Chen, Ming Xu, and Cheng-Zhong Xu. Feddc: Federated learning with non-iid data via local drift decoupling and correction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10112–10121, 2022.

Noa Garcia and George Vogiatzis. How to read paintings: Semantic art understanding with multimodal retrieval. In *ECCV Workshops (2)*, volume 11130 of *Lecture Notes in Computer Science*, pp. 676–691. Springer, 2018.

Zhaochen Guo, Zhixiang Shen, Xuanting Xie, Liangjian Wen, and Zhao Kang. Disentangling homophily and heterophily in multimodal graph clustering. In *Proceedings of the ACM International Conference on Multimedia, MM*, 2025.

Zhiqiang Guo, Jianjun Li, Guohui Li, Chaoyang Wang, Si Shi, and Bin Ruan. Lgmrec: Local and global graph learning for multimodal recommendation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 8454–8462, 2024.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in Neural Information Processing Systems, NeurIPS*, 2017.

Chaoyang He, Keshav Balasubramanian, Emir Ceyani, Carl Yang, Han Xie, Lichao Sun, Lifang He, Liwei Yang, Philip S Yu, Yu Rong, et al. Fedgraphnn: A federated learning benchmark system for graph neural networks. In *ICLR 2021 Workshop on Distributed and Private Machine Learning*, 2021.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*, 2024.

Wenke Huang, Guancheng Wan, Mang Ye, and Bo Du. Federated graph semantic and structural learning, 2024. URL https://arxiv.org/abs/2406.18937.

Bowen Jin, Ziqi Pang, Bingjun Guo, Yu-Xiong Wang, Jiaxuan You, and Jiawei Han. Instructg2i: Synthesizing images from multimodal attributed graphs. *Advances in Neural Information Processing Systems*, 37:117614–117635, 2024.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pp. 5132–5143. PMLR, 2020.

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

Bingjun Li and Sheida Nabavi. A multimodal graph neural network framework for cancer molecular subtype classification. *BMC bioinformatics*, 25(1):27, 2024.

Xunkai Li, Zhengyu Wu, Wentao Zhang, Yinlin Zhu, Rong-Hua Li, and Guoren Wang. Fedgta: Topology-aware averaging for federated graph learning. *Proceedings of the VLDB Endowment*, 2023.

Xunkai Li, Zhengyu Zeng, Di Wu, Yujie Li, Boyu Fang, Mingkai Zhou, Rong-Hua Zhang, and Qing Li. Adafgl: A new paradigm for federated node classification with topology heterogeneity. *arXiv preprint arXiv:2404.09561*, 2024a.

Xunkai Li, Yinlin Zhu, Boyang Pang, Guochen Yan, Yeyu Yan, Zening Li, Zhengyu Wu, Wentao Zhang, Rong-Hua Li, and Guoren Wang. Openfgl: A comprehensive benchmark for federated graph learning. *arXiv preprint arXiv:2408.16288*, 2024b.

Xunkai Li, Yuming Ai, Yinlin Zhu, Haodong Lu, Yi Zhang, Guohao Fu, Bowen Fan, Qiangqiang Dai, Rong-Hua Li, and Guoren Wang. Mm-openfgl: A comprehensive benchmark for multimodal federated graph learning. *arXiv preprint arXiv:2601.22416*, 2026.

László Lovász. Random walks on graphs. *Combinatorics, Paul erdos is eighty*, 2(1-46):4, 1993.

Sijie Mai, Songlong Xing, Jiaxuan He, Ying Zeng, and Haifeng Hu. Multimodal graph for unaligned multimodal sequence analysis via graph convolution and graph pooling. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2):1–24, 2023.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. Pmlr, 2017.

Duong M Nguyen, Trong Nghia Hoang, Thanh Trung Huynh, Quoc Viet Hung Nguyen, and Phi Le Nguyen. Learning reconfigurable representations for multimodal federated learning with missing data. *arXiv preprint arXiv:2510.22880*, 2025.

Manh Duong Nguyen, Trung Thanh Nguyen, Huy Hieu Pham, Trong Nghia Hoang, Phi Le Nguyen, and Thanh Trung Huynh. Fedmac: Tackling partial-modality missing in federated learning with cross-modal aggregation and contrastive regularization. In *2024 22nd International Symposium on Network Computing and Applications (NCA)*, pp. 278–285. IEEE, 2024.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 188–197, 2019.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford infolab, 1999.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision, ICCV*, 2015.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

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 and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Zihan Tan, Guancheng Wan, Wenke Huang, and Mang Ye. Fedssp: Federated graph learning with spectral knowledge and personalized preference. *Advances in Neural Information Processing Systems*, 37:34561–34581, 2024.

Zihan Tan, Suyuan Huang, Guancheng Wan, Wenke Huang, He Li, and Mang Ye. S2fgl: Spatial spectral federated graph learning. *arXiv preprint arXiv:2507.02409*, 2025a.

Zihan Tan, Guancheng Wan, Wenke Huang, He Li, Guibin Zhang, Carl Yang, and Mang Ye. Fedspa : Generalizable federated graph learning under homophily heterogeneity. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15464–15475, 2025b. URL https://api.semanticscholar.org/CorpusID:277057037.

Zhulin Tao, Yinwei Wei, Xiang Wang, Xiangnan He, Xianglin Huang, and Tat-Seng Chua. Mgat: Multimodal graph attention network for recommendation. *Information Processing & Management*, 57(5):102277, 2020.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations, ICLR*, 2018.

Chenxi Wan, Xunkai Li, Yilong Zuo, Haokun Deng, Sihan Li, Bowen Fan, Hongchao Qin, Ronghua Li, and Guoren Wang. Openmag: A comprehensive benchmark for multimodal-attributed graph. *arXiv preprint arXiv:2602.05576*, 2026.

Guancheng Wan, Wenke Huang, and Mang Ye. Federated graph learning under domain shift with generalizable prototypes. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 15429–15437, 2024.

Zhen Wang, Weirui Kuang, Yuexiang Xie, Liuyi Yao, Yaliang Li, Bolin Ding, and Jingren Zhou. Federatedscope-gnn: Towards a unified, comprehensive and efficient package for federated graph learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4110–4120, 2022.

Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM international conference on multimedia*, pp. 1437–1445, 2019.

Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision. *arXiv preprint arXiv:2006.03677*, 2020a.

Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Weinberger. Simplifying graph convolutional networks. In *International Conference on Machine Learning, ICML*, 2019.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24, 2020b.

Han Xie, Jing Ma, Li Xiong, and Carl Yang. Federated graph classification over non-iid graphs. *Advances in neural information processing systems*, 34:18839–18852, 2021.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

Xiao-Yang Liu Yanglet, Yupeng Cao, and Li Deng. Multimodal financial foundation models (mffms): Progress, prospects, and challenges. *arXiv preprint arXiv:2506.01973*, 2025.

Yuhang Yao, Weizhao Jin, Srivatsan Ravi, and Carlee Joe-Wong. Fedgcn: Convergence-communication tradeoffs in federated training of graph convolutional networks. In *Advances in Neural Information Processing Systems*, 2023.

Minji Yoon, Jing Yu Koh, Bryan Hooi, and Ruslan Salakhutdinov. Multimodal graph learning for generative tasks. *arXiv preprint arXiv:2310.07478*, 2023.

Wentao Yu, Shuo Chen, Yongxin Tong, Tianlong Gu, and Chen Gong. Modeling inter-intra heterogeneity for graph federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22236–22244, 2025.

Jiaqi Zhang, Yu Cheng, Yongxin Ni, Yunzhu Pan, Zheng Yuan, Junchen Fu, Youhua Li, Jie Wang, and Fajie Yuan. Ninerec: A benchmark dataset suite for evaluating transferable recommendation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024a.

Ke Zhang, Carl Yang, Xiaoxiao Li, Lichao Sun, and Siu Ming Yiu. Subgraph federated learning with missing neighbor generation. *Advances in neural information processing systems*, 34:6671–6682, 2021.

Ke Zhang, Yuqi Zhu, Xin Xie, Carl Yang, Siu Ming Yiu, and Jiawei Han. Deep efficient private neighbor generation for subgraph federated learning. In *Proceedings of the 2024 SIAM International Conference on Data Mining*, 2024b.

Taolin Zhang, Chengyuan Mai, Yaomin Chang, Chuan Chen, Lin Shu, and Zibin Zheng. Privacy-preserving personalized federated graph learning with ego-graphs. *ACM Transactions on Knowledge Discovery from Data*, 18(2):1–27, 2023.

Haoran Zheng, Renchi Yang, Hongtao Wang, and Jianliang Xu. Cross-contrastive clustering for multimodal attributed graphs with dual graph filtering. *arXiv preprint arXiv:2511.20030*, 2025.

Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81, 2020.

Yinlin Zhu, Xunkai Li, Zhengyu Wu, Di Wu, Miao Hu, and Rong-Hua Li. Fedtad: Topology-aware data-free knowledge distillation for subgraph federated learning. *arXiv preprint arXiv:2404.14061*, 2024.

Yinlin Zhu, Xunkai Li, Jishuo Jia, Miao Hu, Di Wu, and Meikang Qiu. Towards effective federated graph foundation model via mitigating knowledge entanglement, 2025. URL https://arxiv.org/abs/2505.12684.

APPENDIX: TABLE OF CONTENTS

A More Related Works 15

B Additional Experimental Details 15

B.1 Training Protocol 15

B.2 Model Configuration 15

C Efficiency Analysis 16

D Dataset Details 16

E Baseline Details 17

E.1 Isolated Training 17

E.2 Classical FL Methods 17

E.3 Multimodal FL Methods 18

E.4 Unimodal FGL Methods 18

F Theoretical Proofs 19

G Metric Details 20

H Limitations 20

I Environment 20

J Pseudocode 21

A More Related Works

Graph Neural Networks (GNNs). Earlier research on deep graph learning extends convolution to handle graphs Bruna et al. (2013) but comes with notable parameter counts. To this end, GCN Kipf & Welling (2017) simplifies graph convolution by utilizing a 1-order Chebyshev filter to capture local neighborhood information. Moreover, GAT Veličković et al. (2018) adopts graph attention, allowing weighted aggregation. GraphSAGE Hamilton et al. (2017) introduces a variety of learnable aggregation functions for performing message aggregation. Moreover, GIN Xu et al. (2019) aims to preserve structural information maximally and theoretically proves its discriminative power matches the Weisfeiler-Lehman graph isomorphism test. Further details on GNN research can be found in surveys Wu et al. (2020b); Zhou et al. (2020).

Multimodal Federated Learning (MFL). MFL studies federated optimization over multimodal data. PEPSY Nguyen et al. (2025) handles both missing modalities and missing input features by learning client-side data-missing profiles for robust aggregation. FedMVP Che et al. (2024) leverages frozen pre-trained foundation models for cross-modal completion and representation knowledge transfer, with CKA-based importance-aware aggregation. FedMAC Nguyen et al. (2024) aligns heterogeneous modality representations via cross-modal contrastive learning with modality-specific adaptive aggregation. These methods address modality missingness at the data level, but they are not designed for graph-structured multimodal data and assume a uniform downstream task across clients.

B Additional Experimental Details

B.1 Training Protocol

Unless otherwise specified, all reported results are presented as mean$\pm$standard deviation over five random seeds. For each seed, all methods share the same data split and, for robustness experiments, the same realization of sparsity patterns. Baseline implementations use the official released code with the recommended hyperparameter settings, while ablated variants inherit the same configuration with only the corresponding component removed. Federated pre-training runs for $T = 200$ communication rounds with full client participation and $E = 2$ local epochs per round; task-oriented fine-tuning runs for 100 epochs per client.

Perturbation model. For the robustness study (Sec. 5.5), three types of sparsity are injected before training with ratio α, and for a given seed every method receives the identical realization. *Feature sparsity* randomly masks $\alpha\%$ of the feature dimensions of every node in the target modality, setting them to zero. *Topology sparsity* removes $\alpha\%$ of the existing edges uniformly at random. *Label sparsity* removes $\alpha\%$ of the available node labels for fine-tuning, sampled uniformly per client.

Task definitions. We consider two graph-centric tasks: (1) *Node Classification*: assigning category labels to unlabeled nodes; (2) *Link Prediction*: estimating whether an edge (u, v) should exist in $\mathcal{E}^k$; and two modality-centric tasks: (1) *Modality Retrieval*: given a query from one modality (e.g., text), retrieving its corresponding representation in another modality; (2) *Modality Generation*: producing content in a target modality conditioned on a target node along with task instructions and its graph neighborhood, covering Graph-to-Text (G2Text) generation of textual descriptions and Graph-to-Image (G2Image) synthesis of visual content.

B.2 Model Configuration

Each node is associated with text and image features extracted from frozen pre-trained encoders (Sentence-BERT Reimers & Gurevych (2019) or T5 Raffel et al. (2020) for text, ViT Wu et al. (2020a) or DINOv2 Oquab et al. (2023) for images; Qwen2-VL-7B-Instruct Bai et al. (2023) is used for the G2Text/G2Image feature extraction), producing 768-dimensional representations for all datasets. Each modality is encoded by a linear projection (Eq. (1)) followed by low-pass graph filtering (Eq. (2)).

Unless otherwise stated, the default hyperparameters are shared across all datasets: the latent dimension is $d = 128$; the propagation depth is $L = 3$; the smoothing strength is $\alpha = 2.0$ in Eq. (2);

the PageRank damping factor is 0.85 with 20 power iterations; the contrastive temperature is $\tau = 0.5$; the loss weights are $\lambda_m = 0.3$, $\lambda_g = 0.2$, and $\lambda_c = 0.5$, which are normalized to sum to 1; and the differential-privacy configuration is $\delta = 10^{-5}$ with clip bound $C = 1.0$ ($\varepsilon = \infty$ unless the privacy budget is explicitly swept in Sec. 5.7). For the neighbor-level contrastive loss, we simulate 5 random walks of length 10 per node and sample an equal number of negatives. A sensitivity analysis of the main hyperparameters is reported in Sec. 5.6.

Optimization. All models are optimized using Adam with learning rate 10^{-3} and weight decay 10^{-5} under full-batch training. Encoder parameters are aggregated at the server by node-count-weighted averaging, and no server-side model is maintained for the routing outputs.

C Efficiency Analysis

We analyze the computational and communication complexity of FedTCR per federated round. On the client side, multimodal graph encoding costs $\mathcal{O}(|\mathcal{M}| \cdot |\mathcal{V}^k| d_m d)$ for the modality projections and $\mathcal{O}(L \cdot |\mathcal{E}^k| d)$ for the low-pass graph filtering with sparse adjacency multiplications. The topology-aware weights are computed via PageRank power iteration at $\mathcal{O}(|\mathcal{E}^k|)$ per iteration, and only once as a preprocessing step before training. Prototype construction costs $\mathcal{O}(|\bar{\mathcal{M}}| \cdot |\mathcal{V}^k| d)$, and the three contrastive losses are computed over per-client sampled node pairs, sharing the same order as standard cross-contrastive multimodal training. On the server side, parameter aggregation costs $\mathcal{O}(|\mathcal{S}| \cdot |\bar{\mathcal{M}}| d_m d)$, while cross-modal routing requires pairwise cosine similarities among $|\mathcal{S}| \times |\bar{\mathcal{M}}|$ prototypes, i.e., $\mathcal{O}(|\mathcal{S}|^2 |\bar{\mathcal{M}}|^2 d)$, which is negligible compared with client-side training since each prototype is a single d-dimensional vector. In terms of communication, each client uploads the shared encoder parameters $\{\mathbf{\Phi}^{(m)}\}_{m \in \mathcal{M}}$, identical to standard FL methods such as FedAvg, plus $|\bar{\mathcal{M}}|$ prototype vectors, a marginal overhead of $|\bar{\mathcal{M}}| \cdot d$ values per round. The isolated fine-tuning stage involves no further communication. Overall, FedTCR introduces only lightweight overhead over classical FL methods while enabling cross-task and cross-modal collaboration.

D Dataset Details

Detailed statistical information on the datasets is presented in Table 5, where the graph statistics follow the OpenMAG construction Wan et al. (2026), with textual descriptions as follows.

Table 5: Statistics of the experimental datasets.

Datasets	# Modalities	# Nodes	# Edges	# Classes	Domain
Movies	Text, Image	16,672	218,390	20	Movie Network
Grocery	Text, Image	17,074	171,340	20	Recommendation
RedditS	Text, Image	15,894	566,160	20	Social Network
BiliDance	Text, Image	2,307	9,127	-	Entertainment
Toys	Text, Image	20,695	126,886	18	Recommendation
ele-fashion	Text, Image	97,766	199,602	12	Co-purchase Network
Flickr30k	Text, Image	31,783	181,151	-	Image Network
SemArt	Text, Image	21,382	1,216,432	-	Artwork

Movies Ni et al. (2019) is sourced from Amazon's Movies and TV category. Nodes correspond to DVD/Blu-ray products, and edges reflect consumer co-purchasing behavior. Node attributes include textual plot synopses and customer reviews, alongside visual features derived from official cover art.

Grocery Ni et al. (2019); Hou et al. (2024) is sourced from Amazon's Grocery and Gourmet Food category. Nodes correspond to food and household products, and edges reflect co-purchasing behavior. Node attributes include textual product descriptions and visual features from product images. This dataset is used for node classification, where products are categorized into fine-grained grocery categories.

RedditS Desai et al. (2021) is a social network derived from Reddit, where nodes correspond to posts linked by user interaction and community relations. Each node carries textual post content and

associated visual context. This dataset is utilized for link prediction, forecasting potential interaction links between posts.

BiliDance Zhang et al. (2024a) is a video entertainment network collected from the Bilibili platform, where nodes correspond to dance-video content items linked by user co-viewing behavior. Nodes carry textual descriptions and visual cover features. This dataset is utilized for link prediction over content-association structures.

Toys Ni et al. (2019); Hou et al. (2024) is sourced from Amazon's Toys and Games category. Nodes correspond to toy products, and edges reflect co-purchasing behavior. Node attributes include textual product descriptions and visual features from product images. This dataset is utilized for modality retrieval between textual and visual product content.

ele-fashion Ni et al. (2019); Hou et al. (2024) is a heterogeneous graph merging Amazon's Electronics and Fashion categories. Nodes are connected via cross-category co-purchasing links, revealing latent consumer preferences across disparate domains. Features combine technical specs with style descriptions and product imagery. This dataset is utilized for modality retrieval between textual and visual product content.

Flickr30k Plummer et al. (2015) is a canonical image-text reasoning dataset. In the OpenMAG setting Wan et al. (2026), we construct a graph where nodes represent image regions and caption phrases, linked by semantic grounding annotations. This dataset is utilized for Graph-to-Text (G2Text) generation, evaluating the model's ability to generate descriptive captions by traversing grounded visual-textual relationships.

SemArt Garcia & Vogiatzis (2018) is an artwork understanding dataset of fine-art paintings, where nodes correspond to artworks linked by artistic, historical, and semantic relations. Nodes carry visual painting features and textual descriptions of their content and context. This dataset is utilized for Graph-to-Image (G2Image) synthesis, producing visual content conditioned on multimodal graph context.

E Baseline Details

E.1 Isolated Training

MMGCN Wei et al. (2019) introduces a multimodal framework for micro-video recommendation by modeling user preferences across visual, acoustic, and textual channels. It builds separate modality-specific bipartite graphs, captures high-order interactions within each, and fuses them via a structured integration layer, effectively reflecting user-item dynamics in each sensory modality.

MGAT Tao et al. (2020) applies gated attention over parallel multimodal interaction graphs for personalized recommendation. By adaptively weighting different modalities, it disentangles fine-grained user interests and filters out noisy or conflicting signals, enhancing preference modeling robustness.

LGMRec Guo et al. (2024) is a multimodal recommender that models both local and global user interests through graph learning. It separates collaborative and multimodal signals in user embeddings and addresses sparsity in local interest modeling. A local graph embedding module learns collaborative and modality-specific embeddings independently, while a global hypergraph module captures overall dependencies among users and items. Combining these decoupled local and global embeddings improves recommendation accuracy and robustness.

E.2 Classical FL Methods

FedAvg. McMahan et al. (2017) serves as a foundational method in FL, enabling decentralized model training across diverse devices while preserving data privacy. Initiated by a central server that distributes a global model, clients independently execute local updates through stochastic gradient

descent. Subsequently, these updates are aggregated by the server via averaging to refine the global model, with the cycle repeating until convergence.

SCAFFOLD. Karimireddy et al. (2020) employs control variates to mitigate client-drift in FL. Demonstrating significant reductions in communication rounds, Scaffold is resilient to data heterogeneity and client sampling.

FedDC. Gao et al. (2022) is a novel FL algorithm that corrects local drift through lightweight modifications. Each client tracks the deviation between local and global model parameters using an auxiliary variable, enhancing parameter-level consistency.

E.3 Multimodal FL Methods

PEPSY Nguyen et al. (2025) addresses multimodal federated learning with both missing modalities and missing input features by learning client-side data-missing profiles that encode local missing patterns as embedding controls. These profiles are probabilistically aligned and aggregated on the server to reconfigure shared representations toward each client's incomplete data view, enabling robust aggregation and stable performance under severe data incompleteness.

FedMVP Che et al. (2024) addresses modality missing in multimodal federated learning by leveraging frozen pre-trained foundation models for cross-modal completion and representation knowledge transfer. It trains a lightweight joint encoder via multimodal contrastive objectives and performs CKA-based importance-aware aggregation on the server, achieving robust performance under severe modality incompleteness.

FedMAC Nguyen et al. (2024) addresses cross-modal heterogeneity in multimodal federated learning by introducing a modality-aware collaboration framework that aligns heterogeneous modality representations via cross-modal contrastive learning. It further employs modality-specific aggregation with adaptive weighting to mitigate negative transfer across clients with inconsistent modality availability.

E.4 Unimodal FGL Methods

FedSage+ Zhang et al. (2021) extends FedSage to the subgraph federated learning setting by explicitly addressing missing cross-client neighbors. It jointly trains a GraphSAGE classifier with a local missing-neighbor generator that synthesizes potential cross-subgraph neighbors, enabling more complete neighborhood aggregation under federation and improving global generalization without sharing raw graph data.

FedGTA Li et al. (2023) innovatively merges large-scale graph learning with federated graph learning. Clients encode topology and node attributes, compute local smoothing confidence and mixed moments of neighbor features, and then upload these to the server. The server uses this data to perform personalized model aggregation, utilizing local smoothing confidence as weights for effective integration.

FedSPA Tan et al. (2025b) addresses homophily heterogeneity in federated graph learning by explicitly modeling both homophily conflict and homophily bias across clients. It introduces Subgraph Feature Propagation Decoupling (SFPD) to separate homophilic and heterophilic message passing, enabling collaboration under unified homophily levels, and proposes Homophily Bias-Driven Aggregation (HBDA) to adaptively weight client contributions based on spectral and parameter-sensitivity cues, thereby improving global generalization.

FedIIH Yu et al. (2025) addresses heterogeneity in federated graph learning by jointly modeling inter-client and intra-client heterogeneity. It infers subgraph distribution similarities via a hierarchical variational framework from a global perspective, while disentangling local subgraphs into multiple latent factors to enable factor-wise personalized federation, leading to robust collaboration across both homophilic and heterophilic graphs.

FedSSP Tan et al. (2024) proposes a personalized federated graph learning framework that addresses cross-domain structural heterogeneity from a spectral perspective. It shares generic spectral knowledge across clients to mitigate knowledge conflict induced by domain shifts, while retaining client-specific components to preserve personalization.

S2FGL Tan et al. (2025a) tackles subgraph federated graph learning by jointly addressing spatial label-signal disruption and spectral client drift. It reinforces missing label semantics via prototype-based semantic sharing and aligns graph-frequency components across clients to improve robustness and generalization under heterogeneous subgraph distributions.

FedLap Aliakbari et al. (2025) leverages global graph structure in subgraph federated learning via Laplacian smoothing, introducing a structural regularization term that implicitly enforces representation consistency among neighboring nodes without explicit message passing or feature sharing. By operating in the spectral domain, FedLap captures inter-subgraph dependencies while achieving strong privacy guarantees and low communication overhead.

F Theoretical Proofs

Theorem 1 (Unified view of propagation as graph filtering)**.** *Let $\tilde{\mathbf{A}} \in \mathbb{R}^{n\times n}$ be the symmetrically normalized adjacency matrix of an undirected graph and $\mathbf{Z} \in \mathbb{R}^{n\times d}$ be node representations. Consider the general depth-L polynomial graph filter, which can be formulated as follows:*

$$\mathbf{H} = \Big(\sum_{l=0}^{L} \gamma_l \tilde{\mathbf{A}}^l\Big)\mathbf{Z}, \qquad \gamma_l \geq 0,\ \sum_{l=0}^{L} \gamma_l = 1. \tag{12}$$

Then: (i) SGC and the linearized L-layer GCN correspond to the one-hot choice $\gamma_L = 1$; (ii) APPNP-style personalized-PageRank propagation corresponds to the geometric choice $\gamma_l = (1-\beta)\,\beta^l$ with $\beta \in (0,1)$ and $L \to \infty$; (iii) the FedTCR filter of Eq. (2) is exactly the geometric choice with $\beta = \alpha/(\alpha+1)$, i.e., the finite-$L$ truncation of (ii), whose approximation error decays exponentially:

$$\big\|\mathbf{H}_L - \mathbf{H}_\infty\big\|_F \ \leq\ \beta^{L+1}\,\|\mathbf{Z}\|_F. \tag{13}$$

Proof. Since $\tilde{\mathbf{A}}$ is symmetric with spectrum in $[-1,1]$, we have $\|\tilde{\mathbf{A}}\|_2 \leq 1$, so for any $\beta \in (0,1)$ the Neumann series $\sum_{l=0}^{\infty}(\beta\tilde{\mathbf{A}})^l$ converges to $(\mathbf{I}-\beta\tilde{\mathbf{A}})^{-1}$.

(i) With $\gamma_L = 1$ and all other coefficients zero, Eq. (12) reduces to $\mathbf{H} = \tilde{\mathbf{A}}^L\mathbf{Z}$, which is precisely the propagation operator of SGC; an L-layer GCN without intermediate nonlinearities and with shared collapsed weights realizes the same operator as a special case of the general filter.

(ii) With $\gamma_l = (1-\beta)\beta^l$ and $L \to \infty$, the coefficient sequence sums to one, and $\sum_{l=0}^{\infty}(1-\beta)(\beta\tilde{\mathbf{A}})^l = (1-\beta)(\mathbf{I}-\beta\tilde{\mathbf{A}})^{-1}$, which is the personalized-PageRank propagation matrix with teleport probability $1-\beta$, i.e., the exact solution that APPNP approximates with a finite number of power-iteration steps as its propagation matrix.

(iii) Setting $\beta = \alpha/(\alpha+1)$ gives $1-\beta = 1/(\alpha+1)$, so the coefficients of Eq. (2) satisfy $\gamma_l = \frac{1}{\alpha+1}\big(\frac{\alpha}{\alpha+1}\big)^l = (1-\beta)\beta^l$, i.e., FedTCR's filter is the finite-L truncation of (ii). For the error, writing $\mathbf{H}_\infty = (1-\beta)\sum_{l=0}^{\infty}(\beta\tilde{\mathbf{A}})^l\mathbf{Z}$,

$$\begin{aligned}\|\mathbf{H}_L - \mathbf{H}_\infty\|_F &\leq (1-\beta)\sum_{l=L+1}^{\infty}\beta^l\|\tilde{\mathbf{A}}\|_2^l\,\|\mathbf{Z}\|_F \\ &\leq (1-\beta)\,\frac{\beta^{L+1}}{1-\beta}\,\|\mathbf{Z}\|_F = \beta^{L+1}\|\mathbf{Z}\|_F,\end{aligned}$$

where the first inequality applies submultiplicativity of the Frobenius norm and the second uses $\|\tilde{\mathbf{A}}\|_2 \leq 1$. □

G Metric Details

All metrics are computed on each client's held-out split after isolated task-oriented fine-tuning, and are reported as mean±standard deviation over five random seeds (Appendix B.1). Task definitions follow Appendix B.

Node Classification. We report classification accuracy (ACC):

$$\text{ACC} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{y}_i = y_i), \tag{14}$$

where N is the number of test nodes, y_i the ground-truth label, $\hat{y}_i$ the prediction, and $\mathbb{I}(\cdot)$ the indicator function.

Link Prediction. We report the Area Under the ROC Curve (AUC), which measures the probability that a randomly chosen positive (existent) edge receives a higher score than a randomly chosen negative (non-existent) node pair:

$$\text{AUC} = P\big(s(u,v) > s(u',v')\big), \tag{15}$$

where $s(\cdot,\cdot)$ is the model's similarity score, (u,v) is a positive edge, and (u',v') is a sampled negative pair.

Modality Retrieval. We report Recall@1 (R@1), the fraction of queries whose ground-truth cross-modal target is ranked first:

$$\text{R@1} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\mathbb{I}(\text{rank}_i = 1), \tag{16}$$

where Q is the query set and rank_i is the rank of the correct target for query i.

Graph-to-Text (G2Text). Following MMGL Yoon et al. (2023), we report ROUGE-L (R-L), which measures sentence-level recall based on the longest common subsequence between the generated description and the reference:

$$\text{ROUGE-L} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}}. \tag{17}$$

Graph-to-Image (G2Image). Following InstructG2I Jin et al. (2024), we report CLIP-Score (C-S), which quantifies cross-modal semantic consistency between the generated image and its conditioning description via pre-trained CLIP encoders (E_T for text and E_I for images):

$$\text{CLIP-Score}(I,T) = \max\left(100\cdot\cos(E_I(I),E_T(T)),0\right). \tag{18}$$

H Limitations

While FedTCR demonstrates strong empirical performance across diverse FMGL scenarios, several limitations warrant discussion. First, each client distills its modality-specific knowledge into a single prototype vector per modality, which may not fully capture the internal multi-community structure of large-scale MAGs with diverse semantic clusters. Extending to multiple prototypes per modality (e.g., cluster-aware prototypes) could improve expressiveness at the cost of increased communication overhead. Second, our evaluation focuses on MAGs with text and image modalities; the generalizability to other modality types (e.g., audio, video, tabular data) remains to be validated. Third, the two-stage paradigm assumes that the pre-training phase has sufficiently converged before fine-tuning begins, and the potential gap between the task-agnostic objective and specific downstream tasks may limit the adaptability for highly specialized applications. Exploring progressive or continual pre-training strategies could mitigate this issue.

I Environment

All experiments are conducted on a workstation equipped with Intel Xeon Scalable processors and NVIDIA RTX 6000 Ada Generation GPUs with 96 GB of VRAM, supported by 256 GB of system

RAM. The computational environment utilizes CUDA 12.9, while software implementations are developed using Python 3.10.18 and PyTorch 2.8.

J PSEUDOCODE

The complete procedure of FedTCR is presented in Algorithm 1, which consists of two stages: federated task-agnostic pre-training and isolated task-oriented fine-tuning.

Algorithm 1: FedTCR: FMGL with Topology-aware Cross-modal Routing

Input: K clients with MAGs $\{\mathcal{G}^k\}_{k=1}^K$; modality set $\mathcal{M}$; extended modality set $\bar{\mathcal{M}} = \mathcal{M} \cup \{|\mathcal{M}| + 1\}$; communication rounds T; local epochs E; smoothing strength α; propagation depth L; temperature τ; loss weights $\lambda_m, \lambda_g, \lambda_c$

Output: Task-specific predictions for each client

/* Stage 1: Federated Task-agnostic Pre-training */

Initialize modality-specific encoders $\{\mathbf{\Phi}^{(m)}\}_{m\in\mathcal{M}}$

for *each client* $k \in [K]$ ***in parallel*** **do**

  Compute PageRank importance weights $\{\omega_i^k\}_{i\in\mathcal{V}^k}$ via power iteration (Eq. (4))

  Simulate random walks to obtain positive neighbor sets $\{\mathcal{P}_i^k\}_{i\in\mathcal{V}^k}$ and sample negative sets $\{\mathcal{N}_i^k\}_{i\in\mathcal{V}^k}$

for *each round* $t = 1, 2, \ldots, T$ **do**

  Server distributes global encoders $\{\mathbf{\Phi}^{(m)}\}_{m\in\mathcal{M}}$ to all clients

  if $t \geq 2$ **then** Server distributes routed prototypes $\{(\mathbf{P}_{\text{pos}}^{k,(i,j)}, \mathbf{P}_{\text{neg}}^{k,(i,j)})\}_{i,j}$ to each client k

  /* Client-side Local Training */

  for *each client* $k \in [K]$ ***in parallel*** **do**

    for *each local epoch* $e = 1, 2, \ldots, E$ **do**

      /* Multimodal Graph Encoding (Eq. (1), (2)) */

      $\mathbf{Z}^{k,(m)} \leftarrow \text{L2Norm}(\mathbf{X}^{k,(m)}\mathbf{\Phi}^{(m)}), \forall m \in \mathcal{M}$

      $\bar{\mathbf{Z}}^k \leftarrow \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \mathbf{Z}^{k,(m)}$

      $\mathbf{H}^k \leftarrow \frac{1}{\alpha+1}\sum_{l=0}^{L}\left(\frac{\alpha}{\alpha+1}\tilde{\mathbf{A}}^k\right)^l \bar{\mathbf{Z}}^k$

      /* Topology-aware Prototype Construction (Eq. (5)) */

      $\mathbf{P}^{k,(m)} \leftarrow \sum_{i\in\mathcal{V}^k} \omega_i^k \cdot \mathbf{z}_i^{k,(m)}, \forall m \in \bar{\mathcal{M}}$

      /* Local Optimization (Eq. (8), (9), (10)) */

      Compute node-level contrastive loss $\mathcal{L}_m$

      Compute neighbor-level contrastive loss $\mathcal{L}_g$

      if $t \geq 2$ **then** Compute client-level contrastive loss $\mathcal{L}_c$ using routed prototypes

      Update $\{\mathbf{\Phi}^{(m)}\}_{m\in\mathcal{M}}$ via $\nabla(\lambda_m\mathcal{L}_m + \lambda_g\mathcal{L}_g + \lambda_c\mathcal{L}_c)$ (Eq. (11))

    Upload updated encoders $\{\mathbf{\Phi}^{k,(m)}\}_{m\in\mathcal{M}}$ and final prototypes $\{\mathbf{P}^{k,(m)}\}_{m\in\bar{\mathcal{M}}}$ to server

  /* Server-side Aggregation and Routing */

  Aggregate encoders: $\mathbf{\Phi}^{(m)} \leftarrow \sum_{k=1}^{K} \frac{|\mathcal{V}^k|}{N}\mathbf{\Phi}^{k,(m)}, \forall m \in \mathcal{M}$

  /* Cross-modal Prototype Routing (Eq. (7)) */

  for *each client* $a \in [K]$ **do**

    for *each modality pair* $(i, j),\ i \in \bar{\mathcal{M}},\ j \in \bar{\mathcal{M}} \setminus \{i\}$ **do**

      $\mathbf{P}_{\text{pos}}^{a,(i,j)} \leftarrow \hat{\mathbf{P}}^{b^*,(j)}$, where $b^* = \arg\max_{b\neq a} \cos(\hat{\mathbf{P}}^{a,(i)}, \hat{\mathbf{P}}^{b,(j)})$

      $\mathbf{P}_{\text{neg}}^{a,(i,j)} \leftarrow \hat{\mathbf{P}}^{b',(j)}$, where $b' = \arg\min_{b\neq a} \cos(\hat{\mathbf{P}}^{a,(i)}, \hat{\mathbf{P}}^{b,(j)})$

/* Stage 2: Isolated Task-oriented Fine-tuning */

for *each client* $k \in [K]$ ***in parallel*** **do**

  Attach task-specific head g^k to the pre-trained encoder

  Fine-tune $\{\mathbf{\Phi}^{(m)}\}_{m\in\mathcal{M}}$ and g^k on local MAG $\mathcal{G}^k$ with task-specific loss $\mathcal{L}_{\text{task}}$