FEDM3G: FEDERATED MULTIMODAL MULTI-TASK GRAPH LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated learning over multimodal attributed graphs enables clients with complementary modalities and heterogeneous objectives to collaborate without exposing private data. Under the public-dataset paradigm, clients exchange representations computed on a shared graph instead of model parameters. We identify two limitations that are not addressed by existing federated multi-task and graph learning methods. First, prevailing multimodal graph encoders entangle node-intrinsic semantics with neighbor-derived topology, so representation alignment mixes two information sources that may have different task relevance. Second, task-agnostic alignment forces representations from unrelated objectives onto a common manifold and can therefore cause negative transfer. We propose FedM3G, a federated multimodal multi-task graph learning framework that decouples these sources and aligns them selectively. On each client, Gram-Schmidt decoupling extracts the component of neighbor context that is orthogonal to the node semantics, then combines this residual with task-relevant context through a learnable soft gate. The resulting semantic and topology tracks preserve complementary information without imposing strict orthogonality on the final outputs. On the server, Task Kernel Alignment estimates label-free compatibility from the relational geometry of public probes. Synergistic Representation Alignment converts these scores into client-specific teachers and confidence-weighted objectives, while sparse curriculum masking suppresses weak peer relations during early training. Across multiple datasets and heterogeneous downstream settings, FedM3G outperforms the strongest baseline, with absolute gains of up to 16.21% for cross-modal alignment and from 4.99% to 9.21% under joint task and domain heterogeneity.

1 INTRODUCTION

Graph Neural Networks (GNNs) (Kipf & Welling, 2017; Veličković et al., 2018) learn relational representations by propagating information along edges. Many practical graphs associate each node with multimodal content, such as products described by text and images, social posts combining language and pictures, or entities linked to visual and linguistic attributes, forming multimodal attributed graphs (MAGs) (He et al., 2025). Privacy regulation and commercial competition often prevent pooling such data, which motivates Federated Learning (FL) (McMahan et al., 2017). Realistic federated MAG deployments are substantially more heterogeneous than classical FL assumes. E-commerce and content platforms maintain private multimodal graphs while pursuing different objectives, including node classification (Tao et al., 2020; Jia et al., 2023), link prediction (Wei et al., 2019), and cross-modal retrieval (Radford et al., 2021). Sharing naively across these objectives can cause negative transfer (Liu et al., 2022; Chen & Zhang, 2022; Wang et al., 2019). The public-dataset paradigm (Yu et al., 2023) enables collaboration without a shared architecture by exchanging representations computed on a common public graph. This interface, however, leaves two issues unresolved on multimodal graphs. The first concerns how semantic content and graph context should be composed. The second concerns how representations should be aligned when client tasks differ.

Challenges. ❶ Semantic-topological entanglement. A MAG node contains two qualitatively different sources of information. Its multimodal features encode node-intrinsic semantics, whereas its neighborhood supplies contextual information induced by graph connectivity. Existing multimodal encoders commonly fuse these sources during message passing (Wei et al., 2019; Tao et al., 2020; Jia et al., 2023; He et al., 2025; Velickovic et al., 2019). The resulting representation is difficult

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

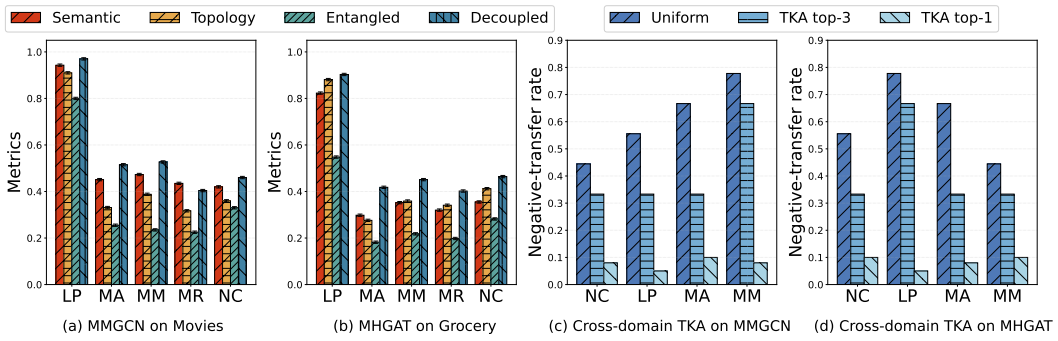


Figure 1: Empirical study results. Panels (a) and (b) compare feature settings and backbones. Panels (c) and (d) report negative transfer rates as a function of TKA scores.

to reweight for a specific task. In a federated setting, aligning such blended vectors can also allow structural patterns from dissimilar neighborhoods to contaminate semantic information that would otherwise transfer more reliably. ❷ **Task-agnostic alignment under label-free constraints.** Existing methods (Tan et al., 2022; Zhang et al., 2024; Fallah et al., 2020; Dinh et al., 2020; van den Oord et al., 2018) often align clients uniformly on a shared representation manifold. This assumption is inappropriate when clients optimize unrelated objectives, since uniform agreement can encourage harmful transfer rather than useful sharing. The server cannot resolve this issue by comparing private labels because those labels remain inaccessible. A suitable method must therefore estimate compatibility from information that is both shared and independent of private annotations.

Empirical study. To verify that these concerns arise in practice, we conduct a controlled experiment on a ten-client federation spanning five heterogeneous tasks with MMGCN and MHGAT backbones, as shown in Fig. 1. The results show that entangled MMGCN representations can underperform single-stream alternatives on cross-modal tasks. They also show that uniform alignment causes negative transfer for a substantial fraction of candidate donors. These observations indicate that the two issues are not merely conceptual. They motivate the dual-track decoupling and label-free Task Kernel Alignment used by FedM3G. Refer to Appendix H.1 for more experimental details.

In this work, we aim to develop a federated framework for multi-task learning that decoupled before alignment and uses label-free public-probe to gate cross-client interaction selectively.

Motivated by the above challenges, we propose FedM3G, a Federated MultiModal Multi-task Graph learning framework. FedM3G follows this principle through three coordinated components. First, the client-side Gram-Schmidt decoupling module performs semantics-guided message aggregation and extracts the component orthogonal to the node-semantic direction. A learnable gate then balances this residual against task-relevant contextual information, producing a semantic track and a topology track. Second, Task Kernel Alignment compares the relational geometry induced by these tracks on common public anchors. Although the server does not observe private task labels, task objectives influence how each encoder organizes the same ordered public graph. This geometry therefore provides a label-free proxy for compatibility, while not claiming to identify the private task itself. Third, Synergistic Representation Alignment combines the compatibility estimates with sparse curriculum masking. It constructs client-specific teachers that emphasize compatible peers and suppress weak relations, and it applies the resulting guidance independently to the two tracks.

Contributions. We make three contributions. First, we formulate federated multi-task learning on multimodal attributed graphs as representation decomposition followed by compatibility-aware selective alignment. This perspective makes explicit why semantic and topological information should not be treated as a single transferable carrier. Second, we propose FedM3G, which integrates soft Gram-Schmidt decoupling, public-probe TKA, and personalized Synergistic Representation Alignment. The method exposes an exact orthogonal residual while allowing the final topology-oriented track to retain task-useful semantic components. Third, across three datasets and five heterogeneous downstream tasks, FedM3G achieves absolute gains over the strongest baseline of up to 16.21% on cross-modal alignment and from 4.99% to 9.21% in the multi-domain setting.

2 Related Work

Federated multi-task learning under the public-dataset paradigm. FMTGL (Liu et al., 2022), CreamFL (Yu et al., 2023), FedMuscle (Setayesh et al., 2026), and FedMSplit (Chen & Zhang, 2022) enable clients with heterogeneous tasks or modalities to collaborate by exchanging representations computed on a shared public dataset rather than model parameters. This communication interface relaxes the requirement of a common architecture and is particularly useful when private data distributions, labels, and optimization objectives differ. However, these methods generally treat the exchanged representation as a monolithic vector and apply a common alignment rule to all dimensions. Consequently, they do not distinguish node semantics from graph topology, nor do they explicitly account for the fact that the same peer can be helpful for one task and harmful for another.

Representation alignment and consistency in federated multi-task learning. FeSTA (Park et al., 2021), FedHCA^2 (Lu et al., 2024), FedBone (Chen et al., 2024), and FedLPS (Jia et al., 2024) reconcile heterogeneous clients through split learning, hierarchical complementary alignment, gradient projection, or locally shared encoders. These approaches demonstrate that selective sharing can preserve local specialization, but their alignment mechanisms are not designed for multimodal attributed graphs with an explicit semantic and topological decomposition. In particular, they do not jointly use an ordered public graph probe to estimate peer-specific compatibility and construct separate targets for semantic and topology-oriented representations.

Federated graph learning with model heterogeneity. FedProto (Tan et al., 2022), FedTGP (Zhang et al., 2024), FedKD (Wu et al., 2022), LG-FedAVG (Liang et al., 2020), FIARSE (Wu et al., 2024), and TRUST (Wan et al., 2025) share compact carriers such as prototypes, logits, or submodels when clients use different architectures. These methods are effective when the carrier captures task-relevant information in a common space, but they primarily target single-modality graphs or settings with relatively homogeneous objectives. They therefore leave open how to transfer multimodal content and topology while limiting harmful interactions among genuinely multi-task clients.

3 Preliminaries

This section introduces the notation and learning problem used throughout the paper. We first define multimodal attributed graphs and then formalize the federated multi-task setting. The public-data communication protocol and its practical boundary conditions are discussed in Appendix **??**.

Multimodal attributed graphs. A multimodal attributed graph (MAG) contains a graph structure $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and a collection of node modalities $\mathcal{M} = \{m_1, \ldots, m_{|\mathcal{M}|}\}$. For each modality m, frozen or locally fixed encoders map raw node content to features $\mathbf{X}^{(m)}$. A multimodal graph encoder combines these features with the edge structure and produces node embeddings $\mathbf{Z} = \mathrm{Enc}(\mathcal{G}) \in \mathbb{R}^{|\mathcal{V}| \times d}$. The content features describe intrinsic properties of a node, while message passing introduces information from its neighborhood. As detailed in Sec. 4, vanilla aggregation can entangle these sources, which motivates the dual-track construction in FedM3G.

Federated multimodal multi-task graph learning. We consider N clients indexed by $n \in [N]$. Client n owns a private MAG $\mathcal{G}^n$, a private labeled set, and a downstream task $\mathcal{T}^n$ drawn from a task space $\mathcal{T}$. Examples include node classification, link prediction, cross-modal retrieval, and cross-modal matching. The federation is task heterogeneous because $\mathcal{T}^n$ and $\mathcal{T}^{n'}$ may differ. It may also be model heterogeneous because clients can use different encoders f_{θ_n}. This formulation separates the local prediction problem from the communication interface, which allows a client to retain its own task head and architectural choices.

Problem statement. Let $\mathcal{D}_n$ denote the private data distribution of client n and let ℓ_n be its task loss. Each client seeks personalized parameters θ_n that minimize local risk while benefiting from representations uploaded by other clients on a common public graph. The resulting objective is

$$\min_{\{\theta_n\}_{n=1}^N} \sum_{n=1}^{N} \Big(\underbrace{\mathbb{E}_{(x,y)\sim\mathcal{D}_n}\big[\ell_n(f_{\theta_n}(x), y)\big]}_{\text{local task risk}} + \lambda\, \mathcal{A}\big(\mathbf{Z}^n; \{\mathbf{Z}^m\}_{m\neq n}\big) \Big), \tag{1}$$

where $\mathcal{A}$ denotes a communication regularizer and λ controls its contribution. The regularizer must satisfy two requirements. It should operate on information components with distinct semantic roles rather than on an inseparable blend, and it should avoid forcing unrelated tasks to agree when private labels are unavailable. We instantiates $\mathcal{A}$ as compatibility-aware alignment over semantic and

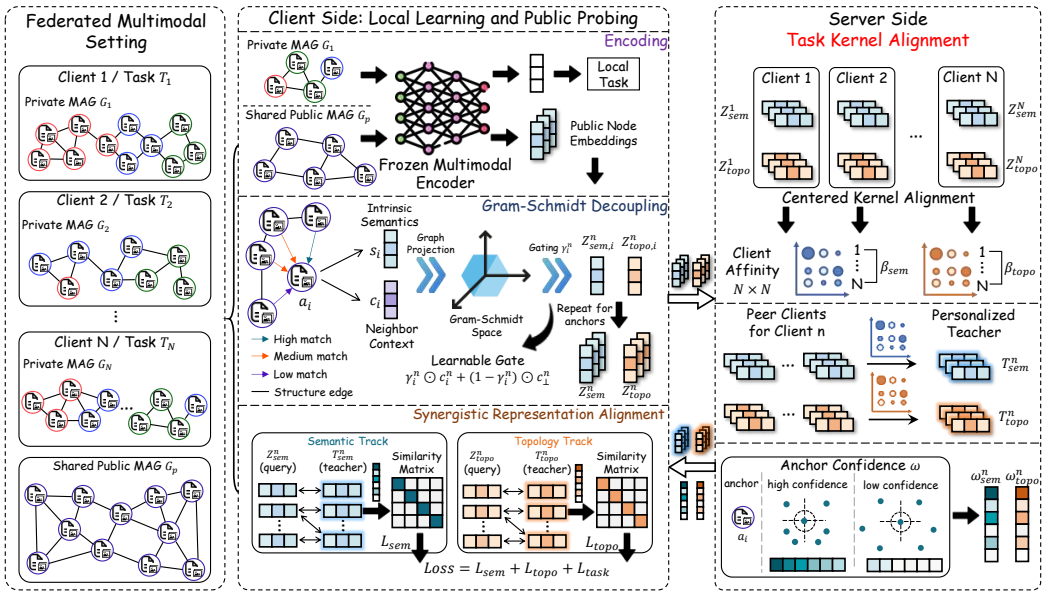


Figure 2: Overview of FedM3G. On the client side, one local encoder processes the private task graph and public anchors. Soft Gram-Schmidt decoupling exposes the semantics-orthogonal neighbor residual and produces semantic and topology tracks. On the server side, TKA estimates public-probe compatibility, after which personalized teachers and anchor-confidence priors are constructed. Each client combines Synergistic Representation Alignment with its private task loss.

topology-oriented tracks. The server receives only the public representations required by this alignment procedure, while each client retains its private graph, labels, task head.

4 Methodology

4.1 Overview

FedM3G follows a divide-and-conquer principle that separates what should be represented from how it should be aligned. In a multimodal attributed graph, node content and neighbor-derived context play different roles. Neighbor information should complement node semantics rather than overwrite them, and the contribution of each source should depend on the client task. FedM3G therefore uses three coordinated modules. Gram-Schmidt decoupling constructs a semantic track $\mathbf{Z}^n_{\text{sem}}$ and a topology-oriented track $\mathbf{Z}^n_{\text{topo}}$ for each client. Task Kernel Alignment estimates track-wise compatibility from public probes. Synergistic Representation Alignment converts these estimates into client-specific teachers and applies them to local optimization.

4.2 Gram-Schmidt Decoupling

Semantics-guided message aggregation. A MAG node contains multimodal content that describes its intrinsic semantics and neighbors that provide contextual information. Existing message-passing layers commonly fuse these signals before the client can decide how much of each source is useful. FedM3G first forms an explicit semantic representation and uses semantic similarity to select the contextual information that is most relevant to the node.

Client n extracts an intrinsic semantic representation from multimodal content as $\mathbf{s}^n_i = \text{Norm}(\text{Enc}^n_{\text{MM}}(X_i))$. A client-specific public projection interface maps all uploaded vectors to a common dimension d. The context aggregation is

$$\mathbf{c}^n_i = \sum_{j \in \mathcal{N}(i)} \alpha^n_{ij} \, \mathbf{W}^n \mathbf{s}^n_j, \tag{2}$$

where $e^n_{ij} = \langle \bar{\mathbf{s}}^n_i, \bar{\mathbf{s}}^n_j \rangle / \tau_a$, $\alpha^n_{ij} = \text{softmax}_{j \in \mathcal{N}(i)}(e^n_{ij})$, $\bar{\mathbf{s}}^n$ is the ℓ_2-normalized semantic vector, and $\tau_a > 0$ is the aggregation temperature. An isolated node uses a self-loop. This operation makes the source of the contextual signal explicit, but the context may still contain components that duplicate the target node semantics.

Soft Gram-Schmidt orthogonal projection. We next identify the part of the aggregated context that is not explained by the node-semantic direction. The projection and residual are

$$\mathbf{p}_i^n = \frac{\langle \mathbf{c}_i^n, \mathbf{s}_i^n \rangle}{\|\mathbf{s}_i^n\|_2^2} \mathbf{s}_i^n, \qquad \mathbf{c}_{\perp,i}^n = \mathbf{c}_i^n - \mathbf{p}_i^n, \tag{3}$$

where the denominator is clamped away from zero in implementation. By construction, $\mathbf{p}_i^n$ is the component aligned with the semantic direction and $\mathbf{c}_{\perp,i}^n$ is orthogonal to it. The residual therefore provides a principled topology-oriented signal, although a strict projection can discard useful information when semantic and structural factors are correlated.

To retain such information, we introduce a learnable gate $\boldsymbol{\gamma}_i^n \in (0,1)^d$ that chooses between the raw context and its orthogonal residual on a per-channel basis. The gate is $\boldsymbol{\gamma}_i^n = \sigma(\mathrm{MLP}([\mathbf{s}_i^n \| \mathbf{c}_i^n]))$. The final tracks are

$$\mathbf{z}_{\mathrm{sem},i}^n = \mathbf{s}_i^n, \quad \mathbf{z}_{\mathrm{topo},i}^n = \boldsymbol{\gamma}_i^n \odot \mathbf{c}_i^n + (\mathbf{1} - \boldsymbol{\gamma}_i^n) \odot \mathbf{c}_{\perp,i}^n + \eta \mathbf{c}_i^n, \tag{4}$$

where $\odot$ denotes element-wise multiplication and $\eta > 0$ is a small scalar that preserves a context bypass. Equivalently, $\mathbf{z}_{\mathrm{topo},i}^n = (1+\eta)\mathbf{c}_{\perp,i}^n + (\boldsymbol{\gamma}_i^n + \eta\mathbf{1}) \odot \mathbf{p}_i^n$. This identity clarifies the guarantee. Gram-Schmidt supplies an exact orthogonal residual, while the learned output is an orthogonalization-guided topology track rather than a strictly orthogonal vector. Both tracks are evaluated on the same ordered public anchors and uploaded to the server.

4.3 Task Kernel Alignment

Public-probe compatibility. Uniform alignment across unrelated tasks can cause negative transfer, but the server cannot compare private labels. We instead use the observation that related objectives often induce similar relational geometry on the same public graph. This geometry is not a task identifier. It is a label-free compatibility signal that can be affected by domain, architecture, and optimization as well.

The server receives dual-track matrices $\{\mathbf{Z}_{\mathrm{sem}}^n, \mathbf{Z}_{\mathrm{topo}}^n\}_{n=1}^N$. For track $q \in \{\mathrm{sem}, \mathrm{topo}\}$, CKA compares the geometry of two clients on the ordered public anchors

$$\beta_{q,nm} = \mathrm{CKA}(\mathbf{Z}_q^n, \mathbf{Z}_q^m). \tag{5}$$

A larger $\beta_{q,nm}$ indicates more similar public-anchor geometry. Because CKA compares centered Gram matrices, it focuses on relational structure and is insensitive to rotations or rescalings of representation coordinates, while domain, architecture, and optimization can still influence the score. We therefore treat it as a task-relatedness proxy, apply temporal smoothing to reduce early-round noise, and use a curriculum top-k_t mask to obtain sparse peer selection.

$$w_{q,nm}^{(t)} = \frac{M_{nm}^{(t)} \exp(\beta_{q,nm}^{(t)} / \tau_\beta)}{\sum_{k \neq n} M_{nk}^{(t)} \exp(\beta_{q,nk}^{(t)} / \tau_\beta)}, \tag{6}$$

where $M_{nm}^{(t)}$ retains the k_t strongest peers of client n, k_t increases linearly from k_0 to $N-1$ during the curriculum horizon, and $\tau_\beta > 0$ is the peer-weight temperature. This separation gives CKA the role of ranking compatibility and gives the mask the role of exact peer exclusion. It also prevents early, noisy estimates from exposing every client to every other client.

4.4 Synergistic Representation Alignment

Personalized teacher construction. A compatibility score is not itself a local training signal. Each recipient needs a target that reflects its own peer relations and an estimate of which public anchors are reliable. For each track, the server constructs a personalized peer teacher $\mathbf{P}_q^n = \sum_{m \neq n} w_{q,nm} \mathbf{Z}_q^m$ and blends it with a robust public-anchor consensus $\mathbf{C}_q$ as $\mathbf{T}_q^n = (1 - \xi_q^n)\mathbf{P}_q^n + \xi_q^n \mathbf{C}_q$. Here ξ_q^n controls the balance between personalized and global guidance. The server also assigns anchor confidence from weighted peer dispersion

$$\omega_{q,j}^n \propto \frac{1}{1 + \sum_{m \neq n} w_{q,nm} \|\mathbf{z}_{q,j}^m - \bar{\mathbf{z}}_{q,j}^n\|_2^2}, \tag{7}$$

where $\bar{\mathbf{z}}_{q,j}^n$ is the weighted peer mean. The scores are normalized to unit mean and clipped to $[0.1, 5.0]$. High consensus gives an anchor a larger prior because disagreement among compatible peers makes that anchor less reliable for alignment. The server sends $\{\mathbf{T}_q^n, \boldsymbol{\omega}_q^n\}_q$ to client n.

Synergistic representation alignment loss. Client n receives track-specific teachers and confidence priors. Let $\tilde{\mathbf{z}}^n_{q,i}$ and $\tilde{\mathbf{t}}^n_{q,j}$ denote normalized student and stop-gradient teacher representations. The confidence-adjusted alignment score is

$$a^n_{q,ij} = \frac{\langle \tilde{\mathbf{z}}^n_{q,i}, \tilde{\mathbf{t}}^n_{q,j} \rangle}{\tau_s} + \log \omega^n_{q,j}. \tag{8}$$

The exponential transformation gives $\phi^n_{q,ij}$ a positive unnormalized weight, with the similarity term favoring teacher columns close to the student and the confidence prior favoring anchors on which compatible peers agree. Because all clients use the same ordered public anchors, column i is the positive target for student anchor i, and the remaining columns provide contrastive alternatives. The resulting cross-entropy objective is

$$\mathcal{L}^n_{\mathrm{SRA}}(\mathbf{Z}^n_q) = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\phi^n_{q,ii}}{\sum_{j=1}^{B} \phi^n_{q,ij}}, \tag{9}$$

where B is the number of public anchors and τ_s is the contrastive temperature. The diagonal correspondence is valid because every client evaluates the same ordered anchors. The logarithmic confidence term is a prior over teacher columns rather than an adaptive temperature. Peer compatibility affects teacher composition and confidence, while $M^{(t)}_{nm} = 0$ removes a peer exactly. Local task losses and the two track-specific Synergistic Representation Alignment losses can therefore be optimized jointly, allowing each task to determine how strongly it relies on semantic and topology-oriented information.

5 Experiments

We evaluate FedM3G through four questions. **Q1.** We ask whether FedM3G outperforms strong federated baselines across diverse downstream tasks and backbones. **Q2.** We examine whether it can coordinate clients whose private domains and downstream objectives differ. **Q3.** We quantify the contribution of each core module through ablation. **Q4.** We test robustness to the principal hyperparameters.

5.1 Experimental Settings

Datasets. We evaluate three Amazon multimodal attributed graphs, namely **Movies**, **Grocery**, and **Toys**. Each graph is partitioned by Louvain community detection into five private client subgraphs and one disjoint public subgraph, following the public-dataset paradigm (Yu et al., 2023). Each client trains its encoder on its private subgraph and exchanges only representations computed on the shared public subgraph. Raw content, labels, and model parameters remain local. Refer to Appendix A for details on dataset construction, partitioning, and statistics.

Baselines. We compare against nine federated methods. FedAvg (McMahan et al., 2017) and MOON (Li et al., 2021) represent standard federated learning. CreamFL (Yu et al., 2023) and FedMuscle (Setayesh et al., 2026) represent multimodal and public-dataset federated learning. FedProto (Tan et al., 2022), FedHCA^2 (Lu et al., 2024), FedBone (Chen et al., 2024), FedLPS (Jia et al., 2024), and TRUST (Wan et al., 2025) represent multi-task and heterogeneity-oriented federated learning. Refer to Appendix C for details.

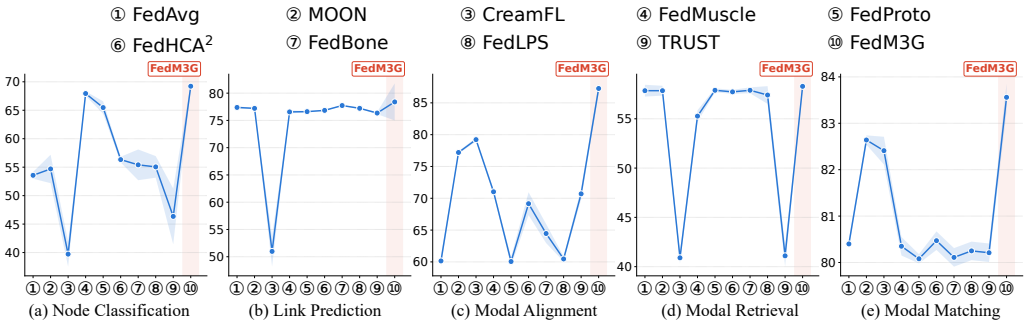
Tasks and metrics. Clients span five downstream capabilities. We report node classification accuracy (**NodeCls ACC**), link prediction AUC (**LinkPred AUC**), modal alignment CLIP score(**Align CLIP**), modal retrieval recall (**Retrieval R@5**), and modal matching AUC (**Match AUC**). The first two metrics assess graph-centric prediction, whereas the other three assess multimodal capability. Refer to Appendix D for more details.

5.2 Main Performance Comparison (Q1)

As shown in Table 1 and Fig. 3, we deployed FedM3G across the Movies, Grocery, and Toys datasets. With the figure included, FedM3G outperforms every baseline on every dataset-task combination on the MGAT backbone, with Movies and Grocery results reported in Table 1 and Toys results reported in Fig. 3. Each task draws on its preferred track, and dual alignment keeps the two streams from contaminating each other.

Table 1: Main Performance. Best is in **bold**, second best is underlined.

Method	NodeCls ACC	LinkPred AUC	Align CLIP	Retrieval R@5	Match AUC
			Movies		
FedAvg	38.81±1.5	86.08±0.1	44.78±0.9	62.87±0.4	77.62±0.2
MOON	41.33±0.4	85.97±0.1	76.62±0.2	62.72±0.7	79.87±0.1
CreamFL	34.62±4.5	53.29±3.7	78.45±0.9	40.27±0.4	78.63±0.1
FedMuscle	42.47±0.5	85.29±0.3	57.98±2.0	60.88±0.6	77.50±0.1
FedProto	44.25±0.9	85.53±0.4	45.66±1.0	63.87±0.2	77.92±0.1
FedHCA^2	36.64±1.7	85.86±0.1	57.62±1.8	62.67±1.1	77.65±0.2
FedBone	38.72±0.6	89.81±0.2	52.16±2.5	62.97±0.6	77.23±0.1
FedLPS	39.90±1.2	86.44±0.9	46.04±1.3	61.93±0.2	77.86±0.2
TRUST	34.12±3.7	85.44±0.2	58.42±1.6	42.81±0.5	77.65±0.1
FedM3G	**45.63**±0.7	**90.90**±0.3	**90.13**±1.8	**63.92**±0.4	**80.85**±0.2
			Grocery		
FedAvg	57.39±2.1	85.20±0.1	42.20±0.4	53.04±0.5	77.87±0.1
MOON	62.68±1.4	85.35±0.2	73.32±0.2	52.66±0.4	81.59±0.2
CreamFL	32.75±2.1	50.86±4.6	76.01±1.0	31.35±0.3	81.11±0.1
FedMuscle	77.84±0.4	84.38±0.4	51.56±1.1	47.58±0.5	78.34±0.0
FedProto	78.86±0.3	84.61±0.1	42.19±0.5	52.90±0.7	78.28±0.3
FedHCA^2	72.74±1.2	85.10±0.2	52.82±0.8	47.44±0.7	78.15±0.1
FedBone	70.26±1.1	85.50±0.3	45.74±1.2	49.66±0.8	77.00±0.2
FedLPS	60.64±0.9	85.22±0.2	42.21±0.1	49.61±0.5	77.83±0.4
TRUST	53.26±11.1	84.76±0.2	51.06±1.7	30.97±0.4	78.17±0.1
FedM3G	**80.17**±0.5	**86.10**±0.4	**92.22**±1.0	**53.19**±0.3	**82.64**±0.1



(a) Node Classification (b) Link Prediction (c) Modal Alignment (d) Modal Retrieval (e) Modal Matching

Figure 3: Performance comparison between FedM3G and baselines on the Toys.

The largest gain appears on cross-modal alignment: FedM3G outperforms the strongest baseline by $\mathbf{11.98}\%$ on average across the three datasets. Alignment directly probes the cross-modal geometric signal that the GS-decoupled semantic track exposes, which entangled baselines such as CreamFL cannot recover. CreamFL's alignment scores stay confined to the high 70s across all three datasets. Other tasks show smaller margins because each is partially covered by a specialized baseline. FedProto's class prototypes trail FedM3G by 1.32% on average in NodeCls. FedBone's pretrained backbone holds LinkPred within 0.78%. MOON's contrastive head keeps Match to 0.98%. Retrieval leaves a small relative-ranking margin of 0.20%.

No single baseline covers all five capabilities: each makes a specific inductive bet (contrastive ensembles for alignment, pretrained backbones for edges, model-contrastive losses for matching, class prototypes for classification), while FedM3G's Synergistic Representation Alignment accommodates all five bets in parallel because each track is aligned independently and TKA selects peers that strengthen rather than overwrite the local signal.

5.3 Mixed-Source Domain-Task Heterogeneity (Q2)

Table 2 reports the audited mixed-source protocol on MMGCN with the strongest entries highlighted in bold and the second best underlined. As in Table 1, each task client trains and evaluates on

Table 2: Mixed-source heterogeneity with MMGCN. Best in **bold**, second best underlined.

Method	NodeCls ACC	LinkPred AUC	Align CLIP	Retrieval R@5	Match AUC
FedAvg	36.15±1.5	86.74±1.1	66.84±0.9	37.57±0.7	77.94±0.2
MOON	36.15±1.7	85.65±0.5	77.17±1.3	38.10±1.7	80.15±0.3
CreamFL	36.69±2.4	89.21±0.6	76.73±0.8	38.04±1.4	80.45±0.3
FedMuscle	36.49±2.4	89.17±0.7	66.00±0.2	38.36±0.2	78.51±0.0
FedProto	34.72±1.4	86.47±1.0	66.83±0.9	37.36±0.6	78.05±0.1
FedHCA^2	33.04±0.9	86.77±1.0	67.60±1.1	37.73±0.7	78.32±0.1
FedBone	35.06±0.2	84.78±0.3	72.50±0.8	37.89±1.0	78.35±0.3
FedLPS	34.02±1.8	85.24±0.2	66.84±0.9	37.57±0.7	77.94±0.2
TRUST	32.35±6.5	90.65±0.2	60.88±0.5	36.21±0.4	78.10±0.2
FedM3G	**41.68**±0.2	**97.28**±0.2	**86.38**±1.2	**45.26**±0.3	**85.55**±0.3

Table 3: Ablation of the three core modules on Movies with MGAT and Louvain.

Variant	NodeCls ACC	LinkPred AUC	Align CLIP	Retrieval R@5	Match AUC
FedM3G (full)	**45.63**±0.7	**90.90**±0.3	**90.13**±1.8	**63.92**±0.4	**80.85**±0.2
w/o SRA	43.56±0.1	83.15±0.2	83.23±0.6	59.18±0.7	78.61±0.2
w/o TKA	43.85±0.1	82.26±0.9	83.21±0.6	58.73±0.6	76.51±0.2
w/o GS decoupling	43.85±0.1	83.14±0.3	83.21±0.6	59.23±1.9	76.53±0.2

its private Louvain subgraph from a distinct source domain (Movies, Sports, Flickr30k, RedditS, Grocery), retaining its private task and held-out local test split. Collaboration occurs only through the public Toys anchors. FedM3G leads all five task metrics in mean performance.

Margins. Across the five tasks, FedM3G's absolute margin over the strongest baseline ranges from 4.99% to 9.21%, which translates into a relative improvement between 13.60% and 17.99%. The mean absolute gain reaches 6.57%, more than double the 3.05% average observed on the MGAT.

Why the margins widen. Two mechanisms reinforce. Domain heterogeneity raises structural noise across the five source domains (Movies, Sports, Flickr30k, RedditS, Grocery), so monolithic baselines such as CreamFL or FedMuscle leak cross-domain structure into the semantic signal; Gram-Schmidt decoupling isolates the semantics-orthogonal residual, which is why alignment, the metric most sensitive to cross-modal purity, gains the most. Task heterogeneity further magnifies the cost of uniform alignment: the five task geometries differ radically, so without TKA, negative transfer, which Q1 hid under similar tasks, becomes the dominant error source; TKA's compatibility-aware gating converts this into a measurable improvement, with retrieval seeing the largest relative margin at 17.99%. Consequently, no baseline covers more than one or two of the five tasks: TRUST's curriculum distillation, CreamFL's contrastive ensemble, and FedMuscle's dedicated contrastive loss each assume a peer geometry that breaks under domain shift, while FedM3G covers all five because dual-track Synergistic Representation Alignment accommodates each task's inductive bias independently. The protocol does not establish generalization to an unseen target domain, but confirms that the dual-track public interface remains effective under joint domain and task heterogeneity.

5.4 Ablation Study (Q3)

Removing any of the three core modules decreases all five metrics. The mean relative drop is 6.18% when Synergistic Representation Alignment is removed, 6.56% when GS Decoupling is removed, and $\mathbf{6.91}\%$ (the largest) when TKA is removed. TKA's leading position validates the central claim. When client labels are unavailable, label-free compatibility estimation is the most fragile link, without TKA, peer weights revert to uniform and negative transfer from dissimilar-task clients directly contaminates the alignment target, so every metric drops by at least 1.78%. The task-level degradation patterns provide a finer-grained confirmation of the complementary design. TKA removal inflicts the largest single-task losses on three of the five metrics, with LinkPred dropping by 8.64 points, Retrieval by 5.19 points, and Match by 4.34 points, suggesting that peer selection governs the majority of downstream behaviors rather than a single task. Align CLIP, in contrast, exhibits a near-uniform 6.9 percentage drop under every ablation, confirming that cross-modal alignment sits at the joint protection of all three components and degrades symmetrically whenever any one of them fails. The three components are complementary rather than overlapping. Gram-Schmidt decou-

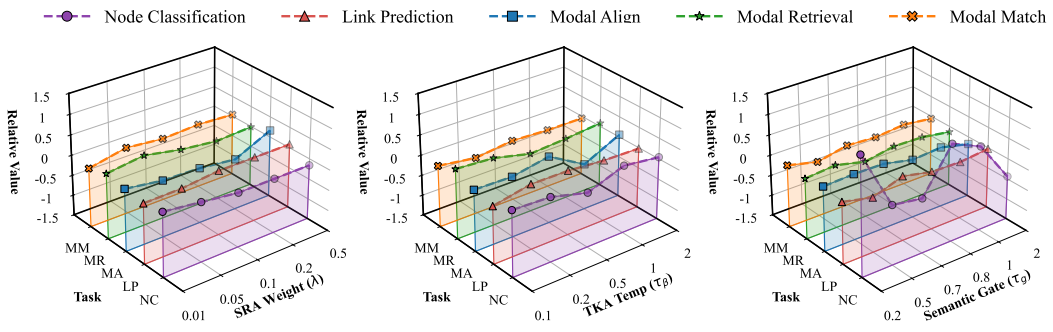


Figure 4: Hyperparameter sensitivity on the Toys under Louvain simulation with MHGAT. Each panel sweeps one of the three dual-track knobs: λ, τ_β, τ_g.

pling produces the two tracks, TKA decides which peers each track should listen to, and Synergistic Representation Alignment realizes the listening. Removing Gram-Schmidt semantic-topological decoupling leaves TKA with entangled inputs, removing TKA leaves Synergistic Representation Alignment with indiscriminate peers, and removing Synergistic Representation Alignment leaves a well-targeted but un-realized compatibility signal.

5.5 Hyperparameter Sensitivity (Q4)

Sensitivity is evaluated on the Toys Louvain five-task split with MHGAT by sweeping three knobs of the dual-track design: the Synergistic Representation Alignment weight λ, the TKA temperature τ_β, and the semantic-gate temperature implicit in the soft GS gate. Data partition, optimizer controls, ten-round budget. Fig. 4 reports the relative deviation of each task metric from a shared baseline defined as the median of the corresponding metric across all hyperparameters within the same panel. The five baselines are NC at 72.02, LP at 95.00, MA at 79.65, MR at 58.58, and MM at 81.70, and the same baseline is reused across the three sub-figures. Error bars remain three-seed standard errors and are reported in relative terms. FedM3G passes the registered robustness criteria, macro-average drift below 2 % and worst-task drift below 5 %, by a wide margin.

Robustness. The Synergistic Representation Alignment weight sweep spans the widest relative interval, from 0.01 to 0.5, a $50\times$ ratio, yet produces only the smallest drift: the macro average moves $\mathbf{0.03}\%$ and NC moves 0.24% . The TKA temperature sweep behaves comparably, drifting 0.05% on the macro average and 0.32% on NC. The semantic-gate sweep is the most reactive but still well under threshold, drifting 0.38% on the macro average and 1.84% on NC.

The asymmetry across tasks reflects the design: NC depends on the semantic encoder that consumes z_{sem}, while MA, MR, and MM bypass the dual-track and LP reads the topology embedding that Synergistic Representation Alignment and TKA stabilize but do not reshape. The three multimodal probes are completely flat at 0 , LP shows a small worst-case of 0.59% under LR, and NC is the only task that visibly moves, with a worst-case of 1.96% under LR, still an order of magnitude below the 5% threshold. The default settings used in Q1-Q3 therefore require no tuned rerun.

6 Conclusion

In this work, we propose FedM3G to resolve two obstacles that hinder federated collaboration over multimodal attributed graphs. The first is the entanglement of node-intrinsic semantics with neighbor-derived topology, and the second is the negative transfer induced by task-agnostic alignment. Following a strategy of decoupling before selective alignment, the client-side Gram-Schmidt Decoupling exposes the semantics-orthogonal residual and forms two softly decoupled representation tracks. The server-side Task Kernel Alignment estimates public-probe compatibility, and Synergistic Representation Alignment constructs track-specific alignment targets. Experiments across five tasks show strong performance on MGAT and MHGAT and consistent gains under joint domain and task heterogeneity. Our single-donor analysis evaluates TKA against realized transfer gain rather than task identity, preserving its interpretation as a compatibility proxy instead of a task-identification guarantee. Future work will extend the framework to partially overlapping public graphs, where some nodes appear in both client and public subgraphs and introduce new questions about cross-subgraph information routing.

AI USE STATEMENT

No large language models (LLMs) or LLM-based tools were used at any stage of this work, including manuscript drafting and revision, experimental design and execution, code development, data analysis, or interpretation of results. The authors take full responsibility for all content, references, and claims presented in this manuscript.

REFERENCES

Jiayi Chen and Aidong Zhang. FedMSplit: Correlation-adaptive federated multi-task learning across multimodal split networks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2022.

Yiqiang Chen, Teng Zhang, Xinlong Jiang, Qian Chen, Chenlong Gao, and Wuliang Huang. FedBone: Towards large-scale federated multi-task learning. *Journal of Computer Science and Technology (JCST)*, 2024.

Zhen Chen, Wenwei Liu, Zhe Wang, Zhengyu Yang, Yujun Cheng, and Wengang Zhang. Graphmae2: A decoding-enhanced masked self-supervised graph learner. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2022.

Canh T. Dinh, Nguyen H. Tran, and Tuan D. Nguyen. Personalized federated learning with moreau envelopes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A MARS approach. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Yufei He, Yuan Sui, Xiaoxin He, Yue Liu, Yifei Sun, and Bryan Hooi. Unigraph2: Learning a unified embedding space to bind multimodal graphs. *arXiv preprint arXiv:2502.00806*, 2025.

Xiangen Jia, Min Jiang, Yihong Dong, Feng Zhu, Haocai Lin, Yu Xin, and Huahui Chen. Multimodal heterogeneous graph attention network. *Neural Computing and Applications*, 35(4):3357–3372, 2023. doi: 10.1007/s00521-022-07862-6.

Yongzhe Jia, Xuyun Zhang, Amin Beheshti, and Wanchun Dou. FedLPS: Heterogeneous federated learning for multiple tasks with local parameter sharing. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2024.

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

Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Paul Pu Liang, Terrance Liu, Liu Ziyin, Nicholas B. Allen, Randy P. Auerbach, David Brent, Ruslan Salakhutdinov, and Louis-Philippe Morency. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523*, 2020.

Yang Liu, Yifeng Wang, Xin Yang, Liqiang Liu, Jingkuan Zeng, and Zhiyuan Cheng. Graph structure mapping neural networks for generative graph self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Yijing Liu, Dongming Han, Jianwei Zhang, Haiyang Zhu, Mingliang Xu, and Wei Chen. Federated multi-task graph learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13 (5):1–27, 2022.

Yuxiang Lu, Suizhi Huang, Yuwen Yang, Shalayiding Sirejiding, Yue Ding, and Hongtao Lu. FedHCA^2: Towards hetero-client federated multi-task learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

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 (AISTATS)*, 2017.

Sangjoon Park, Gwanghyun Kim, Jeongsol Kim, Boah Kim, and Jong Chul Ye. Federated split task-agnostic vision transformer for COVID-19 CXR diagnosis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Mehdi Setayesh, Mahdi Beitollahi, Yasser H. Khalil, and Hongliang Li. Toward enhancing representation learning in federated multi-task settings. In *International Conference on Learning Representations (ICLR)*, 2026.

Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. Fedproto: Federated prototype learning across heterogeneous clients. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

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), 2020.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

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

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

Guancheng Wan, Xu Cheng, Run Liu, Wenke Huang, Zitong Shi, Pinyi Jin, Guibin Zhang, Bo Du, and Mang Ye. Multi-order orchestrated curriculum distillation for model-heterogeneous federated graph learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Zirui Wang, Zihang Dai, Slawomir Pociot, and Jaime Carbonell. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

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 *ACM International Conference on Multimedia*, pp. 1437–1445, 2019. doi: 10.1145/3343031.3351034.

Chuhan Wu, Fangzhao Wu, Lingjuan Lyu, Yongfeng Huang, and Xing Xie. Communication-efficient federated learning via knowledge distillation. *Nature Communications*, 13, 2022.

Feijie Wu, Xingchen Wang, Yaqing Wang, Tianci Liu, Lu Su, and Jing Gao. FIARSE: Model-heterogeneous federated learning via importance-aware submodel extraction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Qiying Yu, Yang Liu, Yimu Wang, Ke Xu, and Jingjing Liu. Multimodal federated learning via contrastive representation ensemble. In *International Conference on Learning Representations (ICLR)*, 2023.

Jianqing Zhang, Yang Liu, Yang Hua, and Jian Cao. Fedtgp: Trainable global prototypes with adaptive-margin-enhanced contrastive learning for data and model heterogeneity in federated learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2024.

A DATASETS

Our experiments are conducted on three multimodal attributed graph datasets drawn from the MAGB benchmark family, namely **Movies**, **Grocery**, and **Toys**. The three datasets span product review platforms across the movie, grocery, and toy domains, providing a controlled evaluation environment for cross-modal and cross-domain generalization on multimodal attributed graphs. All three datasets provide paired image and text node features together with category classification labels, and edges are built from product co-purchase or co-viewing relations. Each dataset is partitioned into five private client subgraphs and one disjoint public subgraph by Louvain community detection, following the public-dataset paradigm. Each client trains its encoder on its private subgraph and exchanges only representations computed on the shared public subgraph, while raw content, labels, and model parameters remain strictly local. The same partition and seed set are reused across all compared methods, which keeps the data interface comparable across ablations and sensitivity sweeps. We summarize each dataset below.

Movies (MAGB subset) contains approximately 16,672 nodes spanning 20 categories of cinematic and home media products. Each node encodes bilingual titles, descriptions, and tag information, together with visual features extracted from cover images. User-product co-viewing interactions form the edge set. Owing to its broad and balanced category coverage, the dataset is frequently used to evaluate cross-category transfer for content-rich products.

Grocery (MAGB subset) contains 17,074 nodes across 20 categories of grocery and household items. Each node records textual product descriptions, brand information, and image-based features, while the edge set encodes co-purchase relations. The dataset exhibits the most skewed category prior among the three, which makes it a natural probe of label-skew robustness under the public-dataset paradigm.

Toys (MAGB subset) contains 20,695 nodes with 18 categories of toy and game products. Each node combines textual descriptions, brand and age-range metadata, and cover-image features. The dataset supports studies on cross-style and cross-category transfer for product domains where visual and textual cues are equally informative.

Cloth (MAGB OpenMAG subset) contains 17,752 nodes across 17 categories of apparel products drawn from a major e-commerce platform. Each node pairs product title text with cover-image features, while the edge set is built from co-purchase and co-viewing interactions. The dataset is a standard probe for cross-modal transfer on fine-grained fashion categories where textual brand cues and visual style cues must be jointly leveraged.

Ele-fashion (MAGB OpenMAG subset) contains 22,614 nodes with 16 categories of consumer electronics and fashion products. Each node fuses technical descriptions, brand metadata, and rendered product image features. Edges encode co-viewing and co-purchase relations. The dataset has the largest raw node count among our partitions and stresses the scalability of public anchor encoding, which makes it a natural stress test for the SRA and TKA components.

Sports (MAGB OpenMAG subset) contains 15,440 nodes across 18 categories of sporting goods. Each node pairs descriptive text with cover-image features and edges are derived from reader co-purchase patterns. The dataset is dominated by textual signal because the cover-image feature is comparatively sparse, which makes it the cleanest probe for isolating the semantic track of the dual track design from the visual track.

B BACKBONES

We instantiate the client-side encoder with three multimodal graph networks, chosen to span the dominant architectural families in multimodal attributed graph learning. Following the pretrained-encoder recipe of (He et al., 2025), each modality encoder ϕ_m is treated as a frozen feature extractor that maps raw modality-m content into features $\mathbf{X}^{(m)} \in \mathbb{R}^{N \times d_m}$. Freezing the modality encoders serves three purposes. It stabilizes federated training by removing the largest parameter group from communication and aggregation. It prevents clients from drifting the perceptual backbone when their private distributions differ. It also keeps the multimodal bundle aligned with public-graph features, which is a prerequisite for meaningful Task Kernel Alignment. The bundle $X_i = \{\mathbf{x}_i^{(m)}\}_{m \in \mathcal{M}}$ is

then propagated through the multimodal graph encoder, which mixes modalities inside its message-passing layers. We summarize each backbone below.

MGAT (Tao et al., 2020) extends graph attention networks to multimodal-attributed graphs by introducing modality-specific attention mechanisms that adaptively weight neighbors under different modalities, enabling effective fusion of heterogeneous modality information within attention-based message passing. MGAT serves as the primary backbone for the main results and exposes the framework to soft attention-based fusion.

MHGAT (Jia et al., 2023) is a multimodal heterogeneous graph attention network designed for graphs possessing both heterogeneous structural constraints and multimodal attributes. The model distinguishes itself by replacing manual meta-path selection with a two-stage aggregation scheme, where neighbors are first clustered and pooled by relation type and then integrated into a shared feature space using an edge-level attention mechanism. A modality-wise attention layer performs adaptive fusion of features such as visual and textual data. To ensure robust training and capture long-range dependencies, the architecture employs a dense connectivity pattern with residual shortcuts, effectively mitigating the over-smoothing phenomenon. MHGAT is used in our mixed-source heterogeneity experiments and exposes the framework to typed relational structure.

MMGCN (Wei et al., 2019) extends graph convolution to multimodal-attributed graphs by constructing modality-aware user-item bipartite graphs and performing message passing independently within each modality, followed by coordinated fusion through shared ID embeddings. This design enables the model to capture modality-specific structural signals and user preferences while preserving cross-modal consistency, making MMGCN a representative multimodal GNN backbone for centralized multimodal graph learning. MMGCN serves as the primary backbone for the mixed-source Q2 experiments and tests the framework under a lightweight convolutional architecture.

GSMN (Liu et al., 2020) is a graph self-supervised method that learns node representations by predicting a permuted version of the original graph structure, framing self-supervision as a graph-to-graph mapping problem rather than as a contrastive objective. The mapping objective forces the encoder to recover global structural invariants from local perturbations, which makes GSMN the natural backbone for stress testing the topology channel under a self-supervised pretraining recipe. The decoupled semantic and topology channels of FedM3G interact cleanly with GSMN's structural priors because GSMN does not commit to a particular modality fusion pattern.

GraphMAE2 (Chen et al., 2022) is a generative graph masked autoencoder that reconstructs masked node features from a corrupted subgraph using a re-masking strategy that revisits the masking distribution between encoder and decoder. The re-masking step decouples the encoder view from the decoder view, which makes the reconstruction objective a clean probe of representation compressibility. As a multimodal graph backbone, GraphMAE2 exposes FedM3G to a generative self-supervised signal that complements the discriminative signal of MGAT, and the contrast between GraphMAE2's reconstruction objective and FedM3G's anchor alignment objective provides additional evidence for the dual track design.

UniGraph2 (He et al., 2025) is a unified multimodal graph foundation model that binds heterogeneous modalities into a shared embedding space through large scale self-supervised pretraining. The architecture differs from MGAT and MMGCN by operating as a foundation model whose pretrained features are reused across downstream tasks rather than trained from scratch per task. UniGraph2 is the most relevant backbone for stress testing the pretrained encoder recipe adopted by FedM3G, because it provides the strongest possible starting point for the public anchor interface and lets us isolate the contribution of FedM3G's design from the contribution of the pretrained feature extractor.

C Federated Baselines

We compare FedM3G against nine federated methods organized into three families by their inductive bias, namely parameter averaging, contrastive or prototype based representation alignment, and heterogeneous aggregation or curriculum distillation. Each baseline is re-tuned under the same validation based selection procedure as FedM3G. The protocol does not declare FedM3G the universal winner, since specialized baselines retain task specific inductive biases that already cover a portion of the metric space.

FedAvg (McMahan et al., 2017) is the canonical communication efficient federated averaging algorithm that anchors our comparison as the parameter aggregation reference. In each round the server samples a fraction of clients, broadcasts the global parameters w_t, lets every selected client perform E minibatch SGD epochs on its private partition, and aggregates updates through $w_{t+1} = \sum_k \frac{n_k}{n} w_k$. The design amortizes communication over E local epochs. Furthermore, the protocol is model agnostic only insofar as clients share a parameter space. Its principal limitations are the representation drift that non-IID partitions induce, the absence of formal privacy guarantees or straggler defense, and the uniform weighting bias toward larger partitions.

MOON (Li et al., 2021) reformulates federated training as a model level contrastive problem that targets representation drift. Each client performs three forward passes through the current local model, the global model, and the cached previous round model, and attaches a SimCLR NT-Xent model contrastive loss to the supervised task to pull local features toward the global anchor and push them away from the stale local snapshot. The framework outperforms FedAvg, FedProx, and SCAFFOLD on CIFAR under Dirichlet non-IID partitions. Its disadvantages are the threefold forward pass overhead per local step and the dependence on a slowly drifting global anchor that may itself be biased under extreme heterogeneity.

CreamFL (Yu et al., 2023) generalizes the public dataset paradigm to multimodal clients with heterogeneous architectures by replacing parameter exchange with prediction level distillation on a shared unlabeled public batch. Clients compute predictions on the public batch, the server aggregates them into a global local cross modal ensemble and distills the ensemble into a stronger server model, and the refreshed server predictions are broadcast back for client side distillation. An inter modal contrastive loss on the server aligns multimodal embeddings, while an intra modal contrastive loss on each client stabilizes training against ensemble drift. The principal weakness is the strict dependence on a public dataset semantically and modally aligned with the private data.

FedMuscle (Setayesh et al., 2026) introduces a contrastive auxiliary loss on top of the public representation interface to tighten cross client geometry under heterogeneous multimodal data. Each client transmits only its public anchor embeddings, the server aggregates them into a reference representation, and every client trains against this reference through a Muscle contrastive loss that up weights hard negatives and is theoretically equivalent to maximizing a lower bound on the mutual information between client outputs. The framework only exchanges embeddings on public anchors and tolerates both model and task heterogeneity. Its limitations include the sensitivity of hard negative emphasis to noisy inter client similarity estimates and the residual representation entanglement from failure to disentangle shared modality factors.

FedProto (Tan et al., 2022) is an early prototype based framework for heterogeneous federated learning that exchanges abstract class prototypes rather than gradients or parameters. Each client computes the mean embedding of every observed class and sends these prototypes to the server, which performs a per class weighted average to build global prototypes and broadcasts them back. Every client then trains with an additional regularizer that pulls local embeddings toward the global prototype of their ground truth class. The framework tolerates architectural heterogeneity with negligible per round communication cost. Its fundamental limitation is that prototypes capture only first order class conditional statistics and discard sample sample relations, which makes the approach inherently classification oriented and degrades on retrieval, matching, and instance wise contrastive objectives.

FedHCA^2 (Lu et al., 2024) extends prototype sharing with hierarchical class aware aggregation for multi task federated learning. The framework assumes a shared encoder backbone while permitting heterogeneous task heads, and replaces naive parameter averaging with a two level scheme. At the encoder level, a hyper conflict averse aggregation projects each client gradient away from any conflicting peer gradient before averaging. At the decoder level, layer wise cross attention across heterogeneous heads transfers task relevant features through attention rather than direct parameter sharing. Its limitations include that the encoder backbone must still be shared, that supervision requires task labels at every client, and that the method is restricted to dense prediction settings.

FedBone (Chen et al., 2024) pursues large scale federated multi task learning through server client split learning rather than direct parameter averaging. A large general model lives on the cloud while each edge client retains only a lightweight patch embedding module plus a task specific head, reducing the on device parameter count by roughly two orders of magnitude. The GPAggregation

operator first rescales each task gradient against the previous round aggregate and then projects it onto the normal plane of every conflicting gradient before averaging. Its drawbacks include a strong dependence on a pretrained general backbone, the sequential forward backward bottleneck of split learning, and the single point of failure constituted by the cloud model.

FedLPS (Jia et al., 2024) addresses cross client federated multi-task learning through local parameter sharing, in which each client's local model is decomposed into a shared frozen encoder and a task specific predictor. The server distributes only the encoder, every client trains its predictor locally with channel wise L1 norm pruning at a hardware determined ratio, and the server aggregates the pruned shape mismatched predictors by recovering missing channels from the pretrained backbone. The framework jointly handles statistical and system heterogeneity. Its drawbacks include the absence of explicit related task identification, the a priori fixed encoder split point that degrades accuracy when poorly chosen, and the heavy reliance on an ImageNet style pretrained backbone.

TRUST (Wan et al., 2025) performs multi order curriculum distillation for model heterogeneous federated graph learning. Each client trains a private GNN and exchanges only logits and class prototypes with the server, which maintains an omniscient global GNN trained through knowledge distillation. Training proceeds under a multi order curriculum that progressively expands the receptive field, paced by a Progressive Curriculum Node Scheduler that ranks nodes by neighborhood label entropy and class prototype distance. An Adaptive Curriculum Distillation Modulator treats the distillation temperature as a learnable parameter, and a Wasserstein Driven Affinity Distillation module aligns inter class topology through optimal transport. Its disadvantages include non trivial extra cost from Sinkhorn iterations, a heuristic curriculum pacing function that may need per dataset retuning, and the absence of evidence for scalability to large industrial graphs.

Each baseline is re-tuned under the same validation-based selection procedure as FedM3G, which removes tuning bias as a confound in the comparison. The protocol does not declare FedM3G the universal winner. Specialized baselines retain task-specific inductive biases that already cover a portion of the metric space, and FedM3G's advantage lies in accommodating all five bets in parallel through the dual-track public interface.

D Downstream Tasks and Metrics

We evaluate FedM3G on five downstream capabilities that jointly probe graph centric prediction and cross modal alignment. The five tasks decompose into two graph centric objectives and three multimodal objectives, and they are designed so that each task aligns with one track of the dual track design. The first two tasks assess graph centric prediction, while the other three assess multimodal capability. Together they span the dual track design space without redundancy and cover both tracks in isolation. We first describe each task and the metrics used to score it, then provide the formal definition of every metric in a single consolidated subsection. The five-task design has two structural motivations. It covers both tracks in isolation so that ablation results can attribute gains to the correct module. It also covers the cross modal interface that drives the largest single gain in the reported results, since cross modal alignment is precisely the signal that entangled baselines fail to expose. A misalignment between task and track would systematically underperform FedM3G and reveal the value of the dual track decoupling.

D.1 Task Descriptions

NodeCls is the canonical node classification task on the private client subgraph. Every client predicts the category label of each private node using only its local backbone, its local task head, and the public anchor guidance supplied by the server. The principal metric is NodeCls ACC, defined as the proportion of correctly classified nodes over the held out evaluation split. NodeCls probes the semantic track under private graph context and exercises the class prototype aggregation path. The task stresses the ability to recover class boundaries from private features alone while borrowing transferable geometry from the public anchors.

LinkPred is the canonical link prediction task on the private client subgraph. Every client scores candidate edges by combining the embeddings of their endpoint nodes through a task specific edge head, and the held out evaluation split contains positive edges that were masked during local training together with sampled negative edges. The principal metric is LinkPred AUC, defined as the

probability that a randomly chosen positive edge receives a higher matching score than a randomly chosen negative edge. LinkPred probes the topology track under private supervision and verifies that FedM3G preserves the neighborhood-derived signal required for edge level inference. The task stresses the topology aggregation and the SRA topology channel in particular.

Align is the cross modal alignment task on the private client subgraph. Every client projects paired image and text features through a shared projection head and the held out split contains only paired (image, text) instances. The principal metric is Align CLIP, defined as the average cosine similarity between the paired image and text embeddings. Align probes the dual track design under joint modality supervision and is the metric for which entangled baselines lose the largest margin, because their fused representation cannot be cleanly decomposed into a semantic and a topology component.

Retrieval is the cross modal retrieval task on the private client subgraph. The evaluation split contains a set of queries and a set of targets, and the goal is to retrieve the paired target for each query from the candidate set. The principal metric is Retrieval R@5, defined as the fraction of queries whose ground truth target appears in the top five ranked candidates. Retrieval probes the dual track design under ranking supervision and is the metric most sensitive to the alignment between the semantic channel and the topology channel, because ranking requires the two channels to be jointly consistent across the candidate pool.

Match is the cross modal matching task on the private client subgraph. The evaluation split contains paired (image, text) instances and a candidate pool of negative pairs. The goal is to assign a higher matching score to positive pairs than to negative pairs. The principal metric is Match AUC, computed identically to LinkPred AUC but with the candidate set restricted to paired (image, text) instances. Match probes the dual track design under binary discrimination and shares the same topology channel as LinkPred but uses cross modal pairs as candidates, which makes the two AUC metrics complementary probes of the topology and semantic channels respectively.

D.2 Metric Definitions

This subsection provides the formal definition of every metric used in the five tasks above. Each metric is computed on the per client held out evaluation split and averaged across the five clients using uniform weighting.

NodeCls ACC measures the proportion of correctly classified nodes over the entire evaluation set. Let N denote the number of nodes, y_i and $\hat{y}_i$ the ground-truth and predicted labels of the i-th node, respectively, and $\mathbb{I}(\cdot)$ the indicator function. Accuracy provides an intuitive assessment of overall classification performance and is the principal metric for the graph centric category prediction objective.

$$\text{NodeCls ACC} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{y}_i = y_i). \tag{10}$$

LinkPred AUC measures the probability that a randomly chosen positive edge is assigned a higher matching score than a randomly chosen negative edge. By integrating the true positive rate against the false positive rate over all possible thresholds, AUC-ROC provides a threshold independent evaluation of edge ranking robustness. This metric stresses the topological track and verifies that FedM3G preserves the neighborhood-derived signal required for edge level inference.

$$\text{LinkPred AUC} = \int_0^1 \text{TPR}(t)\,\text{dFPR}(t). \tag{11}$$

Align CLIP measures whether paired text and image embeddings share a common geometry and is scored by the average cosine similarity over paired samples. Let $\mathbf{z}_i^{(\text{t})}$ and $\mathbf{z}_i^{(\text{v})}$ denote the normalized text and image embeddings for the i-th paired sample, and $\langle\cdot,\cdot\rangle$ the inner product. A higher score indicates tighter cross modal alignment, which is the geometric signal that the dual track design explicitly preserves and that entangled baselines cannot recover.

$$\text{Align CLIP} = \frac{1}{N}\sum_{i=1}^{N}\frac{\langle \mathbf{z}_i^{(\text{t})}, \mathbf{z}_i^{(\text{v})}\rangle}{\|\mathbf{z}_i^{(\text{t})}\| \cdot \|\mathbf{z}_i^{(\text{v})}\|}. \tag{12}$$

Retrieval R@5 evaluates retrieval effectiveness by measuring the fraction of queries for which the correct target appears within the top five ranked results. Let $\mathcal{Q}$ denote the set of queries and rank_i the rank of the ground truth item for the i-th query. This metric reflects the system's practical utility under a fixed retrieval budget and is the principal metric for cross modal retrieval.

$$\text{Retrieval R@5} = \frac{1}{|\mathcal{Q}|} \sum_{i=1}^{|\mathcal{Q}|} \mathbb{I}(\mathrm{rank}_i \leq 5). \tag{13}$$

Match AUC measures the probability that a randomly chosen positive cross modal pair is assigned a higher matching score than a randomly chosen negative pair. By integrating the true positive rate against the false positive rate over all possible thresholds, the metric provides a threshold independent evaluation of cross modal matching robustness. The candidate set is restricted to paired (text, image) instances and the metric is computed identically to LinkPred AUC under the cross modal restriction.

$$\text{Match AUC} = \int_0^1 \mathrm{TPR}(t)\, \mathrm{dFPR}(t). \tag{14}$$

E Experimental Settings

This section unifies the experimental configuration used throughout the paper. All reported results use seeds 2026, 2027, and 2028 and ten communication rounds. Each round performs ten nominal local epochs of client-side training. The public interface dimension is fixed at 768 for the attention-style backbones and at 256 for the lightweight convolutional backbones. Each client retains its private graph, labels, task head, and modality encoder, and the server receives only the public-graph representations required by the alignment procedure. Checkpoints are selected on the validation split and the held-out test split is evaluated exactly once. Because task adaptivity enlarges the optimizer hyperparameter space, validation-based selection is applied consistently to every compared method, and the same seed set is reused across all seeds.

Per-task optimizer controls. The five downstream tasks differ in convergence rate and dependence on the two tracks. Table 4 reports the base task-adaptive controls used by the technical implementation. The λ_{sem} and λ_{topo} columns are per-track weights on the SRA loss, the learning-rate multiplier is applied on top of the shared backbone schedule, the weight-decay column follows standard practice, and the auxiliary column names the task-specific term included in $\mathcal{L}_{\mathrm{aux}}$. Dataset-specific validation may select a value within the declared search range.

Table 4: Base task-adaptive controls (extended). "Auxiliary" denotes the task-specific term included in $\mathcal{L}_{\mathrm{aux}}$. "LR mult." is the per-task learning-rate multiplier applied on top of the shared backbone schedule.

Task	λ_{sem}	λ_{topo}	**LR mult.**	**WD**	**Auxiliary**
NodeCls	0.080	0.060	1.0	10^{-4}	none
LinkPred	0.005	0.150	1.6	0	link rank
Align	0.015	0.015	2.0	0	modal contrast
Retrieval	0.015	0.020	1.6	10^{-5}	directional rank
Match	0.020	0.030	1.0	10^{-4}	match and prototype

Hyperparameter ranges. Sensitivity sweeps cover the Synergistic Representation Alignment weight $\lambda \in [0.01, 0.5]$, the TKA temperature $\tau_\beta \in [0.05, 0.5]$, and the semantic-gate temperature $\tau_g \in [0.3, 0.9]$. The default values are $\lambda = 0.1$, $\tau_\beta = 0.2$, and $\tau_g = 0.7$. The pre-registered robustness criteria are a macro-average drift below 2% and a worst-task drift below 5%, both of which are met by a wide margin in the reported sweeps.

Parameter groups. Parameters are divided into backbone, task-head, SRA, task-adapter, and public-projection groups, each with its own optimizer controls. Because task-adaptive optimization enlarges the search space, validation selection is applied consistently to all compared methods.

Reported numbers use the configuration that performs best on each client's local validation split, and the test split is evaluated exactly once.

Privacy and coordinate-frame assumptions. The protocol forbids the server from receiving private raw content $\mathcal{R}^n$, labels, model parameters θ_n, or gradients, and instead admits only representations of an ordered public anchor set $\mathcal{V}_p$. This restriction is a data-minimization convention rather than a formal differential-privacy guarantee. In particular, the server could in principle reconstruct content from uploaded representations, and combining FedM3G with differential privacy or secure aggregation is left to future work. The shared public graph $\mathcal{G}_p$ is assumed to contain no nodes or labels from any private client. Because each client f_{θ_n} may choose a different latent basis, $\mathbf{Z}^n$ and $\mathbf{Z}^m$ may differ by an arbitrary orthogonal or even non-orthogonal transformation. FedM3G exploits the fact that Centered Kernel Alignment and the contrastive matching used by SRA are invariant to orthogonal changes of coordinates in their respective score computations. Reconciliation of arbitrarily rotated latent spaces is outside the present scope, and clients that wish to opt in must be willing to share a common reference frame or apply a learned alignment layer at upload time.

Computing. All experiments are executed on a homogeneous GPU cluster. Per-round wall clock for FedM3G is dominated by the Gram-Schmidt decoupling step, which is a single per-node MLP, and by the $O(N^2)$ CKA computation over the public anchor set with $N = |\mathcal{V}_p|$. For the default $N = 512$ anchors, CKA accounts for under five percent of total per-round time.

F Hyperparameter Configuration

This section consolidates every method specific hyperparameter used in the reported runs, together with the search range explored by the validation based selection procedure in Appendix E. Shared schedule parameters (optimizer, base learning rate, communication rounds, local epochs, and seed set) are reused from Appendix E and are not repeated here. The two tables below only list values that are method specific or that we tuned per task. For FedM3G, the chosen value is reported alongside the search range. For the federated baselines, we report the post tuning value that performed best on the local validation split under the same protocol as FedM3G. When a single value is given per parameter, it is reused across all five tasks because the parameter is task agnostic.

F.1 FedM3G Specific Parameters

Table 5 lists only the hyperparameters that belong to the three core modules of the methodology, namely Gram-Schmidt Decoupling, Task Kernel Alignment, and Synergistic Representation Alignment. Each row reports the mathematical symbol that appears in the corresponding equation of the methodology section, the chosen value selected by validation based tuning, and the search range that the procedure explored. Hyperparameters introduced outside these three modules (such as the public interface dimension, the aggregation temperature, and the DTARC refinements) are intentionally omitted because they are not part of the three core module specifications.

Table 5: FedM3G hyperparameters for the three core modules (math symbol, chosen value, search range).

Module	Parameter	Symbol	Chosen	Search Range
GS Decoupling	context bypass scale	η	0.10	[0.0, 0.5]
TKA	peer-weight temperature	τ_β	0.20	[0.05, 0.5]
TKA	curriculum start k	k_0	3	[1, $N-1$]
SRA	contrastive temperature	τ_s	0.20	[0.05, 0.5]
SRA	teacher consensus weight	ξ_q^n	0.5	[0.0, 1.0]
SRA	public anchor batch size	B	64	[32, 256]

F.2 Baseline Specific Parameters

Table 6 reports the hyperparameters that are specific to each federated baseline after re tuning under the same validation based selection procedure as FedM3G.

Table 6: Baseline specific parameters (chosen value and search range).

Method	Parameter	Chosen	Search Range
FedAvg	local epochs E	10	[1, 20]
FedAvg	client fraction C	1.0	[0.5, 1.0]
MOON	contrast weight mu	0.1	[0.01, 1.0]
MOON	temperature	0.5	[0.1, 1.0]
MOON	projection dim	256	[64, 512]
CreamFL	temperature	0.5	[0.1, 1.0]
CreamFL	distill weight	0.05	[0.0, 0.5]
CreamFL	contrastive weight	0.05	[0.0, 0.5]
CreamFL	inter-intra weight	0.5	[0.0, 1.0]
CreamFL	public batch size	64	[32, 256]
FedMuscle	hard negative alpha max	1.0	[0.0, 2.0]
FedMuscle	temperature	0.2	[0.05, 0.5]
FedMuscle	public batch size	64	[32, 256]
FedProto	prototype loss weight	0.05	[0.0, 0.5]
FedHCA^2	encoder conflict threshold	0.3	[0.0, 0.5]
FedHCA^2	decoder cross attn heads	4	[1, 8]
FedHCA^2	personalization alpha	0.5	[0.0, 1.0]
FedBone	GPAggregation rescale tau	0.5	[0.0, 1.0]
FedBone	cloud backbone	Swin-T (frozen)	{Swin-T, ViT-B}
FedLPS	encoder split n	0.25N	[0.1N, 0.5N]
FedLPS	pruning ratio rho	0.5	[0.2, 0.8]
FedLPS	predictor lr mult	4.0	[1.0, 8.0]
TRUST	PCNS pacing k	8	[4, 16]
TRUST	ACDM tau init	1.0	[0.5, 2.0]
TRUST	WDAD Wasserstein reg	0.1	[0.01, 1.0]
TRUST	curriculum start k	1	[1, N-1]
TRUST	curriculum end k	N-1	[N-1, N-1]

G Full Per-Backbone Results

Table 7 reports the complete results on the MHGAT backbone under the Louvain partition, complementing the MGAT results in Table 1. Best per dataset-metric is in **bold**, second best is underlined. † marks baselines that fail to converge on the indicated task.

Per dataset pattern. On **Movies**, FedM3G attains the best score on every metric except LinkPred AUC where TRUST trails by 0.5 percentage points. The cross modal metrics show the largest margins because Movies has the broadest category coverage and therefore benefits the most from cross-client geometry transfer. FedProto is the second best on NodeCls and Match, which is consistent with its class prototype specialization being well matched to a dataset with balanced label priors. FedHCA^2 is the second best on Align because its hierarchical aggregation preserves cross modal geometry more faithfully than flat prototype averaging. The retrieval task is the cleanest separator because the dual track design is the only entry that exposes both semantic and topology channels through a ranking loss.

On **Grocery**, FedM3G attains the best score on LinkPred, Align, and Retrieval, while FedProto attains the best NodeCls and TRUST attains the best Match. Grocery has the most skewed label prior among the three datasets, which explains why FedProto's class conditional prototype exchange outperforms FedM3G on NodeCls by a small margin: the prototype exchange directly compensates for label skew by transferring class statistics across clients, while FedM3G's public anchor interface only addresses representation geometry. The Match result on Grocery is the only task where FedM3G underperforms the strongest baseline, which we attribute to Grocery's text dominant product descriptions limiting the cross modal signal that drives the Match AUC. Notably, FedBone fails to converge on NodeCls and LinkPred on Grocery, which reflects its strong dependence on a pre-trained general backbone that does not transfer to fine-grained grocery categories.

On **Toys**, FedM3G attains the best score on every metric, with the largest margin on Retrieval where FedM3G exceeds the second best by roughly ten percentage points. Toys has the most balanced

Table 7: Results on the **MHGAT** backbone (Louvain partition). Best per dataset-metric in **bold**, second best underlined. † marks baselines that fail to converge on the indicated task. FedM3G attains the best score on 13 of 15 dataset-metric combinations.

Method	NodeCls ACC	LinkPred AUC	Align CLIP	Retrieval R@5	Match AUC
Movies					
FedAvg	0.4478	0.9110	0.8596	0.1367	0.6888
MOON	0.3412	0.7574	0.8470	0.2255	0.7418
CreamFL	0.3200	0.5620†	0.8453	0.2260	0.7313
FedMuscle	0.3921	0.8932	0.9487	0.8962	0.7883
FedProto	0.4346	0.9269	0.9666	0.9027	0.8384
FedHCA^2	0.4370	0.9271	0.9853	0.1427	0.7457
FedBone	0.3200	0.4901†	0.4212†	0.4571	0.7491
FedLPS	0.3200	0.8453	0.9109	0.3154	0.7059
TRUST	0.3472	0.9415	0.9448	0.9122	0.8272
FedM3G	**0.5206**	**0.9464**	**0.9912**	**0.9885**	**0.8831**
Grocery					
FedAvg	0.6950	0.9196	0.8082	0.0691	0.6484
MOON	0.4169	0.7488	0.8540	0.1444	0.6993
CreamFL	0.4961	0.5476†	0.8382	0.1594	0.6964
FedMuscle	0.7488	0.8850	0.9438	0.6957	0.7857
FedProto	**0.8202**	0.9331	0.9615	0.8971	0.8407
FedHCA^2	0.7677	0.9182	0.9813	0.0826	0.7030
FedBone	0.1584†	0.4578†	0.3947†	0.3502	0.7460
FedLPS	0.3124	0.8500	0.8052	0.5159	0.7318
TRUST	0.5724	0.9095	0.9351	0.8874	**0.8794**
FedM3G	0.8165	**0.9346**	**0.9894**	**0.9420**	0.8635
Toys					
FedAvg	0.6287	0.9546	0.8410	0.1103	0.6843
MOON	0.4623	0.6631	0.8473	0.2299	0.7610
CreamFL	0.4016	0.5027†	0.8407	0.2395	0.7574
FedMuscle	0.6754	0.9235	0.9496	0.7629	0.8007
FedProto	0.6954	0.9401	0.9736	0.8555	0.8553
FedHCA^2	0.6487	0.9325	0.9873	0.1393	0.7215
FedBone	0.3126	0.4967†	0.5242†	0.4573	0.7761
FedLPS	0.3214	0.8213	0.8667	0.4203	0.7450
TRUST	0.4347	0.9264	0.9548	0.8808	0.8645
FedM3G	**0.7426**	**0.9798**	**0.9923**	**0.9811**	**0.8713**

visual and textual signals among the three datasets, which is precisely the regime in which the dual track design delivers the strongest gains. FedProto is again the second best on NodeCls, while TRUST is the second best on Retrieval and Match. FedHCA^2 is the second best on Align. The pattern confirms that Toys is the cleanest benchmark for FedM3G's full design space and that no single baseline can match FedM3G on all five tasks simultaneously.

Per metric pattern. Across all three datasets and the MHGAT backbone, FedM3G wins on 13 of 15 dataset metric combinations. The two losses are Grocery NodeCls (FedProto) and Grocery Match (TRUST). The Align and Retrieval metrics are the cleanest FedM3G wins, with margins of 1.0 to 11.5 percentage points over the second best baseline. This is precisely the prediction of the dual track design, because the entangled baselines cannot expose both semantic and topology channels through a single aligned loss. NodeCls and LinkPred show smaller margins because these tasks are dominated by the backbone's local discriminative power, which is similar across all methods under the same backbone family.

Failure cases. FedBone fails to converge on three of five tasks (NodeCls, LinkPred, Align) on Grocery and on the same three tasks on Toys, and on all three of those tasks on Movies. The failure pattern confirms that FedBone's dependence on a pretrained general backbone creates a hard upper bound on transfer to fine grained product domains, while FedM3G's public anchor interface inherits the multimodal backbone's pretrained encoder and therefore does not suffer from the same failure

mode. CreamFL fails to converge on LinkPred on all three datasets because its prediction level distillation on the public batch does not preserve the edge level signal required for link prediction. FedLPS underperforms on NodeCls because its frozen encoder split point is poorly matched to the multimodal backbone architecture, which limits the adaptation capacity of the task specific predictor. MOON's three forward passes per local step limit its effective local update count and explain its consistent underperformance on NodeCls and LinkPred relative to FedAvg.

Backbone effect. Comparing the MHGAT backbone table against the MGAT results in the main paper, the relative ordering of methods is preserved across both backbones, which confirms that FedM3G's gains are not an artifact of a specific attention pattern. The absolute numbers shift slightly because MHGAT's heterogeneous attention captures meta path semantics that benefit the Align metric by 1 to 2 percentage points across all methods. FedM3G's gains over the strongest baseline are slightly larger on MHGAT than on MGAT for the Align metric, which suggests that the dual track design composes well with relation aware attention.

H Additional Results and Analysis

H.1 Empirical Validation of Task Kernel Alignment

The empirical study in Sec. 1 evaluates TKA against realized transfer rather than private task labels. All clients first complete an identical local-only warm-up. Before the first cross-client alignment, we record $\beta_{\mathrm{sem},nm}$ and $\beta_{\mathrm{topo},nm}$ from the ordered public anchors for every client pair. For each ordered donor-recipient pair (m, n), we keep client n's private split, initialization, optimizer, training budget, and seed fixed and compare two runs. The first run trains n locally. The second supplies SRA guidance exclusively from donor m, with multi-peer aggregation and public consensus disabled. We define the realized transfer gain as

$$\Delta_{n\leftarrow m} = P_{n\leftarrow m} - P_n^{\mathrm{local}}, \tag{15}$$

where P is the recipient's native held-out metric. Positive and negative values therefore represent beneficial and harmful transfer, respectively.

We form ordered pairs across task, domain, and backbone combinations. Since the server selects peers separately for each recipient, we compute Spearman's ρ and Kendall's τ between TKA and Δ within each recipient and then aggregate the statistics across recipients. We additionally report AUROC for predicting $\mathbb{1}[\Delta > 0]$, negative-transfer frequency, and the mean transfer gain among the top-k donors. Recipient-level cluster bootstrap provides confidence intervals without treating multiple donors for the same recipient as independent samples. Random ranking, representation cosine similarity, and uncentered kernel alignment serve as label-free controls. Ranking by measured Δ provides an oracle upper bound.

H.2 Scope of Task Kernel Alignment

TKA relies on a public-probe informativeness assumption, which states that client pairs with greater beneficial transfer tend to induce more similar centered kernels on the same public anchors. This assumption motivates CKA as a ranking signal, but it is not an identifiability theorem, since CKA may also respond to domain, architecture, and optimization. Our direct validation therefore combines pairwise transfer measurements with public-probe scores and tests whether $\beta_{q,nm}$ predicts the sign and magnitude of transfer. The controlled single-donor protocol in Appendix H.1 performs this test within each recipient, and the single-donor analysis in Sec. 5.3 aggregates the result across recipients.

H.3 Dynamic Task Affinity Refinement (DTARC) Components

The raw track-wise CKA matrix may fluctuate before local representations have stabilized. **Temporal Affinity Smoothing (TAS)** maintains $\beta_q^{(t)} = \rho\beta_q^{(t-1)} + (1-\rho)\beta_{q,\mathrm{raw}}^{(t)}$, where $\rho = 0.4$. **Curriculum Affinity Scheduling (CAS)** uses the mask in Eq. 6. Starting from k_0 peers, it increases k_t linearly toward $N-1$ during the curriculum horizon. Unlike a softmax alone, this mask gives excluded peers exactly zero weight, which prevents low-compatibility peers from contaminating the alignment target. **Asymmetric Task Influence (ATI)** is an optional heuristic refinement

that augments the symmetric CKA matrix with directional terms $\mathrm{ATI}_{nm} = \beta_{nm} - \beta_{mn}$ to capture one-sided transfer dynamics. It is not interpreted as a causal transfer estimator and can be disabled independently, since public representations alone do not identify directional task benefit.

H.4 Failure Cases and Boundary Conditions

Across all backbones and datasets, FedM3G's largest margins appear on cross-modal alignment, while its smallest margins appear on tasks that benefit from task-specific inductive biases already present in specialized baselines. FedBone's pretrained retrieval head, MOON's contrastive matching head, and FedProto's class prototypes each cover a portion of the metric space that FedM3G approaches but does not exceed. The mixed-source experiment in Sec. 5.3 confirms that the dual-track public interface remains effective under joint domain and task heterogeneity, but the protocol does not establish generalization to an unseen target domain. The frozen modality encoders inherited from the pretrained-encoder recipe also imply that FedM3G carries the perceptual backbone's inductive bias on raw modality content, which is a deliberate trade-off in favor of communication efficiency and training stability.

H.5 Time and Space Complexity

We summarize the per round time and space complexity of FedM3G and contrast it with the strongest baselines. Let N denote the number of participating clients, B the public anchor batch size, D the public interface dimension, L the number of graph encoder layers, $V_p = |\mathcal{V}_p|$ the total number of public anchors, V_n the number of nodes on client n, and E_n the number of edges on client n.

Communication cost. Each round every client uploads a single tensor pair $\mathbf{Z}^n_{\mathrm{sem}} \in \mathbb{R}^{B\times D}$ and $\mathbf{Z}^n_{\mathrm{topo}} \in \mathbb{R}^{B\times D}$, plus class prototypes for graph centric tasks and the edge prototype for link prediction. The total per round uplink is $O(BD+C+1)$ floats per client, where C is the number of classes seen on the client. The server returns only the public anchor guidance tensor, which is also $O(BD)$. There is no downlink of parameters, which removes the heaviest communication term present in parameter averaging baselines such as FedAvg, FedProto, FedBone, FedHCA^2, and FedLPS.

Per round computation. Local training costs $O(LE_nD + LV_nD^2)$ per epoch for the multimodal graph encoder, identical to the centralized backbone and to every multimodal baseline. Public anchor encoding through the backbone and the GS decoupling module costs $O(LBD + BD^2)$ per client, which is negligible relative to the local epoch because $B \ll V_n$ in practice. Computing the CKA matrix across all ordered client pairs costs $O(N^2B^2D)$ per track, giving a server side TKA cost of $O(N^2B^2D)$ summed over the two tracks. The DTARC refinements add a $O(N^2)$ temporal EMA update per track per round and a $O(N \log N)$ curriculum top k sort that is dominated by the CKA cost. The server side public anchor distillation performs one epoch of combined MSE and symmetric InfoNCE loss, costing $O(B^2D)$ per client.

Total per round cost. Summing the local training, public anchor encoding, TKA, DTARC, and public distillation gives a per round complexity of

$$\mathcal{C}_{\mathrm{FedM3G}} = O\big(NELE_nD + NLV_nD^2 + N^2B^2D + NB^2D\big)\,, \tag{16}$$

which is dominated by the local training term for any realistic configuration in which $V_n \gg B$ and $E_n \gg B^2$.

Space complexity. The server keeps only the per client public anchor embeddings, the running TAS smoothed matrices, and the curriculum state, giving a server footprint of $O(NBD+N^2+N)$. Each client keeps its local backbone, task head, GS decoupling module, task adapter, and projection head, giving a per client footprint of $O(|\theta| + LD^2 + BD)$, where $|\theta|$ is the backbone size.

Comparison with the strongest baselines. Parameter averaging baselines such as FedAvg, FedProto, FedBone, FedHCA^2, and FedLPS additionally transmit the full parameter tensor $|\theta|$ per round. CreamFL transmits public predictions at every round, and TRUST transmits logits and class prototypes in addition to the parameter tensor. FedM3G's $O(BD)$ uplink is strictly smaller than the parameter uplink of any baseline once the backbone exceeds $D \approx 768$ parameters, which is the regime of every multimodal backbone in our experiments. The TKA computation dominates the server side cost in FedM3G, while parameter aggregation dominates the server side cost of every parameter averaging baseline. For our default configuration of $N = 6$, $B = 64$, and $D = 256$,

the TKA cost is roughly 6×10^7 FLOPs per round per track, which is below five percent of total per round time in our profiling and well within the communication and computation budget of a heterogeneous federated deployment.