LEARNING TO DELETE: TASK-GUIDED MESSAGE-EFFECT INTERVENTION FOR HETEROPHILOUS MULTIMODAL GRAPH LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal graph models propagate multiple forms of node information over a shared topology, yet the utility of an incoming relation can vary with both the receiver and modality. Uniform aggregation cannot express this difference, whereas broad suppression may discard useful content or uniformly shrink neighborhood input. A controlled motivating study further shows that greater connectivity across classes coincides with lower classification performance, with markedly different sensitivity across graphs and host models. We propose **Message Effect Intervention (MEI)**, which learns the task loss contrast induced by removing one exposed modality message from a frozen host. MEI aggregates modality specific estimates into an edge score, applies task specific candidate filtering available at deployment, changes at most one incoming edge bundle per receiver, and rescales retained coefficients to preserve their total mass. Training class prototypes provide the concrete filter for node classification settings. Its fixed message interface supports both the reference CM-MGNN host and compatible multimodal hosts without replacing their representations. The deployed operator admits structural guarantees for recovery, bundle coherence, receiver specific feasibility, coefficient mass preservation, and protected computation paths. On the original task settings, CM-MGNN improves over the strongest displayed baseline by 4.85% in Movies Macro F1, 3.46% in Bili Dance MRR, and 3.51% in CIDEr. MEI also improves every displayed compatible host pair under the fifteen percentage point condition.

1 INTRODUCTION

Multimodal attributed graphs connect topology with node information such as text and images. Their message passing models must nevertheless operate when connected nodes have different labels (Zhu et al., 2020; Platonov et al., 2023). The messages carried by one relation may also support its receiver to different degrees across modalities (Zhu et al., 2025; Yan et al., 2025). Uniform aggregation cannot express this variation, while broad relation suppression may remove useful content and shrink the total neighborhood input available to the receiver at the active aggregation layer.

Existing attention and graph structure methods reweight or revise topology (Veličković et al., 2018; Li et al., 2023b), while multimodal models coordinate multiple views (Guo et al., 2025; Zhu et al., 2026). Their control scores are generally not trained to reflect the receiver loss change caused by removing one exposed modality message. Explanation scores also do not specify a coherent physical relation change, a receiver action limit, or preservation of incoming coefficient mass. Effective intervention therefore requires receiver specific diagnosis, coherent action on the modality bundle of a physical edge, and bounded deployment that separates message composition from a uniform reduction in aggregation scale across all retained messages at the same receiver.

The controlled study in Section 4 makes this need concrete. Classification declines as cross class connectivity increases on Movies and Grocery, yet the severity differs across graphs and host models. This pattern shows that incoming graph information matters to the task, but it neither identifies an individually harmful message nor implies that every cross class relation should be suppressed. A rule based only on cross class status therefore cannot provide the required selectivity. This observation instead makes one research question immediate. *How can a frozen multimodal graph*

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.

model translate message level loss diagnosis into a selective, physically coherent, and scale preserving edge intervention at the deployed replay boundary without replacing the frozen host?

We answer with Message Effect Intervention (MEI). At a fixed interface of a frozen host, MEI learns the task loss contrast induced by removing one modality specific message. It aggregates these estimates into physical edge scores, applies candidate filtering available during deployment, changes at most one edge bundle per receiver, and rescales retained coefficients. Frozen replay grounds diagnosis in the host computation. Bundle action prevents bypass through another modality. Bounded compensated selection limits deployment without treating singleton effects as joint utility. CM MGNN provides the reference realization, while compatible hosts retain their own representations and task context. The analysis establishes structural operator properties. Experiments separately evaluate predictive performance, host reuse, component contributions, and hyperparameter sensitivity.

Our contributions follow this formulation-to-evaluation structure.

- **Problem formalization.** We formulate task generic message diagnosis at a frozen host interface and connect modality resolved singleton contrasts to coherent physical edge actions under receiver wise budget and coefficient mass constraints. The controlled study further exposes graph and host dependent sensitivity to cross class connectivity.
- **Method.** We develop MEI to translate task specific loss contrasts into task aware candidate filtering and compensated bundle replay while preserving the host representations and declared uncontrolled paths at the exposed replay boundary.
- **Evaluation.** On the original task settings, CM MGNN improves over the strongest displayed baseline by 4.85% in Movies Macro F1, 3.46% in Bili Dance MRR, and 3.51% in CIDEr as reported in Tables 1 and 4. MEI also improves every displayed compatible host pair under the fifteen percentage point condition reported in Table 6 and Figure 3.

2 Related Work

Multimodal graph learning. Multimodal graph models combine heterogeneous attributes with relational context (Ektefaie et al., 2023) through fusion, distinct topology views, or task adaptive propagation (Guo et al., 2025; Zhu et al., 2026). Their objective is to construct useful representations from topology and several modalities. MEI addresses a complementary control problem by estimating the task loss change caused by removing one exposed modality message, then translating that diagnosis into a selective relation action at the exposed host interface.

Heterophily aware propagation and structure learning. Methods for heterophilous graphs separate ego and neighbor signals, learn signed propagation, reweight neighbors, or revise connectivity (Zhu et al., 2020; Chien et al., 2021; Veličković et al., 2018; Li et al., 2023b). Their weights or edits need not preserve incoming coefficient mass or the cross modality coherence of a physical relation. MEI instead controls a bounded number of edge bundles and compensates retained coefficients, so its action changes message composition without relying on lower total neighborhood weight.

Graph message attribution and intervention. GNNExplainer and PGExplainer identify influential subgraphs, while GraphMask learns message gates for frozen predictions (Ying et al., 2019; Luo et al., 2020; Schlichtkrull et al., 2021). Counterfactual explainers seek prediction altering edits (Bajaj et al., 2021; Lucic et al., 2022). Their attribution objectives do not themselves define a bounded multimodal edge action with preserved aggregation scale. Because inferred importance depends on the chosen ablation operator (Li & Janson, 2024), a singleton training target cannot be equated with the compensated bundle action. This separation prevents an isolated diagnostic score from being presented as joint deployment utility. More broadly, explanation objectives and protocols shape what scores can establish (Amara et al., 2022). A diagnosis may identify information used by a prediction, while an intervention policy must determine which physical relation can change, how its modalities change together, and how the receiver retains aggregation scale. MEI therefore uses predicted singleton contrasts as diagnostic rankings rather than direct deletion commands. Its candidate set, receiver budget, shared bundle coefficient, and compensation rule jointly determine policy feasibility. Training class prototypes provide only a restrictive filter available without deployment labels (Snell et al., 2017; Ding et al., 2020). The effect estimate supplies task grounded ranking, while compensated bundle replay defines the intervention executed at the receiver.

3 PRELIMINARIES AND PROBLEM FORMULATION

Multimodal graph and task contract. Let $\mathcal{M}$ be the modality set. A multimodal attributed graph (MAG) is $\mathcal{G} = \langle \mathcal{V}, \mathcal{E}, \{\mathbf{X}^{(m)}\}_{m\in\mathcal{M}} \rangle$, where $\mathcal{V}$ is the node set, $\mathcal{E}$ is the directed message edge set, and $\mathbf{X}^{(m)}$ is the modality m feature matrix (Ektefaie et al., 2023; Zhu et al., 2025; Yan et al., 2025). Edge $e = (u, v)$ connects sender u to receiver v. Its exposed bundle $\mathcal{Q}_e = \{\mathbf{q}_{e,m} : m \in \mathcal{M}\}$ contains one message $\mathbf{q}_{e,m}$ per modality. MEI diagnoses these messages separately but acts on the bundle. The incoming set is $\mathcal{E}_v = \{(u, v) \in \mathcal{E}\}$. Self relations remain visible but cannot be selected, preserving the receiver path outside any intervention candidate set (Zhu et al., 2020).

Let $\mathfrak{T} = \{\mathrm{NC}, \mathrm{LP}, \mathrm{MR}, \mathrm{G2T}, \mathrm{G2I}\}$ contain node classification, link prediction, modality retrieval, graph to text generation, and graph to image generation. Task $\tau \in \mathfrak{T}$ has instance set $\mathcal{D}_\tau$. An instance $z \in \mathcal{D}_\tau$ has reference r_z. The frozen host has parameters $\theta^\star$ and output $F_{\theta^\star}(\mathcal{G}; z)$, while ℓ_τ is its scalar loss. The instance loss is

$$\mathcal{L}_\tau(z) = \ell_\tau(F_{\theta^\star}(\mathcal{G}; z), r_z)\,. \tag{1}$$

This definition separates the host loss from evaluation metrics. The reference may be a class, link, modality match, text, or image signal. The active receiver set $\mathcal{R}_\tau(z) \subseteq \mathcal{V}$ contains every node state used to predict z, so a predicted link differs from a message edge considered for intervention.

A compatible host freezes $\theta^\star$ before target generation, supplies the loss and active receivers, and exposes messages at one replayable boundary. Identity coefficients recover its open trace. Only training statistics and deployment context support diagnosis and intervention.

Task specific singleton diagnosis. For $v \in \mathcal{R}_\tau(z)$ and $e = (u, v) \in \mathcal{E}_v$, let $F_{\theta^\star}^{-(e,m)}$ be the frozen replay that removes only $\mathbf{q}_{e,m}$. We define its singleton contrast as

$$\Delta^{(\tau)}_{z,e,m} = \ell_\tau\Big(F_{\theta^\star}^{-(e,m)}(\mathcal{G}; z), r_z\Big) - \ell_\tau(F_{\theta^\star}(\mathcal{G}; z), r_z)\,. \tag{2}$$

A negative value means that isolated removal lowers the same instance loss. This contrast is specific to the task, instance, host, receiver, and modality. It does not establish cause and effect, bundle utility, a joint deletion effect, or a deployment guarantee. The estimator appears in Section 5.

Feasible task aware bundle intervention. Let $\mathcal{D}_{\mathrm{tr}}$ be training information and $\mathcal{I}_{\mathrm{dep}}$ be deployment context. For active receiver v, task adapter Γ_τ constructs candidate set $\mathcal{C}_v \equiv \mathcal{C}_{\tau,z,v}$ as

$$\mathcal{C}_v = \Gamma_\tau(v, z; \mathcal{G}, \mathcal{D}_{\mathrm{tr}}, \mathcal{I}_{\mathrm{dep}}) \subseteq \mathcal{E}_v \setminus \{(v, v)\}. \tag{3}$$

The exclusion protects the receiver path. The adapter makes eligibility task feasible, with class prototypes as one optional realization for node classification.

Let $\widehat{\Delta}^{(\tau)}_{z,e,m}$ be the estimated contrast and $\mathbf{c}_{\tau,z,v}$ be receiver context. Aggregator A_τ defines physical edge score $s_{\tau,z,e}$ as

$$s_{\tau,z,e} = A_\tau\Big(\{\widehat{\Delta}^{(\tau)}_{z,e,m}\}_{m\in\mathcal{M}}, \mathbf{c}_{\tau,z,v}\Big)\,. \tag{4}$$

Aggregation supports one relation action from modality diagnoses. Let B_τ be the receiver budget, $\mathcal{S}_v$ the selected set, $g_{e,m}$ a raw modality coefficient, g_e its shared bundle value, and $\bar{g}_{e,m}$ its compensated value. The intervention contract is

$$\begin{aligned} &\mathcal{S}_v \subseteq \mathcal{C}_v, && |\mathcal{S}_v| \le B_\tau, \\ &g_{e,m} = g_e \quad \forall m \in \mathcal{M}, && \sum_{e\in\mathcal{E}_v} \bar{g}_{e,m} = |\mathcal{E}_v| \quad \forall m \in \mathcal{M}. \end{aligned} \tag{5}$$

Equation 5 collects three coupled requirements. Selection stays inside the task candidate set and receiver budget, each physical edge shares one coefficient across modalities, and compensation preserves incoming coefficient mass for every modality at the receiver.

Learning objective. MEI learns an estimator and policy that turn singleton contrasts into receiver specific bundle interventions without validation or test supervision. Candidate construction remains a task adapter rather than a host requirement. This separation keeps host compatibility distinct from task specific eligibility. The realization appears in Section 5.

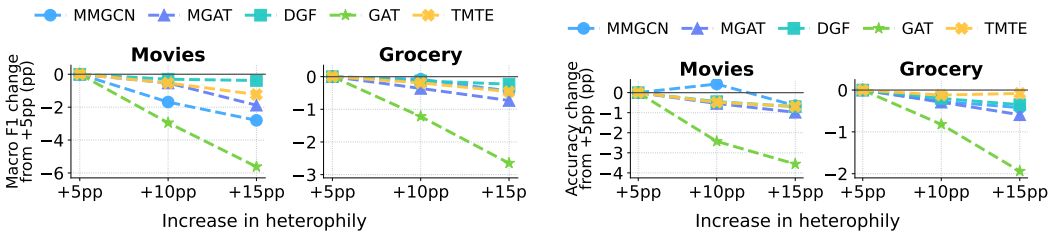


Figure 1: Changes in Macro F1 and Accuracy from the +5pp baseline as heterophily rises. Each panel uses its own scale and supports descriptive association only.

4 Empirical Studies

Research question and controlled setting. How does classification change as the same multimodal nodes receive more input from other classes? On Movies and Grocery (Yan et al., 2025), nested endpoint swaps guided by labels raise the share of nonself edges across classes by five, ten, and fifteen percentage points while preserving node features, edge count, indegree, and outdegree. We retrain GAT, TMTE, MMGCN, MGAT, and DGF (Veličković et al., 2018; Zhu et al., 2026; Hu et al., 2021; Tao et al., 2020; Zheng et al., 2026) at each level. Appendix B details this construction and label boundary, while later appendix sections provide operator proofs, host details, and experimental specifications. This controls cross-class exposure, not higher-order topology or real-world causality.

Observation 1. Broad endpoint sensitivity. From five to fifteen percentage points, Macro F1 falls in all ten dataset and method combinations. Accuracy at fifteen points is also below its five point value throughout. Across two metrics and five classifiers, this shared endpoint deterioration motivates diagnosing prediction-specific incoming information rather than responding to a graph-wide heterophily level. Nonmonotone ten point results mean that endpoint agreement defines neither a stepwise law nor uniformly harmful cross-class relations.

Observation 2. Severity depends on graph and host. Declines are greater on Movies than Grocery, and classifiers follow distinct trajectories under the same sequence of heterophily increments. Thus, a single graph-level response would be poorly calibrated across both datasets and host computations; diagnosis should instead be conditioned on the deployed host and its prediction context. Because edge count, indegree, and outdegree remain fixed, the contrast cannot be attributed to smaller neighborhoods or reduced degree budgets. This pattern is consistent with shifting task relevance, but identifies neither a host mechanism nor a harmful message.

Design implication and boundary. Together, the observations motivate an intervention that is selective, prediction-aware, and tied to the computation it will modify. Varying severity argues against a fixed topology rule, while nonuniform trajectories argue against suppressing every relation across classes. Section 5 therefore diagnoses exposed messages in a frozen host, combines modality-level evidence before acting on a physical edge, and constrains the resulting change at each receiver. This design addresses the study's requirements, but the semisynthetic evidence selects no edge or modality and establishes neither real-graph causality nor deployment benefit; later task-level evaluation tests that benefit across the compatible hosts and tasks considered in this work.

5 Method

The motivating study reveals graph and host dependent sensitivity but not which incoming messages to change. Figure 2 presents MEI's diagnosis, selection, and replay pipeline. We train and freeze an open CM MGNN reference, using singleton replays as estimator targets. At deployment, MEI estimates effects in this frozen context, filters candidates, selects at most one receiver bundle, and replays aggregation with compensated coefficients. The shared boundary aligns diagnosis with action. Training labels supply replay targets and prototypes; scoring and replay on new instances are label-free. Diagnosis remains modality resolved, while action stays coherent across each physical edge selected for controlled replay at the deployed receiver boundary.

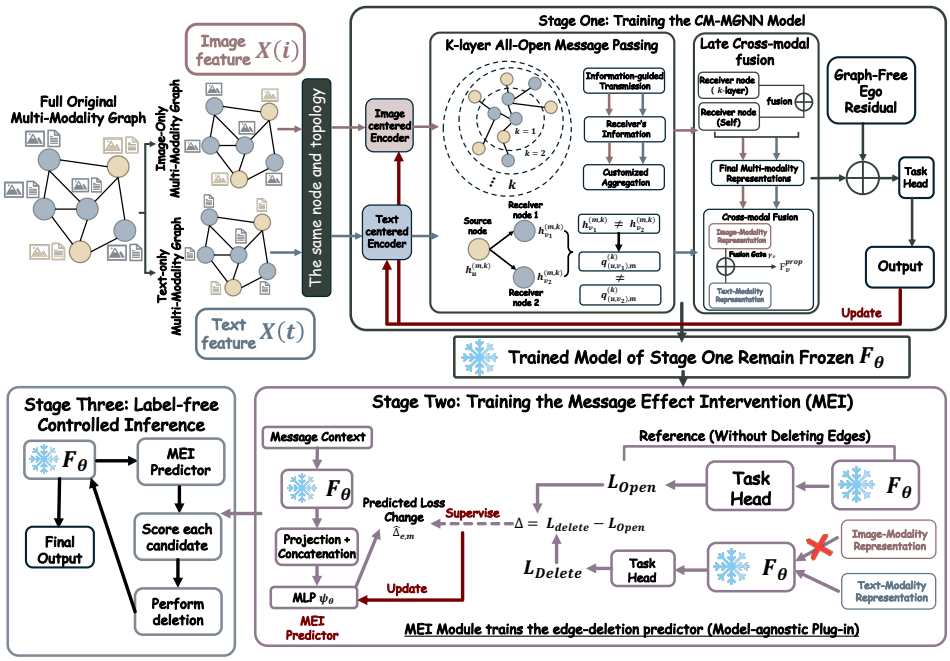


Figure 2: CM MGNN with MEI. The frozen host exposes modality messages. MEI estimates their effects, selects one receiver bundle, and replays the final layer with compensated coefficients.

5.1 Reference CM MGNN Host

Let the host have K message passing layers with index $k \in \{0, \dots, K-1\}$. At layer k, $\mathbf{h}_u^{(m,k)}$ and $\mathbf{h}_v^{(m,k)}$ are sender and receiver modality states, $\omega_{uv}^{(k)}$ is their fixed structural weight, $\varphi_m^{(k)}$ is the message map, and $\|$ denotes concatenation. Let $\mathbf{W}_{\text{self},m}^{(k)}$ and $\mathbf{b}_{\text{self},m}^{(k)}$ be the receiver transform and bias, and let $\rho_m^{(k)}$ be the activation. The exposed contribution and open update form the two line system

$$\mathbf{q}_{e,m}^{(k)} = \omega_{uv}^{(k)}\, \varphi_m^{(k)}\left([\mathbf{h}_u^{(m,k)} \| \mathbf{h}_v^{(m,k)}]\right),$$
$$\mathbf{h}_v^{(m,k+1)} = \rho_m^{(k)}\left(\underbrace{\mathbf{h}_v^{(m,k)} + \mathbf{W}_{\text{self},m}^{(k)}\mathbf{h}_v^{(m,k)} + \mathbf{b}_{\text{self},m}^{(k)}}_{\mathbf{r}_{v,m}^{(k)}} + \sum_{e\in\mathcal{E}_v} \mathbf{q}_{e,m}^{(k)}\right). \tag{6}$$

The underbrace defines the explicit receiver path $\mathbf{r}_{v,m}^{(k)}$, while the preliminary message $\mathbf{q}_{e,m}$ is the exposed final layer instance $\mathbf{q}_{e,m}^{(K-1)}$. Equation 6 therefore separates fixed receiver computation from edge indexed contributions at the replay boundary. The estimator observes the deployed contribution that replay can attenuate, while changing one edge cannot erase receiver information outside the message sum. Structural normalization, fusion, the ego residual, and training are detailed in Appendix E for both the reference and compatible hosts.

5.2 Task Specific Message Effect Estimation

The contrast in Equation 2 is a supervised target. Removing one final layer message from the frozen host records a loss response rather than an internal importance weight. A negative value favors isolated removal but does not establish cause and effect or bundle utility.

The context $\mathbf{c}_{\tau,z,v}$ contains the receiver state and open prediction for node classification or available task context for other adapters. Learned maps $\mathbf{P}_q$, $\mathbf{P}_s$, $\mathbf{P}_r$, and $\mathbf{P}_c$ project the message, sender, receiver, and context. Embeddings $\mathbf{e}_m$ and $\mathbf{e}_{K-1}$ encode modality and layer. Estimator ψ_ϕ with

parameters ϕ predicts $\widehat{\Delta}^{(\tau)}_{z,e,m}$ as

$$\widehat{\Delta}^{(\tau)}_{z,e,m} = \psi_\phi\Big(\mathbf{P}_q\mathbf{q}^{(K-1)}_{e,m} \| \mathbf{P}_s\mathbf{h}^{(m,K-1)}_u \| \mathbf{P}_r\mathbf{h}^{(m,K-1)}_v \| \mathbf{P}_c\mathbf{c}_{\tau,z,v} \| \mathbf{e}_m \| \mathbf{e}_{K-1}\Big). \tag{7}$$

The inputs identify what arrives, which endpoints interact, and why it matters to the prediction. Projections align their dimensions and embeddings distinguish modalities and layers.

Let $\widehat{\Delta}$ and Δ abbreviate predicted and target contrasts. Let Huber be robust regression loss, BCE binary cross entropy, $\mathbb{I}$ the indicator, and $\lambda_{\mathrm{sign}} > 0$ the sign weight. The objective is

$$\mathcal{L}_{\mathrm{MEI}} = \mathrm{Huber}(\widehat{\Delta}, \Delta) + \lambda_{\mathrm{sign}}\,\mathrm{BCE}\Big(\widehat{\Delta}, \mathbb{I}[\Delta > 0]\Big). \tag{8}$$

Huber regression orders contrast magnitudes while limiting extreme targets, and binary cross entropy reinforces whether removal raises or lowers loss. Their combination supports ranking without treating estimates as exact, while task context lets one message affect different predictions. Near zero, sign supervision discourages a small magnitude error from reversing the predicted removal effect.

5.3 Task Specific Candidates and Bundle Intervention

The reference aggregator instantiates the score in Equation 4 by averaging modality estimates on the same edge as

$$s_{\tau,z,e} = \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \widehat{\Delta}^{(\tau)}_{z,e,m}. \tag{9}$$

For fixed τ and z, s_e abbreviates this score. Averaging yields one bundle ranking and prevents modality bypass during subsequent receiver-level intervention and replay.

For training class c, let $\mathcal{V}^c_{\mathrm{tr}}$ be its nodes and $\mathbf{f}_v$ the frozen fused representation. Its prototype $\boldsymbol{\mu}_c$ and nearest prototype assignment a_v are

$$\boldsymbol{\mu}_c = \frac{1}{|\mathcal{V}^c_{\mathrm{tr}}|}\sum_{v\in\mathcal{V}^c_{\mathrm{tr}}} \mathbf{f}_v, \qquad a_v = \arg\max_c \frac{\mathbf{f}_v^\top \boldsymbol{\mu}_c}{\|\mathbf{f}_v\|\,\|\boldsymbol{\mu}_c\|}. \tag{10}$$

Only training labels form prototypes. Later assignments use frozen representations, making disagreement an eligibility cue rather than a predicted label or harm certificate. Let $\tau_{0.1}$ be the tenth percentile of training scores. For node classification, the candidate set in Equation 3 becomes

$$\mathcal{C}_v = \{e = (u, v) \in \mathcal{E}_v : u \neq v,\ a_u \neq a_v,\ s_e \leq \tau_{0.1}\}. \tag{11}$$

The prototype test excludes locally similar relations and the threshold retains strong replay candidates. Thus, the adapter controls eligibility while scores rank singleton responses.

Let $\mathrm{First}_{(s_e,e)}$ return the singleton containing the smallest score with edge index breaking ties, or $\emptyset$ for empty input, and let $D_v = |\mathcal{E}_v|$ be the incoming degree. Selection, shared attenuation, and mass compensation are defined jointly as

$$\begin{aligned} \mathcal{S}_v &= \mathrm{First}_{(s_e,e)}(\mathcal{C}_v), \\ g_e = 1 - \eta\,\mathbb{I}[e \in \mathcal{S}_v], \quad 0 < \eta < 1, &\qquad \bar{g}_{e,m} = \frac{D_v g_e}{\sum_{e'\in\mathcal{E}_v} g_{e'}}. \end{aligned} \tag{12}$$

Equation 12 couples the decision and its weights. The first line returns at most one edge. The second shares g_e across modalities with deployed $\eta = 0.5$ and applies one positive normalizer through $\bar{g}_{e,m}$. Thus, ranking chooses the bundle, η sets attenuation, and normalization preserves relative weights without uniform neighborhood shrinkage at the corresponding receiver.

Let $\widetilde{\mathbf{h}}^{(m,K)}_v$ be the controlled final state. Direct substitution gives the one line replay

$$\widetilde{\mathbf{h}}^{(m,K)}_v = \rho^{(K-1)}_m\left(\mathbf{r}^{(K-1)}_{v,m} + \sum_{e\in\mathcal{E}_v} \bar{g}_{e,m}\mathbf{q}^{(K-1)}_{e,m}\right). \tag{13}$$

Equation 13 changes only incoming coefficients. Compensation alters composition rather than total coefficient mass, while the receiver path, fusion, ego residual, and task head remain fixed. Scoring and replay therefore share one boundary, isolating the selected bundle from unrelated host computation.

Analysis proves recovery, coherent bounded bundle action, coefficient mass preservation, and receiver path separation. Proofs in Appendix C establish neither estimator accuracy nor predictive benefit.

Table 1: Overall node classification and link prediction performance on the original graph.

Tasks	Node Classification				Link Prediction			
Methods	**Movies**		**Grocery**		**DY**		**Bili Dance**	
	Acc.	Macro F1	Acc.	Macro F1	MRR	Hits@3	MRR	Hits@3
MLP	51.02 ± 0.13	40.34 ± 0.93	81.19 ± 0.25	75.21 ± 0.35	62.92 ± 0.86	76.65 ± 1.28	29.34 ± 0.28	34.81 ± 0.42
GCN	52.19 ± 0.41	43.99 ± 0.45	81.59 ± 0.44	71.68 ± 0.40	70.94 ± 0.16	84.90 ± 0.16	38.42 ± 0.28	47.67 ± 0.28
GAT	50.87 ± 0.61	42.63 ± 0.32	81.33 ± 0.31	73.78 ± 0.33	70.45 ± 0.22	83.53 ± 0.29	36.94 ± 0.37	46.98 ± 0.82
DMGC	52.14 ± 0.83	42.15 ± 1.71	82.81 ± 0.52	69.90 ± 0.72	73.06 ± 0.82	89.58 ± 0.71	39.79 ± 2.46	54.43 ± 4.94
DGF	$\mathbf{54.90 \pm 0.32}$	42.10 ± 0.47	83.46 ± 0.17	72.68 ± 0.27	$\underline{76.42 \pm 0.18}$	$\underline{91.65 \pm 0.19}$	$\underline{41.85 \pm 0.13}$	$\underline{57.09 \pm 0.38}$
MMGCN	52.09 ± 1.02	41.71 ± 1.49	81.33 ± 0.25	74.28 ± 1.16	70.51 ± 0.43	83.06 ± 0.56	38.14 ± 0.69	50.56 ± 0.92
MGAT	49.76 ± 0.16	36.87 ± 1.53	83.33 ± 0.25	74.82 ± 0.31	71.66 ± 0.18	86.72 ± 0.31	37.47 ± 0.15	50.12 ± 0.65
LGMRec	52.50 ± 0.24	$\underline{46.58 \pm 0.28}$	$\underline{84.23 \pm 0.05}$	$\mathbf{78.75 \pm 0.06}$	70.54 ± 0.05	86.25 ± 0.18	39.14 ± 0.42	50.96 ± 0.87
UniGraph2	46.88 ± 0.37	30.50 ± 1.55	75.63 ± 1.53	62.81 ± 3.56	65.88 ± 0.60	77.85 ± 0.62	31.29 ± 1.45	41.28 ± 2.63
CM-MGNN (Ours)	$\underline{54.58 \pm 0.20}$	$\mathbf{48.84 \pm 0.32}$	$\mathbf{84.83 \pm 0.14}$	$\underline{78.67 \pm 0.22}$	$\mathbf{78.10 \pm 0.18}$	$\mathbf{93.20 \pm 0.24}$	$\mathbf{43.30 \pm 0.20}$	$\mathbf{59.10 \pm 0.28}$

6 EXPERIMENTS

6.1 EXPERIMENTAL SETTING

We evaluate overall performance (RQ1), response to heterophily pressure (RQ2), reuse across compatible hosts (RQ3), component contributions (RQ4), and hyperparameter sensitivity (RQ5). The seven datasets cover node classification, link prediction, modality retrieval, graph to text, and graph to image generation. Comparisons include the native CM-MGNN host, a feature only MLP, conventional GNNs, enhanced graph and multimodal models, and UniGraph2. Details of the datasets, comparison methods, tasks, and metrics are in the appendix.

6.2 RQ1 OVERALL PERFORMANCE

Answer to RQ1. On the Original condition, CM-MGNN leads six of eight graph task columns. The leading results are Movies Macro F1 (48.84), Grocery Accuracy (84.83), and both metrics on DY (78.10/93.20) and Bili Dance (43.30/59.10). It ranks second on Movies Accuracy (54.58) and Grocery Macro F1 (78.67) (Tables 1 and 4). It also leads T2I MRR, BLEU-4, and CIDEr on the modality tasks and ranks second on I2T MRR. Its native design retains messages from each modality until evaluation at a receiver specific interface. It then changes selected physical edges without removing the receiver path or uniformly reducing neighborhood information. Baselines that aggregate views or edit topology without receiver specific selection supervised by loss do not make this distinction. These comparisons establish the reported pattern. RQ4 examines the individual components.

6.3 RQ2 RESPONSE TO HETEROPHILY PRESSURE

Answer to RQ2. CM-MGNN preserves its relative performance under the evaluated heterophily levels. From the Original graph to the fifteen percentage point condition, Accuracy decreases by 0.20/0.30 points and MRR by 0.42/0.35 points across the two graph task pairs. It leads every displayed graph condition except Grocery at five percentage points, where LGMRec is higher by 0.31 (Tables 2 and 3). At the modality endpoint, it ranks first on T2I MRR, BLEU-4, and CIDEr and second on I2T MRR (Table 5). MEI selects at most one eligible edge per receiver and rescales retained coefficients. Intervention therefore changes message composition without lowering all incoming weight. Uniform aggregation or broad topology changes do not impose this receiver budget. The response is bounded and nonmonotonic. Modality results contain only the fifteen percentage point endpoint.

6.4 RQ3 PLUGIN GENERALITY

Answer to RQ3. MEI improves every displayed augmented baseline pair at the fifteen percentage point condition. Gains across graph task hosts range from 2.4 to 2.8 points. They are 2.8 for GAT, 2.7 for MMGCN, 2.6 for MGAT, and 2.4 for DGF (Table 6). Every modality task MEI curve also lies above its baseline (Figure 3). Each host supplies its own messages, prediction context, and loss contrasts, so MEI ranks and reweights original messages without imposing a shared representation. The unmodified hosts lack this selection based on host specific loss. Results support reuse for the evaluated compatible hosts rather than universal benefit.

Table 2: Node classification performance under heterophily pressure on Movies and Grocery.

Datasets	Movies			Grocery		
Methods	**Accuracy**			**Accuracy**		
	−5pp	−10pp	−15pp	−5pp	−10pp	−15pp
MLP	50.89 ± 0.14	51.96 ± 0.14	50.71 ± 0.15	81.65 ± 0.26	81.82 ± 0.28	84.03 ± 0.29
GCN	49.31 ± 0.43	45.83 ± 0.45	42.16 ± 0.48	79.47 ± 0.46	75.79 ± 0.48	73.09 ± 0.51
GAT	49.12 ± 0.64	45.11 ± 0.67	42.74 ± 0.71	78.97 ± 0.33	76.75 ± 0.34	75.93 ± 0.36
DMGC	49.83 ± 0.87	50.43 ± 0.91	48.79 ± 0.96	79.88 ± 0.55	81.67 ± 0.57	79.19 ± 0.60
DGF	54.26 ± 0.34	53.55 ± 0.35	51.57 ± 0.37	84.04 ± 0.18	81.06 ± 0.19	80.66 ± 0.20
MMGCN	49.98 ± 1.07	48.80 ± 1.12	46.59 ± 1.18	81.81 ± 0.26	79.92 ± 0.28	79.99 ± 0.29
MGAT	49.83 ± 0.17	47.30 ± 0.18	47.15 ± 0.19	80.90 ± 0.26	81.56 ± 0.28	79.62 ± 0.29
LGMRec	51.51 ± 0.25	51.68 ± 0.26	50.82 ± 0.28	$\mathbf{85.04 \pm 0.05}$	84.24 ± 0.06	81.15 ± 0.06
UniGraph2	46.24 ± 0.39	44.29 ± 0.41	43.61 ± 0.43	74.58 ± 1.61	73.17 ± 1.68	74.84 ± 1.77
CM-MGNN (Ours)	$\mathbf{54.52 \pm 0.21}$	$\mathbf{54.45 \pm 0.21}$	$\mathbf{54.38 \pm 0.22}$	84.73 ± 0.14	$\mathbf{84.63 \pm 0.15}$	$\mathbf{84.53 \pm 0.15}$

Table 3: Link prediction performance under heterophily pressure on DY and Bili Dance.

Datasets	DY			Bili Dance		
Methods	**MRR**			**MRR**		
	−5pp	−10pp	−15pp	−5pp	−10pp	−15pp
MLP	59.46 ± 0.91	58.13 ± 0.96	56.07 ± 1.01	27.79 ± 0.30	27.26 ± 0.31	26.08 ± 0.33
GCN	65.03 ± 0.17	59.78 ± 0.18	52.27 ± 0.19	34.66 ± 0.30	31.35 ± 0.31	28.29 ± 0.33
GAT	64.14 ± 0.23	59.39 ± 0.25	54.78 ± 0.26	33.77 ± 0.39	31.43 ± 0.41	29.01 ± 0.44
DMGC	71.27 ± 0.87	70.24 ± 0.92	67.61 ± 0.97	38.27 ± 2.61	38.20 ± 2.76	36.21 ± 2.80
DGF	76.58 ± 0.19	76.28 ± 0.20	72.91 ± 0.21	41.26 ± 0.14	40.84 ± 0.15	41.16 ± 0.15
MMGCN	67.58 ± 0.46	64.03 ± 0.48	60.97 ± 0.51	36.18 ± 0.73	34.26 ± 0.77	33.30 ± 0.81
MGAT	67.94 ± 0.19	65.33 ± 0.20	63.39 ± 0.21	35.75 ± 0.16	35.32 ± 0.17	34.13 ± 0.18
LGMRec	68.33 ± 0.05	67.35 ± 0.06	65.60 ± 0.06	39.39 ± 0.45	38.45 ± 0.47	37.36 ± 0.50
UniGraph2	65.35 ± 0.64	62.19 ± 0.67	60.98 ± 0.71	30.96 ± 1.54	29.53 ± 1.62	29.13 ± 1.71
CM-MGNN (Ours)	$\mathbf{77.95 \pm 0.19}$	$\mathbf{77.82 \pm 0.19}$	$\mathbf{77.68 \pm 0.20}$	$\mathbf{43.18 \pm 0.21}$	$\mathbf{43.07 \pm 0.21}$	$\mathbf{42.95 \pm 0.22}$

Table 4: Overall modality retrieval and graph to text performance.

Tasks	Modality Retrieval		G2Text	
Methods	**Toys**			
	T2I MRR	I2T MRR	BLEU-4	CIDEr
MLP	99.15 ± 0.13	$\mathbf{99.61 \pm 0.15}$	5.77 ± 0.11	39.94 ± 0.65
GCN	94.86 ± 0.05	97.86 ± 0.05	5.63 ± 0.10	39.11 ± 0.55
DGF	93.97 ± 0.05	92.67 ± 0.05	6.82 ± 0.08	43.57 ± 0.48
LGMRec	97.77 ± 0.07	97.37 ± 0.08	5.93 ± 0.09	39.76 ± 0.52
CM-MGNN (Ours)	$\mathbf{99.40 \pm 0.04}$	99.38 ± 0.05	$\mathbf{6.95 \pm 0.07}$	$\mathbf{45.10 \pm 0.42}$

Table 5: Modality task performance under the fifteen percentage point condition.

Tasks	Modality Retrieval		G2Text	
Methods	**Toys T2I MRR**	**Toys I2T MRR**	**BLEU-4**	**CIDEr**
	−15pp	−15pp	−15pp	−15pp
MLP	98.31 ± 0.15	$\mathbf{99.87 \pm 0.17}$	5.92 ± 0.13	38.40 ± 0.77
GCN	93.43 ± 0.06	94.65 ± 0.06	5.06 ± 0.11	34.79 ± 0.65
DGF	93.99 ± 0.06	92.79 ± 0.06	6.65 ± 0.09	43.50 ± 0.57
LGMRec	98.24 ± 0.08	96.81 ± 0.09	5.83 ± 0.10	37.37 ± 0.61
CM-MGNN (Ours)	$\mathbf{99.35 \pm 0.05}$	99.33 ± 0.06	$\mathbf{6.89 \pm 0.08}$	$\mathbf{44.85 \pm 0.50}$

6.5 RQ4 Mechanism of Improvement

Answer to RQ4. Removing each component lowers the metric under the evaluated fifteen percentage point settings. The message effect estimator produces the largest decrease from 0.26 to 0.32 points, followed by the prototype filter from 0.20 to 0.25 and coefficient rescaling from 0.15 to 0.18. This ordering holds for Movies in Figure 4 and Bili Dance in Figure 6. The estimator provides the receiver specific ranking. The prototype filter restricts intervention to eligible edges across classes. Rescaling prevents a side effect on total incoming weight. Removing any element removes its protection. These variants support each component contribution without establishing estimator accuracy or a cause and effect attribution for the gain observed under this controlled ablation protocol.

6.6 RQ5 Hyperparameter Sensitivity

Answer to RQ5. The candidate tail ratio and attenuation strength are the two evaluated hyperparameters. They respectively control the breadth of MEI search among low score edge bundles and the magnitude of the change to a selected edge. The hyperparameter sweep is locally stable in the low to moderate region. From zero to 0.4, the maps span only 84.80 to 84.84 and 78.05 to 78.11, with maxima at the 0.2 and 0.4 setting pair (Figures 4 and 6). The receiver budget and coefficient rescaling keep the effect of small hyperparameter changes limited. Joint settings of 1.0 and 1.0 reduce the metrics to 82.55 and 74.67, consistent with stronger and less selective intervention. This aggressive combination is therefore unsuitable within the evaluated two parameter regime.

Table 6: MEI augmentation of compatible hosts for node classification and link prediction.

Tasks	Node Classification		Link Prediction			
Methods	**Movies**	**Grocery**	**DY**		**Bili Dance**	
	Macro F1	Macro F1	MRR	Hits@3	MRR	Hits@3
GAT	34.99 ± 0.37	67.60 ± 0.38	54.78 ± 0.26	65.26 ± 0.34	29.01 ± 0.44	36.19 ± 0.97
GAT + MEI	$\mathbf{37.79 \pm 0.38}$	$\mathbf{70.40 \pm 0.39}$	$\mathbf{57.58 \pm 0.27}$	$\mathbf{68.06 \pm 0.35}$	$\mathbf{31.81 \pm 0.45}$	$\mathbf{38.99 \pm 0.98}$
MMGCN	38.74 ± 1.73	73.44 ± 1.35	60.97 ± 0.51	74.34 ± 0.66	33.30 ± 0.81	43.43 ± 1.09
MMGCN + MEI	$\mathbf{41.44 \pm 1.74}$	$\mathbf{76.14 \pm 1.36}$	$\mathbf{63.67 \pm 0.52}$	$\mathbf{77.04 \pm 0.67}$	$\mathbf{36.00 \pm 0.82}$	$\mathbf{46.13 \pm 1.10}$
MGAT	34.64 ± 1.77	74.33 ± 0.36	63.39 ± 0.21	78.45 ± 0.37	34.13 ± 0.18	45.17 ± 0.77
MGAT + MEI	$\mathbf{37.24 \pm 1.78}$	$\mathbf{76.93 \pm 0.37}$	$\mathbf{65.99 \pm 0.22}$	$\mathbf{81.05 \pm 0.38}$	$\mathbf{36.73 \pm 0.19}$	$\mathbf{47.77 \pm 0.78}$
DGF	40.40 ± 0.55	70.15 ± 0.31	72.91 ± 0.21	89.43 ± 0.22	41.16 ± 0.15	57.03 ± 0.45
DGF + MEI	$\mathbf{42.80 \pm 0.56}$	$\mathbf{72.55 \pm 0.32}$	$\mathbf{75.31 \pm 0.22}$	$\mathbf{91.83 \pm 0.23}$	$\mathbf{43.56 \pm 0.16}$	$\mathbf{59.43 \pm 0.46}$

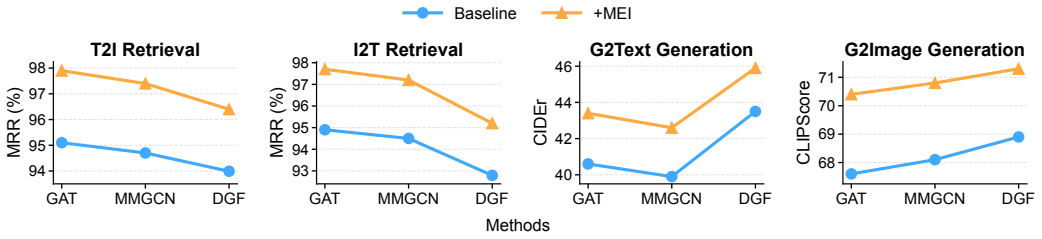


Figure 3: MEI gains at the fifteen percentage point condition for T2I and I2T on Toys, G2Text on Flickr30k, and G2Image on SemArt across the displayed compatible hosts.

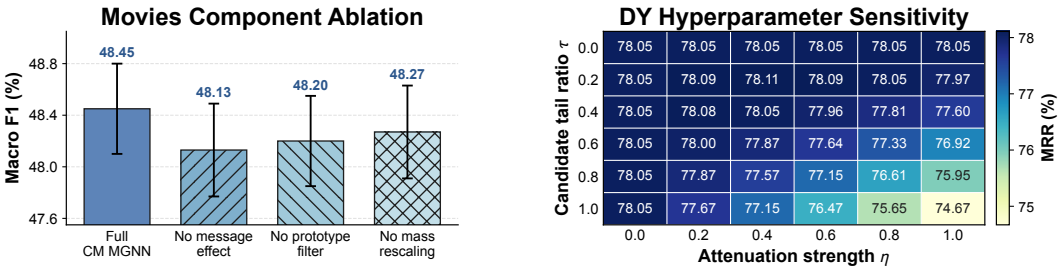


Figure 4: The representative component ablation on Movies appears left, and hyperparameter sensitivity on DY appears right. Complementary results are reported in the appendix.

7 Conclusion

We study selective control of multimodal graph messages whose utility varies by receiver and modality. The controlled stress study shows broad endpoint deterioration as cross-class exposure rises, but its nonuniform trajectories identify neither a universal topology rule nor a harmful edge. This separation between graph-level association and local action motivates prediction-conditioned intervention. At a frozen interface, MEI estimates singleton loss changes, combines diagnoses into a bundle decision, and replays one bundle per receiver with mass compensation. The contrasts rank isolated responses rather than bundle utility. CM MGNN and task adapters define compatible deployment, with labels confined to training prototypes. The operator guarantees fully open recovery, bundle coherence, receiver feasibility, coefficient mass preservation, and receiver path separation. CM MGNN improves over the strongest baselines by 4.85% in Movies Macro F1, 3.46% in Bili Dance MRR, and 3.51% in CIDEr. MEI also improves every displayed compatible host pair under the fifteen percentage point condition. These findings apply only to the evaluated settings and establish neither estimator nor prototype correctness, cause and effect, broader benefit, nor transfer beyond the tested tasks, hosts, and data regimes represented in the present benchmark suite.

Reproducibility Statement

We provide the implementation details of CM-MGNN in the main text and appendix. All experiments use publicly available datasets. Code and configurations of CM-MGNN will be released for reproduction.

AI Use Statement

Generative AI tools were not used in the preparation of this manuscript. All technical contents, method details, and experimental results were designed and verified by the authors.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, pp. 9505–9515, 2018.

Kenza Amara, Zhitao Ying, Zitao Zhang, Zhichao Han, Yang Zhao, Yinan Shan, Ulrik Brandes, Sebastian Schemm, and Ce Zhang. GraphFramEx: Towards systematic evaluation of explainability methods for graph neural networks. In *Proceedings of the First Learning on Graphs Conference*, volume 198 of *Proceedings of Machine Learning Research*, pp. 44:1–44:23, 2022.

Peter M. Aronow and Cyrus Samii. Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics*, 11(4):1912–1947, 2017. doi: 10.1214/16-AOAS1005.

Mohit Bajaj, Lingyang Chu, Zi Yu Xue, Jian Pei, Lanjun Wang, Peter Cho-Ho Lam, and Yong Zhang. Robust counterfactual explanations on graph neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 5644–5655, 2021.

Deyu Bo, Xiao Wang, Chuan Shi, and Huawei Shen. Beyond low-frequency information in graph convolutional networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5): 3950–3957, 2021. doi: 10.1609/aaai.v35i5.16514.

Eli Chien, Jianhao Peng, Pan Li, and Olgica Milenkovic. Adaptive universal generalized PageRank graph neural network. In *International Conference on Learning Representations*, 2021.

Kaize Ding, Jianling Wang, Jundong Li, Kai Shu, Chenghao Liu, and Huan Liu. Graph prototypical networks for few-shot learning on attributed networks. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, pp. 295–304, 2020. doi: 10.1145/3340531.3411922.

Dean Eckles, Brian Karrer, and Johan Ugander. Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference*, 5(1), 2017. doi: 10.1515/jci-2015-0021.

Yasha Ektefaie, George Dasoulas, Ayush Noori, Maha Farhat, and Marinka Zitnik. Multimodal learning with graphs. *Nature Machine Intelligence*, 5:340–350, 2023. doi: 10.1038/s42256-023-00624-6.

Bahare Fatemi, Layla El Asri, and Seyed Mehran Kazemi. SLAPS: Self-supervision improves structure learning for graph neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 22667–22681, 2021.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1263–1272, 2017.

Zhaochen Guo, Zhixiang Shen, Xuanting Xie, Liangjian Wen, and Zhao Kang. Disentangling homophily and heterophily in multimodal graph clustering. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2044–2053, 2025. doi: 10.1145/3746027.3755631.

William L. Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 30, pp. 1024–1034, 2017.

Hans Hao-Hsun Hsu, Yuesong Shen, Christian Tomani, and Daniel Cremers. What makes graph neural networks miscalibrated? In *Advances in Neural Information Processing Systems*, volume 35, pp. 13775–13786, 2022. doi: 10.52202/068431-1001.

Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 5666–5675, 2021. doi: 10.18653/v1/2021.acl-long.440.

Kexin Huang and Marinka Zitnik. Graph meta learning via local subgraphs. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5862–5874, 2020.

Kexin Huang, Ying Jin, Emmanuel Candès, and Jure Leskovec. Uncertainty quantification over graph with conformalized graph neural networks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 26699–26721, 2023. doi: 10.52202/075280-1160.

Michael G. Hudgens and M. Elizabeth Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008. doi: 10.1198/016214508000000292.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30, pp. 5574–5584, 2017.

Sungwon Kim, Junseok Lee, Namkyeong Lee, Wonjoong Kim, Seungyoon Choi, and Chanyoung Park. Task-equivariant graph few-shot learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1120–1131, 2023. doi: 10.1145/3580305.3599515.

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

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6402–6413, 2017.

Lin Lan, Pinghui Wang, Xuefeng Du, Kaikai Song, Jing Tao, and Xiaohong Guan. Node classification on graphs with few-shot novel labels via meta transformed network embedding. In *Advances in Neural Information Processing Systems*, volume 33, pp. 16520–16531, 2020.

Dongyue Li, Haotian Ju, Aneesh Sharma, and Hongyang R. Zhang. Boosting multitask learning on graphs through higher-order task affinities. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1213–1222, 2023a. doi: 10.1145/3580305.3599265.

Maximilian Li and Lucas Janson. Optimal ablation for interpretability. In *Advances in Neural Information Processing Systems*, volume 37, pp. 109233–109282, 2024. doi: 10.52202/079017-3468.

Zhixun Li, Liang Wang, Xin Sun, Yifan Luo, Yanqiao Zhu, Dingshuo Chen, Yingtao Luo, Xiangxin Zhou, Qiang Liu, Shu Wu, Liang Wang, and Jeffrey Yu. GSLB: The graph structure learning benchmark. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, volume 36, pp. 30306–30318, 2023b. doi: 10.52202/075280-1319.

Hao Liu, Jiarui Feng, Lecheng Kong, Ningyue Liang, Dacheng Tao, Yixin Chen, and Muhan Zhang. One for all: Towards training one graph model for all classification tasks. In *International Conference on Learning Representations*, 2024.

Christos Louizos, Max Welling, and Diederik P. Kingma. Learning sparse neural networks through L_0 regularization. In *International Conference on Learning Representations*, 2018.

Sitao Luan, Chenqing Hua, Qincheng Lu, Jiaqi Zhu, Mingde Zhao, Shuyuan Zhang, Xiao-Wen Chang, and Doina Precup. Revisiting heterophily for graph neural networks. In *Advances in Neural Information Processing Systems*, volume 35, pp. 1362–1375, 2022. doi: 10.52202/068431-0100.

Ana Lucic, Maartje A. ter Hoeve, Gabriele Tolomei, Maarten de Rijke, and Fabrizio Silvestri. CF-GNNExplainer: Counterfactual explanations for graph neural networks. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 4499–4511, 2022.

Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized explainer for graph neural network. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19620–19631, 2020.

Xu Luo, Hao Wu, Ji Zhang, Lianli Gao, Jing Xu, and Jingkuan Song. A closer look at few-shot classification again. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23103–23123, 2023.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, pp. 13991–14002, 2019.

Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, and Bo Yang. Geom-GCN: Geometric graph convolutional networks. In *International Conference on Learning Representations*, 2020.

Oleg Platonov, Denis Kuznedelev, Artem Babenko, and Liudmila Prokhorenkova. Characterizing graph datasets for node classification: Homophily–heterophily dichotomy and beyond. In *Advances in Neural Information Processing Systems*, volume 36, pp. 523–548, 2023. doi: 10.52202/075280-0025.

Phillip E. Pope, Soheil Kolouri, Mohammad Rostami, Charles E. Martin, and Heiko Hoffmann. Explainability methods for graph convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10772–10781, 2019. doi: 10.1109/CVPR.2019.01103.

Michael Sejr Schlichtkrull, Nicola De Cao, and Ivan Titov. Interpreting graph neural networks for NLP with differentiable edge masking. In *International Conference on Learning Representations*, 2021.

Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation, 2018. Revised 2019; no peer-reviewed proceedings version found.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4077–4087, 2017.

Maximilian Stadler, Bertrand Charpentier, Simon Geisler, Daniel Zügner, and Stephan Günnemann. Graph posterior network: Bayesian predictive uncertainty for node classification. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18033–18048, 2021.

Xiangguo Sun, Hong Cheng, Jia Li, Bo Liu, and Jihong Guan. All in one: Multi-task prompting for graph neural networks. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2120–2131, 2023. doi: 10.1145/3580305.3599256.

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. doi: 10.1016/j.ipm.2020.102277.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, pp. 2530–2540, 2019.

Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, and Hugo Larochelle. Meta-dataset: A dataset of datasets for learning to learn from few examples. In *International Conference on Learning Representations*, 2020.

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

Minh N. Vu and My T. Thai. PGM-Explainer: Probabilistic graphical model explanations for graph neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12225–12235, 2020.

Song Wang, Chen Chen, and Jundong Li. Graph few-shot learning with task-specific structures. In *Advances in Neural Information Processing Systems*, volume 35, pp. 38925–38936, 2022. doi: 10.52202/068431-2821.

Xiang Wang, Yingxin Wu, An Zhang, Fuli Feng, Xiangnan He, and Tat-Seng Chua. Reinforced causal explainer for graph neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2297–2309, 2023. doi: 10.1109/TPAMI.2022.3170302.

Zehong Wang, Zheyuan Zhang, Chuxu Zhang, and Yanfang Ye. Subgraph pooling: Tackling negative transfer on graphs. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 5153–5161, 2024. doi: 10.24963/ijcai.2024/570.

Yaochen Xie, Sumeet Katariya, Xianfeng Tang, Edward Huang, Nikhil Rao, Karthik Subbian, and Shuiwang Ji. Task-agnostic graph explanations. In *Advances in Neural Information Processing Systems*, volume 35, pp. 12027–12039, 2022. doi: 10.52202/068431-0874.

Hao Yan, Chaozhuo Li, Ruosong Long, Chao Yan, Jianan Zhao, Wenwen Zhuang, Jun Yin, Peiyan Zhang, Weihao Han, Hao Sun, Weiwei Deng, Qi Zhang, Lichao Sun, Xing Xie, and Senzhang Wang. A comprehensive study on text-attributed graphs: Benchmarking and rethinking. In *Advances in Neural Information Processing Systems*, volume 36, pp. 17238–17264, 2023. doi: 10.52202/075280-0755.

Hao Yan, Chaozhuo Li, Jun Yin, Zhigang Yu, Weihao Han, Mingzheng Li, Zhengxin Zeng, Hao Sun, and Senzhang Wang. When graph meets multimodal: Benchmarking and meditating on multimodal attributed graph learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5842–5853, 2025. doi: 10.1145/3711896.3737404.

Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. GNNExplainer: Generating explanations for graph neural networks. In *Advances in Neural Information Processing Systems*, volume 32, pp. 9244–9255, 2019.

Hao Yuan, Haiyang Yu, Jie Wang, Kang Li, and Shuiwang Ji. On explainability of graph neural networks via subgraph explorations. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12241–12252, 2021.

Jiying Zhang, Xi Xiao, Long-Kai Huang, Yu Rong, and Yatao Bian. Fine-tuning graph neural networks via graph topology induced optimal transport. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 3730–3736, 2022. doi: 10.24963/ijcai.2022/518.

Haoran Zheng, Renchi Yang, Hongtao Wang, and Jianliang Xu. Cross-contrastive clustering for multimodal attributed graphs with dual graph filtering. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2020–2030, 2026. doi: 10.1145/3770854.3780236.

Fan Zhou, Chengtai Cao, Kunpeng Zhang, Goce Trajcevski, Ting Zhong, and Ji Geng. Meta-GNN: On few-shot node classification in graph meta-learning. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 2357–2360, 2019. doi: 10.1145/3357384.3358106.

Jing Zhu, Yuhang Zhou, Shengyi Qian, Zhongmou He, Tong Zhao, Neil Shah, and Danai Koutra. Mosaic of modalities: A comprehensive benchmark for multimodal graph learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14215–14224, 2025. doi: 10.1109/CVPR52734.2025.01326.

Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7793–7804, 2020.

Yinlin Zhu, Xunkai Li, Di Wu, Wang Luo, Miao Hu, and Guocong Quan. TMTE: Effective multimodal graph learning with task-aware modality and topology co-evolution, 2026. Accepted by ACM Multimedia 2026; proceedings metadata not yet available.

A Appendix Roadmap and Notation

This appendix expands the main paper's compact presentation of the motivating graph construction, MEI's structural guarantees, the broader literature, host implementation, and empirical evaluation. Specifically, Appendix B defines the controlled graph family used in the motivating study, Appendix C provides complete proofs for the deployed operator, Appendix D develops the literature context for multimodal graph learning and message intervention, Appendix E details the reference CM MGNN and the contract required of a compatible frozen host, and Appendix F documents the datasets, baselines, tasks, metrics, additional analyses, and planned selective intervention audit. This organization separates the semisynthetic construction and formal guarantees from the implementation requirements and empirical protocols used to evaluate MEI. Table 7 collects the core notation used throughout the paper and this appendix.

Table 7: Core notation used in the main text and this appendix. Subscripts and superscripts follow the conventions of Sections 3 and 5.

Symbol	Meaning
$\mathcal{G} = \langle \mathcal{V}, \mathcal{E}, \{\mathbf{X}^{(m)}\}_{m\in\mathcal{M}} \rangle$	Multimodal attributed graph: node set, directed message edge set, and modality feature matrices
$\mathcal{M}$	Modality set
$e = (u, v)$	Directed edge from sender u to receiver v
$\mathcal{E}_v$, $D_v = \lvert\mathcal{E}_v\rvert$	Incoming edge set of receiver v and its size
$\mathbf{h}_v^{(m,k)}$	Modality m state of node v at layer k
$\mathbf{q}_{e,m}^{(k)}$	Exposed message on edge e in modality m at layer k; $\mathbf{q}_{e,m}$ abbreviates the final layer instance
$\mathbf{r}_{v,m}^{(k)}$	Explicit receiver path outside the message sum
$\theta^\star$	Frozen host parameters
$\tau \in \mathfrak{T}$, $z \in \mathcal{D}_\tau$	Task and one of its instances
$\mathcal{L}_\tau(z)$	Scalar loss of the frozen host on instance z
$\Delta_{z,e,m}^{(\tau)}$, $\widehat{\Delta}_{z,e,m}^{(\tau)}$	Singleton deletion contrast and its estimate
$s_{\tau,z,e}$ (s_e)	Physical edge bundle score
$\mathcal{C}_v$, $\mathcal{S}_v$	Candidate set and selected bundle set at receiver v
B_τ	Receiver selection budget
η	Attenuation strength with $0 < \eta < 1$
g_e, $\bar{g}_{e,m}$	Shared raw coefficient and compensated modality coefficient
$\boldsymbol{\mu}_c$, a_v	Class prototype and nearest prototype assignment
h_{het}	Heterophily level of a graph
U_e	Action utility of one physical edge bundle

B Construction of Heterophily Levels

This appendix specifies the graph construction used by the motivating heterophily example. It defines a label guided, semisynthetic topology family. It is not a deployment rule for MEI and does not use labels after the graphs have been constructed.

B.1 Target Levels

Let $\mathcal{G}_0 = (\mathcal{V}, \mathcal{E}_0)$ denote the raw directed graph before the benchmark message transformation, and let $\mathcal{E}_0^- = \{(u, v) \in \mathcal{E}_0 : u \neq v\}$ be its nonself arcs. Given fixed node labels y_v, we define the heterophily level as

$$h_{\mathrm{het}}(\mathcal{G}) = \frac{\lvert\{(u, v) \in \mathcal{E}^- : y_u \neq y_v\}\rvert}{\lvert\mathcal{E}^-\rvert}. \tag{14}$$

For each prescribed increment $\delta \in \{0.05, 0.10, 0.15\}$, the target is $h_{\mathrm{het}}^{(\delta)} = h_{\mathrm{het}}(\mathcal{G}_0) + \delta$. Equivalently, if S_0 is the number of same label nonself arcs in the original graph, construction stops at the nearest

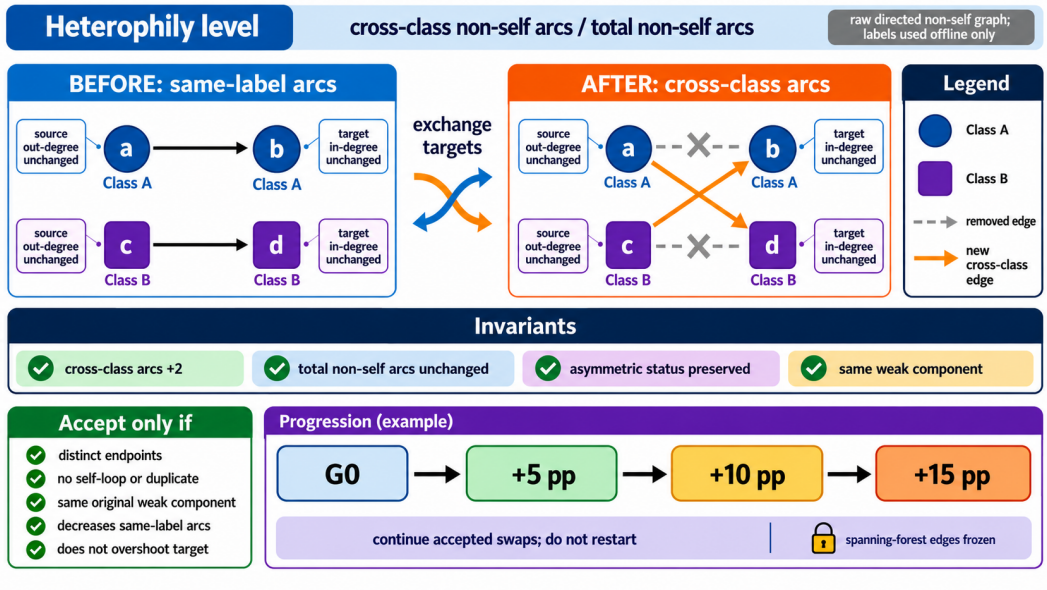


Figure 5: A local asymmetric directed two edge switch. Exchanging two same label arcs' targets produces diagonals across classes while preserving source outdegrees, target indegrees, and other displayed arcs.

attainable count

$$S_\delta = \mathrm{round}\Big[\big(1 - h_{\mathrm{het}}^{(\delta)}\big)|\mathcal{E}_0^-|\Big]. \tag{15}$$

Thus, discretizing a target level can differ from its real valued value by no more than one nonself arc divided by $|\mathcal{E}_0^-|$.

B.2 Label Guided Directed Endpoint Swaps

The construction rewires only the raw directed graph. It separates nonself arcs into reciprocal dyads, for which both (u, v) and (v, u) are present, and asymmetric arcs, for which the reverse arc is absent. The two types are rewired separately so that a reciprocal relation cannot silently become an asymmetric one, or conversely.

For two asymmetric arcs (a, b) and (c, d), an admissible directed two edge switch exchanges their targets:

$$(a, b), (c, d) \quad \longmapsto \quad (a, d), (c, b). \tag{16}$$

This operation preserves the outdegree of a and c and the indegree of b and d. For two reciprocal dyads $\{a, b\}$ and $\{c, d\}$, the procedure pairs the four endpoints again into one of the two alternative dyad pairs, for example $\{a, d\}$ and $\{c, b\}$, and inserts both directions of each new dyad. It therefore preserves the number of reciprocal arcs as well as the indegree and outdegree of every involved node.

An attempted switch is accepted only when all four endpoints are distinct, the selected relations lie in the same original weakly connected component, the new arcs create neither a self loop nor a duplicate, and the swap preserves the reciprocal/asymmetric status required by its branch. A further label condition requires the number of same label arcs to decrease strictly, without passing the remaining target in Equation 15. Rejected proposals leave the graph unchanged. This acceptance rule is the sole use of ground truth labels in the construction.

B.3 Progressive Levels and Preserved Structure

For a fixed construction trace, the $+5$pp graph is generated first, the $+10$pp graph continues from that state, and the $+15$pp graph continues again. The three graphs therefore form nested stages of one progressive topology intervention rather than three unrelated perturbation mechanisms. Candidate relations belonging to a spanning forest of each original weak component are frozen before rewiring. Since that forest remains intact, weak component membership is preserved by construction.

Table 8: Properties controlled by the heterophily level construction. The table describes construction constraints rather than empirical results.

Property	How it is controlled
Node content and labels	The node set, labels, and multimodal features are retained from the original graph. Only raw endpoints are rewired.
Nonself arc count	Every accepted switch removes and introduces the same number of nonself arcs.
Node degrees	Endpoint exchange preserves each involved source's out-degree and each involved destination's indegree.
Reciprocity	Reciprocal dyads and asymmetric arcs are rewired in separate branches, preserving the reciprocal arc count.
Weak components	A spanning forest in every original weak component is frozen, and swaps are restricted within that component.
Graph validity	Proposals producing self loops, duplicate arcs, or an invalid reciprocal/asymmetric relation are rejected.

The construction intentionally controls node content, edge count, degrees, reciprocity, and component membership so that the prescribed change is an increase in neighborhood exposure across classes rather than a change in the available node information or degree distribution. It does not preserve every topological statistic. For example, triangle structure, local mixing patterns, and other higher order properties may change. The resulting family is therefore a label guided semisynthetic stress setting, not an intervention that isolates the effect of heterophily in a real world graph.

B.4 Message Passing Graph and Label Boundary

After raw rewiring, the benchmark message graph is formed by removing raw self loops, adding reverse arcs, and adding exactly one self loop per node. The distinction matters because the level constraint in Equation 14 is imposed on the raw nonself directed graph, whereas the classifier receives the transformed message passing edge set.

Labels are consulted only offline to accept swaps that increase the desired exposure across classes. They are not supplied as node features, graph inputs, or deployment time signals. Consequently, the construction makes no claim that a deployed model observes test labels, that every message across classes is harmful, or that the graph family identifies an effect attributable solely to heterophily.

C Proofs for the Deployed Operator

C.1 Statements of the Operator Guarantees

Proposition C.1 and Lemma C.2 verify exact recovery of the fully open host when MEI selects nothing. Proposition C.3 and Lemma C.4 connect physical edge selection to a shared action across modalities that prevents bypass through an unmodified channel. Proposition C.5 and Lemma C.6 establish bounded selection within the task candidate set and receiver budget. The reference node classification policy permits at most one action per receiver. Proposition C.7 and Lemma C.8 formalize the scalar mass guarantee provided by compensation at each receiver. Proposition C.9 and Lemma C.10 isolate the action to exposed neighbor coefficients in the final layer while retaining the declared receiver and ego paths.

Throughout, we fix the graph, the host trace in evaluation mode, all exposed messages and endpoint states, the prototype assignments, the predicted scores, and the realized edge bundle sets. The controlled replay modifies only the coefficients in the final layer through Equation 13. By the message graph convention in Section 3, every receiver has a self loop message, so $D_v = |\mathcal{E}_v| \geq 1$. Moreover, $0 < \eta < 1$ in Equation 12, so every g_e is strictly positive and the compensation denominator in the same equation is well defined.

Proposition C.1 (Recovery of the fully open host and adapter)**.** *If $\mathcal{S}_v = \emptyset$ for every receiver, then $\bar{g}_{e,m} = 1$ for every edge and modality, and controlled CM MGNN equals its fully open computation. The same identity holds for a compatible frozen host satisfying Section 5.2.*

Lemma C.2 (Identity coefficient bridge)**.** *Under the empty selection premise of Proposition C.1,* $g_e = 1$, $\sum_{e' \in \mathcal{E}_v} g_{e'} = D_v$, *and hence* $\bar{g}_{e,m} = 1$ *for every incoming edge and modality.*

Proposition C.3 (Bundle coherence across modalities)**.** *For each selected physical edge* $e = (u, v)$, *the same raw coefficient* g_e *is applied to every modality.*

Lemma C.4 (Action index without modality)**.** *The deployed selection variable* $\mathcal{S}_v$ *and the raw coefficient* g_e *are indexed by physical edge, and neither introduces an action specific to one modality.*

Proposition C.5 (Receiver budget feasibility)**.** *For every receiver* v *and any valid candidate set for the task* $\mathcal{C}_v \subseteq \mathcal{E}_v \setminus \{(v, v)\}$, *the deployment selection satisfies* $\mathcal{S}_v \subseteq \mathcal{C}_v$ *and* $|\mathcal{S}_v| \leq B_\tau$.

Lemma C.6 (Budget truncation preserves candidates)**.** *Deterministically ordering* $\mathcal{C}_v$ *and retaining at most its first* B_τ *members yields a subset of* $\mathcal{C}_v$ *with cardinality at most* B_τ.

Proposition C.7 (Conservation of receiver coefficient mass)**.** *For every receiver* v *and modality* m,

$$\sum_{e \in \mathcal{E}_v} \bar{g}_{e,m} = D_v. \tag{17}$$

Lemma C.8 (Positive compensation normalizer)**.** *Because each receiver has a self loop and* $0 < \eta < 1$, *the normalizer* $\sum_{e' \in \mathcal{E}_v} g_{e'}$ *is positive.*

Proposition C.9 (Invariance of declared receiver and ego paths)**.** *Changing* $\mathcal{S}_v$ *at the fixed final layer replay boundary does not change* $\mathbf{r}_{v,m}^{(K-1)}$ *in Equation 13 or the graph free ego residual in Equation 19.*

Lemma C.10 (Separation of action support)**.** *At the fixed replay boundary,* $\mathcal{S}_v$ *changes only the coefficients of the exposed neighbor messages. Neither the explicit receiver term nor the graph free ego residual depends on* $\mathcal{S}_v$.

C.2 Exact Recovery of the Fully Open Host

We prove Proposition C.1 and Lemma C.2, both stated in Section C.1.

Proof. Suppose $\mathcal{S}_v = \emptyset$ for every receiver v. Then $\mathbb{I}[e \in \mathcal{S}_v] = 0$ for every incoming edge. The rule in The second line of Equation 12 gives $g_e = 1$. Therefore,

$$\sum_{e' \in \mathcal{E}_v} g_{e'} = |\mathcal{E}_v| = D_v, \qquad \bar{g}_{e,m} = \frac{D_v}{D_v} = 1.$$

Substitution into Equation 13 restores the fully open accumulator in the final layer from Equation 6. Because all subsequent fusion, ego residual, and task head operations are unchanged under the fixed evaluation trace, the complete controlled CM MGNN computation equals its fully open computation. For another compatible frozen host, the same $g_e = 1$ condition invokes the identity requirement in Appendix E.2, which requires reinjection that preserves identity and exact reproduction of the original computation in evaluation mode. □

C.3 Bundle Coherence Across Modalities

We prove Proposition C.3 and Lemma C.4, both stated in Section C.1.

Proof. The policy first forms a single bundle decision for each physical edge $e = (u, v)$, so Equation 12 indexes the raw coefficient by e alone and contains neither a modality index nor a selection variable for one modality. Thus, for every modality $m \in \mathcal{M}$, the message $\mathbf{q}_{e,m}^{(K-1)}$ receives the same g_e. This establishes coherence of the deployed action across the modalities of that physical relation. It does not equate their messages, singleton effects, or task utilities. □

C.4 Receiver Budget Feasibility

We prove Proposition C.5 and Lemma C.6, both stated in Section C.1.

Proof. Fix a receiver v and let the constructor for the task return any valid $\mathcal{C}_v \subseteq \mathcal{E}_v \setminus \{(v, v)\}$. The deployment policy only orders members of $\mathcal{C}_v$ and retains its first $\min\{B_\tau, |\mathcal{C}_v|\}$ members, or none when the candidate set is empty. It therefore cannot introduce an edge outside $\mathcal{C}_v$, so $\mathcal{S}_v \subseteq \mathcal{C}_v$, and its truncation gives $|\mathcal{S}_v| \leq B_\tau$, as required by Equation 5. Because $\mathcal{C}_v$ excludes (v, v), no selected edge is a self loop.

For the concrete node classification realization in Section 5.3, $\mathcal{C}_v = \mathcal{C} \cap \mathcal{E}_v$ contains the physical edges passing the low score tail and prototype mismatch filters. Its deterministic order uses bundle score with edge index as the tie breaker, and the deployed reference policy sets $B_\tau = B = 1$. The general result then specializes to at most one qualified edge per receiver. □

C.5 Conservation of Receiver Coefficient Mass

We prove Proposition C.7 and Lemma C.8, both stated in Section C.1.

Proof. Fix a receiver v and modality m. Since $D_v \geq 1$ and every $g_e \in \{1, 1 - \eta\}$ is positive, the normalizer $Z_v = \sum_{e' \in \mathcal{E}_v} g_{e'}$ is strictly positive. Substituting the compensation term in the second line of Equation 12 and collecting the common factors gives

$$\sum_{e \in \mathcal{E}_v} \bar{g}_{e,m} = \sum_{e \in \mathcal{E}_v} \frac{D_v g_e}{Z_v} = \frac{D_v}{Z_v} \sum_{e \in \mathcal{E}_v} g_e = D_v.$$

The fully open operator assigns unit coefficient to each of the D_v incoming messages, so its scalar coefficient mass is also D_v. The identity concerns only these scalar coefficients. Because the vectors $\mathbf{q}_{e,m}^{(K-1)}$ need not be equal, it does not preserve their vector sum. □

C.6 Invariance of Receiver and Ego Paths

We prove Proposition C.9 and Lemma C.10, both stated in Section C.1.

Proof. Under the fixed final layer trace, changing $\mathcal{S}_v$ changes the indicator in Equation 12 and therefore may change g_e and $\bar{g}_{e,m}$. These coefficients multiply only the exposed neighbor messages in Equation 13. The explicit receiver term $\mathbf{r}_{v,m}^{(K-1)}$ is added outside that weighted sum, and the graph free $\mathbf{f}_v^{\text{ego}}$ in Equation 19 is computed outside graph propagation. Neither term is therefore a function of the selected bundle set at the replay boundary, so both remain unchanged. The controlled propagated state and downstream prediction may nevertheless change, and the result does not recover information already mixed in earlier propagation layers. □

Taken together, these properties show that MEI changes a precisely delimited part of the frozen computation. Empty selection recovers the original host. Nonempty selection remains coherent across the modalities of one physical edge and within each receiver's budget. Compensation avoids a trivial reduction in total incoming coefficient mass, while the declared receiver and ego paths remain outside the action. This distinction clarifies which behavior follows from the operator itself and which behavior must be demonstrated experimentally. In particular, the analysis establishes neither estimator correctness, candidate filter correctness, robustness, nor predictive utility. Those properties remain empirical.

D Extended Related Work

Multimodal Graph Learning. Multimodal graph models combine heterogeneous node attributes with relational context (Ektefaie et al., 2023). Mosaic and MAGB systematize multimodal graph datasets and evaluation settings (Zhu et al., 2025; Yan et al., 2025), while the CS-TAG study isolates the related problem of integrating text semantics and topology at benchmark scale (Yan et al., 2023). Application-specific architectures include MMGCN for multimodal conversational fusion and MGAT for recommendation (Hu et al., 2021; Tao et al., 2020). More recent methods change the graph signal itself: DMGC separates homophily-enhanced and heterophily-aware views (Guo et al., 2025), DGF couples node- and feature-domain filtering with cross-contrastive clustering (Zheng et al., 2026), and TMTE co-evolves modality representations with task-dependent topology (Zhu et al., 2026).

These works motivate topology-aware multimodal learning, but they do not estimate the receiver-loss change caused by removing one exposed modality message. MEI starts from that diagnostic and follows it with a selective physical-edge action rather than another fusion taxonomy.

Message-Passing Foundations. The message-passing abstraction organizes graph computation as learned message, aggregation, and update functions (Gilmer et al., 2017). GCN provides a normalized spectral approximation for semi-supervised node classification (Kipf & Welling, 2017), whereas GraphSAGE learns inductive neighborhood aggregators that can embed unseen nodes (Hamilton et al., 2017). Graph attention replaces fixed aggregation weights with learned, locally normalized coefficients (Veličković et al., 2018). MEI assumes such a host has already exposed its directed modality messages; its contribution is the downstream intervention contract, not a new generic message-passing layer.

Heterophily and Graph Control. Standard aggregation can fail when adjacent nodes carry label-incompatible information, motivating ego–neighbor separation and higher-order aggregation in H2GCN (Zhu et al., 2020). Geom-GCN constructs structural neighborhoods in a latent geometric space to address disassortative graphs (Pei et al., 2020), and FAGCN mixes low- and high-frequency signals through self-gating (Bo et al., 2021). ACM-GCN instead combines aggregation, diversification, and identity channels according to the local regime (Luan et al., 2022), while GPRGNN learns signed generalized PageRank weights (Chien et al., 2021). Dataset characterizations caution that a binary homophily–heterophily label misses differences in label informativeness (Platonov et al., 2023). At the structure level, GSLB compares graph revision methods across effectiveness, robustness, and complexity (Li et al., 2023b), and SLAPS adds self-supervision when task labels alone are insufficient to learn adjacency (Fatemi et al., 2021). MEI differs in both signal and action: its scores approximate frozen-reference task-loss contrasts, while its policy attenuates a bounded set of physical edge bundles and compensates receiver-local coefficient mass.

Message Attribution and Counterfactual Masking. Early graph-specific saliency adaptations transfer gradient and activation-map methods to graph convolutional models (Pope et al., 2019). GNNExplainer optimizes instance-specific subgraphs, whereas PGExplainer learns a parameterized generator that transfers across instances (Ying et al., 2019; Luo et al., 2020). PGM-Explainer represents conditional dependencies among explanatory components (Vu & Thai, 2020), and SubgraphX uses search with Shapley-value scoring to identify important subgraphs (Yuan et al., 2021). TAGE learns task-agnostic explanations for unseen downstream tasks (Xie et al., 2022), while GraphMask learns hard gates over layer-wise messages of a frozen model (Schlichtkrull et al., 2021). Counterfactual methods search for decision-changing graph edits (Bajaj et al., 2021; Lucic et al., 2022); RC-Explainer reinforces causal screening of explanatory subgraphs (Wang et al., 2023). Sparse L_0 gates provide a general differentiable mechanism for discrete model selection (Louizos et al., 2018), but do not by themselves define MEI's receiver-local bundle and compensation constraints. Explanation conclusions remain sensitive to the evaluation frame and to randomization sanity checks (Amara et al., 2022; Adebayo et al., 2018). Because component importance also depends on the ablation operator (Li & Janson, 2024), MEI keeps its singleton target and compensated bundle action conceptually distinct.

Prototype-Based Task Context. Class prototypes summarize labeled examples as representation-space centers and support similarity-based assignment (Snell et al., 2017). Meta-Dataset shows why few-shot evaluation should vary domains, class counts, and shot counts rather than rely on a single episodic template (Triantafillou et al., 2020), and a later controlled study separates the contributions of representation training and test-time adaptation (Luo et al., 2023). On graphs, Meta-GNN learns transferable initializations (Zhou et al., 2019); G-Meta transfers local-subgraph evidence (Huang & Zitnik, 2020); and MetaTNE transforms structural embeddings for novel labels (Lan et al., 2020). Graph Prototypical Networks adapt metric learning to attributed networks (Ding et al., 2020), task-specific-structure learning changes the graph for each episode (Wang et al., 2022), and task-equivariant learning targets transfer from limited meta-tasks (Kim et al., 2023). Broader task transfer can use a single graph model or graph prompts across classification tasks (Liu et al., 2024; Sun et al., 2023); higher-order task affinities and subgraph pooling instead address task interaction and negative transfer (Li et al., 2023a; Wang et al., 2024). Topology-induced optimal transport supplies a graph-aware regularizer for fine-tuning (Zhang et al., 2022). In MEI, prototypes are neither the classifier nor an

intervention rule: they form a training-label-only eligibility filter, while the effect estimator supplies the ranking.

Uncertainty and Deployment Diagnostics. Aleatoric and epistemic uncertainty describe different failure sources (Kendall & Gal, 2017), and deep ensembles provide a practical non-Bayesian estimator of predictive uncertainty (Lakshminarayanan et al., 2017). Large-scale shift experiments show that calibration can degrade away from the training distribution (Ovadia et al., 2019). For graphs, Graph Posterior Networks model Bayesian predictive uncertainty under node dependence (Stadler et al., 2021), while GATS analyzes and corrects graph-specific miscalibration (Hsu et al., 2022). Conformalized GNNs provide marginal coverage guarantees for graph predictions (Huang et al., 2023); under covariate shift, weighted conformal prediction identifies the extra assumptions needed to retain coverage (Tibshirani et al., 2019). These works motivate reporting confidence and shift behavior for MEI, but none turns an uncertainty score into the claimed utility of a compensated bundle action.

Interference and Evaluation Discipline. Potential-outcomes frameworks make explicit that a unit's response may depend on other units' assignments (Hudgens & Halloran, 2008). Exposure mappings and design-based estimands extend this reasoning to general interference (Aronow & Samii, 2017), while network-randomized designs can reduce bias when interference follows observed relations (Eckles et al., 2017). MEI borrows this caution only as terminology: its loss contrast is computed in a fixed trained model and is not a randomized causal estimand. Finally, known sensitivity to data splits, training protocols, and hyperparameter selection in GNN comparisons (Shchur et al., 2018) motivates the frozen-reference and replay audit, but does not establish robustness beyond the reported protocol.

E Reference and Compatible Host Details

E.1 CM MGNN Readout and Reference Training

For edge $e = (u, v)$ from sender u to receiver v with endpoint degrees d_u^{out} and d_v^{in}, the reference host uses

$$\alpha_{uv} = \left(d_u^{\mathrm{out}} d_v^{\mathrm{in}}\right)^{-1/2}, \qquad \omega_{uv}^{(k)} = \begin{cases} \alpha_{uv}(d_v^{\mathrm{in}})^{-1/2}, & k = 0, \\ \alpha_{uv}, & k > 0. \end{cases} \tag{18}$$

For modality m, let $\mathbf{h}_u^{(m,k)}$ and $\mathbf{h}_v^{(m,k)}$ be the endpoint states at layer k and $\varphi_m^{(k)}$ the message map, and $\rho_m^{(k)}$ the activation. The complete message and fully open update equations are given in Section 5.1. Only the raw feature entry layer receives the additional receiver mass factor in Equation 18.

For modality m, an encoder with affine transformation, normalization, GELU, and dropout initializes $\mathbf{h}_v^{(m,0)} = \mathrm{Enc}_m(\mathbf{x}_v^{(m)})$ at a common hidden width. After the fully open propagation in Equation 6, a feature gate combines the image and text states. In parallel, $\mathbf{f}_v^{\mathrm{ego}}$ fuses the encoded modalities without graph propagation. The complete readout is

$$\begin{aligned} \boldsymbol{\gamma}_v &= \sigma\Big(\mathbf{W}_g[\mathbf{h}_v^{(I,K)} \| \mathbf{h}_v^{(T,K)}] + \mathbf{b}_g\Big), \\ \mathbf{f}_v^{\mathrm{prop}} &= \mathrm{LN}\Big(\boldsymbol{\gamma}_v \odot \mathbf{h}_v^{(I,K)} + (\mathbf{1} - \boldsymbol{\gamma}_v) \odot \mathbf{h}_v^{(T,K)}\Big), \\ \mathbf{f}_v &= \mathrm{LN}(\mathbf{f}_v^{\mathrm{prop}} + \sigma(\xi)\mathbf{W}_{\mathrm{ego}}\mathbf{f}_v^{\mathrm{ego}}). \end{aligned} \tag{19}$$

The graph free residual is a declared architectural path. It does not imply that all receiver information remains recoverable after earlier nonlinear transformations. A task head maps $\mathbf{f}_v$ to the task output.

The fully open host is fitted on training instances with its task specific scalar loss. Two projection heads map $\mathbf{f}_v$ toward the frozen image and text feature spaces, where symmetric sampled InfoNCE uses the corresponding node as the positive and other sampled nodes as negatives. The first stage objective is

$$\mathcal{L}_{\mathrm{ref}} = \mathcal{L}_{\mathrm{task}} + \lambda_I \mathcal{L}_{\mathrm{NCE}}^I + \lambda_T \mathcal{L}_{\mathrm{NCE}}^T. \tag{20}$$

These auxiliary terms regularize the shared representation toward both frozen modalities. They neither train nor evaluate the intervention.

E.2 Compatible Host Contract

MEI can be reused only when another frozen multimodal host exposes and resumes the message semantics used for diagnosis and action. Its adapter must satisfy the following requirements:

1. it exposes complete messages specific to each receiver, neighbor, and modality at a fixed boundary;
2. it provides a scalar task loss for each training instance and identifies the receiver representations involved in that instance;
3. it supports deletion and coefficient based reinjection at the same boundary;
4. it provides task context available at deployment while keeping coefficients before intervention independent of the MEI action;
5. it accepts one shared raw coefficient per physical edge bundle and retains any declared receiver contribution outside the bundle coefficients;
6. it reproduces the original computation in evaluation mode exactly under identity coefficients.

Host specific projections may map exposed tensors into the scoring width of MEI, but they do not replace or transform the messages reinjected into the host. Models that irreversibly fuse neighbors or modalities before the controlled boundary, cannot expose a scalar task loss and its active receivers, or cannot perform coherent reinjection fall outside this compatibility class. Candidate construction, including class prototypes when training classes exist, belongs to the task specific policy adapter rather than host compatibility. CM MGNN satisfies the contract directly. The contract establishes architectural reuse. Whether MEI improves a compatible host remains an empirical question.

F Experimental Details

F.1 Datasets

All evaluated datasets provide textual and visual attributes for every node. Grocery is derived from Amazon's Grocery and Gourmet Food category. Its nodes are food or household products, edges are complementary purchases from "joint purchase" records, product titles and nutritional descriptions are text, and packaging images are visual attributes. Movies comes from Amazon's Movies and TV category, where DVD or Blu-ray products are connected through joint purchasing behavior. Plot synopses and customer reviews provide text, and cover art provides visual attributes. Both datasets are used for node classification.

DY is built from Douyin short videos, with shared interaction edges such as videos liked by the same user. Captions and hashtags provide text, and video frames provide visual attributes. Bili Dance contains dance performances or tutorial videos linked by shared viewing behavior. Descriptions and hashtags provide text, and dance keyframes provide visual attributes. Both datasets are used for link prediction. Toys is drawn from Amazon's Toys and Games category, in which joint purchasing relations connect product nodes with textual product specifications and age recommendations and visual product photos. It is used for modality retrieval. Flickr30k is an image and text reasoning dataset whose OpenMAG graph links image regions and caption phrases through semantic grounding annotations, and it is used for graph to text generation. SemArt is built from the Web Gallery of Art. Its artworks are linked by attributes such as artist, period, or school, with titles and historical comments as text and painting images as visual attributes. It is used for graph to image generation.

F.2 Baselines

We use an MLP as a feature only comparator. GCN is a topology focused baseline that aggregates information from immediate neighbors with a first order graph convolution, and GAT learns adaptive attention coefficients over neighbors. DMGC decomposes the graph into cross modality homophily enhanced and modality specific heterophily aware graphs, combining low pass and high pass filtering under self supervised alignment. DGF uses dual graph filtering to denoise multimodal representations, together with three way contrastive learning across modalities, neighborhoods, and communities.

Table 9: Statistics of the datasets used in the experiments. None in the label column denotes a task without a node class label count. Every dataset uses a 60%/20%/20% train/validation/test split.

Dataset	Domain	Nodes	Edges	Labels
Grocery	Electronic Commerce	17,074	171,340	20
Movies	Electronic Commerce	16,672	218,390	20
Toys	Electronic Commerce	20,695	126,886	18
Bili Dance	Video Recommendation	2,307	9,127	None
DY	Video Recommendation	8,299	35,627	None
SemArt	Art Networks	21,382	1,216,432	None
Flickr30k	Image Networks	31,783	181,151	None

MMGCN constructs a separate interaction graph for each modality and fuses modality specific, structural, and shared representations. MGAT adaptively weights modalities and neighbors with gated attention. LGMRec combines a local graph embedding module that disentangles collaborative and modality related information with a global hypergraph module for dependency modeling. UniGraph2 aligns frozen modality specific features in a unified embedding space using a Mixture of Experts module before injecting graph structure with a GNN.

F.3 Tasks

Node Classification. Given a node and its multimodal attributes and graph context, the model predicts its discrete class label. We evaluate this task on Movies and Grocery.

Link Prediction. Given a source node and candidate destinations, the model ranks whether a candidate relation is present. We evaluate this task on DY and Bili Dance.

Modality Retrieval. The model ranks candidates in the opposite modality for a query item. Text to image (T2I) retrieval uses text as the query, and image to text (I2T) retrieval uses image as the query. We evaluate both directions on Toys.

Graph Based Generation. Graph to text (G2Text) generates a textual description based on graph information and is evaluated on Flickr30k. Graph to image (G2Image) generates an image based on graph information and is evaluated on SemArt.

F.4 Metrics

Higher values indicate better performance for all reported metrics. Accuracy is the proportion of nodes assigned the correct class. Macro F1 is the unweighted mean of the per class F1 scores and therefore gives each class equal weight. Mean reciprocal rank (MRR) averages the reciprocal rank of the correct target in a candidate list, whereas Hits@3 is the proportion of cases in which that target is ranked among the top three candidates. We use MRR and Hits@3 for link prediction, and MRR for modality retrieval.

For G2Text, BLEU-4 measures four gram overlap with reference text, and CIDEr measures consensus with reference descriptions using weighted n gram matching. For G2Image, CLIPScore measures alignment between the generated image and its textual input signal in a shared image and text representation space.

F.5 Additional Ablation and Hyperparameter Results

The component ablation and hyperparameter sensitivity analyses are summarized in Figure 6.

F.6 Selective Intervention Audit

This planned audit evaluates the deployed physical edge bundle action rather than the fidelity of the modality resolved singleton targets. In particular, it asks whether the bundles selected by MEI yield larger receiver loss reductions under final layer compensated attenuation than equally budgeted, less

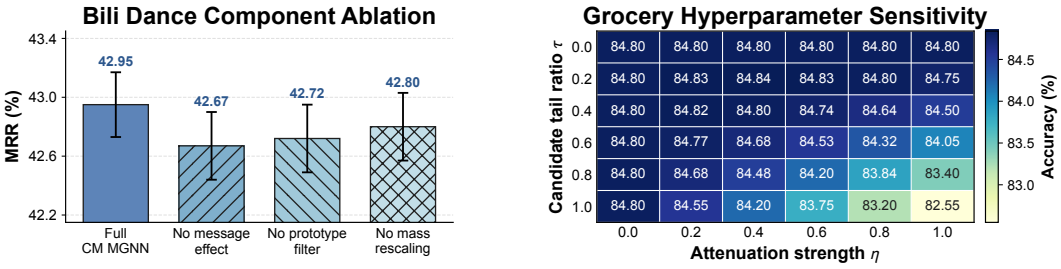


Figure 6: The additional component ablation on Bili Dance appears left, and hyperparameter sensitivity on Grocery appears right.

selective alternatives. It therefore does not treat the singleton deletion effect in Equation 2 as the utility of the bundle action defined by Equation 12.

Audit Setting and Policies. The audit is defined for node classification on the -15pp heterophily pressure graphs of Movies and Grocery. It freezes the fully open reference, the global lower 10% score tail, the candidate physical edge bundles, and the final layer replay boundary. For each receiver selected by the deployed MEI policy, all compared policies attenuate exactly one nonself bundle, with $B = 1$, $\eta = 0.5$, and receiver local coefficient mass compensation fixed as in Equations 5 and 12.

MEI uses its score tail and prototype mismatch filters. A random policy draws one bundle from the same receiver specific candidate set with low scores, whereas a policy using only scores selects by score without the prototype mismatch filter. An offline oracle chooses the bundle with the largest realized action utility from that same candidate set. The oracle is a diagnostic upper bound only and is not a deployment policy. Test labels are used only after all policies have selected their bundles to measure receiver loss. They are never used for scoring, candidate filtering, prototype assignment, or edge selection.

Metrics. For an eligible physical edge bundle $e = (u, v)$, let L_v^{open} be the frozen fully open receiver loss and let $L_v^{\text{att}(e)}$ be the loss after attenuating e with the same compensated replay used at deployment. The action utility is

$$U_e = L_v^{\text{open}} - L_v^{\text{att}(e)}. \tag{21}$$

Thus, larger U_e is better. The primary measures are mean action utility and the positive utility rate, namely the fraction of selected bundles with $U_e > 0$. We additionally report the selective gain relative to the matched random policy and, when the denominator is positive, the oracle normalized utility $(\overline{U}_{\text{MEI}} - \overline{U}_{\text{random}})/(\overline{U}_{\text{oracle}} - \overline{U}_{\text{random}})$. Macro F1 after complete policy deployment is a secondary task level measure. These metrics are specified here as an audit protocol. No audit outcomes, statistical claims, or generalization claims across tasks are reported.