Self-Supervised Transferability-Sensitive Multimodal Heterogeneous Graph Learning

1st Khalil BACHIRI
ETIS Laboratory, ENSEA, UMR8051, CNRS
CY Cergy Paris Université
Cergy, France
LaMSN - La Maison des Sciences Numériques, USPN
khalil.bachiri@cyu.fr

2nd Maria MALEK
ETIS Laboratory, ENSEA, UMR8051, CNRS
CY Cergy Paris Université
Cergy, France
maria.malek@cyu.fr

3rd Ali YAHYAOUY
L3IA Laboratory
Sidi Mohamed Ben Abdellah University
Fez, Morocco
LaMSN - La Maison des Sciences Numériques, USPN
ali.yahyaouy@usmba.ac.ma

***Abstract*—Multimodal heterogeneous graphs are increasingly used to model complex systems combining diverse entity types, relations, and data modalities. While multimodal graph learning has shown strong empirical performance, existing methods often rely on uniform cross-modal alignment or indiscriminate fusion, which can induce negative transfer when modalities are imbalanced, sparse, or semantically incompatible. In this paper, we propose a self-supervised, transferability-sensitive framework for multimodal heterogeneous graph neural networks (S^2T-MMHGNN). Instead of enforcing symmetric alignment, our approach leverages directional predictive self-supervision to explicitly estimate inter-modality transferability and uses this signal to regulate multimodal fusion. This design selectively amplifies complementary modalities while suppressing unreliable cross-modal interactions. We provide theoretical analysis establishing collapse avoidance and monotone fusion behavior under increasing modality incompatibility. Extensive experiments on five real-world benchmarks demonstrate consistent improvements over state-of-the-art heterogeneous, multimodal, and self-supervised baselines, particularly under noisy and sparse multimodal settings.**

***Index Terms*—Multimodal Learning, Heterogeneous Graphs Learning, Self-Supervised Learning, Transferability Modeling, Negative Transfer Mitigation, Graph Neural Networks, Multimodal Fusion, Statistical Learning**

I. Introduction

Heterogeneous graphs provide a natural abstraction for modeling complex real-world systems with multiple node and relation types, such as recommender platforms, review networks, and multimedia knowledge graphs. In modern applications, these graphs are rarely purely structural: nodes and edges are increasingly associated with multiple modalities, each capturing complementary aspects of the underlying semantics. Learning effective representations from such multimodal heterogeneous graphs is therefore a key challenge for a wide range of downstream tasks, including node classification, ranking, recommendation, and retrieval.

Graph representation learning has evolved from random-walk and proximity-preserving embeddings [1], [2] to graph neural networks based on neighborhood aggregation and message passing. In heterogeneous graphs, models such as HAN [3] exploit type-aware attention and meta-path semantics, while lightweight architectures like LightGCN [4] offer scalable solutions for interaction graphs. Nevertheless, most heterogeneous Graph Neural Networks (GNNs) remain largely structure-centric, either overlooking rich multimodal information or treating it as static node attributes without explicitly modeling cross-modal reliability or compatibility.

To exploit multimodal information, a dominant line of work attaches modality encoders and applies fusion strategies such as early fusion, late fusion, attention-based fusion, or gated fusion [5], [6]. In recommendation and multimedia graph learning, this has led to effective multimodal graph baselines, including VBPR [7], MMGCN [8], MGAT [9], and LATTICE [10], as well as more recent robust or target-aware fusion designs such as FREEDOM [11] and MONET [12] and recent work has also shown that combining multimodal graph learning with sequential modeling can better capture semantic and temporal dependencies in video recommendation [13]. Despite these advances, a key assumption is often left implicit: modalities are treated as mutually beneficial and broadly compatible, so that aligning or combining them should monotonically improve representations.

In practice, multimodal observations are often imbalanced, sparse, and semantically mismatched, so that uniformly aligning or indiscriminately fusing modalities can induce negative transfer, where noisy or incompatible signals degrade representations. Self-supervised learning (SSL) has proven effective for leveraging unlabeled graph data through mutual-information maximization [14], [15], contrastive learning [16]–[18], and predictive modeling [19], [20]. In multimodal settings, SSL

has also been adopted to align modalities, as in BM3 [21], SLMRec [22] and mm-hgnn [23] heterogeneity-aware multimodal fusion HGNN. However, most existing SSL formulations enforce uniform cross-modal alignment, which can exacerbate negative transfer when modalities are weakly correlated or incompatible. However, existing multimodal SSL objectives typically treat alignment as the goal, whereas our goal is to estimate transferability and use it to control cross-modal interaction.

In this paper, we propose S^{2}T-MMHGNN, a transferability-sensitive self-supervised framework for multimodal heterogeneous graphs that uses self-supervision as a principled tool to estimate and regulate inter-modality transferability. By modeling directional cross-modal predictability rather than enforcing symmetric alignment, our approach derives a transferability-based incompatibility signal that directly controls multimodal fusion, amplifying complementary modalities while suppressing unreliable transfers. This selective, heterogeneity-aware design addresses the limitations of uniform alignment and is supported by theoretical guarantees on collapse avoidance and monotone fusion behavior.

II. Related Work

A. Self-Supervised Learning on Graphs

Self-supervised learning (SSL) on graphs aims to learn representations from unlabeled data by exploiting structural and attribute regularities. Early methods preserve proximity via random walks [1], [2], matrix factorization [24], or autoencoders [25]. GNN-based SSL is dominated by mutual-information maximization (DGI, InfoGraph) [14], [15], graph contrastive learning with augmented views (GraphCL, JOAO) [16]–[18], and predictive approaches such as masked modeling and latent graph prediction [19], [20]. Most methods enforce that representations from different modalities should be invariant or aligned, assuming that modalities are always consistent and mutually beneficial.

B. Multimodal Graph Learning and Fusion

Multimodal graph learning extends GNNs to graphs with multi-source node or edge attributes. Typical pipelines rely on modality-specific encoders followed by early, late, or intermediate (attention/gated) fusion [8]. Graph-based models exploit relational structure to propagate multimodal signals [10], [26], while recent work incorporates richer semantic or hypergraph structures [11], [27]. However, most approaches implicitly assume that aggregating or aligning modalities is always beneficial, overlooking redundancy and semantic conflict. Topological perspectives have also been explored to enhance multimodal graph learning by capturing higher-order structural dependencies and improving robustness under sparsity [28]. Multimodal fusion integrates multiple modalities into unified representations via additive, gated, attention-based, or higher-order interaction mechanisms [5], [6]. Tensor and polynomial fusion explicitly model cross-modal interactions [29], [30] but suffer from high computational complexity. Alignment-based methods, including CCA and contrastive learning [31], improve cross-modal consistency but may induce negative transfer when modalities differ in quality or semantics, motivating selective and heterogeneity-aware fusion. Together, these limitations highlight the need for representation learning frameworks that jointly account for graph heterogeneity, modality-specific semantics, and controlled cross-modal information transfer.

III. Method

We propose self-supervised transferability-sensitive for multimodal heterogeneous graph neural networks (S^{2}T-MMHGNN), a principled framework that extends multimodal heterogeneous graph learning with an explicit, self-supervised modeling of inter-modality transferability. The central thesis of this work is that self-supervision should not enforce uniform alignment across modalities, but should instead quantify, regulate, and selectively control cross-modal information transfer in the presence of structural and observational heterogeneity, thereby preventing negative transfer when modalities are incompatible. Fig. 1 provides a high-level overview of the architecture, including modality-wise graph propagation, self-supervised transferability estimation, and transferability-aware fusion.

For each modality stream, we learn a graph-aware representation using modality-wise heterogeneous propagation. Then, rather than aligning modalities symmetrically, we learn directional cross-modal predictors and define a transferability gap that measures how reliably one modality can explain another relative to the intrinsic difficulty of the target modality. Finally, we incorporate this signal into a transferability-aware fusion mechanism that with a no-negative-transfer guarantee under mild conditions incompatible transfers.

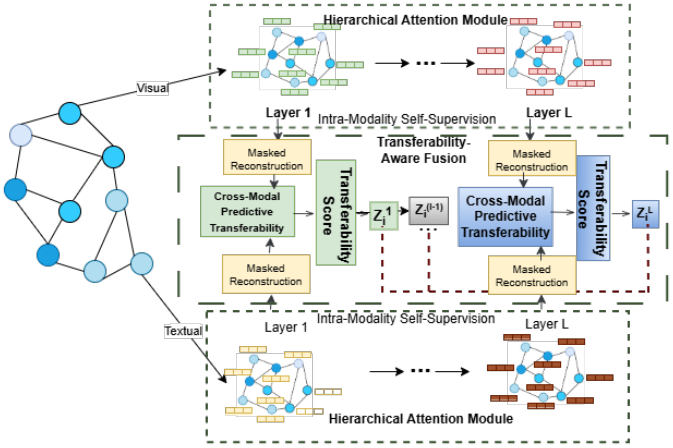


Fig. 1: Self-Supervised Multimodal Architecture

A. Problem Setup

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \phi, \varphi)$ be a heterogeneous graph, where $\phi : \mathcal{V} \to \mathcal{A}$ maps nodes to types and $\varphi : \mathcal{E} \to \mathcal{R}$ maps edges to relation types. Each node $v \in \mathcal{V}$ has a subset of observed modalities $\mathcal{M}_v \subseteq \mathcal{M}$, and for each $m \in \mathcal{M}_v$ we observe

$X_v^{(m)}$. We denote the node set where modality m is observed as $\mathcal{V}_m = \{v \in \mathcal{V} : m \in \mathcal{M}_v\}$ and the co-observed set for modalities (m_1, m_2) as $\mathcal{V}_{m_1,m_2} = \mathcal{V}_{m_1} \cap \mathcal{V}_{m_2}$.
We view the observed modalities as noisy and possibly incomplete measurements of shared, task-relevant semantics S_v. Formally, each modality is generated conditionally on S_v: $X_v^{(m)} \sim p_m(\cdot \mid S_v), m \in \mathcal{M}_v$. This assumption merely motivates learning representations robust to heterogeneous relations, missing modalities, and negative transfer.

B. Modality-Wise Heterogeneous Graph Propagation

To preserve modality semantics while leveraging relation heterogeneity, we perform heterogeneous message passing independently per modality. For layer $l = 0, \ldots, L-1$ and modality m:

$$H_v^{(m,l+1)} = \sigma\left(\sum_{r\in\mathcal{R}} \sum_{u\in\mathcal{N}_r(v)} \frac{1}{c_{v,r}} W_r^{(l,m)} H_u^{(m,l)}\right), \quad (1)$$

where $\mathcal{N}_r(v) = \{u : (u, v) \in \mathcal{E}, \varphi(u, v) = r\}$ is the r-typed neighborhood, $c_{v,r}$ is a normalization factor $|\mathcal{N}_r(v)|$, $W_r^{(l,m)} \in \mathbb{R}^{d\times d}$ is relation-specific and modality-specific. We use residual connections to mitigate over-smoothing:

$$\tilde{H}_v^{(m,l+1)} = H_v^{(m,l+1)} + H_v^{(m,l)}. \quad (2)$$

After L layers, the final modality-wise graph embedding is:

$$\tilde{H}_v^{(m)} := \tilde{H}_v^{(m,L)}. \quad (3)$$

Eq. (1) avoids early cross-modal entanglement: each modality stream learns a graph-aware summary of its own observations before any fusion. This is crucial because inter-modality transferability is not uniform across node types, relations, and observational regimes.

C. Intra-Modality Self-Supervision

We impose an intra-modality objective that makes each modality pathway informative on its own. We adopt masked reconstruction per modality:

$$\mathcal{L}_{\text{intra}} = \sum_{m\in\mathcal{M}} \mathbb{E}_{v\in\mathcal{V}_m}\left[\ell_{\text{rec}}^{(m)}\Big(X_v^{(m)}, D_m(\tilde{H}_v^{(m)})\Big)\right], \quad (4)$$

where D_m is a modality-specific decoder/head and $\ell_{\text{rec}}^{(m)}$ is modality-appropriate. This loss plays two roles:

- Collapse prevention: it excludes trivial constant solutions for modality streams.
- Intrinsic difficulty baseline: $\mathcal{L}_{\text{intra}}^{(m)}$ (the m-term in Eq. (4)) quantifies how hard it is to learn modality m without help from other modalities, which will be used to define transferability.

For clarity, we denote the modality-specific intra loss:

$$\mathcal{L}_{\text{intra}}^{(m)} := \mathbb{E}_{v\in\mathcal{V}_m}\left[\ell_{\text{rec}}^{(m)}(X_v^{(m)}, D_m(\tilde{H}_v^{(m)}))\right]. \quad (5)$$

D. Directional Cross-Modality Predictive Self-Supervision

Instead of symmetric alignment (which can induce negative transfer), we learn directional prediction to measure how informative one modality is about another. For each ordered pair (m_1, m_2) with $m_1 \neq m_2$, we learn a predictor $G_{m_1\to m_2}$ and define:

$$\mathcal{L}_{\text{pred}}^{(m_1\to m_2)} = \mathbb{E}_{v\in\mathcal{V}_{m_1,m_2}}\left[\left\|G_{m_1\to m_2}(\tilde{H}_v^{(m_1)}) - \text{sg}\Big(\tilde{H}_v^{(m_2)}\Big)\right\|_2^2\right], \quad (6)$$

where $\text{sg}(\cdot)$ is stop-gradient. Stop-gradient prevents representational collapse/shortcutting and makes the prediction error a reliable transferability signal. If m_1 can reliably predict m_2 for co-observed nodes, then the two modalities share substantial semantic overlap on the current data distribution. If prediction is consistently difficult, forcing alignment would likely create negative transfer.

E. Transferability and Heterogeneity Modeling

We now formalize how cross-modal transferability is quantified and how modality heterogeneity is modeled within S^{2}T-MMHGNN. The objective is to derive a principled signal that reflects whether information transfer between two modalities is beneficial or potentially harmful, and to expose this signal explicitly for downstream fusion.

We measure transferability from modality m_1 to modality m_2 by comparing the difficulty of predicting the target modality from another modality against the intrinsic difficulty of modeling the target modality itself. This leads to the following definition:

$$T(m_1 \to m_2) = \mathcal{L}_{\text{pred}}^{(m_1\to m_2)} - \mathcal{L}_{\text{intra}}^{(m_2)}. \quad (7)$$

A small value of $T(m_1 \to m_2)$ indicates that representations from modality m_1 are nearly as informative as direct supervision from modality m_2, while larger values indicate unreliable transfer and a high risk of negative transfer if the modalities are fused indiscriminately.

Since transferability is inherently directional, we aggregate the two directions into a symmetric heterogeneity measure that reflects mutual incompatibility:

$$d(m_1, m_2) = T(m_1 \to m_2) + T(m_2 \to m_1), \quad (8)$$

and use its non-negative stabilized form

$$d_+(m_1, m_2) = \max\{0, d(m_1, m_2)\}, \quad (9)$$

which serves as the core signal for regulating multimodal fusion. Large values of $d(m_1, m_2)$ correspond to weak mutual predictability and high modality heterogeneity, whereas values close to zero indicate redundancy or strong semantic overlap.

The intra-modality reconstruction objective further ensures that each modality stream remains informative on its own. In particular, collapsed representations that map all nodes to a constant embedding cannot minimize the reconstruction loss. Indeed, if $\tilde{H}_v^{(m)} \equiv c_m$ for all v, the decoder output becomes constant across nodes, while reconstruction targets

remain node-dependent due to stochastic corruption and non-degenerate modality signals. Consequently, the gradient

$$\nabla_{\tilde{H}^{(m)}} \mathcal{L}_{\text{intra}}^{(m)} = \mathbb{E}_{v \in \mathcal{V}_m} \Big[\nabla_h D_m(c_m)^\top \, \nabla_{\hat{x}} \ell_{\text{rec}}^{(m)} \Big(X_v^{(m)}, D_m(c_m) \Big) \Big] \tag{10}$$

cannot vanish, excluding collapsed solutions and ensuring meaningful modality-specific representations. When $\lambda_1 > 0$ in the final objective, this property directly extends to the full training loss.

Under squared loss and Bayes-optimal predictors, the transferability gap admits a statistical interpretation in terms of conditional unpredictability:

$$T(m_1 \to m_2) = \mathbb{E}\Big[\mathrm{Var}\big(\tilde{H}^{(m_2)} \mid \tilde{H}^{(m_1)}\big)\Big] - \mathbb{E}\Big[\mathrm{Var}\big(\tilde{H}^{(m_2)} \mid X^{(m_2)}\big)\Big]. \tag{11}$$

This expression shows that transferability quantifies how much uncertainty about modality m_2 remains after conditioning on modality m_1, relative to the uncertainty intrinsic to modality m_2 itself.

Finally, under δ-optimal training and finite-capacity predictors, the transferability gap concentrates around a Bayes-risk difference:

$$T(m_1 \to m_2) \;\approx\; R^\star_{m_1 \to m_2} - R^\star_{m_2 | X^{(m_2)}}, \tag{12}$$

up to approximation and optimization errors, yielding the bounds

$$T(m_1 \to m_2) \;\le\; \big(R^\star_{m_1 \to m_2} - R^\star_{m_2 | X^{(m_2)}}\big) + \varepsilon_{m_1 \to m_2} + \delta, \tag{13}$$

$$T(m_1 \to m_2) \;\ge\; \big(R^\star_{m_1 \to m_2} - R^\star_{m_2 | X^{(m_2)}}\big) - \varepsilon_{m_2,\text{intra}} - \delta. \tag{14}$$

These relations justify the use of $d_+(m_1, m_2)$ as a principled heterogeneity signal and motivate its role in transferability-aware fusion.

F. Transferability-sensitive Fusion (Negative Transfer Control)

The final step of S^{2}T-MMHGNN consists in fusing modality-specific representations while explicitly controlling negative transfer. The core principle is simple: modalities that are weakly transferable should contribute less to the fused representation. To this end, we leverage the incompatibility score d_+ introduced in Section III-E as a direct regulator of multimodal fusion.

For a node v with observed modalities $\mathcal{M}_v$, each modality embedding $\tilde{H}_v^{(m)}$ is assigned a fusion weight that balances two factors: (i) an intra-modality quality score reflecting the reliability of the modality itself, and (ii) a heterogeneity penalty that suppresses modalities incompatible with the others. We first define a transferability-penalized softmax fusion rule. Let $s(\cdot)$ denote a learnable quality function. The fusion weights are given by:

$$\alpha_v^{(m)} = \frac{\exp\Big(s(\tilde{H}_v^{(m)}) - \beta \sum_{m' \in \mathcal{M}_v \setminus \{m\}} d_+(m, m')\Big)}{\sum_{k \in \mathcal{M}_v} \exp\Big(s(\tilde{H}_v^{(k)}) - \beta \sum_{m' \in \mathcal{M}_v \setminus \{k\}} d_+(k, m')\Big)}, \tag{15}$$

where $\beta > 0$ controls the strength of transferability regularization. This formulation ensures that increasing modality incompatibility directly reduces the corresponding fusion weight.

An equivalent and more explicit formulation decomposes fusion into a quality term and a heterogeneity penalty. Using $s(\tilde{H}) = a^\top \sigma(W\tilde{H} + b)$, we define:

$$\omega_v^{(m)} = a^\top \sigma\Big(W\,\tilde{H}_v^{(m)} + b\Big) - \beta \, \frac{1}{|\mathcal{M}_v| - 1} \sum_{\substack{m' \in \mathcal{M}_v \\ m' \neq m}} d_+(m, m'), \tag{16}$$

followed by softmax normalization:

$$\alpha_v^{(m)} = \frac{\exp(\omega_v^{(m)})}{\sum_{k \in \mathcal{M}_v} \exp(\omega_v^{(k)})}, \qquad Z_v = \sum_{m \in \mathcal{M}_v} \alpha_v^{(m)} \, \tilde{H}_v^{(m)}. \tag{17}$$

Eqs. (15) and (17) correspond to equivalent parameterizations of the same fusion mechanism.

This design induces a monotonic behavior: increasing incompatibility scores $d_+(m, m')$ strictly decreases the corresponding fusion score $\omega_v^{(m)}$, and therefore cannot increase the normalized weight $\alpha_v^{(m)}$ due to the monotonicity of the softmax function. As a result, incompatible modalities are automatically down-weighted without requiring hard exclusion. Beyond interpretability, this fusion rule provides an explicit safeguard against negative transfer. Let $q(\cdot)$ denote the task predictor with Lipschitz constant K, and let the task loss $\ell_{\text{task}}(\cdot, \cdot)$ be L-Lipschitz. Let $m^\star$ denote the best single modality in expected task risk. For the fused representation

$$Z_v = \sum_{m \in \mathcal{M}_v} \alpha_v^{(m)} \, \tilde{H}_v^{(m)},$$

the expected task loss satisfies:

$$\begin{aligned} \mathbb{E}[\ell_{\text{task}}(y, q(Z_v))] \;\le\; & \mathbb{E}\Big[\ell_{\text{task}}\Big(y, q(\tilde{H}_v^{(m^\star)})\Big)\Big] \\ & + \; LK\,\mathbb{E}\left[\sum_{m \neq m^\star} \alpha_v^{(m)} \, \big\|\tilde{H}_v^{(m)} - \tilde{H}_v^{(m^\star)}\big\|\right]. \end{aligned} \tag{18}$$

By Lipschitz continuity,

$$\ell_{\text{task}}(y, q(Z_v)) \le \ell_{\text{task}}(y, q(\tilde{H}_v^{(m^\star)})) + LK \, \|Z_v - \tilde{H}_v^{(m^\star)}\|, \tag{19}$$

and using the convex structure of Z_v,

$$\begin{aligned} \Big\|Z_v - \tilde{H}_v^{(m^\star)}\Big\| &= \left\| \sum_{m \neq m^\star} \alpha_v^{(m)} \Big(\tilde{H}_v^{(m)} - \tilde{H}_v^{(m^\star)}\Big) \right\| \\ &\le \sum_{m \neq m^\star} \alpha_v^{(m)} \Big\|\tilde{H}_v^{(m)} - \tilde{H}_v^{(m^\star)}\Big\|. \end{aligned} \tag{20}$$

As incompatibility scores $d_+(m, m^\star)$ increase, the corresponding fusion weights $\alpha_v^{(m)}$ vanish, causing the excess term in Eq. (18) to converge to zero. Thus, transferability-aware fusion guarantees that combining modalities cannot asymptotically perform worse than relying on the best individual modality, while still exploiting complementary information when transferability is high.

G. Task Objective and Final Loss

Let $\mathcal{L}_{\text{task}}$ denote the downstream task loss (classification, link prediction, ranking). The final objective is:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_1 \mathcal{L}_{\text{intra}} + \lambda_2 \sum_{m_1 \neq m_2} \mathcal{L}_{\text{pred}}^{(m_1 \to m_2)}. \tag{21}$$

Unlike alignment-based SSL, S^{2}T-MMHGNN selectively regulates cross-modal interaction based on estimated transferability.

The overall time complexity is $\mathcal{O}\big(L|\mathcal{M}||\mathcal{E}|d + |\mathcal{M}|^2|\mathcal{V}|d\big)$, where the second term accounts for directional cross-modality prediction. Since $|\mathcal{M}|$ is small in practice, the complexity remains comparable to existing multimodal HGNNs.

IV. Experimental Evaluation

In this section, we present a comprehensive experimental evaluation of the proposed S^{2}T-MMHGNN framework. The primary objective of these experiments is to assess whether transferability-sensitive self-supervision can improve multimodal representation learning on heterogeneous graphs, while effectively mitigating negative transfer that commonly arises when uniform cross-modal alignment is enforced in the presence of modality incompatibility.

We evaluate our method S^{2}T-MMHGNN on five widely used real-world benchmarks: IMDB, Amazon Appliances, Amazon Electronics, Amazon Clothing, and Yelp. These datasets exhibit (i) rich node and relation heterogeneity, (ii) diverse multimodal observations, and (iii) varying degrees of inter-modality redundancy and incompatibility. Such characteristics make them particularly suitable for stress-testing the transferability modeling and fusion mechanisms introduced in Section III. Unless stated otherwise, we expand the label space to the top 15-20 most frequent categories per dataset. This setting increases semantic granularity and results in a more challenging multi-class node classification task, enabling a more rigorous evaluation of representation quality.

A. Datasets

Table I summarizes the key statistics of the datasets used in our experiments, including graph size, relational diversity, and the number of target classes.

TABLE I: Summary statistics of the datasets used in our experiments.

Dataset	#Nodes	#Edges	#Rel. Types	#Classes
IMDB	18,620	89,512	8	20
Amazon Appliances	22,134	208,417	8	20
Amazon Electronics	24,803	234,619	9	15
Amazon Clothing	27,645	248,903	9	15
Yelp	41,791	515,200	8	15

a) Evaluation metrics.: We evaluate performance using three complementary metrics: Macro-F1, Micro-F1, and ROC-AUC. Macro-F1 emphasizes balanced performance across classes and is robust to class imbalance, which is prevalent in real-world multimodal datasets. Micro-F1 reflects overall classification accuracy. ROC-AUC measures ranking confidence and is particularly informative for assessing the quality of learned representations and the effectiveness of transferability-aware fusion. All reported results correspond to the mean and standard deviation over five independent runs.

B. Baselines

We compare S^{2}T-MMHGNN against strong and representative baselines spanning heterogeneous graph learning, multimodal fusion, and self-supervised multimodal alignment.

(i) Non-fusion heterogeneous graph models: These methods focus primarily on structural modeling and do not explicitly fuse multimodal signals. **HAN (WWW'19)** [3]: meta-path attention for heterogeneous information networks. **LightGCN (SIGIR'20)** [4]: simplified GCN for user-item interaction graphs.

(ii) Multimodal fusion graph models: VBPR (AAAI'16) [7]: additive fusion of visual features. **MMGCN (MM'20)** [8]: multimodal GCN with late fusion. **MGAT (IJIPM'21)** [9]: attention-based early fusion with graph propagation. **LATTICE (MM'21)** [10]: latent structure mining across modalities. **FREEDOM (MM'23)** [11]: gated fusion of denoised and frozen graph views. **MONET (WSDM'24)** [12]: modality-aware GCN with target-aware attention. **BM3 (WWW'23)** [21]: bootstrap-based multimodal contrastive learning. **SLMRec (TMM'23)** [22]: contrastive SSL on multimodal graphs. **MM-HGNN (IJCIS'25)** [23]: heterogeneity-aware multimodal fusion HGNN.

C. Experimental Setup

All models are trained under identical experimental conditions to ensure fairness. We fix the embedding dimension to $d = 64$ and use $L = 4$ graph layers. Dropout is selected from $[0.5, 0.6]$, and optimization is performed using Adam with early stopping on the validation set. Learning rates are tuned in $\{10^{-3}, 5 \times 10^{-4}, 10^{-4}\}$ and weight decay in $\{10^{-5}, 5 \times 10^{-5}, 10^{-4}\}$.

For S^{2}T-MMHGNN, we set $\lambda_1 = 1$ and $\lambda_2 = 0.5$ to balance intra-modality and predictive self-supervision, and tune the fusion regularization coefficient β in $\{0.1, 0.5, 1.0\}$. We observe stable performance across a wide range of these values, with graceful degradation when transferability regularization is removed ($\beta = 0$). All experiments are conducted on a single GPU (NVIDIA GTX 1080 Ti), and results are averaged over five random seeds.

D. Performance Comparison and Analysis

Table II summarizes node classification results on five multimodal heterogeneous benchmarks. We report Macro-F1, Micro-F1, and ROC-AUC, capturing complementary aspects

TABLE II: Node classification results. Best results are highlighted.

Model	IMDB			Appliances			Electronics			Clothing			Yelp		
	Macro	Micro	AUC	Macro	Micro	AUC	Macro	Micro	AUC	Macro	Micro	AUC	Macro	Micro	AUC
HAN	0.62	0.65	0.68	0.53	0.59	0.62	0.49	0.42	0.55	0.60	0.62	0.66	0.54	0.59	0.62
LightGCN	0.59	0.64	0.63	0.60	0.62	0.61	0.67	0.57	0.61	0.50	0.54	0.59	0.51	0.57	0.60
VBPR	0.66	0.67	0.72	0.60	0.62	0.66	0.68	0.70	0.74	0.66	0.69	0.62	0.54	0.54	0.56
MMGCN	0.69	0.69	0.75	0.73	0.76	0.77	0.71	0.75	0.66	0.70	0.74	0.75	0.61	0.64	0.61
MGAT	0.71	0.70	0.75	0.73	0.78	0.77	0.72	0.76	0.60	0.66	0.71	0.77	0.62	0.63	0.64
LATTICE	0.80	0.83	0.86	0.81	0.83	0.80	0.80	0.82	0.85	0.72	0.76	0.77	0.75	0.68	0.71
BM3	0.65	0.66	0.77	0.78	0.82	0.83	0.77	0.79	0.80	0.76	0.80	0.82	0.76	0.79	0.76
SLMRec	0.81	0.85	0.84	0.80	0.83	0.85	0.79	0.82	0.84	0.78	0.81	0.83	0.82	0.80	0.86
FREEDOM	0.83	0.85	0.87	0.85	0.89	0.91	0.84	0.88	0.89	0.83	0.87	0.88	0.83	0.83	0.85
MONET	0.81	0.83	0.86	0.84	0.88	0.90	0.83	0.87	0.89	0.82	0.86	0.88	0.84	0.87	0.86
MM-HGNN	0.89	0.88	0.90	0.87	0.93	0.94	0.88	0.91	0.93	0.87	0.91	0.92	0.87	0.89	0.92
S^{2}T-MMHGNN (Ours)	**0.91**	**0.90**	**0.92**	**0.88**	**0.95**	**0.96**	**0.89**	**0.93**	**0.96**	**0.88**	**0.93**	**0.94**	**0.89**	**0.90**	**0.94**

of predictive quality: Macro-F1 emphasizes robustness under class imbalance, Micro-F1 reflects overall instance-level accuracy, and AUC measures ranking confidence and margin separation. Across all datasets and metrics, our approach achieves the best performance, establishing a consistent improvement over heterogeneous graph baselines, multimodal fusion methods, and contrastive alignment-based self-supervised models.

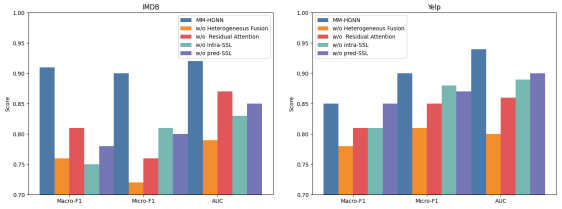
Compared to structure-centric baselines (HAN, LightGCN), S^{2}T-MMHGNN yields large gains on every dataset, confirming that graph topology alone is insufficient when semantics are distributed across text, images, and attributes. For instance, on IMDB, S^{2}T-MMHGNN improves Macro-F1 from 0.62 (HAN) to 0.91, highlighting the necessity of multimodal modeling for fine-grained semantic discrimination. Relative to its direct predecessor MM-HGNN, our approach provides systematic improvements on all benchmarks, with the most pronounced gains in ROC-AUC, indicating more separable and better-calibrated representations. In particular, on Amazon Electronics, our approach increases AUC from 0.93 to 0.96 (+0.03), suggesting that transferability-sensitive regularization improves decision margins under noisy and partially incompatible modalities. Consistent (though smaller) gains are also observed on IMDB (Macro-F1 $0.89 \rightarrow 0.91$), Appliances (Micro-F1 $0.93 \rightarrow 0.94$), Clothing (Macro-F1 $0.87 \rightarrow 0.88$), and Yelp (AUC $0.92 \rightarrow 0.94$).

Early/late fusion methods (VBPR, MMGCN, MGAT) improve over graph-only baselines but remain consistently behind stronger multimodal graph models (LATTICE, FREEDOM, MONET), reflecting the limitations of rigid fusion under modality sparsity and noise. Even for advanced fusion mechanisms (e.g., gated or target-aware attention), the absence of an explicit estimate of transfer reliability can lead to propagating noisy or weakly related signals. By contrast, S^{2}T-MMHGNN explicitly models directional transferability and uses it to regulate fusion, thereby mitigating negative transfer rather than relying on a fixed notion of complementarity. Contrastive SSL baselines (BM3, SLMRec) achieve strong performance, validating the importance of self-supervised signals in multimodal learning. However, their objective typically encourages uniform cross-modal alignment, which can be suboptimal when modalities are weakly correlated, missing, or semantically misaligned. S^{2}T-MMHGNN consistently outperforms these methods, especially in ROC-AUC (e.g., 0.96 on Electronics and 0.94 on Yelp), supporting our core thesis: selective transferability regulation is preferable to universal alignment in multimodal heterogeneous graphs.

The advantage of our method is particularly evident on datasets with higher modality sparsity or noise (Appliances, Electronics, Yelp), where avoiding harmful transfer is critical. On Yelp, where text dominates and auxiliary modalities can be inconsistent, S^{2}T-MMHGNN attains the best AUC (0.94), indicating robust ranking under heterogeneous relational structure and noisy observations.

E. Ablation Study

We conduct an ablation study to quantify the contribution of each major component of S^{2}T-MMHGNN. Specifically, we evaluate the full model against four variants: (i) w/o heterogeneous fusion, (ii) w/o residual attention, (iii) w/o intra-modality SSL, and (iv) w/o predictive (directional) SSL. Results on IMDB and Yelp are reported in Figure 2.



(a) Ablation results on IMDB dataset (b) Ablation results on Yelp dataset

Fig. 2: Quantitative ablation analysis of S^{2}T-MMHGNN. The figures show the performance impact of removing key components.

Removing the heterogeneous fusion module results in the largest performance degradation across all metrics and datasets. On IMDB, Macro-F1 and AUC drop sharply, while on Yelp the degradation is particularly evident in Micro-F1 and AUC. This empirically confirms that naïve aggregation of modalities in heterogeneous graphs leads to negative transfer, and validates our central claim that cross-modal interactions must be explicitly regulated through transferability estimation.

Eliminating residual attention consistently degrades performance, with a stronger effect on IMDB 2a. This indicates that residual connections are crucial for stabilizing deep heterogeneous message passing and mitigating over-smoothing in semantically rich graphs. Without intra-modality SSL, Macro-F1 drops noticeably, especially on Yelp 2b. This suggests that intra-modality self-supervision plays a critical role in preserving modality-specific semantics and preventing representation collapse, particularly under class imbalance and modality sparsity. Removing predictive (directional) SSL results in a consistent performance drop across all settings. Although fusion remains active, it becomes uninformed by transferability signals, demonstrating that predictive SSL is the key mechanism enabling transferability-aware fusion and explicit negative transfer control.

V. Conclusion

In this paper, we introduced S^2T-MMHGNN, a transferability-sensitive self-supervised framework for multimodal heterogeneous graph learning that explicitly addresses negative transfer arising from modality incompatibility. Unlike existing approaches that rely on uniform cross-modal alignment, our method leverages directional predictive self-supervision to estimate inter-modality transferability and uses this signal to regulate multimodal fusion in a selective and heterogeneity-aware manner. We provided theoretical analysis showing that the proposed formulation prevents representation collapse and enforces monotone down-weighting of incompatible modalities. Extensive experiments on five real-world benchmarks demonstrate consistent improvements over strong heterogeneous, multimodal, and self-supervised baselines, particularly under modality sparsity and noise. Overall, our approach establishes a principled connection between self-supervision and controlled cross-modal information transfer, opening new directions for robust and interpretable multimodal graph learning. While S^2T-MMHGNN explicitly controls negative transfer through self-supervised transferability estimation, this signal is inherently data- and objective-dependent and may vary across distributions or training regimes. Addressing adaptive calibration of transferability scores and extending the framework to dynamic or evolving modalities constitute promising directions for future work.

References

[1] B. Perozzi, R. Al-Rfou, and S. Skiena, "DeepWalk: Online Learning of Social Representations," Mar. 2014. [Online]. Available: https://arxiv.org/abs/1403.6652v2

[2] A. Grover and J. Leskovec, "node2vec: Scalable Feature Learning for Networks," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 855–864. [Online]. Available: https://doi.org/10.1145/2939672.2939754

[3] X. Wang, H. Ji, C. Shi, B. Wang, Y. Ye, P. Cui, and P. S. Yu, "Heterogeneous Graph Attention Network," in *The World Wide Web Conference*, ser. WWW '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 2022–2032. [Online]. Available: https://doi.org/10.1145/3308558.3313562

[4] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang, "LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 639–648. [Online]. Available: https://doi.org/10.1145/3397271.3401063

[5] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, Jan. 2009. [Online]. Available: https://ieeexplore.ieee.org/document/4468714

[6] J. Arevalo, T. Solorio, M. Montes-y Gómez, and F. A. González, "Gated Multimodal Units for Information Fusion," Feb. 2017, arXiv:1702.01992 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1702.01992

[7] R. He and J. McAuley, "VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, Feb. 2016, number: 1. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/9973

[8] Y. Wei, X. Wang, L. Nie, X. He, R. Hong, and T.-S. Chua, "MMGCN: Multi-modal Graph Convolution Network for Personalized Recommendation of Micro-video," in *Proceedings of the 27th ACM International Conference on Multimedia*, ser. MM '19. New York, NY, USA: Association for Computing Machinery, Oct. 2019, pp. 1437–1445. [Online]. Available: https://doi.org/10.1145/3343031.3351034

[9] Z. Tao, Y. Wei, X. Wang, X. He, X. Huang, and T.-S. Chua, "MGAT: Multimodal Graph Attention Network for Recommendation," *Information Processing & Management*, vol. 57, no. 5, p. 102277, Sep. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306457320300182

[10] J. Zhang, Y. Zhu, Q. Liu, S. Wu, S. Wang, and L. Wang, "Mining Latent Structures for Multimedia Recommendation," in *Proceedings of the 29th ACM International Conference on Multimedia*, ser. MM '21. New York, NY, USA: Association for Computing Machinery, Oct. 2021, pp. 3872–3880. [Online]. Available: https://doi.org/10.1145/3474085.3475259

[11] X. Zhou and Z. Shen, "A Tale of Two Graphs: Freezing and Denoising Graph Structures for Multimodal Recommendation," in *Proceedings of the 31st ACM International Conference on Multimedia*, ser. MM '23. New York, NY, USA: Association for Computing Machinery, Oct. 2023, pp. 935–943. [Online]. Available: https://dl.acm.org/doi/10.1145/3581783.3611943

[12] Y. Kim, T. Kim, W.-Y. Shin, and S.-W. Kim, "MONET: Modality-Embracing Graph Convolutional Network and Target-Aware Attention for Multimedia Recommendation," in *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, ser. WSDM '24. New York, NY, USA: Association for Computing Machinery, Mar. 2024, pp. 332–340. [Online]. Available: https://dl.acm.org/doi/10.1145/3616855.3635817

[13] K. Bachiri, A. Yahyaouy, M. Malek, and N. Rogovschi, "Multimodal graph learning and sequential modeling for video recommendation," in *2025 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2025, pp. 1–8.

[14] W. Zhao, G. Xu, Z. Cui, S. Luo, C. Long, and T. Zhang, "Deep Graph Structural Infomax," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 4, pp. 4920–4928, Jun. 2023. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/25618

[15] F.-Y. Sun, J. Hoffmann, V. Verma, and J. Tang, "InfoGraph: Unsupervised and Semi-supervised Graph-Level Representation Learning via Mutual Information Maximization," Jan. 2020, arXiv:1908.01000 [cs]. [Online]. Available: http://arxiv.org/abs/1908.01000

[16] Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, and Y. Shen, "Graph contrastive learning with augmentations," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20. Curran Associates Inc., 2020, pp. 5812–5823.

[17] Y. You, T. Chen, Y. Shen, and Z. Wang, "Graph contrastive learning automated," in *Proceedings of the 38th International Conference on Machine Learning*. PMLR, Jul. 2021, iSSN: 2640-3498. [Online]. Available: https://proceedings.mlr.press/v139/you21a.html

[18] H. Chen, X. Wang, Z. Zhang, H. Li, L. Feng, and W. Zhu, "AutoGFM: Automated graph foundation model with adaptive architecture customization," in *Proceedings of the 42nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Singh, M. Fazel, D. Hsu, S. Lacoste-Julien, F. Berkenkamp, T. Maharaj, K. Wagstaff, and J. Zhu, Eds., vol.

267. PMLR, 13–19 Jul 2025, pp. 9295–9315. [Online]. Available: https://proceedings.mlr.press/v267/chen25bp.html

[19] Z. Hou, X. Liu, Y. Cen, Y. Dong, H. Yang, C. Wang, and J. Tang, "GraphMAE: Self-Supervised Masked Graph Autoencoders," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ser. KDD '22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 594–604. [Online]. Available: https://doi.org/10.1145/3534678.3539321

[20] Y. Xie, Z. Xu, and S. Ji, "Self-Supervised Representation Learning via Latent Graph Prediction," in *Proceedings of the 39th International Conference on Machine Learning*. PMLR, Jun. 2022, pp. 24 460–24 477, iSSN: 2640-3498. [Online]. Available: https://proceedings.mlr.press/v162/xie22e.html

[21] X. Zhou, H. Zhou, Y. Liu, Z. Zeng, C. Miao, P. Wang, Y. You, and F. Jiang, "Bootstrap Latent Representations for Multi-modal Recommendation," in *Proceedings of the ACM Web Conference 2023*, ser. WWW '23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 845–854. [Online]. Available: https://dl.acm.org/doi/10.1145/3543507.3583251

[22] Z. Tao, X. Liu, Y. Xia, X. Wang, L. Yang, X. Huang, and T.-S. Chua, "Self-Supervised Learning for Multimedia Recommendation," *IEEE Transactions on Multimedia*, vol. 25, pp. 5107–5116, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/9811387

[23] K. Bachiri, A. Yahyaouy, M. Malek, and N. Rogovschi, "Mm-hgnn: Multimodal representation learning heterogeneous graph neural network," *International Journal of Computational Intelligence Systems*, vol. 18, no. 1, p. 178, 2025.

[24] S. Cao, W. Lu, and Q. Xu, "GraRep: Learning Graph Representations with Global Structural Information," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, ser. CIKM '15. New York, NY, USA: Association for Computing Machinery, Oct. 2015, pp. 891–900. [Online]. Available: https://doi.org/10.1145/2806416.2806512

[25] S. Cao, W. Lu, and Xu, "Deep neural networks for learning graph representations," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, Feb. 2016. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/10179

[26] Q. Huang, J. Yu, J. Wu, and B. Wang, "Heterogeneous Graph Attention Networks for Early Detection of Rumors on Twitter," in *2020 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2020, pp. 1–8, iSSN: 2161-4407. [Online]. Available: https://ieeexplore.ieee.org/document/9207582

[27] Z. Guo, J. Li, G. Li, C. Wang, S. Shi, and B. Ruan, "LGMRec: local and global graph learning for multimodal recommendation," in *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'24/IAAI'24/EAAI'24, vol. 38. AAAI Press, 2024, pp. 8454–8462. [Online]. Available: https://doi.org/10.1609/aaai.v38i8.28688

[28] K. Bachiri, A. Yahyaouy, M. Malek, and N. Rogovschi, "Topological data analysis and graph-based learning for multimodal recommendation," *IEEE Access*, vol. 13, pp. 108 934–108 954, 2025.

[29] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor Fusion Network for Multimodal Sentiment Analysis," Jul. 2017. [Online]. Available: https://arxiv.org/abs/1707.07250v1

[30] M. Hou, J. Tang, J. Zhang, W. Kong, and Q. Zhao, "Deep Multimodal Multilinear Fusion with High-order Polynomial Pooling," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://papers.nips.cc/paper_files/paper/2019/hash/f56d8183992b6c54c92c16a8519a6e2b-Abstract.html

[31] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, Feb. 2019. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8269806