Multimodal Graph Representation Learning with Dynamic Information Pathways

Xiaobin Hong[1], Mingkai Lin[1*], Xiaoli Wang[2], Chaoqun Wang[3], Wenzhong Li[1*]

[1]State Key Laboratory for Novel Software Technology, Nanjing University
[2]Nanjing Forest University, College of Information Science and Technology
[3]Zhejiang Provincial Seaport Investment & Operation Group Co. Ltd
xiaobinhong@smail.nju.edu.cn; mingkai@nju.edu.cn; lwz@nju.edu.cn

Abstract

Multimodal graphs, where nodes contain heterogeneous features such as images and text, are increasingly common in real-world applications. Effectively learning on such graphs requires both adaptive intra-modal message passing and efficient inter-modal aggregation. However, most existing approaches to multimodal graph learning are typically extended from conventional graph neural networks and rely on static structures or dense attention, which limit flexibility and expressive node embedding learning. In this paper, we propose a novel multimodal graph representation learning framework with Dynamic information Pathways (DiP). By introducing modality-specific pseudo nodes, DiP enables dynamic message routing within each modality via proximity-guided pseudo-node interactions and captures inter-modality dependence through efficient information pathways in a shared state space. This design achieves adaptive, expressive, and sparse message propagation across modalities with linear complexity. We conduct the link prediction and node classification tasks to evaluate performance and carry out full experimental analyses. Extensive experiments across multiple benchmarks demonstrate that DiP consistently outperforms baselines.

Introduction

Graphs are powerful abstractions for modeling relational and structural dependencies across a wide range of domains, including social networks (Zhang et al. 2022; Sharma et al. 2024), recommendation systems (Yang et al. 2024; Wang et al. 2024), and biological interactions (Valous et al. 2024; Ma et al. 2023). In many real-world applications, graph nodes are enriched with multimodal attributes, such as textual descriptions and visual content, leading to the emergence of Multimodal Graphs (MMGs) (Yan et al. 2024; Zhu et al. 2024). As illustrated in Figure 1, a typical MMG from a recommendation system represents items as nodes annotated with both images and textual metadata, while edges encode complex semantic or behavioral relationships among them. The inherent heterogeneity of MMGs, characterized by diverse feature modalities and intricate inter-node dependencies, makes them a compelling representation for downstream tasks. such as recommendation (Wei et al. 2019), knowledge

Figure 1: A multimodal ego-graph example from a recommendation system, where its node is attributed with multimodal raw data (i.e., image and text), and the link indicates the complex relations between nodes.

discovery (Zhang et al. 2020; Zhao et al. 2022), and scene understanding (Lee et al. 2023; Wang et al. 2025). To effectively leverage such heterogeneous information, models must not only propagate messages within each modality in a context-aware manner but also enable efficient and meaningful communication across modalities.

While multimodal learning (Zong 2024) has garnered significant attention in recent years, research on multimodal graph representation learning remains relatively nascent. Most existing approaches to multimodal graph learning typically extend conventional graph neural networks (GNNs, like GCN (Kipf 2016) and GAT (Veličković et al. 2017)) by stacking modality-specific encoders with static graph convolutions (Wei et al. 2019) or cross-modal attention mechanisms (Tao et al. 2020). While such architectures offer a straightforward extension of unimodal GNNs, they face significant limitations when applied to realistic, large-scale, and semantically complex multimodal graphs. First, there exists a fundamental **misalignment in information granularity across modalities**. Visual data often encodes fine-grained, instance-level cues such as spatial layout or object parts, whereas textual descriptions tend to abstract high-level semantic concepts. This granularity gap complicates alignment, as a direct fusion of such heterogeneous features often leads to semantic dilution or misinterpretation, especially when

*Mingkai Lin and Wenzhong Li are co-corresponding authors.

applied uniformly across all nodes. Second, most prior methods **rely on static structure-based aggregation**, either predefined by heuristics or constructed from fixed similarity measures. Such static topology fails to capture dynamic and context-aware dependencies between nodes, and cannot adapt to task-specific signals. Consequently, these rigid structures often lead to known issues such as *over-smoothing* (Oono 2019) and *over-squashing* (Di Giovanni et al. 2023a,b). Third, previous strategies exploit **modal-agnostic fusion**, such as feature concatenation or shared cross-modal attention after modal-independent learning, overlook the complementary nature of different modalities during both local and global aggregation. Without explicitly modeling modality-aware interactions and routing dynamics, these methods fail to fully exploit the expressiveness of MMGs.

To address these challenges, we propose a novel multimodal graph representation learning method with Dynamic information Pathways (termed DiP). DiP introduces modality-specific pseudo nodes as lightweight, dynamic intermediaries enabling flexible, efficient, and scalable multimodal graph learning. The core idea is to decouple the complexity of node-level interactions by introducing two information pathways: (i) **Intra-modal diffusion pathway**: Each modal is equipped with a set of learnable pseudo nodes that mediate the message diffusion through a shared proximity-based attention mechanism. This allows the model to construct flexible message passing pathways that adapt to task-specific context, overcoming the rigidity of static graph structures. (ii) **Inter-modal aggregation pathway**: Instead of directly modeling dense cross-modal node interactions, DiP restricts inter-modality communication to pseudo-to-pseudo interactions. Pseudo nodes from different modalities interact in a shared state space using dynamic proximity, enabling expressive and complementary information fusion at significantly reduced computational cost. This design enables adaptive routing, heterogeneity-aware fusion, and sparse computation in a unified framework. The overall complexity scales linearly with node number, ensuring high scalability. We instantiate DiP with a recurrent architecture and shared message update function, improving parameter efficiency and generalizability. As a result, DiP supports expressive multimodal reasoning while maintaining linear complexity in graph size, making it suitable for large-scale applications. Experiments on various multimodal graph tasks (i.e., link prediction and node classification) show that DiP consistently outperforms existing MMG methods, particularly when modality relationships are complex or dynamically shifting. Our main contributions can be summarized as follows:

- We propose DiP, a novel framework for multimodal graph representation learning, which enables adaptive, efficient, and scalable message propagation via learnable dynamic information pathways.
- We design a multimodal message passing system that constructs dynamic intra- and inter-modality pathways, leading to expressive and context-aware node embeddings.
- We conduct extensive experiments on multiple downstream tasks and provide a comprehensive analysis to demonstrate the effectiveness and efficiency of DiP.

Related Work

Multimodal Graph Learning

Multimodal Graphs (MMGs) indicate that the nodes are associated with multimodal attributes, such as texts and images, which have been widely featured in real-life scenarios. Prior research in multimodal graph learning has largely focused on domain-specific applications such as knowledge graphs (Chen et al. 2022; Zeng et al. 2023), molecular structures (Jin et al. 2018), and brain networks (Wang et al. 2023). These models are often tightly coupled with particular tasks and rely on handcrafted designs or domain expertise, which limits their ability to generalize across different graph structures, modalities, or learning objectives. More recently, MMGL (Yoon et al. 2023) attempts to bridge this gap by employing foundation models from multiple modalities on multimodal graphs. However, its focus remains constrained to generative settings, leaving the challenge of building transferable, task-agnostic representations for multimodal graphs largely unaddressed. UniGraph2 (He et al. 2025) performs large-scale pretraining on visually grounded heterogeneous graphs using modality-specific encoders and MoE fusion. While these methods provide valuable insights into multimodal fusion, they typically use static intra-modality structures or dense fusion strategies, lacking the adaptability and efficiency necessary for dynamic multimodal graphs.

Graph Message Passing

Message passing is the fundamental paradigm of GNNs, which aggregates the information from neighbors and updates node states (Kipf 2016; Hamilton 2017; Hong et al. 2024). Many graph learning methods enhance expressiveness by approximating filters with parameterized polynomials (Chien et al. 2020; Gasteiger 2019), but are limited to low-order terms due to computational cost, leading to local aggregation and issues like over-smoothing (Li 2018; Oono 2019; Lin et al. 2024) and over-squashing (Alon and Yahav 2020; Topping et al. 2021). To address these problems, recent works introduce auxiliary structures—such as edge shortcuts (Hong et al. 2021a; Gutteridge et al. 2023), graph pooling (Gao and Ji 2019; Ranjan 2020; Hong et al. 2021b), and pseudo nodes (Liu et al. 2022; Shirzad et al. 2023) to decouple message passing from the input topology. While edge shortcut methods dynamically rewire graphs for improved multi-hop communication, pseudo nodes offer global context but are often implemented with static, uniform connections that hinder adaptive message propagation (Shirzad et al. 2023). In this work, we revisit pseudo nodes with a dynamic and task-aware design that promotes efficient and adaptive global communication.

Methodology

As shown in Figure 2, our DiP uses the frozen modality encoders to embed the raw modality data, followed by L-step multimodal message passing that models adaptive intra- and inter-modal pathways through measurable node relations in a shared space. The resulting node embeddings are used for downstream tasks. We first define the dynamic pathway construction, where modality-specific pseudo nodes re-

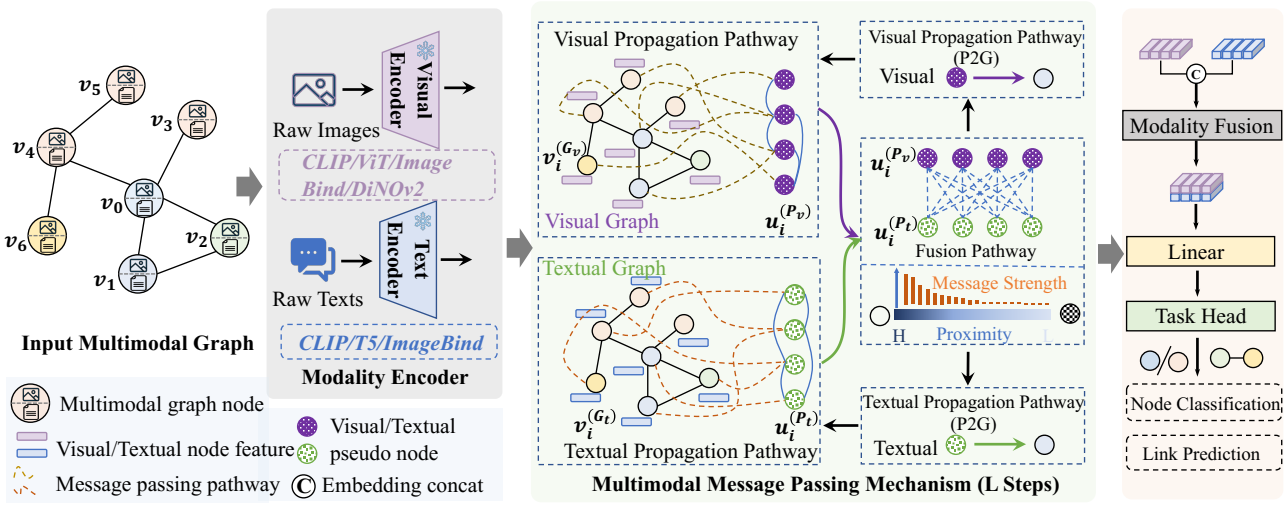


Figure 2: The overview framework of DiP. DiP first encodes the raw images and texts with frozen modality encoders for multimodal graph nodes. The recursive L-steps multimodal message passing mechanism is the key component of DiP, which consists of the *Intra-Modal Diffusion Pathway* and *Inter-Modal Aggregation Pathway* modules, which output expressiveness node representations incorporating the intra- and inter-modal message passing. Finally, the learned multimodal embeddings are fed to the task heads for link prediction and node classification training.

duce computational cost and enable cross-modal and cross-neighborhood communication. We then detail how DiP learns node displacements to facilitate adaptive message passing.

Preliminaries

Notations. We define a multimodal graph as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathrm{X})$, where $\mathcal{V} = \{v_1, \cdots, v_n\}$ is the node set with multimodal raw data. Following the benchmark (Zhu et al. 2025), we focus on image-text pair graphs, denoting visual and text nodes as $\mathcal{V}^{(v)}$ and $\mathcal{V}^{(t)}$. $\mathcal{E} = \{e_{v_i,v_j} | v_j \in \mathcal{N}(v_i)\}$ represents edges, with $\mathcal{N}(v_i)$ the one-hop neighbors of v_i. $\mathrm{X} = [\mathrm{X}^{(v)}, \mathrm{X}^{(t)}] = \{\mathrm{x}_{v_1}, \cdots, \mathrm{x}_{v_n}\}$ denotes multimodal node attributes. For each modality, we introduce pseudo nodes $\mathcal{P}^{(v)}$ and $\mathcal{P}^{(t)}$ with tunable sizes n_{p_v} and n_{p_t} to enable adaptive multimodal message passing.

Modality Encoder. We employ modality encoders to project raw images and texts into latent spaces, yielding visual features $\mathrm{X}^{(v)} \in \mathbb{R}^{n \times d_v}$ and textual features $\mathrm{X}^{(t)} \in \mathbb{R}^{n \times d_t}$. For the visual encoder E_v, we adopt CLIP (Radford et al. 2021), ViT (Dosovitskiy et al. 2020), DINOv2 (Oquab et al. 2023), and ImageBind (Girdhar et al. 2023), covering supervised, self-supervised, and joint vision-language paradigms. ImageBind further supports embedding diverse modalities (e.g., audio, video) into a shared space for generalized multimodal learning. For the text encoder E_t, we use CLIP, T5 (Raffel et al. 2020), and ImageBind.

Dynamic Pathway Construction

Conventional GNNs rely on static graph structures, limiting message passing to local neighborhoods. To enable global interactions, some works introduce pseudo nodes as intermediaries; however, they usually learn pairwise edge weights between nodes and pseudo nodes (Gilmer et al. 2017; Liu et al. 2022), leading to parameter growth linear in graph size. To enhance scalability, we draw inspiration from DyN (Pei and Wang 2023) and N^2 (Sun et al. 2024), which model neuronal interactions via spatially parameterized functional connections instead of individualized edge weights. These methods improve parameter efficiency using a shared path integral function conditioned on spatial coordinates, allowing dynamic modulation of information flow. Following this analogy, we treat graph nodes as neurons and introduce a unified state space $\mathcal{S} \in \mathbb{R}^{d_s}$, where each node and pseudo node has a learnable state embedding encoding both modality-specific features and local topology. A shared metric function over $\mathcal{S}$ computes proximity between nodes and pseudo nodes, constructing dynamic pathways without per-edge parameters.

Given a pseudo node set $\mathcal{P} = \{u_1, \cdots, u_p\}$[1], we embed them in $\mathcal{S}$ as learnable parameters $\mathrm{H} = \{\mathrm{h}_{u_1}, \cdots, \mathrm{h}_{u_p}\}^\top \in \mathbb{R}^{n_p \times d_s}$. These pseudo nodes interact with modality graph nodes for adaptive message passing through dynamic pathways. Encoded node features are projected into the common space $\mathcal{S}$ as $\mathrm{Z} = \{\mathrm{z}_{v_1}, \cdots, \mathrm{z}_{v_n}\} = f(\mathrm{X}) \in \mathbb{R}^{n \times d_s}$, where $f : \mathbb{R}^d \mapsto \mathcal{S}$ is permutation equivariant. To capture complex relations, we approximate non-linear interactions with a multi-channel path integral, projecting each node into τ channels for proximity computation, analogous to multi-head attention (Vaswani et al. 2017):

$$\phi(\mathrm{v}_i, \mathrm{v}_j) = \sum_{t=1}^{\tau} \lambda_t \, \mathrm{z}_{tv_i}^\top \mathrm{z}_{tv_j}, \quad \mathrm{z}_{tv} = \sigma(\mathrm{z}_t), \tag{1}$$

where λ_t is a learnable weight, $\sigma(\cdot)$ is a non-linear function with linear mapping followed by $\mathtt{LeakyReLU}(\cdot)$, and τ is the channel scale (following N^2, we set $\tau = 8$).

[1]For simplicity, modal subscripts are omitted.

Multimodal Message Passing Mechanism

We propose a Multimodal Message Passing Mechanism that jointly models intra-modal diffusion and inter-modal aggregation for effective node representation learning. As shown in Figure 2, our framework consists of two core components: the **Intra-Modal Diffusion Pathway** and the **Inter-Modal Aggregation Pathway**. Within each modality, we introduce pseudo nodes as latent mediators to enable scalable and expressive message propagation. Specifically, the intra-modal pathway includes both graph-to-pseudo (G2P) and pseudo-to-graph (P2G) routes, allowing pseudo nodes to gather and redistribute information across the entire graph. For inter-modal interaction, modality-specific pseudo nodes act as bridges that exchange information between different modalities. The updated pseudo node representations are then propagated back to their respective graph nodes to refine the node embeddings in a modality-aware and adaptive manner. For clarity, we omit modality-specific subscripts in intra-modal formulations, as the visual and textual branches adopt symmetric architectures.

Intra-Modal Diffusion Pathway (G2P). This pathway aims to capture the visual/text in-modality patterns, where pseudo nodes serve as global proxies. Following the common practice in GNNs (Kipf 2016), we interpret the interaction between embedded nodes as message passing. To achieve this, both graph node v and pseudo node u learn their message $\mathrm{m}_v, \mathrm{m}_u \in \mathbb{R}^d$ to be passed in the state space $\mathcal{S}$.The messages of graph nodes are initialized with node features, and learnable parameters construct pseudo nodes. We first perform local message passing ($\texttt{LocalMP}(\mathrm{M}, \mathcal{E})$) for graph nodes by exchanging their message among the ego-neighborhood:

$$\mathrm{m}_v = \psi(\mathrm{z}_v, \frac{1}{|\mathcal{N}(v)|} \sum_{v_j \in \mathcal{N}(v)} \mathrm{z}_{v_j}), \tag{2}$$

where $\psi(\cdot)$ denotes message aggregation function. The topology-coupled message passing encodes the local structure into node states, and we then use the pseudo nodes to aggregate the global modality patterns. Given graph node states $\mathrm{Z} \in \mathbb{R}^{n \times d_s}$, the pseudo node states $\mathrm{H} \in \mathbb{R}^{n_p \times d_s}$, and graph node message $\mathrm{M}^G = (\mathrm{m}_{v_1}, \cdots, \mathrm{m}_{v_n})^\top \in \mathbb{R}^{n \times d}$, the global message-passing process can be formulated as:

$$\mathrm{D} = \mathrm{W}^{GP}\mathrm{M}^G, \quad \mathrm{W}^{GP}_{ij} = \phi(\mathrm{H}_{i,\cdot}, \mathrm{Z}_{j,\cdot}), \tag{3}$$

$$\hat{\mathrm{D}} = \mathrm{W}^{PP}\mathrm{D}, \Delta\mathrm{H} = \sigma(\hat{\mathrm{D}}), \quad \mathrm{W}^{PP}_{ij} = \phi(\mathrm{H}_{i,\cdot}, \mathrm{H}_{j,\cdot}), \tag{4}$$

$$\hat{\mathrm{M}}^G = \mathrm{W}^{PG}\sigma(\hat{\mathrm{D}}), \mathrm{W}^{PG}_{i,j} = \phi(\mathrm{Z}_{i,\cdot}, [\mathrm{H} + \Delta\mathrm{H}]_{j,\cdot}), \tag{5}$$

where $\mathrm{W}^{GP} \in \mathbb{R}^{n_p \times n}$ denotes the edge weight matrix form graph nodes to pseudo nodes calculated by Eq. 1, W^{PP} and W^{PG} follow the similar definition. The Eq. 3, 4, and 5 perform the message diffusion, refinement, and aggregation process. We compile them as global message passing function($\texttt{GlobMP}(\cdot)$). Note that the complexity of our global message passing in $\mathcal{O}(\tau n n_p)$, with $\tau n_p \ll n$, significantly lower than $\mathcal{O}(n^2)$ in the dense scenario.

Intra-modality G2P pathway performs both local and global message passing. At the local level, graph nodes aggregate their message and update the node states:

$$\hat{\mathrm{Z}}^{(l)} = \mathrm{Z}^{(l-1)} + \sigma(\mathrm{M}^{(l)}), \mathrm{M}^{(l)} = \texttt{LocalMP}(\mathrm{M}||\mathrm{Z}, \mathcal{E}). \tag{6}$$

At the global level, we update the node states and messages according to global modality patterns:

$$\hat{\mathrm{M}}^G, \Delta\mathrm{H} = \texttt{GlobMP}(\mathrm{H}, \mathrm{M}^G, \mathrm{Z}),\ \tilde{\mathrm{Z}}^{(l)} = \hat{\mathrm{Z}}^{(l)} + \sigma(\hat{\mathrm{M}}^G),$$
$$\tilde{\mathrm{M}}^{(l)} = \mathrm{M}^{(l-1)} + \hat{\mathrm{M}}^G,\ \hat{\mathrm{H}}^{(l)} = \mathrm{H}^{(l-1)} + \Delta\mathrm{H}. \tag{7}$$

Inter-Modal Aggregation Pathway aims to enable inter-modality message passing by constructing cross-modal interactions. Given the updated visual and text modality pseudo node states $\hat{\mathrm{H}}^{(l)}_v \in \mathbb{R}^{n_{p_v} \times d_s}, \hat{\mathrm{H}}^{(l)}_t \in \mathbb{R}^{n_{p_t} \times d_s}$ from the intra-modality pathway, we use the proximity measurement to construct the adaptive message pathway for capturing cross-modal patterns:

$$\tilde{\mathrm{H}}^{(l)}_v = \hat{\mathrm{H}}^{(l)}_v + \mathrm{W}^{tv}\hat{\mathrm{H}}^{(l)}_t, \quad \mathrm{W}^{tv} = \phi(\hat{\mathrm{H}}^{(l)}_t, \hat{\mathrm{H}}^{(l)}_v), \tag{8}$$

$$\tilde{\mathrm{H}}^{(l)}_t = \hat{\mathrm{H}}^{(l)}_t + \mathrm{W}^{vt}\hat{\mathrm{H}}^{(l)}_v, \quad \mathrm{W}^{vt} = \phi(\hat{\mathrm{H}}^{(l)}_v, \hat{\mathrm{H}}^{(l)}_t). \tag{9}$$

Intra-Modal Diffusion Pathway (P2G) propagates the updated pseudo node states from P2P pathway into the graph nodes through global message passing across modalities. This enables the in-modal node states to perceive cross-modal information and enhances the expressive power of node embedding. Given the pseudo node states $\tilde{\mathrm{H}}^{(l)} \in \mathbb{R}^{n_p \times d_s}$, the graph node states $\tilde{\mathrm{Z}}^{(l)} \in \mathbb{R}^{n \times d_s}$, and node messages $\tilde{\mathrm{M}}^{(l)} \in \mathbb{R}^{n \times d}$, we perform adaptive global message passing:

$$\mathrm{M}^G, \Delta\mathrm{H} = \texttt{GlobMP}(\tilde{\mathrm{H}}^{(l)}, \tilde{\mathrm{M}}^{(l)}, \tilde{\mathrm{Z}}^{(l)}),\ \mathrm{Z}^{(l)} = \tilde{\mathrm{Z}}^{(l)} + \sigma(\mathrm{M}^G),$$
$$\mathrm{M}^{(l)} = \tilde{\mathrm{M}}^{(l)} + \mathrm{M}^G, \quad \mathrm{H}^{(l)} = \tilde{\mathrm{H}}^{(l)} + \Delta\mathrm{H}. \tag{10}$$

DiP updates the states of the embedded nodes recursively with a single recurrent layer, and the associated parameters are shared across steps. After recursive L-steps multimodal message passing system, the graph nodes receive their final states $\mathrm{Z}^{(L)}_v, \mathrm{Z}^{(L)}_t \in \mathbb{R}^{n \times d_s}$ for visual and text modality. The learned node states are then sent to the downstream task heads for objective output.

Modality Fusion and Task Head

This module receives the final-layer representations from different modalities and integrates them for downstream tasks. Specifically, we perform modality fusion via straightforward concatenation followed by a linear projection:

$$\mathrm{Z}^{(L)} = g(\texttt{Concat}[\mathrm{Z}^{(L)}_v, \mathrm{Z}^{(L)}_t]), \tag{11}$$

where $g : \mathbb{R}^{2d_s} \mapsto \mathbb{R}^d$ is a linear projector. In this study, we evaluate DiP in multimodal graph link prediction and node classification tasks. We attach a lightweight task head atop the fused representation $\mathrm{Z}^{(L)}$ to support different learning objectives. For node classification, we apply a softmax classifier (implemented by a 2-layer MLP) to predict the label of each node based on its embedding. For link prediction, we estimate the likelihood of an edge between two nodes using the inner product of their embeddings, followed by a sigmoid activation. These tasks are jointly used to assess the representational quality and generalization ability of our model in multimodal graph scenarios.

Encoder	Model	Amazon-Sports			Amazon-Cloth			Goodreads-LP		
		MRR↑	**Hit**@1 ↑	**Hit**@10 ↑	**MRR**↑	**Hit**@1 ↑	**Hit**@10 ↑	**MRR**↑	**Hit**@1 ↑	**Hit**@10 ↑
E_v = CLIP E_t = CLIP	MLP	28.22 ±0.09	14.54 ±0.16	59.40 ±0.08	21.10 ±0.04	10.70 ±0.03	42.77 ±0.05	11.03 ±0.06	4.87 ±0.04	21.61 ±0.11
	GCN	31.38 ±0.08	16.58 ±0.13	66.14 ±0.08	22.28 ±0.05	11.83 ±0.04	43.52 ±0.10	25.34 ±0.06	13.81 ±0.12	50.36 ±0.14
	SAGE	33.83 ±0.08	17.57 ±0.14	71.90 ±0.07	24.58 ±0.18	12.16 ±0.11	51.12 ±0.09	44.10 ±1.37	32.32 ±1.38	69.07 ±1.19
	BUDDY	31.55 ±0.13	15.05 ±0.43	70.92 ±0.25	23.44 ±0.26	11.06 ±0.20	51.08 ±0.50	43.25 ±0.23	31.84 ±0.35	67.93 ±0.03
	MMGCN	31.96 ±0.10	16.35 ±0.11	68.46 ±0.08	22.20 ±0.05	10.76 ±0.1	46.62 ±0.12	31.84 ±0.09	18.63 ±0.31	59.85 ±0.19
	MGAT	27.56 ±0.30	13.55 ±0.29	60.21 ±0.21	21.38 ±0.23	10.39 ±0.22	44.60 ±0.36	44.75 ±1.23	**34.53 ±1.48**	62.81 ±0.64
	UniGraph2	31.61 ±0.14	15.72 ±0.28	65.52 ±0.36	23.58 ±0.63	12.54 ±0.26	49.65 ±0.28	29.68 ±0.52	21.75 ±0.22	58.71 ±0.36
	DiP (ours)	**34.26 ±0.31**	**18.45 ±0.27**	**72.54 ±0.71**	**25.19 ±0.06**	**14.26 ±0.17**	**52.09 ±0.35**	**45.13 ±0.24**	34.47 ±0.28	**70.49 ±0.62**
E_v = ViT E_t = T5	MLP	24.81 ±0.05	11.63 ±0.05	54.78 ±0.04	17.65 ±0.06	8.14 ±0.04	36.77 ±0.06	11.10 ±0.17	4.84 ±0.15	21.94 ±0.24
	GCN	30.83 ±0.07	16.31 ±0.08	64.76 ±0.15	21.60 ±0.05	11.37 ±0.03	42.29 ±0.14	26.50 ±0.10	14.86 ±0.08	51.54 ±0.14
	SAGE	32.01 ±0.10	15.94 ±0.17	69.84 ±0.21	23.11 ±0.05	11.10 ±0.04	48.89 ±0.09	44.79 ±0.18	33.11 ±0.21	69.43 ±0.18
	BUDDY	30.41 ±0.40	14.11 ±0.28	69.55 ±0.80	22.82 ±0.19	10.24 ±0.12	51.04 ±0.39	43.18 ±0.53	31.73 ±0.54	67.89 ±0.57
	MMGCN	30.33 ±0.03	15.01 ±0.05	66.41 ±0.11	19.45 ±0.34	9.22 ±0.20	40.49 ±0.61	31.11 ±0.25	19.30 ±0.45	56.24 ±0.19
	MGAT	30.15 ±0.34	15.28 ±0.34	64.84 ±0.41	20.59 ±0.41	9.79 ±0.30	43.44 ±0.76	35.26 ±1.21	35.23 ±1.62	62.90 ±1.89
	UniGraph2	32.17 ±0.09	16.80 ±0.14	67.81 ±0.29	22.68 ±0.72	12.03 ±0.20	48.17 ±0.25	28.52 ±0.08	22.01 ±0.29	57.13 ±0.17
	DiP (ours)	**33.74 ±0.15**	**18.21 ±0.09**	**70.53 ±0.26**	**25.16 ±0.18**	**13.05 ±0.17**	**51.86 ±0.63**	**45.18 ±0.24**	**35.29 ±0.34**	**70.34 ±0.23**
E_v = ImageBind E_t = ImageBind	MLP	30.45 ±0.14	15.91 ±0.10	64.10 ±0.07	22.18 ±0.02	11.42 ±0.04	44.86 ±0.06	7.73 ±0.06	3.37 ±0.07	13.26 ±0.03
	GCN	31.67 ±0.09	17.07 ±0.14	65.61 ±0.10	22.81 ±0.03	12.27 ±0.05	44.28 ±0.09	27.56 ±1.26	14.31 ±1.37	57.25 ±0.52
	SAGE	34.32 ±0.11	17.87 ±0.23	73.04 ±0.15	25.20 ±0.09	12.63 ±0.05	52.53 ±0.21	34.61 ±0.43	23.82 ±0.51	56.67 ±0.21
	BUDDY	33.02 ±0.44	17.61 ±0.43	69.17 ±0.43	24.35 ±0.24	12.05 ±0.46	51.44 ±0.87	41.56 ±0.61	29.89 ±0.91	67.41 ±0.05
	MMGCN	31.74 ±0.21	16.45 ±0.13	67.39 ±0.74	24.72 ±0.19	12.47 ±0.09	51.32 ±0.56	26.32 ±0.23	16.05 ±0.22	46.37 ±0.66
	MGAT	30.15 ±0.12	15.50 ±0.05	64.20 ±0.43	22.13 ±0.27	10.96 ±0.15	45.84 ±0.57	34.77 ±0.49	34.95 ±0.61	62.51 ±0.47
	UniGraph2	32.35 ±0.06	15.28 ±0.30	67.83 ±0.15	24.37 ±0.29	11.75 ±0.23	50.21 ±0.06	32.43 ±0.57	26.55 ±0.21	59.77 ±0.38
	DiP (ours)	**35.16 ±0.53**	**19.57 ±0.14**	**74.23 ±0.27**	**26.18 ±0.19**	**14.08 ±0.61**	**54.17 ±0.24**	**46.13 ±0.12**	**36.07 ±0.24**	**71.06 ±0.04**
E_v = DINOv2 E_t = T5	MLP	24.81 ±0.16	11.62 ±0.18	54.97 ±0.22	17.53 ±0.11	8.07 ±0.09	36.53 ±0.26	10.28 ±0.04	4.49 ±0.05	19.86 ±0.03
	GCN	30.42 ±0.02	16.02 ±0.03	64.02 ±0.06	21.19 ±0.08	11.09 ±0.06	41.46 ±0.16	28.21 ±1.12	15.11 ±1.06	57.94 ±0.95
	SAGE	32.20 ±0.12	16.19 ±0.20	69.98 ±0.32	22.98 ±0.01	11.12 ±0.04	48.28 ±0.11	**45.61 ±0.22**	34.01 ±0.27	70.01 ±0.11
	BUDDY	30.02 ±0.34	13.78 ±0.19	69.18 ±0.67	22.95 ±0.06	10.45 ±0.09	50.87 ±0.61	43.25 ±0.13	31.77 ±0.33	68.08 ±0.42
	MMGCN	30.04 ±0.27	14.98 ±0.07	64.56 ±0.56	21.77 ±0.23	10.47 ±0.12	45.81 ±0.52	27.64 ±0.95	16.21 ±0.65	51.46 ±1.71
	MGAT	28.91 ±0.09	14.47 ±0.18	62.11 ±0.22	21.42 ±0.13	10.38 ±0.13	44.11 ±0.50	74.89 ±1.46	64.70 ±1.98	92.92 ±0.41
	UniGraph2	31.08 ±0.16	16.04 ±0.32	66.73 ±0.30	22.86 ±0.15	12.14 ±0.06	49.78 ±0.42	29.74 ±0.62	22.68 ±0.18	59.45 ±0.28
	DiP (ours)	**33.47 ±0.25**	**17.93 ±0.12**	**70.42 ±0.06**	**24.35 ±0.19**	**13.42 ±0.26**	**51.44 ±0.08**	44.76 ±0.27	**34.75 ±0.65**	**71.42 ±0.28**

Table 1: Link prediction results on Amazon-Sports, Amazon-Cloth, and Goodreads-LP. **Bold** is the best result.

Experiments

Experiments Setup

We evaluate DiP on two fundamental graph tasks (i.e., link prediction and node classification) in comparison with a topology-free method (i.e., MLP), three conventional GNNs (i.e., GCN (Kipf 2016), SAGE (Hamilton 2017), and BUDDY (Chamberlain et al. 2022)), and three multimodal GNNs (i.e., MMGCN (Wei et al. 2019), MGAT (Tao et al. 2020), and UniGraph2 (He et al. 2025)). Experiments are conducted on five real-world multimodal graph datasets (Three for link prediction, i.e., Amozon-Sports, Amazon-Cloth, and Goodreads-LP. Two for node classification, i.e., Ele-Fashion, Goodreads-NC), which are proposed by `MM-GRAPH` (Zhu et al. 2025). For evaluation metrics, we report Mean Reciprocal Rank (MRR), Hits@10, and Hits@1, the three most common-used evaluation metrics for link prediction and accuracy for node classification (Li et al. 2023). We implement DiP using PyTorch 2.4.0 and CUDA 12.2. All experiments are conducted on 4*Tesla V100-SXM2-32GB GPUs.

Multimodal Graph Link Prediction

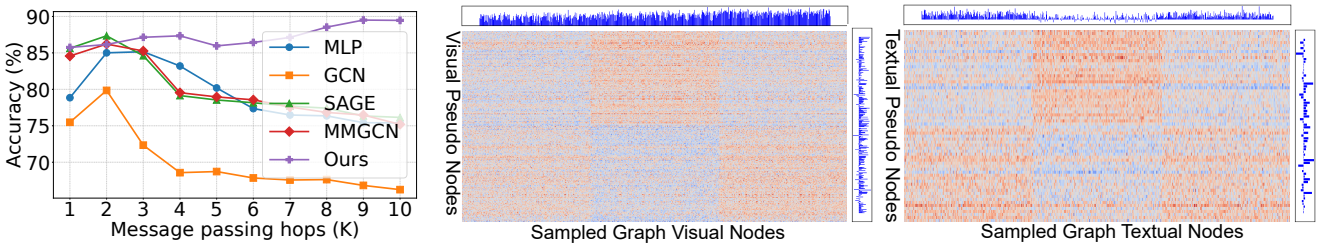
The detailed link prediction results are presented in Table 1, covering a diverse range of vision-language encoder combinations (E_v, E_t), including CLIP, ViT-T5, ImageBind-ImageBind, and DINOv2-T5. All metrics are reported as mean ± standard deviation over ten runs to ensure statistical reliability. As shown, our method consistently achieves state-of-the-art performance across all datasets and encoder configurations. In particular, it surpasses all baselines in MRR, Hit@1, and Hit@10 on Amazon-Sports, Amazon-Cloth, and Goodreads-LP, demonstrating both effectiveness and robustness. The most significant gains appear on Goodreads-LP, where our method exceeds the best baseline (e.g., BUDDY) by up to **+2.88** in MRR and **+5.79** in Hit@10, highlighting its strength in handling long-tail distributions and sparse relational structures. Comparable improvements are also observed under various encoder settings, suggesting that our design generalizes well to different visual and textual feature spaces. These results collectively confirm the effectiveness of our adaptive, modality-aware message passing mechanism in capturing intricate intra- and inter-modal dependencies, leading to more expressive and reliable multimodal graph representations.

Multimodal Graph Node Classification

Table 2 reports node classification accuracy on Ele-Fashion and Goodreads-NC across various vision-language encoder settings. Our method consistently achieves the best performance under all configurations, outperforming both unimodal (e.g., MLP, GCN, SAGE) and multimodal baselines (e.g., MMGCN, MMGAT, UniGraph2). On Ele-Fashion, our model achieves up to **89.50%** accuracy with ImageBind encoder, surpassing the strongest baseline by **+2.28%**. On the

Dataset	E_v	E_t	MLP	GCN	SAGE	MMGCN	MMGAT	UniGraph2	DiP (ours)
Ele-Fashion	CLIP	CLIP	85.16 ±0.03	79.83 ±0.03	87.10 ±0.02	86.10 ±0.50	84.66 ±0.29	85.93 ±0.15	**87.93 ±0.04**
	ViT	T5	84.98 ±0.05	79.63 ±0.07	84.41 ±0.09	82.39 ±0.30	84.01 ±0.08	85.47 ±0.20	**87.19 ±0.06**
	ImageBind	ImageBind	88.73 ±0.01	80.35 ±0.02	87.71 ±0.13	86.21 ±0.84	86.12 ±0.08	86.52 ±0.31	**89.50 ±0.07**
	DINOv2	T5	84.87 ±0.01	79.37 ±0.04	85.31 ±0.09	85.53 ±0.33	84.54 ±0.27	86.18 ±0.09	**87.46 ±0.05**
Goodreads-NC	CLIP	CLIP	72.29 ±0.02	81.61 ±0.01	83.30 ±0.02	83.29 ±0.20	76.48 ±0.59	82.56 ±0.35	**85.37 ±0.12**
	ViT	T5	67.82 ±0.02	81.67 ±0.01	83.30 ±0.02	81.85 ±0.22	75.43 ±0.76	81.93 ±0.41	**86.64 ±0.07**
	ImageBind	ImageBind	58.75 ±0.05	78.91 ±0.04	80.39 ±0.21	80.58 ±1.08	69.45 ±6.25	80.52 ±0.08	**83.21 ±0.13**
	DINOv2	T5	68.83 ±0.03	81.71 ±0.03	82.99 ±0.08	82.44 ±0.11	74.98 ±1.23	81.63 ±0.21	**84.32 ±0.15**

Table 2: Node classification results on Goodreads-NC and Ele-Fashion. **Bold** indicates the best performance.



(a) Over-smoothing effects

(b) The dynamic information pathways, path($\mathcal{V}^{(v)}, \mathcal{P}^{(v)}$) and ($\mathcal{V}^{(t)}, \mathcal{P}^{(t)}$)

Figure 3: (a) DiP with adaptive pathways can maintain the performance as the model depth increases. (b) Message passing pathways. The proximities between sampled graph nodes and pseudo nodes among two modalities. Some pseudo-nodes show link activation (bright rows), which may be a clustering pattern of nodes from different classes.

Ablation	Ele-Fashion		Amazon-Sports		
	Acc↑	**F1**↑	**MRR**↑	**Hit**@1 ↑	**Hit**@10 ↑
w/o. $\mathcal{P}^{(v)}$	84.62	78.48	31.24	17.26	69.85
w/o. $\mathcal{P}^{(t)}$	85.34	77.68	31.59	17.42	70.16
w/o. Local.	84.62	77.81	30.26	16.85	61.34
w/o. Global.	83.16	76.45	29.48	16.37	67.48
w/o. $\mathcal{P}^{(v)} \leftrightarrow \mathcal{P}^{(t)}$	82.47	74.16	30.17	16.03	65.27
Ours	**87.88**	**82.06**	**34.26**	**18.45**	**72.54**

Table 3: The ablation studies of modules in DiP.

Model	Ele-Fashion		Amazon-Sports	
	Time(s)	Mem(MB)	Time(s)	Mem(MB)
GCN	**0.253**	1605.20	**208.527**	1962.37
SAGE	0.326	1312.31	224.361	2043.06
MMGCN	1.237	2030.45	346.217	2570.52
MGAT	1.524	2340.30	382.425	2803.25
Ours	0.531	**462.34**	213.134	**660.24**

Table 4: The model complexity analysis. Time (s) and memory (MB) consumed per epoch, the lower is better.

more challenging Goodreads-NC, which features noisier and sparser labels, our method still leads across all encoder settings, achieving up to **85.37%** in CLIP encoder and maintaining an average gain of **+2.35%** over the best-performing baselines. These results demonstrate the generalizability and robustness of our approach to node classification tasks.

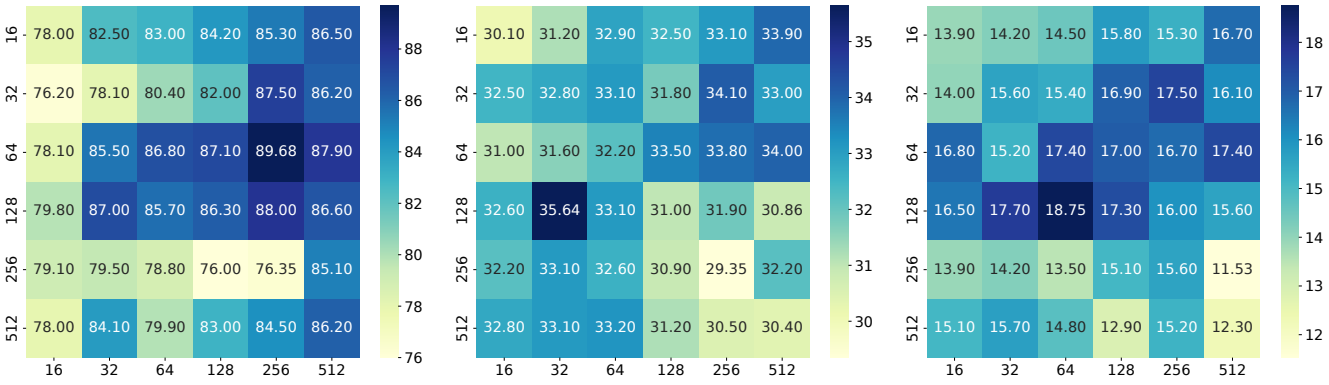
Ablation Study

We conduct ablation experiments to assess the contribution of each module in DiP. In Table 3, "w/o $\mathcal{P}^{(v)}$" and "w/o $\mathcal{P}^{(t)}$" remove visual and textual pseudo nodes; "w/o Local" and "w/o Global" disable local and global message passing; and "$\mathcal{P}^{(v)} \leftrightarrow \mathcal{P}^{(t)}$" removes cross-modal pseudo node interaction. From the ablation results, we have the following observations: (1) Message passing improves representation learning via structural context, confirming the benefit of relational modeling. (2) Pseudo nodes enhance intra-modal interaction. (3) Cross-modal communication complements embedding expressiveness. These findings confirm that each component of DiP is crucial for achieving strong and robust performance.

Model Analysis

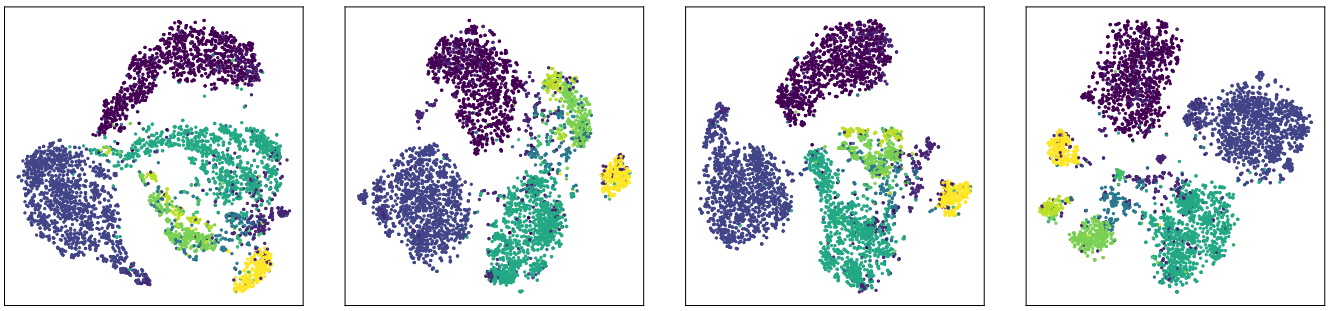
Tackling Over-Smoothing. DiP alleviates over-smoothing by decoupling message passing from the input topology via adaptive pseudo nodes. Dynamic P2P connections enable diverse cross-modal aggregation, while the G2P subsystem preserves modality-specific signals across layers. This design supports deeper propagation without representation collapse. As shown in Figure 3(a), DiP maintains higher Dirichlet energy than static baselines, verifying its ability to retain discriminative features and mitigate over-smoothing.

Message Passing Pathways. To understand the adaptive routing behavior of DiP, we visualize the similarity between sampled graph nodes and modality-specific pseudo nodes in Figure 3(b). Notably, certain pseudo nodes exhibit strong activation patterns (bright rows), suggesting that they serve as hubs aggregating information from semantically related nodes across modalities. This clustering behavior indicates that DiP dynamically organizes message pathways based on latent structure rather than static topology.



(a) (n_{p_v}, n_{p_t}) on Ele-Fashion (Accuracy) (b) (n_{p_v}, n_{p_t}) on Amazon-Sports (MRR) (c) (n_{p_v}, n_{p_t}) on Amazon-Sports (Hit@1)

Figure 4: Ablation on the number of visual (n_{p_v}) and textual (n_{p_t}) pseudo nodes.



(a) SAGE (b) MMGCN (c) MGAT (d) DiP

Figure 5: The T-SNE plots for node embedding from Ele-Fashion dataset.

Complexity Analysis. DiP performs adaptive message passing on multimodal graphs without dense pairwise modeling, with complexity $\mathcal{O}(\tau n n_p)$, where $n_p = \max(n_{p_v}, n_{p_t}) \ll n$, notably lower than uniform dense approaches. Empirical results in Table 4 show that our time complexity is comparable to efficient GNNs (e.g., GCN and SAGE), while memory overhead is significantly lower. These findings demonstrate that DiP achieves an effective balance between scalability and expressive power.

Pseudo Node Numbers. We perform a parameter search over the number of visual (n_{p_v}) and textual (n_{p_t}) pseudo nodes. As shown in Figure 4, only a subset of pseudo nodes exhibit strong activation, while others remain underutilized. This indicates that using too many pseudo nodes may introduce redundancy, whereas too few may limit the model's expressiveness. We therefore select the optimal number of pseudo nodes based on validation performance to balance efficiency and representation capacity.

Visualization

To qualitatively assess the expressiveness of node embeddings, we visualize the learned representations of different models using t-SNE. As shown in Figure 5, multimodal GNNs produce more structured and discriminative embeddings compared to their unimodal counterparts, particularly for nodes associated with multiple modalities. Notably, our proposed method yields superior category separability, with clearer decision boundaries and reduced embedding overlap. This suggests that DiP better captures modality-specific semantics while promoting cross-modal alignment, which is especially beneficial for distinguishing classes with ambiguous or fuzzy boundaries. These visualization results align with our quantitative gains and provide further evidence for the effectiveness of our multimodal message passing design.

Conclusion

In this work, we propose DiP, a novel pseudo node-enabled framework for multimodal graph representation learning with dynamic information pathways. By decoupling message propagation from fixed graph topology and leveraging modality-aware pseudo nodes, DiP effectively captures both intra- and inter-modal dependencies while maintaining scalability. Extensive experiments on link prediction and node classification tasks demonstrate its superiority over strong baselines in both performance and efficiency. Our analysis further shows that DiP mitigates over-smoothing and adapts message pathways dynamically, offering a promising direction for future research on structured multimodal learning.

Acknowledgments

This work was supported by Basic Research Program of Jiangsu (Grant No. BK20251198), the Natural Science Foundation of Jiangsu Province (Grant No. BK20222003), the National Natural Science Foundation of China (Grant Nos. 62502201, 62572236), Postdoctoral Fellowship Program of CPSF (Grant No. GZC20251067), Jiangsu Funding Program for Excellent Postdoctoral Talent, the Collaborative Innovation Center of Novel Software Technology and Industrialization, and the Sino-German Institutes of Social Computing.

References

Alon, U.; and Yahav, E. 2020. On the bottleneck of graph neural networks and its practical implications. *arXiv preprint arXiv:2006.05205*.

Chamberlain, B. P.; Shirobokov, S.; Rossi, E.; Frasca, F.; Markovich, T.; Hammerla, N.; Bronstein, M. M.; and Hansmire, M. 2022. Graph neural networks for link prediction with subgraph sketching. *arXiv preprint arXiv:2209.15486*.

Chen, X.; Zhang, N.; Li, L.; Deng, S.; Tan, C.; Xu, C.; Huang, F.; Si, L.; and Chen, H. 2022. Hybrid transformer with multi-level fusion for multimodal knowledge graph completion. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, 904–915.

Chien, E.; Peng, J.; Li, P.; and Milenkovic, O. 2020. Adaptive universal generalized pagerank graph neural network. *arXiv preprint arXiv:2006.07988*.

Di Giovanni, F.; Giusti, L.; Barbero, F.; Luise, G.; Lio, P.; and Bronstein, M. M. 2023a. On over-squashing in message passing neural networks: The impact of width, depth, and topology. In *International conference on machine learning*, 7865–7885. PMLR.

Di Giovanni, F.; Rusch, T. K.; Bronstein, M. M.; Deac, A.; Lackenby, M.; Mishra, S.; and Veličković, P. 2023b. How does over-squashing affect the power of GNNs? *arXiv preprint arXiv:2306.03589*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Gao, H.; and Ji, S. 2019. Graph u-nets. In *international conference on machine learning*, 2083–2092. PMLR.

Gasteiger, W. S. G. S., Johannes. 2019. Diffusion improves graph learning. *Advances in neural information processing systems*, 32.

Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; and Dahl, G. E. 2017. Neural message passing for quantum chemistry. In *International conference on machine learning*, 1263–1272. PMLR.

Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15180–15190.

Gutteridge, B.; Dong, X.; Bronstein, M. M.; and Di Giovanni, F. 2023. Drew: Dynamically rewired message passing with delay. In *International Conference on Machine Learning*, 12252–12267. PMLR.

Hamilton, Y. Z. L. J., Will. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

He, Y.; Sui, Y.; He, X.; Liu, Y.; Sun, Y.; and Hooi, B. 2025. Unigraph2: Learning a unified embedding space to bind multimodal graphs. In *Proceedings of the ACM on Web Conference 2025*, 1759–1770.

Hong, X.; Li, W.; Wang, C.; Lin, M.; and Lu, S. 2024. Label attentive distillation for GNN-based graph classification. In *Proceedings of the AAAI conference on artificial intelligence*, 8499–8507.

Hong, X.; Zhang, T.; Cui, Z.; Huang, Y.; Shen, P.; Li, S.; and Yang, J. 2021a. Graph game embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 7711–7720.

Hong, X.; Zhang, T.; Cui, Z.; and Yang, J. 2021b. Variational gridded graph convolution network for node classification. *IEEE/CAA Journal of Automatica Sinica*, 8(10): 1697–1708.

Jin, W.; Yang, K.; Barzilay, R.; and Jaakkola, T. 2018. Learning multimodal graph-to-graph translation for molecular optimization. *arXiv preprint arXiv:1812.01070*.

Kipf, W. M., Thomas N. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Lee, J.; Chung, C.; Lee, H.; Jo, S.; and Whang, J. 2023. VISTA: Visual-textual knowledge graph representation learning. In *Findings of the association for computational linguistics: EMNLP 2023*, 7314–7328.

Li, H. Z. W. X.-M., Qimai. 2018. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Li, J.; Shomer, H.; Mao, H.; Zeng, S.; Ma, Y.; Shah, N.; Tang, J.; and Yin, D. 2023. Evaluating graph neural networks for link prediction: Current pitfalls and new benchmarking. *Advances in Neural Information Processing Systems*, 36: 3853–3866.

Lin, M.; Li, W.; Hong, X.; and Lu, S. 2024. Scalable multi-source pre-training for graph neural networks. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 1292–1301.

Liu, X.; Cheng, J.; Song, Y.; and Jiang, X. 2022. Boosting graph structure learning with dummy nodes. In *International conference on machine learning*, 13704–13716. PMLR.

Ma, A.; Wang, X.; Li, J.; Wang, C.; Xiao, T.; Liu, Y.; Cheng, H.; Wang, J.; Li, Y.; Chang, Y.; et al. 2023. Single-cell biological network inference using a heterogeneous graph transformer. *Nature Communications*, 14(1): 964.

Oono, S. T., Kenta. 2019. Graph neural networks exponentially lose expressive power for node classification. *arXiv preprint arXiv:1905.10947*.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Pei, Z.; and Wang, S. 2023. Dynamics-inspired neuromorphic visual representation learning. In *International Conference on Machine Learning*, 27521–27541. PMLR.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

Ranjan, S. S. T. P., Ekagra. 2020. Asap: Adaptive structure aware pooling for learning hierarchical graph representations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 5470–5477.

Sharma, K.; Lee, Y.-C.; Nambi, S.; Salian, A.; Shah, S.; Kim, S.-W.; and Kumar, S. 2024. A survey of graph neural networks for social recommender systems. *ACM Computing Surveys*, 56(10): 1–34.

Shirzad, H.; Velingker, A.; Venkatachalam, B.; Sutherland, D. J.; and Sinop, A. K. 2023. Exphormer: Sparse transformers for graphs. In *International Conference on Machine Learning*, 31613–31632. PMLR.

Sun, J.; Yang, C.; Ji, X.; Huang, Q.; and Wang, S. 2024. Towards Dynamic Message Passing on Graphs. *arXiv preprint arXiv:2410.23686*.

Tao, Z.; Wei, Y.; Wang, X.; He, X.; Huang, X.; and Chua, T.-S. 2020. Mgat: Multimodal graph attention network for recommendation. *Information Processing & Management*, 57(5): 102277.

Topping, J.; Di Giovanni, F.; Chamberlain, B. P.; Dong, X.; and Bronstein, M. M. 2021. Understanding over-squashing and bottlenecks on graphs via curvature. *arXiv preprint arXiv:2111.14522*.

Valous, N. A.; Popp, F.; Zörnig, I.; Jäger, D.; and Charoentong, P. 2024. Graph machine learning for integrated multi-omics analysis. *British journal of cancer*, 131(2): 205–211.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Wang, B.; Chen, J.; Li, C.; Zhou, S.; Shi, Q.; Gao, Y.; Feng, Y.; Chen, C.; and Wang, C. 2024. Distributionally robust graph-based recommendation system. In *Proceedings of the ACM Web Conference 2024*, 3777–3788.

Wang, C.; Hong, X.; Li, W.; and Zhang, R. 2025. Semantic-Supervised Spatial-Temporal Fusion for LiDAR-based 3D Object Detection. *arXiv preprint arXiv:2503.10579*.

Wang, M.; Shao, W.; Huang, S.; and Zhang, D. 2023. Hypergraph-regularized multimodal learning by graph diffusion for imaging genetics based alzheimer's disease diagnosis. *Medical image analysis*, 89: 102883.

Wei, Y.; Wang, X.; Nie, L.; He, X.; Hong, R.; and Chua, T.-S. 2019. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM international conference on multimedia*, 1437–1445.

Yan, H.; Li, C.; Yu, Z.; Yin, J.; Liu, R.; Zhang, P.; Han, W.; Li, M.; Zeng, Z.; Sun, H.; et al. 2024. When Graph meets Multimodal: Benchmarking on Multimodal Attributed Graphs Learning. *arXiv preprint arXiv:2410.09132*.

Yang, Y.; Wu, L.; Wang, Z.; He, Z.; Hong, R.; and Wang, M. 2024. Graph bottlenecked social recommendation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3853–3862.

Yoon, M.; Koh, J. Y.; Hooi, B.; and Salakhutdinov, R. 2023. Multimodal graph learning for generative tasks. *arXiv preprint arXiv:2310.07478*.

Zeng, Y.; Jin, Q.; Bao, T.; and Li, W. 2023. Multi-modal knowledge hypergraph for diverse image retrieval. In *Proceedings of the AAAI conference on artificial intelligence*, 3376–3383.

Zhang, X.; Zhang, T.; Hong, X.; Cui, Z.; and Yang, J. 2020. Graph wasserstein correlation analysis for movie retrieval. In *European Conference on Computer Vision*, 424–439. Springer.

Zhang, Y.; Gao, S.; Pei, J.; and Huang, H. 2022. Improving social network embedding via new second-order continuous graph neural networks. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 2515–2523.

Zhao, Y.; Cai, X.; Wu, Y.; Zhang, H.; Zhang, Y.; Zhao, G.; and Jiang, N. 2022. Mose: Modality split and ensemble for multimodal knowledge graph completion. *arXiv preprint arXiv:2210.08821*.

Zhu, J.; Zhou, Y.; Qian, S.; He, Z.; Zhao, T.; Shah, N.; and Koutra, D. 2024. Multimodal graph benchmark. *arXiv preprint arXiv:2406.16321*.

Zhu, J.; Zhou, Y.; Qian, S.; He, Z.; Zhao, T.; Shah, N.; and Koutra, D. 2025. Mosaic of modalities: A comprehensive benchmark for multimodal graph learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 14215–14224.

Zong, M. A. O. H.-T., Yongshuo. 2024. Self-supervised multimodal learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.