SCIENCE CHINA
Information Sciences

• POSITION PAPER •

November 2025, Vol. 68, Iss. 11, 213101:1–213101:20
https://doi.org/10.1007/s11432-025-4627-3

Towards multimodal graph large language model

Xin WANG[1,2], Zeyang ZHANG[1], Linxin XIAO[1], Haibo CHEN[1], Chendi GE[1] & Wenwu ZHU[1,2*]

[1]*Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China*
[2]*Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China*

Received 11 June 2025/Revised 19 September 2025/Accepted 13 October 2025/Published online 27 October 2025

Abstract Multimodal graphs, which integrate diverse multimodal features and relations, are ubiquitous in real-world applications. However, existing multimodal graph learning methods are typically trained from scratch for specific graph data and tasks, failing to generalize across various multimodal graph data and tasks. To bridge this gap, we explore the potential of multimodal graph large language models (MG-LLM) to unify and generalize across diverse multimodal graph data and tasks. We propose a unified framework of multimodal graph data, tasks, and models, discovering the inherent multi-granularity and multi-scale characteristics in multimodal graphs. Specifically, we present five key desired characteristics for MG-LLM: (1) unified space for multimodal structures and attributes, (2) capability of handling diverse multimodal graph tasks, (3) multimodal graph in-context learning, (4) multimodal graph interaction with natural language, and (5) multimodal graph reasoning. We then elaborate on the key challenges, review existing literature, and highlight promising future research directions towards realizing these ambitious characteristics. Finally, we summarize existing multimodal graph datasets pertinent for model training. We believe this paper can contribute to the ongoing advancement of the research towards MG-LLM for generalization across multimodal graph data and tasks.

Keywords multimodal graph, large language model, foundation model, graph machine learning, multimodality

Citation Wang X, Zhang Z Y, Xiao L X, et al. Towards multimodal graph large language model. Sci China Inf Sci, 2025, 68(11): 213101, https://doi.org/10.1007/s11432-025-4627-3

1 Introduction

Multimodal graphs, which integrate features from diverse modalities such as text, image, audio, and video, as well as capture the complex intra-model and inter-modal relations, are becoming increasingly ubiquitous in real-world applications. From social networks [1] and e-commerce [2] platforms to scientific discovery in biomedicine and materials science [3, 4], these complex data structures offer a richer, more holistic representation of interconnected entities than traditional unimodal graphs, which unlock more opportunities for advanced analytics, reasoning, and generation capabilities over multimodal information.

However, multimodal graph learning currently faces a significant issue: existing methods are predominantly designed for specific tasks on particular types of graphs. This specialization often limits their applicability, preventing them from generalizing effectively across the vast diversity of multimodal graph data and tasks encountered in practice. This lack of universality necessitates constant redesign and retraining for new scenarios, hindering the development of truly versatile and scalable solutions.

To bridge this gap, we explore the potential of multimodal graph large language models (MG-LLM). Inspired by the remarkable success of large language models (LLMs) in unifying diverse natural language tasks, we propose that MG-LLM can serve as a powerful paradigm to unify and generalize across the complex landscape of multimodal graph data and tasks. Our exploration begins by establishing a unified framework for multimodal graph data, tasks, and models, which uncovers the inherent characteristics of multimodal graphs, i.e., multi-granularity and multi-scale.

Specifically, we highlight that multimodal graphs inherently exhibit multi-granularity, organizing information from fine-grained features like pixels and words to coarse-grained concepts such as entire images or documents, along with diverse structural complexities. This leads to multi-scale characteristics in

* Corresponding author (email: wwzhu@tsinghua.edu.cn)

multimodal graph tasks, where inputs and outputs can vary dramatically in their scope, from individual nodes to entire graph structures.

Building upon this foundational understanding, we articulate five key desired characteristics towards MG-LLM.

• Unified space for multimodal structures and attributes. The ability to align and represent diverse multimodal features and relations within a single unified embedding space, capable of handling highly irregular and continuous information.

• Ability of handling diverse multimodal graph tasks. The capacity to frame and solve all multimodal graph tasks, from traditional discriminative problems like node classification to emerging generative tasks such as multimodal content generation, under a unified generative modeling paradigm.

• Multimodal graph in-context learning. The capability of performing novel tasks by leveraging a limited number of multimodal graph examples provided directly within the prompt, without requiring explicit model fine-tuning.

• Multimodal graph interaction with natural language. The possibility of enabling users to query, edit, generate, and reason about complex multimodal graph-structured knowledge using intuitive natural language, bridging the gap between human language and structured data.

• Multimodal graph reasoning. The proficiency in performing complex multi-hop, cross-modal reasoning, including analogical inference, by seamlessly combining information from various modalities and relational structures.

While the vision of MG-LLMis ambitious, realizing these characteristics presents significant challenges, ranging from developing unified multimodal graph vocabularies and tokenization schemes to multimodal graph architectures capable of large-scale pretraining. This paper delves into these key challenges, reviews existing research that moves towards this paradigm, and outlines promising future research directions to accelerate the development of MG-LLM for generalizing across diverse multimodal graph data and tasks. Finally, we summarize existing multimodal graph datasets that could be useful for the training and evaluation of such models. This work aims to foster progress towards a new era of multimodal graph intelligence.

Our main contributions are summarized as follows.

• We explore the potential of MG-LLM to unify and generalize across diverse multimodal graph data and tasks, aiming for universal generalization across diverse multimodal graph data and tasks. We systematically discuss the potential of MG-LLM, for the first time, to the best of our knowledge.

• We present a unified framework for understanding multimodal graph data, tasks, and models, highlighting their inherent multi-granularity and multi-scale characteristics for designing MG-LLM.

• We propose five essential characteristics that MG-LLM should possess, coupled with detailed discussions of challenges and future directions, thereby setting a clear research roadmap. We also summarize existing multimodal graph datasets and tasks for developing MG-LLM.

The domain of multimodal learning on graphs is rapidly advancing. We distinguish our work from several related lines of research. (1) Compared to surveys on multimodal graph learning [5, 6], which primarily catalog existing techniques, our work introduces a novel conceptual framework and a forward-looking vision for a unified MG-LLM. (2) Unlike existing GraphLLMs [7, 8], which mainly focus on adapting unimodal graph data for LLMs, we address the more complex challenge of natively handling graphs with rich multimodal attributes. (3) Distinct from general-purpose Omni-MLLMs [9], we argue for a specialized paradigm that deeply integrates graph-structured reasoning rather than treating graphs as just another input modality. By identifying the inherent characteristics of multimodal graphs and the desired characteristics for MG-LLM, we chart a new research roadmap towards developing multimodal graph large language models, thereby complementing surveys of the existing state-of-the-art by looking towards a next-generation paradigm.

2 Towards a unified view of multimodal graph data, task, and model

In this section, we introduce a unified framework of multimodal graph data, task, and model, and remark on the inherent characteristics in multimodal graphs, serving as a foundation to discuss the desired characteristics of MG-LLM. The overall framework is shown in Figure 1.

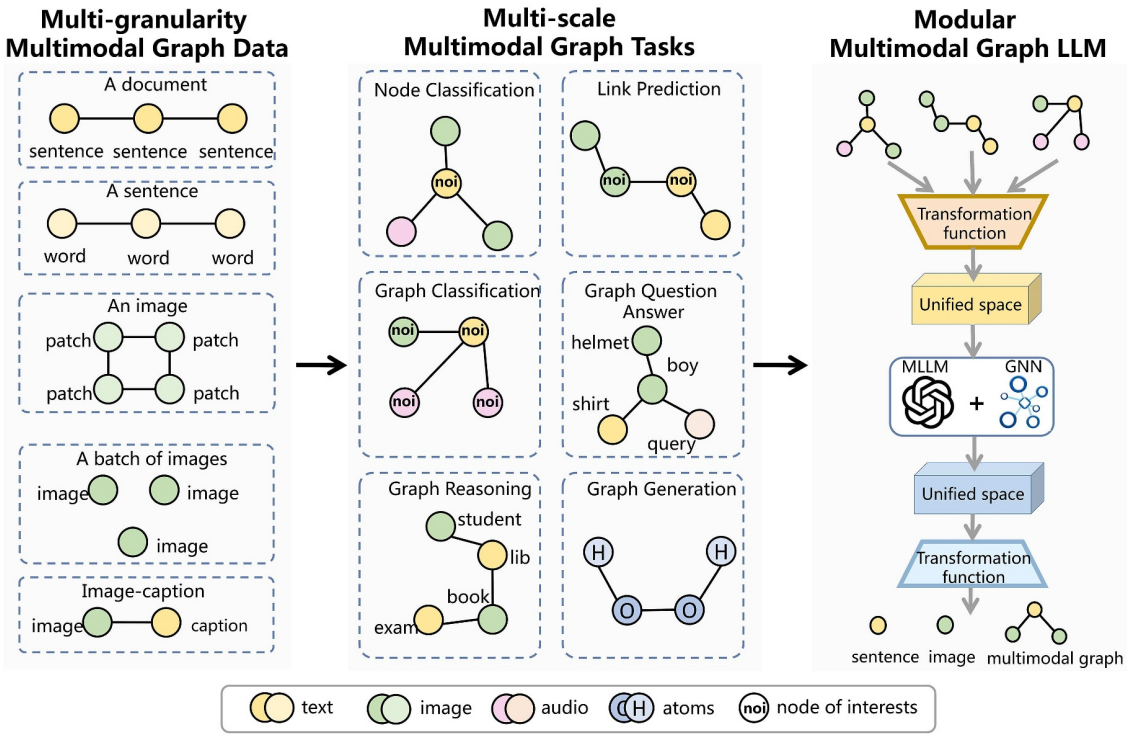

Figure 1 (Color online) Unified view of multimodal graph data, tasks, and models towards MG-LLM.

2.1 Unified formulation of multimodal graph data

In this section, we define multimodal graphs, extending standard graphs with diverse node and edge modalities. Then we outline three decomposable types (feature-, node-, graph-level), and their versatility in representing various data forms, from single instances to full datasets. Moreover, we give remarks on the challenges of indecomposability and multi-granularity for building effective MG-LLM.

Graph. A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consists of a finite set of vertices $\mathcal{V} = \{v_1, v_2, \ldots, v_n\}$ and a set of edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$, each edge being an ordered pair of vertices denoting a directed relation between them.

Multimodal graph. A modality is a distinct type or source of information associated with nodes or edges. Let $\mathcal{M} = \{1, 2, \ldots, M\}$ denote the set of all modalities. We define a mapping set $\mathcal{F} = \{\mathcal{F}_m\}_{m=1}^{M}$ for all M modalities. For each $m \in \mathcal{M}$, the mapping $\mathcal{F}_m$ maps from the modality-specific node feature space $\mathcal{V}$ to a shared representation space $\mathcal{X}$, such that $\mathcal{F}_m : \mathcal{V} \to \mathcal{X}$. The space $\mathcal{X}$ serves as a unified embedding space for all modalities. Similarly, we define a mapping $\mathcal{F}_m$ from the modality-specific edge feature space $\mathcal{E}$ to the shared representation space $\mathcal{X}$, such that $\mathcal{F}_m : \mathcal{E} \to \mathcal{X}$. For simplicity, we reuse $\mathcal{F}_m$ as the modality-m map for both nodes and edges. We likewise leave out multimodal edges in the formulation, even though extending them would be straightforward. The features in different modalities could be texts, images, audios, and videos. A multimodal graph can be defined by the quadruple $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{F}, \mathcal{M})$. To make the formulation more concrete, a running example is shown in Figure 2.

Special cases of decomposable multimodal graphs. By instantiation of the modality set and feature mapping, we could obtain several classic types of multimodal graphs which are ubiquitous in real-world applications [6]. These types of multimodal graphs share the same assumption that they can be decomposed by modality from different perspectives, i.e., feature, node, and graph.

- Feature-level multimodal graph, where the features of nodes or edges come from different modalities, i.e., $\mathcal{G} = \bigcup_{m=1}^{M} \mathcal{G}_m = \bigcup_{m=1}^{M} (\mathcal{V}, \mathcal{E}, \mathcal{F}_m, \{m\})$, where m representing one modality in M modalities. The feature of node v could be represented as $\mathbf{x}(v) = \bigoplus_{m=1}^{M} \mathcal{F}_m(v)$. For example, on an e-commerce product graph [10], each node has the feature of product title and image, i.e., $\mathbf{x}(v) = \bigoplus(\mathcal{F}_{\text{text}}(v), \mathcal{F}_{\text{image}}(v))$.
- Node-level multimodal graph, where the nodes or edges come from different modalities, while each node or edge has unimodal features, i.e., $\mathcal{G} = \bigcup_{m=1}^{M} \mathcal{G}_m = \bigcup_{m=1}^{M} (\mathcal{V}_m, \mathcal{E}, \mathcal{F}_m, \{m\})$, where m representing one modality in M modalities, and $\mathcal{V}_i \bigcap \mathcal{V}_j = \emptyset$. For example, on a multimodal knowledge base [11], each node might be either an image or a textual description.

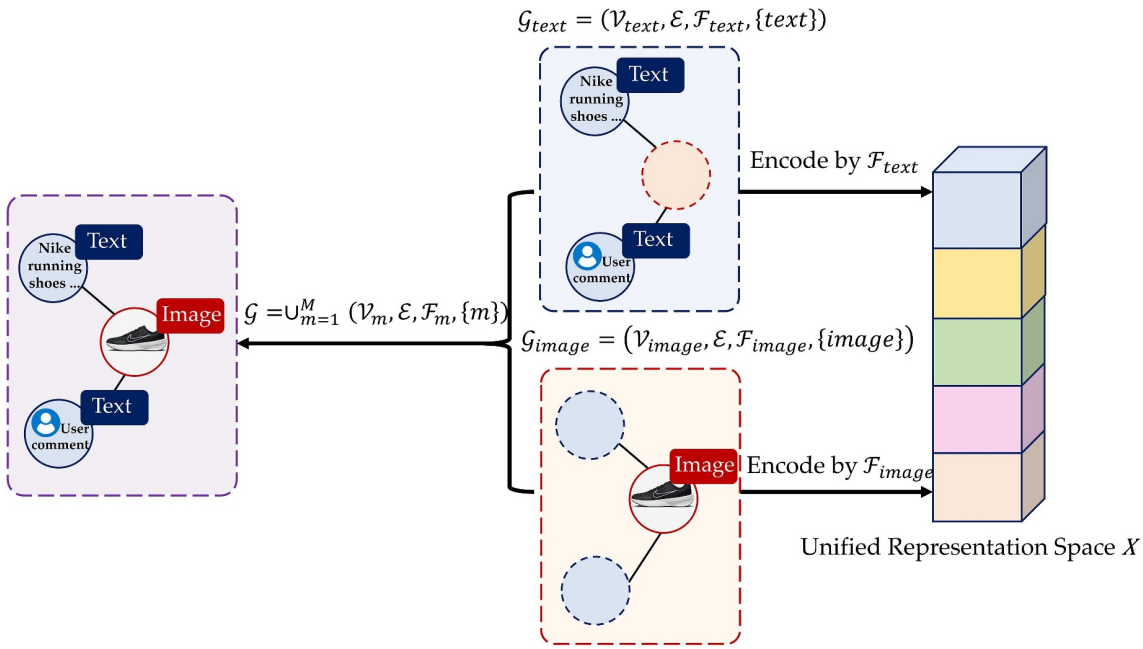

Figure 2 (Color online) An illustrative diagram and a running example of a multimodal graph, taking a small product graph with text and image nodes as an example.

• Graph-level multimodal graph, where the graphs come from different modalities, while each graph has unimodal features, i.e., $\mathcal{G} = \bigcup_{m=1}^{M} \mathcal{G}_m = \bigcup_{m=1}^{M} (\mathcal{V}_m, \mathcal{E}_m, \mathcal{X}_m, \{m\})$, where m representing one modality in M modalities, and $\mathcal{E}_i \bigcap \mathcal{E}_j = \emptyset$. For example, on a multimodal question answering graph [12], we may have a graph with images, a graph with texts, etc.

Remark 1 (Indecomposable characteristics)**.** Although practitioners may model their data with the aforementioned decomposable multimodal graphs for convenience, most multimodal graphs in real-world scenarios, with nodes and edges having features from various modalities, may not be easily decomposable to several uni-modal subgraphs. For instance, in a multimodal graph where the text node says 'The Transformer was incredible!', the image node shows Optimus Prime (a central robot character from the Transformers movie series), and the knowledge node links to the movie Transformers, only joint reasoning over all three nodes can resolve the ambiguity and correctly interpret 'Transformer' as a film character rather than a neural network architecture. Due to the indecomposable characteristics, established multimodal fusion techniques in other multimodal fields [13] may fail to flexibly solve multimodal graph problems, calling for the need of native modeling of multimodal graph data in multimodal graph large language models.

Special cases of multimodal graph instances. The versatility of multimodal graphs allows them to represent not only complex inter-modal relationships but also instances or datasets composed of single or multiple modalities as special cases, e.g., instances of texts, images, audios, videos, etc, or pairs of image-captions, text-audios, etc. Here are examples of representing single-modal instances.

• A text sequence can be represented by a multimodal graph with a single text-attributed node, i.e., $\mathcal{G} = (\mathcal{V} = \{v\}, \mathcal{E} = \emptyset, \{\mathcal{F}_{\text{text}}\}, \{\text{text}\})$, where $\mathcal{F}_{\text{text}}$ is the text feature mapping function.

• A text sequence, more granularly, can be represented by a multimodal graph where each word is a node and sequential or semantic connections form edges, i.e., $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \{\mathcal{F}_{\text{word}}\}, \{\text{word}\})$, where $\mathcal{V} = \{v_1, \ldots, v_L\}, \mathcal{E} = \{(v_i, v_{i+1})\}_{i=1}^{L-1}$, L is the length of the text sequence, and $\mathcal{F}_{\text{word}}$ is the word feature mapping function.

• An image can be represented as a multimodal graph where pixels are nodes and their grid-like inter-connections form edges, i.e., an $H \times W$ image can be $\mathcal{G} = \left(\mathcal{V} = \{v_{ij}\}_{i=1,j=1}^{H,W}, \mathcal{E}_{\text{grid}}, \{\mathcal{F}_{\text{pixel}}\}, \{\text{pixel}\}\right)$, where $\mathcal{E}_{\text{grid}}$ represents grid-like inter-connections, and $\mathcal{F}_{\text{pixel}}$ is the pixel feature mapping function.

Beyond individual instances, multimodal graphs can efficiently represent entire datasets. Here are some examples.

• A batch of images can be represented as a multimodal graph with several image-attributed nodes without any edges, i.e., $\mathcal{G} = (\mathcal{V} = \{v_1, \ldots, v_K\}, \mathcal{E} = \emptyset, \{\mathcal{F}_{\text{image}}\}, \{\text{image}\})$, where $\mathcal{F}_{\text{image}}$ is the image

feature mapping function, and K is the number of images.

• An image-captioning dataset can be represented as a multimodal graph where edges connect image-attributed nodes to their corresponding text-attributed caption nodes, i.e., for K image-caption pairs, $\mathcal{G} = (\mathcal{V}_{\text{image}} \cup \mathcal{V}_{\text{text}}, \mathcal{E}_{\text{image-text}}, \{\mathcal{F}_{\text{image}}, \mathcal{F}_{\text{text}}\}, \{\text{image, text}\})$, where $\mathcal{V}_{\text{image}} = \{v_1, \ldots, v_K\}$ are image nodes, $\mathcal{V}_{\text{text}} = \{u_1, \ldots, u_K\}$ are caption nodes, $\mathcal{E}_{\text{image-text}} = \{(v_k, u_k) \mid 1 \leqslant k \leqslant K\}$ are edges connecting images to their captions, and $\mathcal{F}_{\text{image}}$ and $\mathcal{F}_{\text{text}}$ are the image and text feature mapping functions, respectively.

This ability to abstract various data forms into a unified graph structure underscores the expressive power of multimodal graphs.

Remark 2 (Multi-granularity characteristics)**.** Multimodal graphs inherently possess the ability to organize data with multi-granularity across modalities, features, and structures. However, this capability is a double-edged sword. While they can represent vast amounts of information, they also introduce significant challenges for models to process them flexibly. Unlike other domains that feature units of roughly uniform granularity, such as word tokens in natural language processing (NLP) or image pixels in computer vision (CV), multimodal graphs often contain units ranging from fine-grained features (e.g., pixels, words) to coarse-grained concepts (e.g., full images, entire documents). To build an effective MG-LLM capable of flexibly handling information at diverse granularities on multimodal graphs, it may be necessary to design a unified multimodal graph vocabulary and tokenizer for learning multimodal graph representations in a shared space.

2.2 Generative modeling of multimodal graph tasks

In this section, we can frame, through generative modeling, that all multimodal graph tasks are multimodal graph generation. Due to the inherent multi-granularity characteristics of multimodal graphs, we can unify several classical discriminative tasks and emerging generative tasks under a single generative perspective, which can bring advantages of unified task forms, types, and interfaces. Suppose MG-LLM learns a conditional probability distribution to generate an output multimodal graph $\mathcal{G}_{\text{out}}$ given an input $\mathcal{G}_{\text{in}}$. Formally, the objective is to model

$$P(\mathcal{G}_{\text{out}} | \mathcal{G}_{\text{in}}; \Theta), \tag{1}$$

where Θ are the MG-LLM's parameters. Various multimodal graph tasks can be redefined generatively.

• **Multimodal node classification (NC)** aims to take a multimodal ego-graph centered around a target node as input, and generate a multimodal graph representing the predicted class, i.e., $P(\mathcal{G}_{\text{class}} \mid \mathcal{G}_v)$, where $\mathcal{G}_v = (\mathcal{V}_v, \mathcal{E}_v, \mathcal{F}_v, \mathcal{M}_v)$ is a subgraph centered at node $v \in \mathcal{V}$ and $\mathcal{G}_{\text{class}} = (\{v\}, \emptyset, \{\mathcal{F}_m\}, \{m\})$ is an output graph where $\mathcal{F}_m(v)$ encodes the class label (e.g., text or image).

• **Multimodal link prediction (LP)** takes a multimodal subgraph containing two endpoint nodes and their local neighborhood, and outputs a multimodal graph indicating the link's existence or properties, i.e., optimizing the objective $P(\mathcal{G}_{\text{link}} \mid \mathcal{G}_{(u,w)})$, where $\mathcal{G}_{(u,w)} = (\mathcal{V}_{(u,w)}, \mathcal{E}_{(u,w)}, \mathcal{F}_{(u,w)}, \mathcal{M}_{(u,w)})$ is a subgraph with nodes $u, w \in \mathcal{V}$ and their neighborhood, and $\mathcal{G}_{\text{link}} = (\{v\}, \emptyset, \{\mathcal{F}_{(m)}\}, \{m\})$ is an output graph where v's feature $\mathcal{F}_m(v)$ encodes link existence or type.

• **Multimodal graph classification (GC)** takes the entire multimodal graph as input and generates a multimodal graph representing the graph's overall class or category, i.e., optimizing the objective $P(\mathcal{G}_{\text{class}} \mid \mathcal{G})$, where $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{F}, \mathcal{M})$ is the input graph and $\mathcal{G}_{\text{class}} = (\{v\}, \emptyset, \{\mathcal{F}_m\}, \{m\})$ is an output graph where v's feature $\mathcal{F}_m(v)$ describes the predicted class.

• **Multimodal graph question answering (GQA)** aims to generate an answer based on a multimodal graph $\mathcal{G}$ and a text-attributed query node v_Q, i.e., optimizing the objective $P(\mathcal{G}_{\text{answer}} \mid \mathcal{G}_{\text{Q}})$, where $\mathcal{G}_{\text{Q}} = (\mathcal{V} \cup \{v_Q\}, \mathcal{E} \cup \mathcal{E}_Q, \mathcal{X} \cup \{\mathcal{F}_{\text{text}}(v_Q)\}, \mathcal{M} \cup \{\text{text}\})$ is the graph augmented with v_Q and potential edges $\mathcal{E}_Q$ and $\mathcal{G}_{\text{answer}}$ is the generated answer graph (e.g., a text/image node or a subgraph).

• **Multimodal graph reasoning (GR)** extends GQA with complex multi-hop reasoning. The generated output $\mathcal{G}_{\text{reasoning}}$ may embody complex logical structures or a chain of thought, i.e., optimizing the objective $P(\mathcal{G}_{\text{reasoning}} \mid \mathcal{G}_{\text{Q}})$, where $\mathcal{G}_{\text{Q}}$ includes the graph and query, and $\mathcal{G}_{\text{reasoning}}$ encapsulates the reasoning result, which could be the thinking process like chain-of-thoughts or graph-of-thoughts.

• **Multimodal graph text generation (TG)** utilizes multimodal graph information to generate coherent text sequences, i.e., optimizing the objective $P(\mathcal{G}_{\text{text}} | \mathcal{G})$, where $\mathcal{G}_{\text{text}}$ is the generated text, such as a summary of a group of papers cited by each other or a new git patch based on correlated git commits.

• **Multimodal graph image generation (IG)** aims to generate novel images, where a multimodal graph with textual descriptions, structured data, or other modal inputs can serve as a basis, i.e., optimizing the objective $P(\mathcal{G}_{\text{image}} | \mathcal{G})$, where $P(\mathcal{G}_{\text{image}})$ is the generated image, such as a descriptive image of the

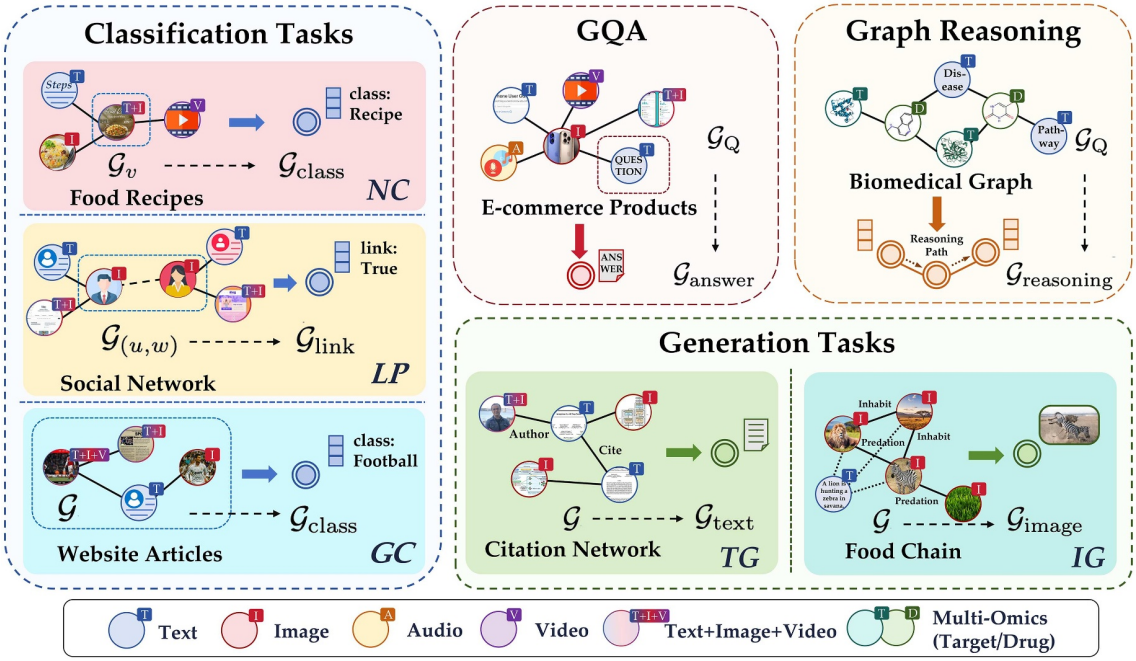

Figure 3 (Color online) Concrete working example of generative modeling across different tasks.

food chain based on a multimodal graph of ecosystems or a novel-style painting based on a multimodal graph of artist networks.

We provide concrete applications for the aforementioned tasks under a generative perspective, as illustrated in Figure 3. Further details on the datasets for these tasks are available in Section 4.

Remark 3 (Multi-scale characteristics)**.** This generative paradigm offers a powerful and flexible framework for modeling diverse multimodal graph tasks with a unified interface. Since multimodal graphs inherently represent multi-granular information, ranging from unimodal instances and bi-modal instances to entire multimodal datasets, the resulting input and output spaces are exceptionally versatile, accommodating a wide array of tasks. This versatility, however, introduces multi-scale characteristics for multimodal graph tasks: the input and output graphs can differ significantly in scope. For example, a task might take an entire graph as input but require only a single node or path as output (as in a GQA task), or vice versa for other potential tasks. This disparity in scales and granularity poses critical challenges for unified task modeling and task prompt design in MG-LLM, as the scales and semantic levels of inputs and outputs can vary widely.

2.3 Unified view of multimodal graph models

In this section, we provide a unified view on current multimodal graph models by first proposing a transformation function. Then, we bring together the two main categories that are moving towards MG-LLM, and we will briefly overview these approaches.

Transformation function. We first propose a transformation function, denoted as $\mathcal{T}$, which maps one multimodal graph to another. This transformation often involves a reduction in the number of modalities and the size of the graph, thereby facilitating processing by models or aligning the output with desired modalities and formats. This transformation function $\mathcal{T}$ can be parameter-free (e.g., designed via heuristic rules) or parametric (learned using neural networks), i.e., $\mathcal{T}_\theta$ with the parameters θ, which we omit for brevity.

For instance, in the input space, $\mathcal{T}$ can convert an entire multimodal graph into text. This could involve image captioning for image features, speech-to-text conversion for audio features, and textualizing edges into XML-like languages for structures. Consequently, a complex multimodal graph is transformed into a simplified multimodal graph where nodes primarily possess text attributes. Formally, given an input multimodal graph $\mathcal{G}$, the transformation $\mathcal{T}$ could yield $\mathcal{G}' = (\mathcal{V}', \mathcal{E}', \mathcal{F}', \{\text{text}\})$, where $\mathcal{F}'$ is the text feature mapping function.

Similarly, in the output space, $\mathcal{T}$ can summarize a multimodal graph into a single label, a document, or an image. An image output, for example, might not solely originate from extracting a single image-attributed node but could also involve rendering an entire multimodal graph into a coherent visual representation (e.g., visualizing a family tree).

In this view, a multimodal graph model can be seen as first transforming the input multimodal graph via a transformation, then modeling it, and finally transforming it again into the desired output. This can be formally expressed as

$$\mathcal{G}_{\text{out}} = \mathcal{T}_{\text{out}}(\phi_\theta(\mathcal{T}_{\text{int}}(\mathcal{G}_{\text{in}}))), \tag{2}$$

where $\mathcal{G}_{\text{in}}$ is the input multimodal graph, $\mathcal{T}_{\text{in}}$ is the input transformation, ϕ_θ is the core multimodal graph model, $\mathcal{T}_{\text{out}}$ is the final transformation, and $\mathcal{G}_{\text{out}}$ is the generated output multimodal graph. This generalized framework highlights the crucial role of these transformations in aligning diverse multimodal graph data with the model's capabilities and desired output formats.

Multimodal graph neural networks. The transformation function $\mathcal{T}$ is typically considered a parametric function. Here, information from each modality is initially mapped into a learned representation via trained modality-specific encoders. Subsequently, a graph neural network (GNN) leverages its message-passing mechanism to learn from these representations and derive the required labels for downstream tasks. For instance, multimodal graph convolution networks (MGCNs) [14–16] utilize the learned multimodal representation to form an adjacency matrix, and multimodal graph attention networks (MGATs) fuse information from different modalities by assigning different attention weights to each node [17–19]. The primary advantage of these approaches lies in the MGNN's ability to explicitly utilize structural information within the graph. However, a significant drawback is the lack of flexible input and output spaces, which complicates the development of unified foundation models for multimodal graphs. Furthermore, this late fusion function can lead to substantial information loss, making it challenging to capture fine-grained modal interactions.

Graph large language models. GraphLLMs often employ different strategies for the transformation function $\mathcal{T}$. (1) Some studies utilize a non-parametric transformation function. For instance, they might describe the entire graph as text, which is then processed by an LLM [20–26]. Alternatively, they might transform the graph into image-text pairs to be processed by a vision-language model (VLM). The advantage of these methods is their ability to leverage the flexible input and output spaces of LLMs or VLMs, enabling them to handle a wide range of tasks. However, they are heavily dependent on the inherent capabilities of the underlying LLM or VLM. Describing a complex graph entirely in text can lead to very long contexts, making comprehension difficult for the model. Moreover, certain modal information may be inherently challenging to textualize (e.g., an image converted to text can suffer significant information loss). (2) Other studies employ a parametric transformation function. For instance, LLaGA [7] employs a parametric projector to transform graph data into structured sequences, which are then embedded into the token space and fed into a large language model (LLM) for further processing. This mapping enables LLMs to effectively handle graph-structured data, enhancing their versatility, generalizability, and interpretability. GOFA [8] interleaves randomly initialized GNN layers within a frozen pre-trained LLM, organically combining semantic and structural modeling capabilities. This design leverages the GNN's strength in processing graph structures alongside the LLM's generative and reasoning abilities. It is important to note that current GraphLLMs rarely address multimodal graph problems directly.

Remark 4 (Modular characteristics)**.** For building a native MG-LLM, the transformation function $\mathcal{T}$ might ideally be an identity mapping, i.e., the input multimodal graph is directly fed into the primitive MG-LLM without any information loss. To achieve this ambitious goal, we might have to pretrain the model on extremely large multimodal graphs, e.g., the entire internet, so that sufficient pairwise data across modalities could directly empower the model with comprehensive knowledge of various modalities and graph structures. However, this monolithic approach might be impractical in the near future due to the considerable computational expenses and data acquisition challenges. One possible solution to circumvent this limitation might be building a modular multimodal graph LLM, which integrates various parameterized modules designed for specific functions to understand as well as generate both the structures and multimodal information within multimodal graphs. This modularity could allow for more efficient training and flexible adaptation to diverse multimodal graph tasks. It is important, however, to distinguish this from a generic Omni-MLLM with a graph plugin, as an ideal modular MG-LLM would still feature deeply co-designed components.

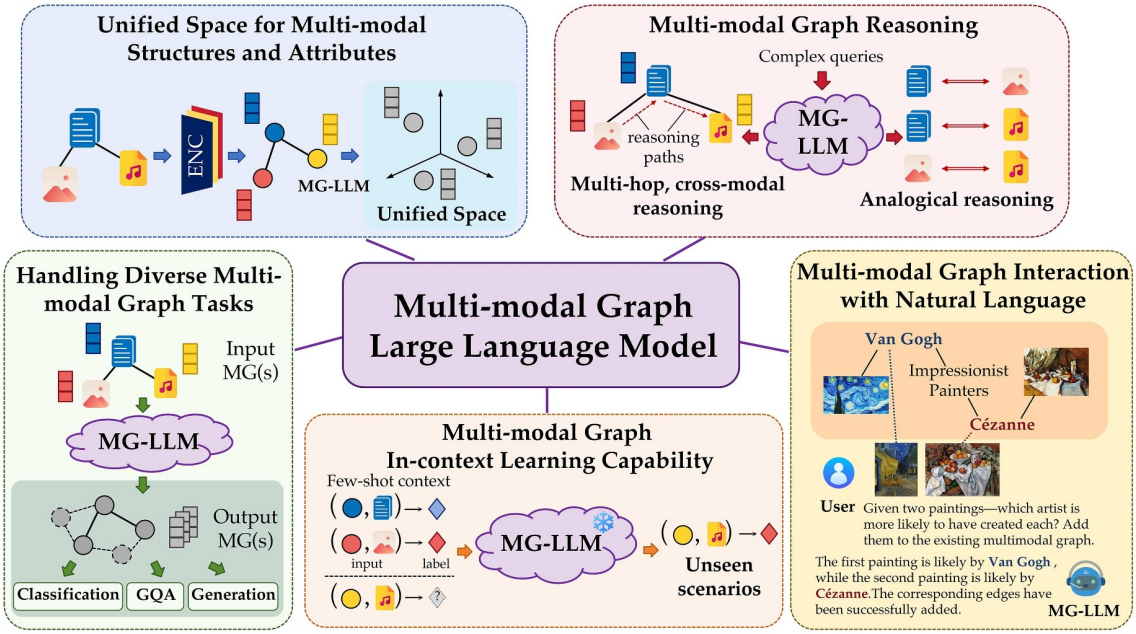

Figure 4 (Color online) Key characteristics towards MG-LLM.

3 Towards multimodal graph large language models

In this section, we will delve into the essential characteristics that a multimodal graph large language model should possess, the core challenges in achieving these characteristics, current relevant approaches, and potential future research directions. The key characteristics are illustrated in Figure 4.

3.1 Unified space for multimodal structures and attributes

Desired characteristics. MG-LLM is envisioned to effectively process information across diverse domains and a wide range of data modalities [6]. Real-world applications in healthcare, finance, scientific discovery, and social media often present highly heterogeneous structures and data types [27]. Therefore, MG-LLM should possess strong domain transferability and the capacity for broad representation generalization. A key characteristic of MG-LLM is the ability to align diverse multimodal features and relations within a truly unified representation space. This requires developing a comprehensive multimodal graph vocabulary capable of capturing highly irregular, multi-level, and continuous structural and attribute information. Such a space should facilitate the learning of transferable patterns that generalize across heterogeneous data domains. The goal is to create a seamless integration where information from text, images, audio, video, and structured graph topologies can be jointly processed and understood. This unified space should minimize redundancy while maximizing information retention and the faithful representation of relational semantics, enabling robust domain transferability and flexible interpretation of various data inputs.

Key challenges. One of the fundamental challenges in building MG-LLM is the inherent heterogeneity of data domains. Data often originates from various sources with different formats and structures, such as biomedical data (e.g., proteins, drugs) [28], social networks (e.g., text, images) [29], and multimodal knowledge graphs (e.g., language, images, temporal data) [30]. These domains exhibit distinct properties, including categorical, continuous, or structured data types [31]. This heterogeneity poses a significant obstacle in designing a unified model capable of efficiently integrating and processing such diverse data. Despite recent efforts to develop foundational models for graph data [32], these approaches remain limited to a few domains. Models [33,34] still fall short of providing a truly universal multimodal graph encoding scheme, which hinders generalization to broader and more diverse application scenarios. Furthermore, the multi-granularity of nodes and structures presents a significant challenge. In multimodal graphs, nodes represent entities that may belong to different modalities (e.g., images, text, molecules) with various granularities, while edges capture relationships between these nodes. Effectively modeling these heterogeneous nodes and their connections, which often have varying structures, requires novel approaches for

multimodal embedding and multimodal graph structure learning to better accommodate the complexity of diverse node types and relations. The highly irregular, multi-level, and potentially continuous nature of multi-graph vocabularies further complicates the creation of a unified representational space, leading to potential issues like redundancy and information loss during the integration of multimodal features and relations.

Relevant literture. Efforts towards achieving domain transferability and generalizable representations in graph learning have led to the development of foundation models and pre-training strategies for graph data [32–34]. Specifically for multimodal graphs, research has begun exploring how to learn domain-invariant representations that can generalize across different knowledge graphs and multimodal networks [35]. The growing interest in prompt-driven and instruction-based paradigms, largely influenced by the success of large language models, has also spurred work in adapting these approaches for unified data processing within structured and multimodal contexts [36–38]. While these studies represent significant steps, they often grapple with unifying the vastly different granularities and semantic levels inherent in multimodal graph data, or they rely on transformation functions that might incur information loss.

Future directions. To achieve a truly unified space for multimodal structures and attributes, several critical future directions emerge. Firstly, there is a pressing need to develop novel multimodal graph vocabulary and tokenization schemes that can flexibly represent information ranging from fine-grained features (e.g., pixels, words) to coarse-grained concepts (e.g., full images, entire documents), as well as complex graph topologies. This involves moving beyond simple concatenations or late fusion, aiming for an early and deep integration of multimodal signals. Secondly, research should focus on designing architectures capable of learning genuinely transferable patterns across the highly irregular, multi-level, and continuous nature of real-world multimodal graphs. This might involve exploring advanced graph neural network designs combined with foundation models that can process diverse modalities natively. Thirdly, strategies to mitigate redundancy and information loss during multimodal feature and relation integration are crucial. This could involve attention mechanisms that dynamically weigh modality contributions or latent space learning that preserves critical inter-modal dependencies. Finally, developing benchmark datasets specifically designed to evaluate the effectiveness of models in this unified multimodal graph space, across diverse domains, will be essential to drive progress in this field. This foundational work will be instrumental in realizing the vision of native modeling capable of operating directly on rich, multimodal graph data without substantial information loss.

3.2 Handling diverse multimodal graph tasks

Desired characteristics. A multimodal graph large language model should be adept at handling a vast array of tasks, moving beyond traditional discriminative objectives to embrace a unified generative paradigm. As previously outlined, the ability to frame all multimodal graph tasks as multimodal graph generation is a key characteristic. This necessitates a model capable of treating diverse problems, such as multimodal node classification, link prediction, graph classification, question answering, reasoning, and even multimodal content generation (text or image), as transformations from an input multimodal graph to an output multimodal graph. The model should demonstrate flexibility in its input and output spaces, seamlessly adapting to tasks that operate on different granularities, from fine-grained node features to entire graph structures. Furthermore, the model should support prompt-driven or instruction-based learning, allowing for versatile task adaptation and generalization to new, unseen multimodal graph scenarios through natural language commands or structured prompts.

Key challenges. A fundamental challenge in developing such models lies in the multi-scale characteristics of multimodal graph tasks. As highlighted in the generative modeling section, the input and output graphs can differ significantly in scope and granularity. For instance, a task might take a comprehensive multimodal graph as input but require only a single predicted node's attribute as output (as in node classification or a specific graph question answering task) [39]. Conversely, other tasks might demand the generation of an entire sub-graph or even a new multimodal graph from a simple input. This disparity in scales complicates unified task modeling and especially task prompt design, as the semantic levels of inputs and outputs vary widely [40]. Beyond structural scale, the inherent scales of different modalities also pose challenges; text inputs can vary greatly in length, images in size, and videos in temporal duration. Integrating and generating information across these vastly different intrinsic modal scales, alongside varying graph structural complexities, requires sophisticated mechanisms to prevent information loss or redundancy and maintain coherence across the entire representation [41].

Relevant literture. Efforts to address the diversity of graph tasks have led to the development of multi-task graph foundation models, which aim to provide a single framework for various graph-related tasks [42–45]. Inspired by the success of large language models, research has also begun to explore prompt-driven and instruction-based paradigms for graph and multimodal contexts. These approaches leverage the flexibility of language models to adapt to different downstream tasks by formulating them as sequence-to-sequence or graph-to-sequence problems [46, 47]. Current GraphLLMs or vision-language models, by virtue of their flexible input and output capabilities, can handle some multimodal graph tasks by first transforming the graph into a text-centric or image-text paired representation. While these methods demonstrate promising abilities in task generalization and leveraging pre-trained knowledge, they often face limitations when directly handling the intricate multi-scale nature and inherent graph structures, sometimes relying on transformation functions that may incur information loss or struggle with extremely long or complex contexts.

Future directions. To effectively handle diverse multimodal graph tasks, future research should prioritize developing novel approaches for unified task modeling that natively account for the varying scales and granularities of both input and output multimodal graphs. This involves designing architectures that can process fine-grained features (e.g., pixels, words) alongside coarse-grained concepts (e.g., full images, entire documents) and complex graph topologies in a seamless manner. Another critical direction is the investigation of adaptive prompting mechanisms that can dynamically adjust to the task's specific scale and the required output granularity, moving beyond generic prompts. Furthermore, attention should be given to extending generative modeling techniques to truly open-set multimodal graph generation, where the model can synthesize novel graph structures or content (e.g., images, texts, audios) that are not limited to predefined sets. This requires a deeper integration of multimodal understanding with generative capabilities, aiming for models that can operate directly on rich multimodal graph information without relying on lossy intermediate transformations.

3.3 Multimodal graph in-context learning capability

Desired characteristics. A central aspiration for MG-LLMis to exhibit robust in-context learning (ICL) capabilities. This involves the model's ability to solve novel tasks by conditioning on a limited number of graph-anchored examples provided directly within the prompt, without requiring explicit weight updates or fine-tuning. Similar to how large language models learn from demonstrations in text, MG-LLM should infer underlying patterns and generalize to unseen multimodal graph scenarios. This necessitates a model that can interpret and leverage diverse multimodal features and complex graph structures present in the few-shot examples to inform its predictions or generations for new queries. Ultimately, achieving this capability relies on effective generative pretraining on large-scale, paired multimodal graph data, coupled with an architectural design and self-supervised learning objectives that facilitate flexible transfer and scaling.

Key challenges. Extending in-context learning to graph-based multimodal contexts presents significant challenges that are not encountered in plain text domains. Graphs inherently possess a variable topology, long-range dependencies, and a non-sequential structure [6,48], making it difficult to define what constitutes an effective context window for ICL. Unlike a linear sequence of tokens, the 'neighborhood' or 'context' around a graph element can be complex and multifaceted. Furthermore, encoding the rich relational priors and intricate inter-modal connections compactly for consumption by models, especially transformer architectures that are often designed for sequential data, remains a non-trivial task. The diversity of modalities within a graph (e.g., text, images, audio associated with nodes or edges) further complicates the unified representation and contextual understanding necessary for effective in-context learning.

Relevant literture. Inspired by the success of large language models in in-context learning [49], researchers have begun exploring methods to imbue this capability into graph-based models. Strategies for graph ICL often involve linearization strategies, such as converting graphs or subgraphs into sequences of triples or using template filling to represent graph information in a text-like format [36]. Another approach involves the use of sampled subgraph prompts, where relevant subgraphs are selected to serve as examples for the model [50]. Hybrid architectures have also emerged, combining the strengths of pretrained GNNs for structural encoding with autoregressive decoders, which are adept at sequence generation and ICL [51]. Specific methods like AskGNN [50] and retrieval-augmented transformers adapted for graphs [52] demonstrate that carefully selecting relevant subgraphs and aligning them with language

tokens can enhance few-shot learning within graph constraints, pointing towards the importance of context retrieval and integration. Progress has also been made in aligning graph information with textual prompts for language models [53, 54].

Future directions. To fully unlock the multimodal graph in-context learning capability, several critical future directions should be explored. A key area is the development of more sophisticated multimodal graph tokenization schemes that can capture the inherent multi-granularity of graph entities (from fine-grained features to coarse-grained concepts) and their complex inter-modal relations in a way that is amenable to in-context learning. This would involve designing tokenizers that can flexibly abstract both structural and attribute information across modalities. Furthermore, research should focus on architectures that can intrinsically process irregular graph structures and multimodal information without significant information loss from linearizing or simplifying the graph. This might involve novel graph-specific attention mechanisms or prompt-based techniques that can dynamically integrate and reason over graph-anchored examples. The integration of retrieval mechanisms that can efficiently fetch relevant subgraphs or multimodal contexts for ICL, especially from vast multimodal knowledge graphs, will also be crucial. Ultimately, continued advancement in large-scale generative pretraining on diverse and rich multimodal graph datasets, coupled with self-supervised objectives tailored for graph understanding and generation, will be essential for models to acquire robust and transferable in-context learning abilities.

3.4 Natural multimodal graph interaction

Desired characteristics. An essential goal of MG-LLMis to enable natural language-based interaction over structured and multimodal data. Users should be able to query, edit, and reason about graph-structured knowledge using plain language, without needing to learn formal query languages like SPARQL or Cypher. This requires the MG-LLM to accurately map natural language inputs to graph traversal operations, complex reasoning chains, or precise node and edge modifications. Furthermore, the model should support rich, multi-turn dialogue, allowing for clarification and refinement of user intentions. A crucial aspect is visual grounding within graphs, enabling the model to align natural language descriptions with visual elements present within the graph context, thereby facilitating a more intuitive understanding of multimodal information. Beyond querying, the model should also be capable of graph-based summarization and generation of new graph structures or content in response to natural language commands. Ultimately, the desired characteristic is to provide a seamless, intuitive, and highly interactive interface between human users and complex multimodal graph data, allowing for direct understanding, editing, reasoning, and generation of information, bridging the gap from unstructured human input to structured graph knowledge, while aligning with human values and intentions.

Key challenges. One of the fundamental challenges in achieving natural multimodal graph interaction lies in the inherent semantic gap between the ambiguity and flexibility of natural language and the precise, structured nature of graph data. Natural language expressions can be vague or underspecified, making it difficult for a model to infer the user's exact intention for graph operations, especially when dealing with heterogeneous multimodal features [5]. Mapping these vague intentions to concrete graph traversals, edits, or reasoning chains is non-trivial. Furthermore, supporting multimodal interactions introduces complexities, as the model must seamlessly interpret queries that might refer to text, image, audio, or video attributes within the graph, and potentially generate responses in a desired modality [55]. The ability to understand, edit, reason, and generate content within a multimodal graph poses distinct challenges; for instance, editing an image-attributed node based on a textual command requires sophisticated cross-modal understanding and generative capabilities [56]. Ensuring consistency and avoiding unintended side effects during graph modifications initiated by natural language commands is another significant hurdle. Finally, scaling such interactive capabilities to complex, real-world multimodal graphs, such as those representing industry traffic patterns, molecular structures, or visual relational graphs, presents significant challenges in terms of computational efficiency and maintaining accuracy across diverse domains [57].

Relevant literture. Recent advancements in natural language interfaces for structured data have laid the foundational groundwork for multimodal graph interaction. Efforts in conversational knowledge graphs [58, 59] have explored how to enable multi-turn dialogue over knowledge bases, allowing users to progressively refine their queries. Similarly, research on natural language interfaces for databases (NLIDB) [60] focuses on translating natural language questions into structured query languages, a precursor to graph traversal and modification. With the rise of multimodal large language models, there

has been increasing interest in visual grounding, where models align language descriptions with visual elements in complex scenes or contexts [61–63]. These techniques are directly relevant for grounding natural language queries within image or video attributed graph nodes. Graph-based generative tasks, such as graph summarization [53, 64], have also emerged, demonstrating the ability to condense graph information into coherent text. Moreover, the integration of instruction-tuned language models with symbolic reasoning modules [65, 66] represents a promising direction for building more robust dialogue-centric MG-LLM, enabling them to leverage both pattern matching and logical deduction for complex graph interactions.

Future directions. To fully realize natural multimodal graph interaction, several critical future directions need exploration. Firstly, developing more sophisticated methods for disambiguating vague or underspecified natural language intentions and aligning them precisely with multimodal graph structures and attributes is crucial. This could involve interactive clarification dialogues where the model asks follow-up questions to refine its understanding. Secondly, research should focus on robust mechanisms for human feedback integration during the interaction process, allowing users to correct model interpretations or outputs, thereby continually improving the model's understanding and alignment with human values. This iterative feedback loop is essential for adapting models to new domains and user preferences. Thirdly, advancing generative capabilities to enable not only querying but also complex graph editing and generation through natural language commands is vital. This includes the ability to modify existing multimodal nodes and edges, add new entities, or even generate entire subgraphs based on high-level instructions, with applications in areas like molecular design or urban planning. Finally, scaling these interaction paradigms to dynamic and evolving multimodal graphs that represent real-world phenomena (e.g., real-time industry traffic, evolving scientific knowledge graphs) will require innovations in efficient graph indexing, retrieval, and incremental updates to maintain responsiveness and accuracy during natural language-driven interactions.

3.5 Multimodal graph reasoning

Desired characteristics. An MG-LLM should be capable of multi-hop, cross-modal reasoning. For example, the model should answer a complex query by combining clues from text and images through multiple inferential hops. Recent benchmarks like MultiModalQA [67] show that solving such cross-modal multi-hop questions remains challenging. Models must jointly reason over different modalities and knowledge sources to succeed on these tasks. Another desired capability is analogical inference across modalities, where the model draws structural comparisons between, for instance, an image pair and a text pair. Analogical reasoning is a fundamental aspect of human cognition. Initial studies suggest large models have some analogical ability. However, most prior work on analogies is single-modal. Multimodal analogical reasoning is still in its early stages. Early efforts on multimodal analogies over knowledge graphs (e.g., the MARS benchmark [68]) illustrate both the potential and the difficulty of this skill. For instance, demonstrate that even advanced multimodal LLMs struggle with visual analogies unless special prompting or training is provided. This result underscores the importance of analogical inference as a future MG-LLM capability [69].

Key challenges. Building a MG-LLM poses several major challenges. First, modality alignment is difficult. The model must align and fuse information from heterogeneous sources such as images, text, and graphs into a coherent representation. Without explicit alignment mechanisms, an image's contents may not correctly map to textual concepts, impeding reasoning. Techniques like contrastive image-text pre-training (e.g., CLIP) are often used to partially address this problem by embedding modalities in a shared space [70]. Recent MG-LLM approaches include dedicated alignment modules for vision and language. For example, MR-MKG [55] employs a cross-modal alignment module to optimize image and text correspondences within a multimodal knowledge graph. Second, factual consistency remains a critical issue. Multimodal LLMs are prone to hallucination and may produce inconsistent answers that conflict with factual knowledge. This problem worsens when the model must recall external knowledge, such as from a graph that it was not pre-trained on. Recent work has highlighted these hallucination problems and proposed evaluation benchmarks (e.g., MHaluBench [71]) and detection frameworks to reduce them. Indeed, MR-MKG was motivated by the observation that vanilla LLMs often fabricate details about images due to missing visual knowledge and injects a multimodal knowledge graph to ground the model in reality [55]. A third challenge is the fragility of current processing pipelines. Many models operate in stages or rely on external tools, and errors in early steps can cascade. Ref. [72] noted that fixed sub-models

in current systems make them unable to recover from intermediate mistakes. Improving robustness and feedback mechanisms is therefore an important challenge.

Relevant literture. Several initial approaches to MG-LLM have been proposed to address these challenges. A common strategy is to integrate a knowledge graph (KG) or graph neural network into the multimodal pipeline to better handle structured, relational information. For example, MR-MKG [55] augments a vision-language model with a multimodal knowledge graph that contains nodes and relations spanning text and images. It uses a relation-aware graph neural network to encode the MMKG and injects these representations into the LLM to improve reasoning. Another line of work focuses on graph construction and alignment between modalities. MAIL [61] constructs a scene graph from image objects and a concept graph from external knowledge, aligning them via shared entities and fusing them through a pseudo-siamese graph neural network. This enables reasoning over a combined multimodal graph and has shown strong results in knowledge-intensive visual QA. Beyond QA, graph-enhanced multimodal models have been applied to other domains. For example, Choi et al. [73] proposed a model for healthcare that injects patient-specific graphs into an LLM and uses GNN-based message passing to align clinical text, lab results, and images. These methods highlight the benefit of structured reasoning but also reveal engineering complexity, as each task may require tailored graph construction and alignment strategies [74].

Future directions. Future research on multimodal graph reasoning should tackle several ambitious goals. First, novel graph representation strategies tailored explicitly for multimodal contexts could be developed, going beyond current embedding approaches to represent complex interactions between modalities more intuitively. Second, dynamically constructed multimodal graphs that adapt in real-time to the context or queries presented to the model may enhance reasoning efficiency and accuracy. Additionally, exploring scalable inference techniques specifically designed for large and dense multimodal graphs is essential to overcoming the context-length limitations of current models. Finally, there is a significant opportunity to advance explainability by designing methods that produce interpretable reasoning paths within multimodal graph structures, enabling users to better understand and trust the model's decision-making process.

3.6 Discussion on scalability and computational efficiency

Computational efficiency. Compared with unimodal LLMs, MG-LLMs face significantly higher computational demands due to the need to jointly encode complex graph structures and diverse multimodal signals. Large-scale pretraining introduces massive parameter counts and memory usage, making naive extensions of existing LLM or GNN architectures impractical [75, 76]. To improve efficiency, several strategies can be adopted: (1) parameter sharing across modalities to reduce redundancy (as explored in unified multi-domain models [75]), (2) modular architectures that enable different sub-modules to specialize on particular modalities or functions (e.g., separate expert components for each modality [77]), and (3) sparse or sampled attention mechanisms to focus computation on the most relevant subgraphs and modalities [76, 78]. For inference, pruning redundant tokens and layers (using efficient transformer techniques [78]), designing lightweight multimodal graph tokenizers, and incorporating retrieval-augmented mechanisms [52] are promising directions to lower latency without sacrificing accuracy.

Scalability. Real-world multimodal graphs often involve millions of nodes and edges, coupled with highly heterogeneous modalities and relations. Scaling MG-LLMs to such large graphs requires both algorithmic and system-level innovations. On the algorithmic side, graph sampling methods (e.g., neighbor sampling [79] and subgraph sampling [80]) and hierarchical modeling (e.g., differentiable graph pooling [81]) can make training and inference feasible on large datasets by reducing the effective graph size per batch. On the system side, distributed training pipelines and memory-efficient representations (e.g., sparse adjacency matrices or quantized features) are indispensable to handle billion-scale graphs. Frameworks like DistDGL [82] demonstrate hybrid CPU-GPU training to scale GNNs to extremely large graphs. Furthermore, scalability must also account for modality and task diversity: MG-LLMs should adapt to varying input granularities and output structures while maintaining consistent generalization. Progressive scaling strategies and curriculum learning techniques [83] may help stabilize training as the model gradually expands to increasingly large and diverse multimodal graph corpora.

Deployment strategies. Even if an MG-LLM can be trained successfully, deployment in real-world scenarios poses additional constraints. Many applications (such as biomedical analysis or recommendation) impose strict latency requirements and often operate under hardware limitations. To make deployment feasible, model compression and distillation can reduce MG-LLMs into lighter domain-specific

variants that retain core capabilities. For example, knowledge distillation transfers knowledge from a large teacher model to a smaller student model [84], and has been effective in compressing deep models without major performance loss. Quantization and pruning techniques can further improve inference speed and memory footprint (e.g., 8-bit quantization and weight pruning as in deep compression [85]), enabling deployment on resource-limited edge devices. Domain-adapted MG-LLMs (for example, a variant specialized for molecular graphs in chemistry [86] or for urban spatial graphs in planning) are another strategy to balance generality and efficiency by focusing on a narrower range of modalities/tasks. Finally, hybrid deployment pipelines can be employed, where heavy multimodal graph computations are offloaded to powerful servers (cloud) while lightweight modules run on clients (edge devices). Such split-computing approaches (akin to the Neurosurgeon framework for splitting DNN workloads [87]) achieve a practical compromise between responsiveness and accuracy, ensuring that MG-LLM-based solutions can meet real-time constraints in production environments.

4 Multimodal graph datasets

Recent years have witnessed the emergence of multimodal graph learning datasets, which enrich traditional graph structures by incorporating image, text, video, audio, and multi-omics data. These datasets facilitate more challenging and realistic graph learning tasks by providing heterogeneous node and edge attributes. In this review, we categorize representative benchmarks by task type and summarize their scale, modalities and domains. Statistics can be summarized in Table 1. The datasets can also be grouped based on the origin of their domains, reflecting the types of real-world data they are derived from. Social network-related datasets originate from user interactions such as e-commerce activities, book recommendations, online articles, and urban information, which capture patterns of social behavior and digital connectivity. Knowledge graph datasets stem from structured repositories of knowledge, including multimodal knowledge bases, artistic relationships, biomedical repositories, and recipe step data, emphasizing their foundation in curated domain-specific resources. Scene graph datasets, on the other hand, originate from visual scenes where objects and their relationships are explicitly annotated, making them distinct in their focus on spatial and semantic structure within images, and are mainly used for visual graph question answering tasks. Such a categorization is summarized in Table 2.

4.1 Node classification

Node classification benchmarks evaluate a model's ability to predict node labels when each node carries multimodal attributes. ELE Fashion [10] is a medium-scale e-commerce product graph comprising approximately 97800 product nodes, each being annotated with a product title and a high-resolution image. Books NC [10] includes roughly 685300 book nodes, each with cover images and descriptions, and is annotated for ten book categories. G2MF-Urban [88] is an urban planning graph with about 100000 street nodes and 2000000 edges, where nodes incorporate overhead imagery and point of interest (POI) text for functional zone classification. In the biomedical domain, the Pan-Cancer Atlas [89] comprises approximately 11286 tumor samples across 33 cancer types, profiled by multiple omics assays (mRNA, miRNA, DNA methylation, proteomics, CNV), and is used for pan-cancer molecular subtype classification; each sample is modeled as a node whose features are the concatenated multimodal measurements, with edges encoding biological relations. Similarly, TCGA-BRCA [90] contains around 1084 breast tumor samples assayed on six platforms (genomic, epigenomic, transcriptomic, proteomic) and supports breast cancer subtype classification, also modeling each sample as a node with concatenated multimodal measurements and edges encoding biological relations. Additionally, the OMG-NAS framework [27] evaluates two real-world out-of-distribution benchmarks: the Tencent graph from WeChat official accounts, which includes 8000 article nodes and 60000 user-view edges, with each node carrying head images and titles; and the Amazon review graph, featuring 100000 review nodes and 300000 co-review edges, combining product images and textual feedback.

4.2 Link prediction

Link prediction datasets require inferring missing or future edges in graphs with multimodal node features. Books LP [10] comprises approximately 636500 book nodes, each with cover images and descriptions, used for link inference. The Sports CoPurchase and Cloth CoPurchase datasets [10] are co-purchase graphs

Table 1 Overview of tasks and their corresponding multimodal graph datasets. NC: node classification; LP: link prediction; GC: graph classification; GQA: graph question answering; GR: graph reasoning; TG: text generation; IG: image generation.

Task	Dataset	Modalities	Scale	Domain
NC	ELE Fashion [10]	Text+Vision	98k nodes, 20k edges	E-commerce products
	Books NC [10]	Text+Vision	684k nodes, 7M edges	Book recommendation
	G2MF-Urban [88]	Text+Vision	100k nodes, 2M edges	Urban planning
	Pan-Cancer Atlas [89]	Multi-omics	11k samples from 33 cancer types	Biomedical repository
	TCGA-BRCA [90]	Multi-omics	1084 breast tumor samples	Biomedical repository
	OMG-NAS Tencent [27]	Text+Vision	8k nodes, 60k edges	Website articles
	OMG-NAS Amazon [27]	Text+Vision	100k nodes, 300k edges	E-commerce products
LP	Books LP [10]	Text+Vision	64k nodes, 3437k edges	Book recommendation
	Sports CoPurchase [10]	Text+Vision	50k nodes, 25k edges	E-commerce products
	Cloth CoPurchase [10]	Text+Vision	126k nodes, 951k edges	E-commerce products
	HyperGCL-Ecomm [91]	Text+Vision	1M edges, 500M edges	E-commerce products
	VTKG-I&C [92]	Text+Vision	130 entities, 842 triples	Multimodal KGs
	TIVA-KG [93]	Text+Vision+Audio	50k entities, 200k triples	Multimodal KGs
GC	OMG-NAS Recipe [27]	Text+Vision	20k nodes, 160k edges	Food recipes
	Large-RG [17]	Text+Vision	500k nodes	Food recipes
GQA	GQA [12]	Text+Vision	113k images, 23M questions	Scene graphs
	CLEVR [94]	Text+Vision	100k images, 1M questions	Scene graphs
	SceneGraph-VQA [95]	Text+Vision	50k images	Scene graphs
GR	MARS&MarKG [68]	Text+Vision	34k triples, 13k questions	Multimodal KGs
	FB-ING-TXT [96]	Text+Vision	6k entities	Multimodal KGs
	WN9-ING-TXT [96]	Text+Vision	12k entities	Multimodal KGs
TG	VRF [97]	Text+Vision	200 recipes, 89 actions	Food recipes
	MS Recipe Corpus [98]	Text+Vision	4k dishes, 150k recipes	Food recipes
	Richpedia [11]	Text+Vision	3M entities	Multimodal KGs
IG	ART500K [99]	Text+Vision	311k nodes, 643M edges	Artwork relationships
	Amazon Coview [100]	Text+Vision	178k nodes, 3M edges	E-commerce products
	Goodreads [100]	Text+Vision	93k nodes, 637k edges	Book recommendation

Table 2 Datasets categorized by their origin, domain, and examples.

Origin	Domain	Datasets
Social network	E-commerce products	ELE Fashion [10], OMG-NAS Amazon [27], Sports CoPurchase [10], Cloth CoPurchase [10], HyperGCL-Ecomm [91], Amazon Coview [100]
	Book recommendation	Books NC [10], Books LP [10], Goodreads [100]
	Website articles	OMG-NAS Tencent [27]
	Urban planning	G2MF-Urban [88]
Knowledge graph	Multimodal KGs	VTKG-I&C [92], TIVA-KG [93], MARS&MarKG [68], FB-ING-TXT [96], WN9-ING-TXT [96], Richpedia [11]
	Artwork relationships	ART500K [99]
	Biomedical repository	Pan-Cancer Atlas [89], TCGA-BRCA [90]
	Food recipes	OMG-NAS Recipe [27], Large-RG [17], VRF [97], MS Recipe Corpus [98]
Scene graph	Scene graphs	GQA [12], CLEVR [94], SceneGraph-VQA [95]

containing 50250 and 125839 product nodes respectively, with each node annotated with product titles and images. HyperGCL-Ecomm [91] is a hypergraph dataset with 1000000 product nodes possessing multimodal features and 500000000 behavioral edges. VTKG-I and VTKG-C [92] present two commonsense knowledge graphs (KGs), each with approximately 130 entities and 842 triples, where entities are associated with images and text, designed for knowledge graph completion tasks. TIVA KG [93] is a quad-modal knowledge graph containing roughly 50000 entities and 200000 triples, which combines text, image, video, and audio modalities for completion tasks.

4.3 Graph classification

Graph classification datasets learn holistic representations for entire graphs with heterogeneous node and edge attributes. In the context of the OMG-NAS framework [27], a Recipe graph is utilized, comprised of approximately 20000 recipe nodes and 160000 ingredient/instruction edges, where images are partitioned into 16×16 patch nodes and text into word nodes. Separately, the Large-RG dataset [17] models a culinary graph containing over 500000 recipe nodes linked by ingredient edges, enriched with image and textual attributes.

4.4 Visual graph QA

Visual reasoning datasets convert images into scene graphs paired with compositional questions to assess structured inference. GQA [12] contains 113018 real-world images, 22.7 million questions, and scene graphs annotated with functional programs. CLEVR [94] is a synthetic visual question answering benchmark consisting of 100000 rendered RGB scenes paired with approximately 853000 automatically generated question-answer pairs. Each scene is accompanied by a detailed scene graph that encodes object attributes and spatial relations, and every question is mapped to a functional program specifying the multi-step reasoning required, with the dataset designed to minimize biases and provide exhaustive annotations. SceneGraph-VQA [95] comprises 50000 scene graphs with QA pairs, combining object-relationship hierarchies, image regions, and textual questions.

4.5 Graph reasoning

Multimodal graph reasoning datasets integrate heterogeneous information sources to support complex reasoning tasks. MARS and MarKG [68] serve as benchmark datasets for multimodal analogical reasoning. MARS contains 10685 training, 1228 validation, and 1415 test instances, where each task instance is a visual-textual analogical quadruple requiring the prediction of a missing entity. MarKG is a supporting knowledge graph for MARS, containing 11292 entities and 192 relations, with entities enriched by 76424 images along with textual and visual descriptions. Furthermore, two datasets, WN9-IMG-TXT and FB-IMG-TXT [96], are widely adopted multimodal knowledge graph benchmarks. WN9-IMG-TXT contains 6555 entities, while FB-IMG-TXT contains 11757 entities. In both datasets, each entity is associated with three modalities: structural graph information, images, and text.

4.6 Text generation

These datasets evaluate alignment or generation of text conditioned on multimodal workflows. Visual Recipe Flow [97] annotates 200 recipes with before-and-after image pairs for each action and is grounded in a recipe-flow graph designed for stepwise text generation. The Microsoft Multimodal Aligned Recipe Corpus [98] contains approximately 150000 text-video alignments across 4262 dishes, structured with cross-modal graphs between recipe steps and corresponding video segments for description tasks. Richpedia [11] models a multimodal knowledge base containing over 1000000 entities, where relations exist between KG textual entities and image entities, among image entities, or between image entities and values such as pixel information. This dataset is designed to support applications such as semantic search and text generation.

4.7 Image generation

Image generation utilizing multimodal graphs aims to synthesize visual content by leveraging the interconnected textual and visual information inherent in these complex network structures. For example, ART500K [99] curates an artwork domain with 311288 creations and 643 million interconnections reflecting shared artists or styles, each accompanied by textual and visual information. Amazon Coview [100] charts an e-commerce product landscape where 178890 items are linked by 3 million connections derived from concurrent Browse patterns, each item possessing textual and visual descriptions. Lastly, Goodreads [100] forms a book recommendation network of 93475 literary works interconnected by 637210 relationships that highlight their comparability, with each book entry including textual and visual elements.

4.8 Summary

In summary, these datasets could be useful for evaluating multimodal graph-learning methods across diverse tasks, scales, and fusion mechanisms, thereby laying a groundwork for MG-LLM. Despite this progress, the current volume of multimodal graph datasets remains significantly smaller than that used for pre-training large language models. This disparity highlights the urgent need for the community to develop more effective methods for data collection and utilization. Furthermore, many existing multimodal graph tasks are discriminative, and future efforts might be devoted to more generative multimodal graph tasks to advance the evaluation and design of MG-LLM.

5 Related work

5.1 Graph representation learning

GNNs such as the graph convolutional network (GCN) [101], graph attention network (GAT) [102], and graph isomorphism network (GIN) [103] have become foundational for learning representations on graph-structured data. These models aggregate information over node neighborhoods to enable tasks like node classification and link prediction. However, recent surveys note several persistent challenges: scaling GNNs to massive graphs [104], ensuring robustness to adversarial or noisy inputs [105], and integrating multimodal data into graph models [5] remain open problems. To address graph-specific modeling in a more general framework, recent work has begun to incorporate LLMs into graph learning. For example, LLaGA [7] reformulates graphs as token sequences for an LLM, GOFA [8] interleaves GNN layers within a pretrained LLM, and GraphRAG [106] augments retrieval-augmented generation with knowledge graphs. These GraphLLM approaches explicitly leverage graph topology in their design. In contrast, general multimodal LLMs (e.g., GPT-4 [107]) are built for text and image inputs and do not directly encode graph structure. In this paper, we propose MG-LLM, a novel model that integrates graph structure with multimodal information modeling to overcome existing limitations in unified representation of multimodal structures and attributes, handling diverse multimodal graph tasks, enabling in-context learning, and supporting multimodal graph reasoning.

5.2 Large language models and multimodal large language models

The field of artificial intelligence has been fundamentally reshaped by the advent of LLMs like the GPT series [49], LLaMA [108], and Qwen [109], built upon the transformer architecture [110] and scaling principles [49]. This success has expanded the frontier to multimodal large language models (MLLMs) such as the seminal open-source LLaVA [38], GPT-4V [111], Qwen-VL [112], InternVL [113], and GLM-4.5V [114]. The trend has further pushed towards Omni-MLLMs that handle arbitrary modalities [9], with prominent examples including the closed-source model GPT-4o [115] and Gemini 2.5 Pro [116] and open-source efforts like One-LLM [117] and CoDi-2 [118]. A parallel development is the rise of LLMs with agentic capabilities for tool use and planning, exemplified by models like Kimi K2 [119] and GLM-4.5 [120]. Despite these significant advancements, a common limitation persists: current models are primarily designed for sequential data and lack a native mechanism for reasoning over explicit, complex relational structures. This highlights a critical gap in handling multimodal graph topology, which our proposed MG-LLM aims to address.

6 Conclusion

In this paper, we aim to address the generalization limits of current multimodal graph neural networks by proposing MG-LLM. We introduce a unified framework, highlighting the inherent characteristics of multimodal graphs, and define five essential capabilities for MG-LLM, ranging from unified data representation to complex reasoning. By discussing challenges, related research, and future directions, this work aims to contribute to the multimodal graph community and accelerate the development of versatile, general-purpose MG-LLM.

Acknowledgements This work was supported by National Key Research and Development Program of China (Grant No. 2023YFF1205001), National Natural Science Foundation of China (Grant No. 62222209), Beijing National Research Center for Information Science and Technology (Grant No. BNR2023TD03006), and Beijing Key Lab of Networked Multimedia.

References

1 Bhattacharyya S, Yang S, Wang J Z. A heterogeneous multimodal graph learning framework for recognizing user emotions in social networks. In: Proceedings of the 12th International Conference on Affective Computing and Intelligent Interaction (ACII), 2024. 327–336
2 Wang X, Wang C, Li L, et al. Fashionklip: enhancing e-commerce image-text retrieval with fashion multi-modal conceptual knowledge graph. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023. 149–158
3 Marini N, Marchesin S, Wodzinski M, et al. Multimodal representations of biomedical knowledge from limited training whole slide images and reports using deep learning. Med Image Anal, 2024, 97: 103303
4 Ye Y, Ren J, Wang S, et al. Construction and application of materials knowledge graph in multidisciplinary materials science via large language model. In: Proceedings of Advances in Neural Information Processing Systems, 2024. 56878–56897
5 Ektefaie Y, Dasoulas G, Noori A, et al. Multimodal learning with graphs. Nat Mach Intell, 2023, 5: 340–350
6 Peng C, He J, Xia F. Learning on multimodal graphs: a survey. ArXiv:2402.05322
7 Chen R, Zhao T, Jaiswal A, et al. LLAGA: large language and graph assistant. ArXiv:2402.08170
8 Kong L, Feng J, Liu H, et al. Gofa: a generative one-for-all model for joint graph language modeling. ArXiv:2407.09709
9 Jiang S, Liang J, Wang J, et al. From specific-MLLMs to OMNI-MLLMs: a survey on MLLMs aligned with multi-modalities. In: Proceedings of Findings of the Association for Computational Linguistics, 2025. 8617–8652
10 Zhu J, Zhou Y, Qian S, et al. Mosaic of modalities: a comprehensive benchmark for multimodal graph learning. ArXiv:2406.16321
11 Wang M, Qi G, Wang H, et al. Richpedia: a comprehensive multi-modal knowledge graph. In: Proceedings of Semantic Technology. Springer International Publishing, 2020. 130–145
12 Hudson D A, Manning C D. GQA: a new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019. 6693–6702
13 Zhao F, Zhang C, Geng B. Deep multimodal data fusion. ACM Comput Surv, 2024, 56: 1–36
14 Wei Y, Wang X, Nie L, et al. MMGCN: multi-modal graph convolution network for personalized recommendation of micro-video. In: Proceedings of the 27th ACM International Conference on Multimedia, 2019. 1437–1445
15 Song X, Zhou F, Frangi A F, et al. Multicenter and multichannel pooling GCN for early AD diagnosis based on dual-modality fused brain network. IEEE Trans Med Imag, 2023, 42: 354–367
16 Zeng Y, Jin Q, Bao T, et al. Multi-modal knowledge hypergraph for diverse image retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2023. 3376–3383
17 Tian Y, Zhang C, Guo Z, et al. Recipe2vec: multi-modal recipe representation learning with graph neural networks. In: Proceedings of the 31st International Joint Conference on Artificial Intelligence, 2022. 3473–3479
18 He X, Wang X. Multimodal graph transformer for multimodal question answering. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2023. 189–200
19 Cai H, Gao Y, Liu M. Graph transformer geometric learning of brain networks using multimodal MR images for brain age estimation. IEEE Trans Med Imag, 2023, 42: 456–466
20 Wang H, Feng S, He T, et al. Can language models solve graph problems in natural language? In: Proceedings of Advances in Neural Information Processing Systems, 2024
21 Guo J, Du L, Liu H, et al. GPT4graph: can large language models understand graph structured data? An empirical evaluation and benchmarking. ArXiv:2305.15066
22 Zhao H, Liu S, Chang M, et al. GIMLET: a unified graph-text model for instruction-based molecule zero-shot learning. In: Proceedings of Advances in Neural Information Processing Systems, 2023. 5850–5887
23 Liu P, Ren Y, Tao J, et al. GIT-Mol: a multi-modal large language model for molecular science with graph, image, and text. Comput Biol Med, 2024, 171: 108073
24 Liu C, Wu B. Evaluating large language models on graphs: performance insights and comparative analysis. ArXiv:2308.11224
25 Fatemi B, Halcrow J, Perozzi B. Talk like a graph: encoding graphs for large language models. ArXiv:2310.04560
26 Hu Y, Zhang Z, Zhao L. Beyond text: a deep dive into large language models' ability on understanding graph data. ArXiv:2310.04944
27 Cai J, Wang X, Li H, et al. Multimodal graph neural architecture search under distribution shifts. AAAI, 2024, 38: 8227–8235
28 Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics, 2018, 34: i457–i466
29 Wu Z, Pan S, Chen F, et al. A comprehensive survey on graph neural networks. IEEE Trans Neural Netw Learn Syst, 2020, 32: 4–24
30 Wang X, Gao T, Zhu Z, et al. KEPLER: a unified model for knowledge embedding and pre-trained language representation. Trans Assoc Comput Linguist, 2021, 9: 176–194
31 Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res, 2014, 15: 1929–1958
32 Mao H, Chen Z, Tang W, et al. Position: graph foundation models are already here. In: Proceedings of the 41st International Conference on Machine Learning, 2024
33 Liu H, Feng J, Kong L, et al. One for all: towards training one graph model for all classification tasks. In: Proceedings of the 12th International Conference on Learning Representations, 2024
34 Galkin M, Yuan X, Mostafa H, et al. Towards foundation models for knowledge graph reasoning. In: Proceedings of the 12th International Conference on Learning Representations, 2024
35 Zhu Y, Shi H, Wang X, et al. GraphCLIP: enhancing transferability in graph foundation models for text-attributed graphs. ArXiv:2410.10329
36 Ye R, Zhang C, Wang R, et al. Language is all a graph needs. In: Proceedings of Findings of the Association for Computational Linguistics, 2024. 1955–1973
37 Cao H, Liu Z, Lu X, et al. Instructmol: multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. In: Proceedings of the 31st International Conference on Computational Linguistics, 2025. 354–379
38 Liu H, Li C, Wu Q, et al. Visual instruction tuning. In: Proceedings of Advances in Neural Information Processing Systems, 2023. 34892–34916
39 Yoon M, Koh J Y, Hooi B, et al. Multimodal graph learning for generative tasks. ArXiv:2310.07478
40 Sun X, Cheng H, Li J, et al. All in one: multi-task prompting for graph neural networks. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2023. 2120–2131

41 Zhang C, Yang Z, He X, et al. Multimodal intelligence: representation learning, information fusion, and applications. IEEE J Sel Top Signal Process, 2020, 14: 478–493
42 Qiu J, Chen Q, Dong Y, et al. GCC: graph contrastive coding for graph neural network pre-training. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020. 1150–1160
43 Sun M, Zhou K, He X, et al. GPPT: graph pre-training and prompt tuning to generalize graph neural networks. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022. 1717–1727
44 Liu Z, Yu X, Fang Y, et al. Graphprompt: unifying pre-training and downstream tasks for graph neural networks. In: Proceedings of the ACM Web Conference, 2023. 417–428
45 Yan Y, Zhang P, Fang Z, et al. Inductive graph alignment prompt: bridging the GAP between graph pre-training and inductive fine-tuning from spectral perspective. In: Proceedings of the ACM on Web Conference, 2024. 4328–4339
46 Yu X, Zhou C, Fang Y, et al. Multigprompt for multi-task pre-training and prompting on graphs. In: Proceedings of the ACM on Web Conference, 2024. 515–526
47 Zhao H, Chen A, Sun X, et al. All in one and one for all: a simple yet effective method towards cross-domain graph pretraining. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024. 4443–4454
48 Bronstein M M, Bruna J, LeCun Y, et al. Geometric deep learning: going beyond Euclidean data. IEEE Signal Process Mag, 2017, 34: 18–42
49 Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. In: Proceedings of Advances in Neural Information Processing Systems, 2020. 1877–1901
50 Sun R, Dai H, Yu A W. Does GNN pretraining help molecular representation? In: Proceedings of Advances in Neural Information Processing Systems, 2022. 12096–12109
51 Huang X, Han K, Yang Y, et al. Can GNN be good adapter for LLMs? In: Proceedings of the ACM Web Conference, 2024. 893–904
52 Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Proceedings of Advances in Neural Information Processing Systems, 2020. 9459–9474
53 Tang J, Yang Y, Wei W, et al. GraphGPT: graph instruction tuning for large language models. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2024. 491–500
54 Yu S, Wang Y, Li R, et al. Graph2text or graph2token: a perspective of large language models for graph learning. ArXiv:2501.01124
55 Lee J, Wang Y, Li J, et al. Multimodal reasoning with multimodal knowledge graph. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 2024. 10767–10782
56 Chai Z, Zhang T, Wu L, et al. GraphLLM: boosting graph reasoning ability of large language model. ArXiv:2310.05845
57 Li R, Jiang H. Graph-to-vision: multi-graph understanding and reasoning using vision-language models. ArXiv:2503.21435
58 Xia F, Li B, Weng Y, et al. Lingyi: medical conversational question answering system based on multi-modal knowledge graphs. ArXiv:2204.09220
59 Huang Y, Shi L, Liu A, et al. Evaluating and enhancing large language models for conversational reasoning on knowledge graphs. ArXiv:2312.11282
60 Liu M, Xu J. NLI4DB: a systematic review of natural language interfaces for databases. ArXiv:2503.02435
61 Dong J, Zhang Q, Zhou H, et al. Modality-aware integration with large language models for knowledge-based visual question answering. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 2024. 2417–2429
62 Huang N, Deshpande Y R, Liu Y, et al. Endowing language models with multimodal knowledge graph representations. ArXiv:2206.13163
63 Chen S, Li B. Multi-modal dynamic graph transformer for visual grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 15534–15543
64 Edge D, Trinh H, Cheng N, et al. From local to global: a graph RAG approach to query-focused summarization. ArXiv:2404.16130
65 Creswell A, Shanahan M, Higgins I. Selection-inference: exploiting large language models for interpretable logical reasoning. In: Proceedings of the 11th International Conference on Learning Representations, 2023
66 Wei J, Hou L, Lampinen A K, et al. Symbol tuning improves in-context learning in language models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2023
67 Talmor A, Yoran O, Catav A, et al. Multimodal{qa}: complex question answering over text, tables and images. In: Proceedings of International Conference on Learning Representations, 2021
68 Zhang N, Li L, Chen X, et al. Multimodal analogical reasoning over knowledge graphs. In: Proceedings of the 11th International Conference on Learning Representations, 2023
69 Guo D, Cao C, Yuan F, et al. Can multimodal large language model think analogically? ArXiv:2411.01307
70 Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. In: Proceedings of International Conference on Machine Learning, 2021. 8748–8763
71 Chen X, Wang C, Xue Y, et al. Unified hallucination detection for multimodal large language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 2024. 3235–3252
72 Liu X, LI R, Ji W, et al. Towards robust multi-modal reasoning via model selection. In: Proceedings of the 12th International Conference on Learning Representations, 2024
73 Choi I, Yun S, Xin J, et al. Multimodal graph-LLM: leveraging graph-enhanced LLMs for multimodal healthcare predictions. In: Proceedings of ICLR Conference, 2025
74 Chen X, Zhang N, Li L, et al. 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, 2022. 904–915
75 Kaiser L, Gomez A N, Shazeer N, et al. One model to learn them all. ArXiv:1706.05137
76 Zaheer M, Guruganesh G, et al. Big bird: transformers for longer sequences. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS), 2020. 17283–17297
77 Jaegle A, Gimeno F, et al. Perceiver: general perception with iterative attention. In: Proceedings of the 38th International Conference on Machine Learning (ICML), 2021. 4651–4664
78 Tay Y, Dehghani M, Bahri D, et al. Efficient transformers: a survey. ArXiv:2009.06732
79 Hamilton W L, Ying R, Leskovec J. Inductive representation learning on large graphs. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS), 2017. 1024–1034
80 Zeng H, Zhou H, Srivastava A, et al. Graphsaint: graph sampling based inductive learning method. In: Proceedings of International Conference on Learning Representations (ICLR), 2020
81 Ying R, You J, Morris C, et al. Hierarchical graph representation learning with differentiable pooling. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS), 2018. 4800–4810
82 Zheng D, Song X, Yang C, et al. Distributed hybrid CPU and GPU training for graph neural networks on billion-scale graphs. ArXiv:2010.05337
83 Bengio Y, Louradour J, Collobert R, et al. Curriculum learning. In: Proceedings of the 26th International Conference on Machine Learning (ICML), 2009. 41–48

84 Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. ArXiv:1503.02531
85 Han S, Mao H, Dally W J. Deep compression: compressing deep neural networks with pruning, trained quantization and Huffman coding. In: Proceedings of the 4th International Conference on Learning Representations (ICLR), 2016
86 Ying C, Cai T, Luo S, et al. Do transformers really perform bad for graph representation? In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS), 2021. 28877–28888
87 Kang Y, Hauswald J, Gao C, et al. Neurosurgeon: collaborative intelligence between the cloud and mobile edge. In: Proceedings of the 22nd International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2017. 615–629
88 Tao Y, Liu W, Chen J, et al. A graph-based multimodal data fusion framework for identifying urban functional zone. Int J Appl Earth Obs GeoInf, 2025, 136: 104353
89 Hoadley K A, Yau C, Hinoue T, et al. Cell-of-origin patterns dominate the molecular classification of 10000 tumors from 33 types of cancer. Cell, 2018, 173: 291–304
90 Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature, 2012, 490: 61–70
91 Saifuddin K M, Ji S, Akbas E. HyperGCL: multi-modal graph contrastive learning via learnable hypergraph views. ArXiv:2502.13277
92 Lee J, Chung C, Lee H, et al. Vista: visual-textual knowledge graph representation learning. In: Proceedings of Findings of the Association for Computational Linguistics, 2023. 7314–7328
93 Wang X, Meng B, Chen H, et al. TIVA-KG: a multimodal knowledge graph with text, image, video and audio. In: Proceedings of the 31st ACM International Conference on Multimedia, 2023. 2391–2399
94 Johnson J, Hariharan B, van der Maaten L, et al. CLEVR: a diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017
95 Alonso I, Salaberria A, Azkune G, et al. Vision-language models struggle to align entities across modalities. ArXiv:2503.03854
96 Mousselly-Sergieh H, Botschen T, Gurevych I, et al. A multimodal translation-based approach for knowledge graph representation learning. In: Proceedings of the 7th Joint Conference on Lexical and Computational Semantics, 2018. 225–234
97 Shirai K, Hashimoto A, Nishimura T, et al. Visual recipe flow: a dataset for learning visual state changes of objects with recipe flows. In: Proceedings of the 29th International Conference on Computational Linguistics, 2022. 3570–3577
98 Lin A, Rao S, Celikyilmaz A, et al. A recipe for creating multimodal aligned datasets for sequential tasks. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020. 4871–4884
99 Fang Y, Jin B, Shen J, et al. GraphGPT-O: synergistic multimodal comprehension and generation on graphs. ArXiv:2502.11925
100 Jin B, Pang Z, Guo B, et al. Instructg2I: synthesizing images from multimodal attributed graphs. ArXiv:2410.07157
101 Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks. In: Proceedings of International Conference on Learning Representations (ICLR), 2017
102 Veličkovi'c P, Cucurull G, Casanova A, et al. Graph attention networks. In: Proceedings of International Conference on Learning Representations (ICLR), 2018
103 Xu K, Hu W, Leskovec J, et al. How powerful are graph neural networks? In: Proceedings of International Conference on Learning Representations (ICLR), 2019
104 Zhang S, Sohrabizadeh A, Wan C, et al. A survey on graph neural network acceleration: algorithms, systems, and customized hardware. ArXiv:2306.14052
105 Dai E, Zhao T, Zhu H, et al. A comprehensive survey on trustworthy graph neural networks: privacy, robustness, fairness, and explainability. Mach Intell Res, 2024, 21: 1011–1061
106 Han H, Wang Y, Shomer H, et al. Retrieval-augmented generation with graphs (graphrag). ArXiv:2501.00309
107 OpenAI. GPT-4 technical report. ArXiv:2303.08774
108 Grattafiori A, Dubey A, Jauhri A, et al. The llama 3 herd of models. ArXiv:2407.21783
109 Yang A, Li A, Yang B, et al. QWEN3 technical report. ArXiv:2505.09388
110 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of Advances in Neural Information Processing Systems, 2017
111 OpenAI. GPT-4V(ision) system card. 2023. https://openai.com/index/gpt-4v-systemcard/
112 Bai S, Chen K, Liu X, et al. Qwen2. 5-vl technical report. ArXiv:2502.13923
113 Wang W, Gao Z, Gu L, et al. InternVL3.5: advancing open-source multimodal models in versatility, reasoning, and efficiency. ArXiv:2508.18265
114 Team V, Hong W, Yu W, et al. GLM-4.5V and GLM-4.1V-thinking: towards versatile multimodal reasoning with scalable reinforcement learning. ArXiv:2507.01006
115 Hurst A, Lerer A, Goucher A P, et al. GPT-4O system card. ArXiv:2410.21276
116 Comanici G, Bieber E, Schaekermann M, et al. Gemini 2.5: pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. ArXiv:2507.06261
117 Han J, Gong K, Zhang Y, et al. OneLLM: one framework to align all modalities with language. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 26584–26595
118 Tang Z, Yang Z, Khademi M, et al. Codi-2: in-context interleaved and interactive any-to-any generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 27425–27434
119 Team K, Bai Y, Bao Y, et al. KIMI K2: open agentic intelligence. ArXiv:2507.20534
120 Zeng A, Lv X, Zheng Q, et al. GLM-4.5: agentic, reasoning, and coding (ARC) foundation models. ArXiv:2508.06471