Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Survey Track

When Vision Meets Graphs: A Survey on Graph Reasoning and Learning

Xinjian Zhao[1], **Wei Pang**[1], **Zhixuan Yu**[1], **Xiangru Jian**[2], **Xiaozhuang Song**[1], **Yaoyao Xu**[1], **Zhongkai Xue**[1], **Dingshuo Chen**[3], **Shu Wu**[3], **Philip Torr**[4] and **Tianshu Yu**[†,1]
[1]The Chinese University of Hong Kong, Shenzhen
[2]University of Waterloo
[3]Institute of Automation, Chinese Academy of Sciences
[4]University of Oxford

Abstract

Graphs are a fundamental data structure underlying many problems in the natural and social sciences. Over the past decade, Graph Neural Networks (GNNs) have dominated graph machine learning, supported by solid theoretical foundations. Yet scientists often understand graph structure through vision: chemists read molecular diagrams and social scientists inspect network visualizations. Despite decades of work on graph visualization, most graph learning pipelines still treat graphs purely as symbolic structures, rarely leveraging the visual form of graphs. We argue that this gap deserves renewed attention in the era of powerful vision and vision-language models. This survey provides a first systematic overview of the emerging area we term *vision meets graphs*, which treats visual depictions of graphs as first-class inputs for reasoning and learning. We organize existing work into three threads. *Vision for Graph Reasoning* studies how models can use visual depictions of graphs to understand structure and carry out multi-step reasoning. *Vision for Graph Learning* explores how visual features can complement or augment graph encoders beyond known limitations of message passing. *Scientific Graphs* examines domains where standardized depiction conventions support both reasoning and learning. Our goal is to clarify what current methods can and cannot do, and to outline a path toward foundation models that perceive and reason about graphs as scientists do.

1 Introduction

Graphs are a fundamental data structure in science and engineering. They describe how atoms form molecules, how people form communities, and how components form circuits. A central question across these domains is how to understand and reason about graph structure.

Human scientists have long relied on vision to answer this question. Chemists read molecular diagrams, social scientists inspect network visualizations, and engineers reason with circuit graphs. These visual conventions encode accumulated knowledge about which spatial arrangements make structural patterns easy to perceive. Notably, tasks that are theoretically hard for GNNs, such as testing biconnectivity [Zhang *et al.*, 2023] or distinguishing certain non-isomorphic graphs [Wang and Zhang, 2024; Horn *et al.*, 2022], often become almost trivial once a graph is visualized appropriately. Decades of graph visualization research have formalized these layout principles [Tamassia, 2013], and scientific disciplines have developed domain-specific depiction standards tightly integrated into expert workflows.

Graph machine learning, however, has largely bypassed this visual tradition. GNNs operate directly on adjacency structure, learning representations through message passing [Khoshraftar and An, 2024]. LLMs serialize graphs into text to tackle multi-step reasoning [Wang *et al.*, 2023; Xu *et al.*, 2025]. Both approaches have driven substantial progress, yet neither sees the visual form of graphs that scientists see. We argue that this gap deserves renewed attention. Recent work suggests that visual representations encode structural information qualitatively different from adjacency-based encodings. Vision encoders pretrained only on natural images can outperform GNNs on tasks requiring global structure understanding [Zhao *et al.*, 2025a]. The rise of vision-language models (VLMs) has prompted researchers to revisit this visual channel [Zhang *et al.*, 2024b]. Work has emerged on benchmarks for reasoning over graph images [Wei *et al.*, 2024; Li *et al.*, 2024; Zhu *et al.*, 2025], methods that incorporate visual features into graph representations [Wei *et al.*, 2025], and scientific applications leveraging domain-specific depiction conventions [Li *et al.*, 2025; Zhao *et al.*, 2025c]. This growing body of research spans multiple communities but has developed largely in isolation.

This survey provides a first systematic overview of this emerging area, which we term *vision meets graphs*. We introduce the Rendering-Perception-Inference (RPI) framework as a diagnostic lens, and organize work into three threads: **Vision for Graph Reasoning** (§3), **Vision for Graph Learning** (§4), and **Scientific Graphs** (§5). Fig. 2 summarizes our taxonomy of the literature, while Fig. 3 illustrates representative input–model–output pipelines for each paradigm. We conclude by discussing open challenges and outlining how this line of research can contribute to broader visions such

[†] Corresponding author: yutianshu@cuhk.edu.cn

as an AI social scientist [Chen *et al.*, 2025] that extracts and analyzes networks from scientific literature, and an AI scientist [Ghafarollahi and Buehler, 2025] that reasons with scientific graphs as human experts do.

2 Background

In this section, we introduce notation and review how graphs are represented for GNNs, for LLMs, and for vision models.

2.1 Graphs and Their Representations

A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consists of a node set $\mathcal{V}$ with $|\mathcal{V}| = n$ and an edge set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. Nodes may carry attributes $\mathbf{X} \in \mathbb{R}^{n \times d}$, and edges may also have features or weights. For machine learning, this discrete structure must be converted into a representation suitable for computation. GNNs operate directly on $\mathcal{G}$, learning node embeddings through neighborhood aggregation:

$$\mathbf{h}_v^{(l+1)} = \phi\left(\mathbf{h}_v^{(l)}, \bigoplus_{u \in \mathcal{N}(v)} \psi(\mathbf{h}_u^{(l)}, \mathbf{h}_v^{(l)}, \mathbf{e}_{uv})\right), \quad (1)$$

where $\mathcal{N}(v)$ denotes the neighbors of node v, $\bigoplus$ is a permutation-invariant aggregation, and ϕ, ψ are learnable functions. This message-passing paradigm respects topology but aggregates locally, which leads to well-known expressiveness limitations [Zhang *et al.*, 2024a].

LLMs take a different approach, serializing $\mathcal{G}$ into a token sequence $\mathcal{S}$ by a serialization function $\mathcal{T} : \mathcal{G} \to \mathcal{S}$ for language-based reasoning. The common formats include edge lists, adjacency descriptions, and structured templates [Wang *et al.*, 2023; Luo *et al.*, 2024; Xu *et al.*, 2025]. A third option, the focus of this survey, is to render $\mathcal{G}$ as an image $\mathbf{I}$ via a visualization function. This visual representation makes graph structure accessible to vision models, but introduces variability through layout and styling choices. We formalize this representation in the next subsection.

2.2 Graph Visualization

Graph visualization maps abstract topology to a perceptible form. We formalize this as a rendering function:

$$\mathcal{R} : (\mathcal{G}, \mathbf{P}, \boldsymbol{\theta}) \to \mathbf{I}, \quad (2)$$

where $\mathbf{P} \in \mathbb{R}^{n \times 2}$ denotes node positions, and $\boldsymbol{\theta}$ denotes visual encoding parameters such as colors, shapes, and line styles. The output $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$ is an image. Different choices of $\mathbf{P}$ and $\boldsymbol{\theta}$ yield visually distinct depictions of the same graph, a property central to robustness probes.

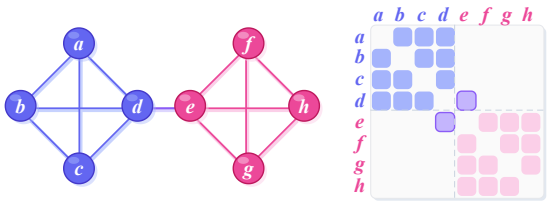

Figure 1: Two common graph visualization paradigms. Left: node-link diagrams. Right: a matrix view of the same graph.

Two common visualization paradigms dominate practice (as shown in Fig. 1). *Node-link diagrams* represent nodes as geometric primitives and edges as connecting lines. Layout algorithms determine $\mathbf{P}$ by optimizing criteria such as edge length uniformity, node separation, and physical energy; common families include force-directed, spectral, and hierarchical methods [Tamassia, 2013]. This paradigm offers intuitive correspondence to network structure but suffers from clutter in dense graphs. *Matrix views* arrange nodes along rows and columns, encoding edges as filled cells. Reordering algorithms permute rows and columns to reveal block structure [Behrisch *et al.*, 2016]. Matrix views eliminate edge crossings but make path tracing less direct.

Scientific domains often adopt conventions that blend these paradigms. Molecular skeletal diagrams follow node-link conventions with chemical semantics encoded through bond types. Protein contact maps use the matrix paradigm, with cells indicating residue proximity. These conventions, refined over decades, constrain $\boldsymbol{\theta}$ and simplify the visual vocabulary that models must interpret.

RPI view: rendering, perception, and inference. In this survey, we use the RPI view as a diagnostic lens. ***Rendering*** maps a discrete graph and visualization choices to images, $\mathbf{I} = \mathcal{R}(\mathcal{G}, \mathbf{P}, \boldsymbol{\theta})$, and determines which geometric and stylistic cues are available. ***Perception*** concerns establishing structural correspondences from pixels, such as identifying nodes, tracing edges, associating labels, and resolving endpoints under crossings, occlusion, and stylistic variation. Perception does not require explicitly reconstructing the full adjacency matrix; many systems rely on partial or implicit structure as long as it supports the downstream objective. ***Inference*** produces task outputs from the perceived structure, ranging from *predictive inference* for classification and regression to *multi-step reasoning* for structural question answering and complex scientific reasoning. Throughout this survey, inference serves as the umbrella term; we reserve reasoning for tasks requiring multi-step computation over graph structure. We use the RPI view to localize where methods intervene and where failures arise, without treating it as a taxonomy.

3 Vision for Graph Reasoning

Graph reasoning is a nontrivial skill central to scientific research, requiring powerful structure perception and multi-step compositional reasoning capability. For foundation models, graph reasoning therefore serves as a particularly revealing testbed. Typical graph reasoning tasks include structural understanding (e.g., counting, connectivity), algorithmic reasoning (e.g., shortest path, topological ordering), combinatorial problems, and multi-step reasoning over social-science graph [Wang *et al.*, 2023; Li *et al.*, 2024; Xu *et al.*, 2025; Wei *et al.*, 2024; Ai *et al.*, 2024].

From a human-centered perspective, graphs are rarely understood as raw adjacency. Instead, we translate graphs into external representations that are easier to perceive or manipulate, such as textual descriptions, symbolic forms (e.g., matrices), and visual depictions. Early LLM-era work naturally began with text, serializing graphs into sequences for prompting or instruction tuning [Wang *et al.*, 2023; Ren *et al.*, 2024].

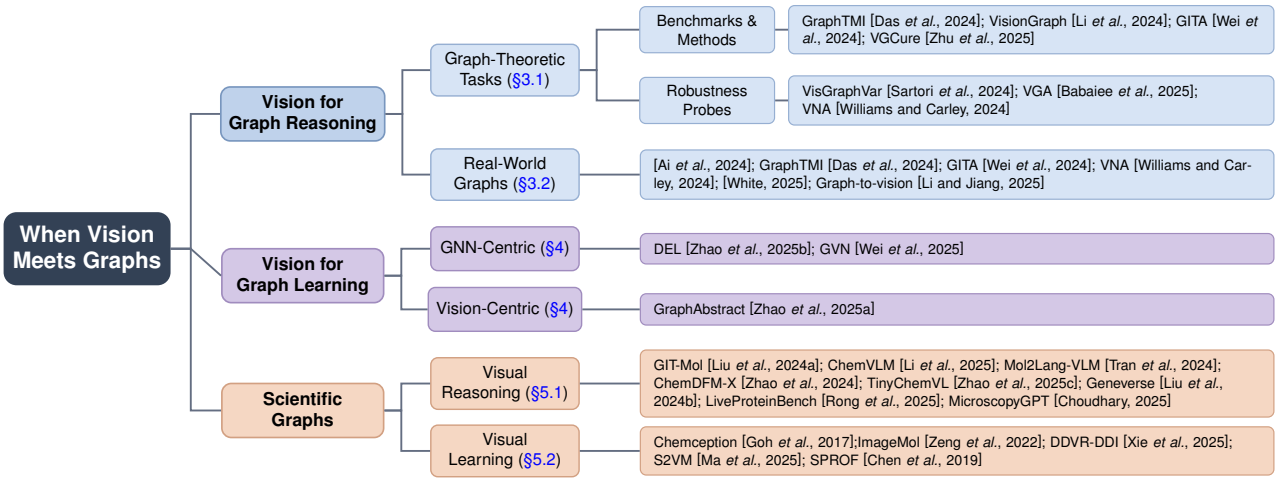

Figure 2: Taxonomy of *When Vision Meets Graphs*. We organize the literature into three threads: graph reasoning, graph learning, and scientific graphs. Representative works are listed for each category.

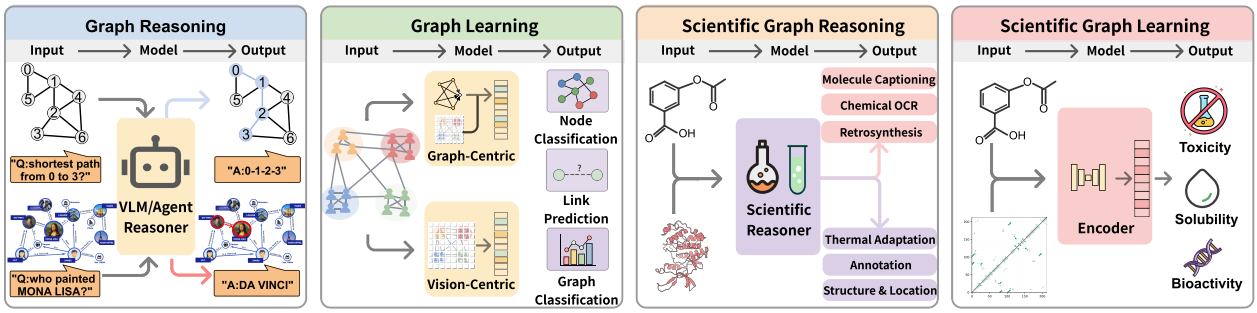

Figure 3: Illustrative pipelines for each paradigm. **Graph Reasoning** (§3): rendered node-link diagrams serve as input to VLMs for structural QA and knowledge-grounded queries. **Graph Learning** (§4): visualizations provide features that either augment GNN representations or feed directly into vision encoders. **Reasoning and Learning for Scientific Graphs** (§5): VLMs interpret molecular depictions for structured outputs, while visual encoders extract representations for property prediction.

This line produced valuable benchmarks and methods, and also revealed some limitations: reasoning performance varies with serialization format [Xu *et al.*, 2025], and sequential encoding does not guarantee invariance to graph symmetries such as node reindexing or edge reordering [Herbst *et al.*, 2025]. Rendered graphs offer a different trade-off. They can expose global structure through spatial cues, yet they introduce variability through layout and style. In this section, we review the efforts that probe whether models can robustly extract usable structure from pixels and then perform reliable reasoning on top of it, first in controlled graph-theoretic settings (§3.1) and then in real-world graphs (§3.2) that additionally require semantic comprehension.

3.1 Graph-Theoretic Tasks

A primary line of work evaluates VLMs on classical graph-theoretic problems rendered as images. Beyond reporting accuracy, these benchmarks are insightful because they expose where performance is gated, whether the model fails to read the graph structure from images, or fails to carry out multi-step reasoning after structure is perceptible.

Early work showed that vision is a viable channel for graph reasoning. GraphTMI [Das *et al.*, 2024] found that multimodal inputs consistently improve performance over text-only baselines. This suggests that visual renderings make certain structural cues easier to perceive than text serialization, even when the reasoning mechanism is otherwise similar. At the same time, the remaining gap indicates that simply adding an image does not remove the need for reliable graph perception and reasoning. VisionGraph [Li *et al.*, 2024] pushed further, introducing a benchmark spanning eight graph-theoretic task types, including connectivity, shortest path, and cycle detection, and it also includes basic topology perception probes for node and edge recognition. It shows that the dominant failure mode in the zero-shot setting is structure perception: models frequently miss nodes or misbind edge endpoints, while supervised fine-tuning substantially improves both node and edge recognition. This indicates that performance is bounded by perceptual grounding, since downstream multi-step reasoning becomes much more reliable once the perceived structure is stable. To reduce error cascades, VisionGraph also proposes Description-Program-Reasoning (DPR), an agent pipeline that first produces an explicit structural description and then executes algorithm-aware code for multi-step reasoning. By externalizing intermediate structure, DPR makes downstream reason-

ing more controllable and reduces error cascades. GITA [Wei *et al.*, 2024] scales evaluation with a larger dataset for vision-language graph reasoning. A central observation is that combining visual and textual modalities outperforms either alone, with each excelling at different task types. This complementarity is informative: visual inputs tend to stabilize global topology cues, while text can help with label semantics and instruction following. GITA also proposes rendering-level augmentation, varying layout, node shape, and edge thickness, which improves generalization by expanding the rendering distribution seen during training.

VGCure [Zhu *et al.*, 2025] probes deeper into complex graph reasoning, broadening evaluation to 22 tasks spanning node, edge, and graph levels. Results confirm that accuracy degrades sharply as queries require more precise relational binding, such as the nested relation query and relation analogy query. This highlights a coupling effect: small perception errors, such as a single missed endpoint, can invalidate an entire multi-step reasoning trace. To address this, VGCure proposes MCDGraph, a structure-aware fine-tuning framework using masked graph infilling, contrastive graph discrimination, and graph description generation. These objectives strengthen structure reading without relying on task-specific labels, which is especially relevant when new tasks are introduced but the perception problem remains the same.

Several additional benchmarks probe robustness and invariance. VisGraphVar [Sartori *et al.*, 2024] systematically varies layouts, styles, and imperfections such as node overlap, showing that accuracy is highly sensitive to such perturbations. This indicates that many methods implicitly assume a narrow rendering distribution, and improvements may partly reflect overfitting to a particular visualization style. Visual Graph Arena (VGA) [Babaiee *et al.*, 2025] tests invariance by asking whether models can recognize isomorphic graphs across different visualizations. The gap between human and model performance suggests that learning layout-invariant structure from pixels remains challenging. The VNA Benchmark [Williams and Carley, 2024] evaluates basic network analysis tasks, including maximum degree identification, triad balance, and component counting. These studies suggest that current models still struggle with reasoning. However, all three benchmarks use relatively sparse and simple visual styling. When visualization quality is limited, models may fail at perception before reasoning even begins.

3.2 Real-World Graphs

Beyond graph-theoretic reasoning (§3.1), real-world graphs couple domain semantics with graph visualization. We discuss two families: *social-science graphs* and *general-purpose visual graphs* such as flowcharts.

Social-science graphs. Social science relies heavily on graph visualizations: citation and collaboration networks reveal the structure of scientific communities, and system maps such as causal loop diagrams encode causal hypotheses in policy research. Social scientists extract insights from these visualizations. However, some graphs exist as structured data but are communicated through visualizations, while others appear only as figures in papers and reports and must be extracted from images before analysis. For the first case, several benchmarks in §3.1 evaluate whether VLMs can analyze social-science graphs as researchers do. GITA evaluates on real-world social-science graphs, including PolBlogs graph and citation graphs; GraphTMI tests on citation networks; VNA uses synthetic graphs to test network analysis primitives commonly used in social science, including triad balance and component counting. Notably, GITA reports large improvements from visual inputs on PolBlogs, suggesting that vision may be particularly useful for social-science graph analysis. For the second case, [White, 2025] evaluates whether VLMs can extract node and edge lists from system-map images (e.g., Causal loop diagrams, fuzzy cognitive maps) used in policy research, finding that node identification is relatively reliable but edge recovery (especially direction and polarity) remains a bottleneck. Social-science graph reasoning has received less attention than graph-theoretic tasks, but we believe it deserves more. There are many potential applications; for example, an AI social scientist that automates science-of-science research would need to analyze citation and collaboration networks [Chen *et al.*, 2025].

General-purpose visual graphs. Flowcharts, mind maps, route maps, and organizational charts convey rich semantic and visual cues through their rendering conventions, but their topologies are typically simple. This shifts the bottleneck toward perception: Optical Character Recognition (OCR), symbol recognition, and binding text to the correct nodes and connectors. [Ai *et al.*, 2024] constructs instruction-following data from web-crawled images across multiple graph types, showing that perception-based supervision improves downstream reasoning. Graph-to-Vision [Li and Jiang, 2025] evaluates multi-graph joint reasoning and separates parsing quality from reasoning consistency, helping localize whether failures arise from visual graph reading, semantic interpretation, or subsequent inference.

3.3 Challenges and Future Directions

Rendering: controllable rendering distributions. Rendered graphs offer structural cues that are difficult to preserve under text serialization, but they also introduce a distribution over layouts and styles. Many benchmarks adopt simplified visualization schemes to isolate topology, which is useful for diagnosis, yet it narrows the rendering distribution and can limit how far conclusions generalize to tool-generated or human-facing diagrams. Moreover, minimal renderings can inadvertently make perception harder by weakening endpoint cues, reducing node edge contrast, or removing textual anchors, so failures may mix up perceptual difficulty with downstream reasoning limits. A promising direction is to build benchmarks with controllable rendering distributions spanning both perception-friendly and stress-test regimes, and reporting rendering parameters to allow comparison across studies. Visualization research also offers useful tools here, including layout objectives, transparency, and emphasis techniques that reduce ambiguity in dense regions [Tamassia, 2013; Meng *et al.*, 2025].

Perception: robust structure reading under ambiguity. Many failures labeled as reasoning errors originate earlier, during structure reading. Models must localize nodes, follow edges, resolve endpoints at crossings, and integrate text

labels, often under overlap and stylistic variation. These challenges mirror spatial reasoning problems in natural images, where models must establish object correspondences under occlusion and depth ambiguity. Progress in that domain shows that training on large-scale synthetic spatial VQA tasks can substantially improve spatial understanding [Chen *et al.*, 2024]; designing analogous synthetic tasks for graph-specific ambiguities, such as endpoint disambiguation and crossing resolution, is a promising direction.

Inference: agentic pipelines and thinking with images. Beyond perception, the key challenge is to make multi-step inference controllable and checkable. Agent-based pipelines that separate structure extraction from algorithm execution offer one path forward [Li *et al.*, 2024], with broader work on visual programming showing how code generation can orchestrate perception modules [Surís *et al.*, 2023]. When reasoning spans multiple graphs from heterogeneous sources (e.g., system maps alongside social networks), selectively routing each graph to a suitable perception encoder [Zhang *et al.*, 2025] may help. A complementary direction, explored in recent work on visual reasoning [Hu *et al.*, 2024; Qin *et al.*, 2025; Su *et al.*, 2025], is to treat images as manipulable intermediate states rather than static inputs. For graphs, this suggests operations like highlighting visited nodes, zooming into ambiguous regions, and local re-rendering to support multi-step reasoning while keeping each step inspectable. These capabilities point toward an ***AI social scientist*** that uses visual intermediate states to extract and verify network structure from figures, and integrates with external analysis tools.

4 Vision for Graph Learning

In the previous section, we focused on VLM for graph reasoning, where a vision encoder and a language model work together. For many core graph tasks, however, practitioners still rely on representation learning with numeric outputs rather than text generation. Problems such as link prediction and graph classification are usually solved by encoders followed by simple predictors. This raises a basic question: can a vision encoder learn useful graph representations from rendered images? This section reviews early work that treats graph visualizations as inputs for representation learning.

GNNs have been known to have expressiveness limitations [Zhang *et al.*, 2024a]. For example, they are bounded by the Weisfeiler–Lehman test [Wang and Zhang, 2024] and cannot detect properties such as graph biconnectivity that simple algorithms can compute [Zhang *et al.*, 2023]. These limitations have often been visualized in past work to present to readers: "Look, these two graphs are not the same, but our advanced GNNs cannot distinguish them." This observation suggests that graph visualizations encode certain information that message passing fails to capture, motivating the use of visual representations for graph learning. Here, the dominant objective is predictive inference, and perception is often implicit and optimized end-to-end. A model may never emit an explicit intermediate graph, yet it still must establish correspondences between pixels and the underlying structural regularities that support prediction. This is why rendering choices, such as the layout distribution and visual encoding, can matter even when the training objective is standard classification or regression.

Vision-Enhanced GNNs. Rather than replacing GNNs entirely, one line of work incorporates visual priors from graph visualizations to compensate for GNN limitations. GVN [Wei *et al.*, 2025] renders subgraph visualizations centered on target edges and extracts features using a pretrained ResNet encoder. Analysis shows these visual features help distinguish links that 1-WL-bounded GNNs cannot. DEL [Zhao *et al.*, 2025b] takes a different approach. It models graph visualizations as samples from a Boltzmann distribution, generates multiple visualizations, and extracts hand-crafted features from them to feed into GNNs. Hand-crafted features can be seen as the simplest form of visual encoder, which reduces ambiguity but discards much information. Even so, these features provide theoretical advantages, helping distinguish non-isomorphic graphs that 1-WL-bounded GNNs cannot. Both methods suggest that graph visualizations carry structural signals that complement message passing, though neither offers a purely visual solution.

Vision Encoders for Graph-Level Tasks. A more direct approach uses vision encoders to process graph visualizations as standalone inputs. Recent work [Zhao *et al.*, 2025a] shows that on graph-level tasks requiring global structure understanding, vision encoders pretrained on natural images can outperform GNNs. This suggests that spatial arrangements from visualization algorithms encode patterns that local message passing cannot capture. It also shows that vision encoders can serve as graph encoders, which could enable their use as backbones in graph-specialized VLMs. However, current evidence remains limited to graph-level tasks. Whether vision-based approaches can match or exceed GNNs on node-level and edge-level tasks, where local structure matters more, remains an open question. Furthermore, all existing methods rely on vision encoders pretrained on natural images, which lack inductive biases for graph-specific patterns such as centrality, motifs, or community structure.

These methods have been evaluated on datasets spanning citation networks, social networks, and biological graphs, demonstrating applicability across both social science and natural science domains.

4.1 Challenges and Future Directions

Rendering: theoretical understanding of what survives into pixels. Graph visualization is a structured but lossy channel. Different graphs can map to visually similar images, and the same graph can look different under different layouts and styles. For graph learning, an important theoretical question is which structural properties are preserved, amplified, or erased by a rendering distribution, and when those properties are learnable by standard vision encoders. Progress here would turn rendering from an implementation detail into a design variable, enabling principled choices of layouts, multi-view renderings, or task-aware depiction rules that improve expressiveness.

Perception: graph-native visual pretraining and invariance. Existing results are largely based on vision encoders pretrained on natural images, yet graph visualizations are symbolic and rule-governed. A promising direction is graph-

native visual pretraining that explicitly targets invariance central to graph learning, such as robustness to node reindexing, layout changes, and cosmetic style shifts, while learning sensitivity to true structural differences such as motifs and global connectivity patterns. Well-chosen pretraining objectives can also clarify what the encoder is actually reading, reducing the risk that gains come from dataset-specific rendering artifacts.

Inference: moving beyond graph-level prediction and isolating when vision helps. Evidence for purely visual encoders is powerful on graph-level tasks that reward global pattern recognition; it remains unclear whether similar benefits extend to node-level and edge-level prediction, where locality, semantics, and calibration matter. The next step is to identify which task families and graph characteristics benefit most from visual cues, and whether hybrids provide new information or mainly act as regularizers. Such studies would also inform evaluation design, separating improvements due to better structure reading from improvements due to downstream prediction.

5 Reasoning and Learning for Scientific Graphs

Scientific domains present a distinctive setting for vision-based graph reasoning and learning. Molecular diagrams, protein contact maps, and reaction schemes all follow standardized conventions refined over decades of scientific practice. These conventions constrain both layout and visual encoding (the $\mathbf{P}$ and $\boldsymbol{\theta}$ of §2.2), resulting in near-canonical structure-to-pixel mappings that encode domain knowledge while dramatically reducing perceptual ambiguity. The same visual vocabulary that enables expert reasoning thus becomes accessible to visual models, opening a path toward foundation models that interpret scientific structure as scientists do.

Two domains have received the most attention: molecular structures and proteins. Molecular depictions follow node-link diagrams with atoms as nodes and bonds as edges, arranged according to well-established chemical drawing rules that chemists use daily in papers, textbooks, and laboratory notebooks. Protein structures, by contrast, are often represented either as rendered 3D structure images or as contact and distance maps, which are matrix views where entry (i, j) encodes spatial proximity between residues i and j. These two domains thus exemplify both major visualization paradigms introduced in §2.2, providing natural testbeds for vision-based methods. This section examines how these conventions support both reasoning over rendered structures (§5.1) and representation learning from visual inputs (§5.2).

5.1 Vision for Scientific Graph Reasoning

VLMs and MLLMs have opened new possibilities for scientific graph reasoning. These models can directly interpret visual depictions of scientific graphs and generate natural language responses, from describing properties to extracting symbolic representations. Early work focuses on perception and question answering, but the broader goal is to enable reasoning over scientific structures, such as predicting properties or understanding reactions. This subsection surveys such reasoning approaches.

The molecular domain has seen the most activity, with research spanning tasks from focused applications to ambitious unifying frameworks. A single molecule admits three coupled representations: a molecular graph $\mathcal{G}$, a symbolic string form such as SMILES, and a 2D depiction image $\mathbf{I}$ rendered under long-standing conventions. The ability to map between these representations is the perceptual foundation for scientific graph reasoning.

Early efforts primarily target perception: aligning rendered graphs with symbolic forms. GIT-Mol aligns image, graph, and text modalities for molecule captioning and image-to-SMILES conversion [Liu *et al.*, 2024a], while Mol2Lang-VLM focuses on generating natural language descriptions from molecular images [Tran *et al.*, 2024]. More recent work moves beyond perception to multi-step reasoning. ChemVLM [Li *et al.*, 2025] trains on a bilingual multimodal dataset of molecular structures, reactions, and chemistry examinations, with evaluation suites spanning structure recognition, multi-step reasoning, and property prediction. Strong performance across these tasks, substantially outperforming general-purpose VLMs, demonstrates that domain-focused training can yield models proficient in both perception and chemical reasoning. ChemDFM-X [Zhao *et al.*, 2024] extends the notion of rendering beyond images to include five chemical modalities (2D graphs, 3D conformations, images, mass spectra, and infrared spectra), and demonstrates cross-modal collaboration where spectra provide structural hints while symbolic representations constrain invalid options. These results lend support to the premise that visual conventions encoding chemical knowledge are learnable, offering concrete progress toward foundation models that interpret scientific structure as scientists do.

As work in this area expands, the efficiency and evaluation have drawn more attention. TinyChemVL observes that molecular images contain large uninformative backgrounds and proposes visual token reduction to improve efficiency, alongside reaction-level benchmarks that test more complex reasoning [Zhao *et al.*, 2025c]. MolVision provides a systematic evaluation across multiple VLMs, finding that visual information alone is insufficient: competitive performance requires multimodal fusion with textual context [Adak *et al.*, 2025].

Protein structure reasoning presents a contrasting picture. While molecular VLMs and MLLMs have grown rapidly, only a handful of studies have explored whether similar approaches transfer to proteins. Recent efforts include Geneverse, which finetunes LLaVA on protein structure images rendered from AlphaFold predictions for protein function inference [Liu *et al.*, 2024b]. LiveProteinBench evaluates whether adding multi-view structural images improves LLM performance on protein property and function prediction, and reports that structural inputs often fail to help and can even degrade performance relative to sequence-only baselines [Rong *et al.*, 2025]. This negative result is informative rather than discouraging. One plausible explanation is a perception and alignment gap: if the model is not trained to reliably read structural cues from these renderings, adding images can act as nuisance variation and burden cross-modal fusion, especially when the language or sequence channel already pro-

vides strong signals. A second possibility is a rendering adequacy gap: common 2D views may not preserve the functionally relevant 3D information required by the inference target. Disentangling these failure modes remains an open challenge and motivates depiction-aware training and evaluation that explicitly separates depiction reading, fusion quality, and downstream scientific inference.

Visual reasoning also extends to materials science. Scanning transmission electron microscopy (STEM) images can be viewed as physically grounded renderings whose appearance is governed by an instrument-specific image formation process. MicroscopyGPT [Choudhary, 2025] shows that VLMs can predict crystallographic descriptors from such renderings, including lattice parameters, space-group information, and atomic coordinates, providing early evidence of perceptually grounded reasoning and suggesting that more complex multi-step inference may follow.

5.2 Vision for Scientific Graph Learning

Besides reasoning, another line of work learns visual representations of scientific graphs for prediction tasks. Rather than generating language, these methods train encoders to produce embeddings that capture structural and chemical information, which are then fed to task-specific predictors.

For molecular graphs, early work applies visual encodes directly to 2D molecular images for property prediction [Goh *et al.*, 2017]. ImageMol [Zeng *et al.*, 2022] improves substantially by pretraining on millions of unlabeled molecular images with chemically motivated objectives such as structure clustering and mask-based contrastive learning. Beyond single-molecule prediction, visual representations naturally extend to relational tasks involving multiple molecules. For drug-drug interaction prediction, DDVR-DDI fuses two drug depictions into a single image to capture potential interaction interfaces [Xie *et al.*, 2025], while S^2VM blends local visual fragments from drug pairs to learn interaction-relevant embeddings [Ma *et al.*, 2025].

For proteins, contact and distance maps encode pairwise residue proximities as matrices, providing a complementary view that avoids the edge crossings and node occlusion common in node-link diagrams. CNNs have been applied to such matrix representations for sequence profile prediction [Chen *et al.*, 2019] and residue-level architecture segmentation [Eguchi and Huang, 2020]. Compared to molecules, however, visual representation learning for proteins remains underexplored. This asymmetry may help explain the reasoning failures reported in §5.1: without first establishing that visual encoders can reliably extract structural cues from protein renderings, multimodal reasoning has no solid perceptual foundation to build on. Protein-specific visual pretraining, as ImageMol provided for molecules, may be necessary before VLMs and MLLMs can effectively understand and reason over protein structures.

5.3 Challenges and Future Directions

Rendering: scientific graph visualization as foundational infrastructure. Scientific graphs are produced by established tools (e.g., RDKit, PyMOL) following conventions that evolved primarily for human readers. Yet the design and iteration of these tools has rarely been treated as a research problem in its own right. As rendering is the first stage of the pipeline, limitations here propagate downstream: a poorly rendered structure leaves perception and inference with no solid ground to build on. Systematic benchmarks that evaluate how rendering choices affect model comprehension would help identify bottlenecks and guide tool improvements. Such infrastructure is a prerequisite for any AI scientist who must reliably read scientific figures at scale.

Perception: domain-specific visual pretraining. §3 and §4 identified perception as a recurring bottleneck: models must reliably read graph structure before reasoning or learning can proceed. Scientific graphs offer a partial advantage, as standardized depiction conventions reduce layout and style variation. Yet domain-specific visual pretraining remains essential. Molecules benefit from large-scale efforts like ImageMol, while proteins lack an equivalent foundation, which may explain the negative results in §5.1. Many other scientific domains with standardized conventions, such as metabolic pathways and gene regulatory networks, remain unexplored and would benefit from similar investments.

Inference: from exam-style QA to scientific workflows. Current evaluations often emphasize exam-style question answering, but practical scientific workflows demand more (e.g., synthesis planning, pathway analysis, or materials property prediction): models must produce extracted structures that respect domain constraints. These constraints are graph-structured (e.g., valence rules, regulatory directionality, or crystal symmetry) and should be checked by external tools (e.g., rule checkers or simulators), not just language reasoning. Achieving the broader vision of an AI scientist calls for benchmarks that require models to output explicit graph structures, invoke domain tools for validation, and trace errors back to rendering or perception when results fail.

6 Conclusion and Outlook

This survey has offered the first unified treatment of *vision meets graphs*, introducing a diagnostic framework organized around rendering, perception, and inference. Across all three threads, perception emerges as a recurring bottleneck: models must reliably extract graph structure from pixels before reasoning or learning can proceed. Looking ahead, we see opportunities in graph-native visual pretraining, active visual reasoning where models manipulate rendered graphs during inference, and tighter integration with domain tools for structure verification. Scientific domains offer promising testbeds, as standardized depiction conventions reduce perceptual ambiguity while practical workflows provide meaningful evaluation. Beyond social and natural science, engineering domains with established visual conventions remain open for exploration. These directions point toward broader visions such as an AI social scientist and an AI scientist that read and reason over graphs as humans do.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (grant No. 92470113).

References

[Adak *et al.*, 2025] Deepan Adak, Yogesh Singh Rawat, and Shruti Vyas. Molvision: Molecular property prediction with vision language models. *arXiv preprint arXiv:2507.03283*, 2025.

[Ai *et al.*, 2024] Qihang Ai, Jiafan Li, Jincheng Dai, Jianwu Zhou, Lemao Liu, Haiyun Jiang, and Shuming Shi. Advancement in graph understanding: A multimodal benchmark and fine-tuning of vision-language models. In *ACL*, pages 7485–7501, 2024.

[Babaiee *et al.*, 2025] Zahra Babaiee, Peyman M Kiasari, Daniela Rus, and Radu Grosu. Visual graph arena: Evaluating visual conceptualization of vision and multimodal large language models. *ICML*, 2025.

[Behrisch *et al.*, 2016] Michael Behrisch, Benjamin Bach, Nathalie Henry Riche, Tobias Schreck, and Jean-Daniel Fekete. Matrix reordering methods for table and network visualization. In *Computer Graphics Forum*, 2016.

[Chen *et al.*, 2019] Sheng Chen, Zhe Sun, Lihua Lin, Zifeng Liu, Xun Liu, Yutian Chong, Yutong Lu, Huiying Zhao, and Yuedong Yang. To improve protein sequence profile prediction through image captioning on pairwise residue distance map. *JCIM*, 60(1):391–399, 2019.

[Chen *et al.*, 2024] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *CVPR*, 2024.

[Chen *et al.*, 2025] Renqi Chen, Haoyang Su, Shixiang Tang, Zhenfei Yin, Qi Wu, Hui Li, Ye Sun, Nanqing Dong, Wanli Ouyang, and Philip Torr. Ai-driven automation can become the foundation of next-era science of science research. *arXiv preprint arXiv:2505.12039*, 2025.

[Choudhary, 2025] Kamal Choudhary. Microscopygpt: Generating atomic-structure captions from microscopy images of 2d materials with vision-language transformers. *The Journal of Physical Chemistry Letters*, 2025.

[Das *et al.*, 2024] Debarati Das, Ishaan Gupta, Jaideep Srivastava, and Dongyeop Kang. Which modality should i use-text, motif, or image?: Understanding graphs with large language models. In *Findings of the NAACL*, 2024.

[Eguchi and Huang, 2020] Raphael R Eguchi and Po-Ssu Huang. Multi-scale structural analysis of proteins by deep semantic segmentation. *Bioinformatics*, 2020.

[Ghafarollahi and Buehler, 2025] Alireza Ghafarollahi and Markus J Buehler. Sciagents: automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*, 37(22):2413523, 2025.

[Goh *et al.*, 2017] Garrett B Goh, Charles Siegel, Abhinav Vishnu, Nathan O Hodas, and Nathan Baker. Chemception: a deep neural network with minimal chemistry knowledge matches the performance of expert-developed qsar/qspr models. *arXiv preprint arXiv:1706.06689*, 2017.

[Herbst *et al.*, 2025] Daniel Herbst, Lea Karbevska, Divyanshu Kumar, Akanksha Ahuja, Fatemeh Gholamzadeh Nasrabadi, and Fabrizio Frasca. Lost in serialization: Invariance and generalization of llm graph reasoners. *arXiv preprint arXiv:2511.10234*, 2025.

[Horn *et al.*, 2022] Max Horn, Edward De Brouwer, Michael Moor, Yves Moreau, Bastian Rieck, and Karsten Borgwardt. Topological graph neural networks. *ICLR*, 2022.

[Hu *et al.*, 2024] Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *NeurIPS*, 2024.

[Khoshraftar and An, 2024] Shima Khoshraftar and Aijun An. A survey on graph representation learning methods. *ACM Transactions on Intelligent Systems and Technology*, 15(1):1–55, 2024.

[Li and Jiang, 2025] Ruizhou Li and Haiyun Jiang. Graph-to-vision: Multi-graph understanding and reasoning using vision-language models. *arXiv preprint arXiv:2503.21435*, 2025.

[Li *et al.*, 2024] Yunxin Li, Baotian Hu, Haoyuan Shi, Wei Wang, Longyue Wang, and Min Zhang. Visiongraph: Leveraging large multimodal models for graph theory problems in visual context. *ICML*, 2024.

[Li *et al.*, 2025] Junxian Li, Di Zhang, Xunzhi Wang, Zeying Hao, Jingdi Lei, Qian Tan, Cai Zhou, Wei Liu, Yaotian Yang, Xinrui Xiong, et al. Chemvlm: Exploring the power of multimodal large language models in chemistry area. In *AAAI*, 2025.

[Liu *et al.*, 2024a] Pengfei Liu, Yiming Ren, Jun Tao, and Zhixiang Ren. Git-mol: A multi-modal large language model for molecular science with graph, image, and text. *Computers in biology and medicine*, 171:108073, 2024.

[Liu *et al.*, 2024b] Tianyu Liu, Yijia Xiao, Xiao Luo, Hua Xu, Wenjin Zheng, and Hongyu Zhao. Geneverse: A collection of open-source multimodal large language models for genomic and proteomic research. In *Findings of the EMNLP*, pages 4819–4836, 2024.

[Luo *et al.*, 2024] Zihan Luo, Xiran Song, Hong Huang, Jianxun Lian, Chenhao Zhang, Jinqi Jiang, Xing Xie, and Hai Jin. Graphinstruct: Empowering large language models with graph understanding and reasoning capability. *arXiv preprint arXiv:2403.04483*, 2024.

[Ma *et al.*, 2025] Tengfei Ma, Kun Chen, Yongsheng Zang, Yujie Chen, Xuanbai Ren, Bosheng Song, hongxin xiang, Yiping Liu, and xiangxiang Zeng. Self-supervised blending structural context of visual molecules for robust drug interaction prediction. In *NeurIPS*, 2025.

[Meng *et al.*, 2025] Zhiyuan Meng, Yunpeng Yang, Qiong Zeng, Kecheng Lu, Lin Lu, Changhe Tu, Fumeng Yang, and Yunhai Wang. Seeing through the overlap: The impact of color and opacity on depth order perception in visualization. In *CHI*, pages 1–14, 2025.

[Qin *et al.*, 2025] Yiming Qin, Bomin Wei, Jiaxin Ge, Konstantinos Kallidromitis, Stephanie Fu, Trevor Darrell, and Xudong Wang. Chain-of-visual-thought: Teaching vlms to

see and think better with continuous visual tokens. *arXiv preprint arXiv:2511.19418*, 2025.

[Ren *et al.*, 2024] Xubin Ren, Jiabin Tang, Dawei Yin, Nitesh Chawla, and Chao Huang. A survey of large language models for graphs. In *KDD*, 2024.

[Rong *et al.*, 2025] Dingyi Rong, Zijian Chen, Qi Jia, Kaiwei Zhang, Haotian Lu, Guangtao Zhai, and Ning Liu. Liveproteinbench: A contamination-free benchmark for assessing models' specialized capabilities in protein science. *arXiv preprint arXiv:2512.22257*, 2025.

[Sartori *et al.*, 2024] Camilo Chacón Sartori, Christian Blum, and Filippo Bistaffa. Visgraphvar: A benchmark generator for assessing variability in graph analysis using large vision-language models. *arXiv preprint arXiv:2411.14832*, 2024.

[Su *et al.*, 2025] Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*, 2025.

[Surís *et al.*, 2023] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *CVPR*, 2023.

[Tamassia, 2013] Roberto Tamassia. *Handbook of graph drawing and visualization*. CRC press, 2013.

[Tran *et al.*, 2024] Duong Tran, Nhat Truong Pham, Nguyen Nguyen, and Balachandran Manavalan. Mol2lang-vlm: Vision-and text-guided generative pre-trained language models for advancing molecule captioning through multimodal fusion. In *Proceedings of the 1st Workshop on Language+ Molecules (L+ M 2024)*, 2024.

[Wang and Zhang, 2024] Yanbo Wang and Muhan Zhang. An empirical study of realized gnn expressiveness. *ICML*, 2024.

[Wang *et al.*, 2023] Heng Wang, Shangbin Feng, Tianxing He, Zhaoxuan Tan, Xiaochuang Han, and Yulia Tsvetkov. Can language models solve graph problems in natural language? *NeurIPS*, 2023.

[Wei *et al.*, 2024] Yanbin Wei, Shuai Fu, Weisen Jiang, Zejian Zhang, Zhixiong Zeng, Qi Wu, James Kwok, and Yu Zhang. Gita: Graph to visual and textual integration for vision-language graph reasoning. *NeurIPS*, 2024.

[Wei *et al.*, 2025] Yanbin Wei, Xuehao Wang, Zhan Zhuang, Yang Chen, Shuhao Chen, Yulong Zhang, Yu Zhang, and James Kwok. Open your eyes: Vision enhances message passing neural networks in link prediction. *ICML*, 2025.

[White, 2025] Jordan White. Using vision-language models to extract network data from images of system maps. Technical report, Institute for New Economic Thinking at the Oxford Martin School, University, 2025.

[Williams and Carley, 2024] Evan M Williams and Kathleen M Carley. Multimodal llms struggle with basic visual network analysis: a vna benchmark. *arXiv preprint arXiv:2405.06634*, 2024.

[Xie *et al.*, 2025] Lingxuan Xie, Tengfei Ma, Yuqin He, Yiping Liu, and Xiangxiang Zeng. Predicting drug–drug interaction via dual-drug visual representation. *JCIM*, 2025.

[Xu *et al.*, 2025] Hao Xu, Xiangru Jian, Xinjian Zhao, Wei Pang, Chao Zhang, Suyuchen Wang, Qixin Zhang, Zhengyuan Dong, Joao Monteiro, Bang Liu, et al. Graphomni: A comprehensive and extendable benchmark framework for large language models on graph-theoretic tasks. *arXiv preprint arXiv:2504.12764*, 2025.

[Zeng *et al.*, 2022] Xiangxiang Zeng, Hongxin Xiang, Linhui Yu, Jianmin Wang, Kenli Li, Ruth Nussinov, and Feixiong Cheng. Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework. *Nature Machine Intelligence*, 4(11):1004–1016, 2022.

[Zhang *et al.*, 2023] Bohang Zhang, Shengjie Luo, Liwei Wang, and Di He. Rethinking the expressive power of gnns via graph biconnectivity. *ICLR*, 2023.

[Zhang *et al.*, 2024a] Bingxu Zhang, Changjun Fan, Shixuan Liu, Kuihua Huang, Xiang Zhao, Jincai Huang, and Zhong Liu. The expressive power of graph neural networks: A survey. *IEEE TKDE*, 2024.

[Zhang *et al.*, 2024b] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE TPAMI*, 2024.

[Zhang *et al.*, 2025] Tianyu Zhang, Suyuchen Wang, Chao Wang, Juan Rodriguez, Ahmed Masry, Xiangru Jian, Yoshua Bengio, and Perouz Taslakian. Scope: Selective cross-modal orchestration of visual perception experts. *arXiv preprint arXiv:2510.12974*, 2025.

[Zhao *et al.*, 2024] Zihan Zhao, Bo Chen, Jingpiao Li, Lu Chen, Liyang Wen, Pengyu Wang, Zichen Zhu, Danyang Zhang, Yansi Li, Zhongyang Dai, et al. Chemdfm-x: towards large multimodal model for chemistry. *Science China Information Sciences*, 2024.

[Zhao *et al.*, 2025a] Xinjian Zhao, Wei Pang, Zhongkai Xue, Xiangru Jian, Lei Zhang, Yaoyao Xu, Xiaozhuang Song, Shu Wu, and Tianshu Yu. The underappreciated power of vision models for graph structural understanding. *arXiv preprint arXiv:2510.24788*, 2025.

[Zhao *et al.*, 2025b] Xinjian Zhao, Chaolong Ying, Yaoyao Xu, and Tianshu Yu. Graph learning with distributional edge layouts. In *KDD*, pages 2055–2066, 2025.

[Zhao *et al.*, 2025c] Xuanle Zhao, Shuxin Zeng, Xinyuan Cai, Xiang Cheng, Duzhen Zhang, Xiuyi Chen, and Bo Xu. Tinychemvl: Advancing chemical vision-language models via efficient visual token reduction and complex reaction tasks. *arXiv preprint arXiv:2511.06283*, 2025.

[Zhu *et al.*, 2025] Yingjie Zhu, Xuefeng Bai, Kehai Chen, Yang Xiang, Jun Yu, and Min Zhang. Benchmarking and improving large vision-language models for fundamental visual graph understanding and reasoning. In *ACL*, 2025.