Mario: Multimodal Graph Reasoning with Large Language Models

Supplementary Material

1. Appendix

1.1. Dataset Details

Statistics and Introduction. The detailed statistics of the datasets we used are shown in Table 1. In these datasets, nodes represent individual products or posts, and edges denote relationships such as co-purchase or co-comment interactions between products or posts. Each node is assigned a label corresponding to its category. Every node is enriched with two modalities: textual attributes, such as product titles, descriptions, or post content, and visual attributes extracted from associated product or post images. Unlike conventional image–text benchmarks where captions are written to explicitly describe the visual content, here the two modalities are only loosely coupled and often contain complementary or even disjoint information. For example, a clothing item may have a textual description that focuses on material and fit (e.g., "soft cotton hoodie with relaxed, oversized silhouette, ideal for fall weather") while its image emphasizes color, style, and brand logos that are never mentioned in the text. Conversely, the text may include attributes such as size range, discount information, or user-targeted marketing slogans that are not visually observable.
Data Splits. For the node classification task, we adopt a standardized 6:2:2 split into training, validation, and testing for Mario and all the baselines. For the link prediction task, the training, validation, and test sets consist of 3,000, 2,000, and 1,000 edges, respectively, for training and evaluation.

Table 1. Dataset statistics across multiple MMG datasets.

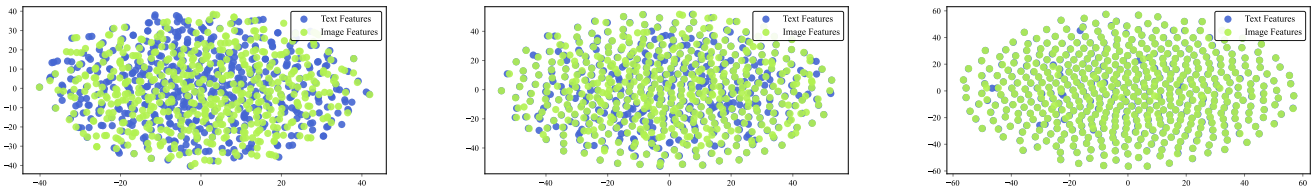
Dataset	Domain	# Nodes	# Edges	# Classes
Movies	E-commerce	16,672	109,195	20
Toys	E-commerce	20,694	63,443	18
CDs	E-commerce	36,272	844,878	15
Arts	E-commerce	28,195	197,428	7
Reddit(S)	Social Media	15,894	566,160	20
Goodreads	Literature	685,294	7,235,084	11

1.2. Experiment Details

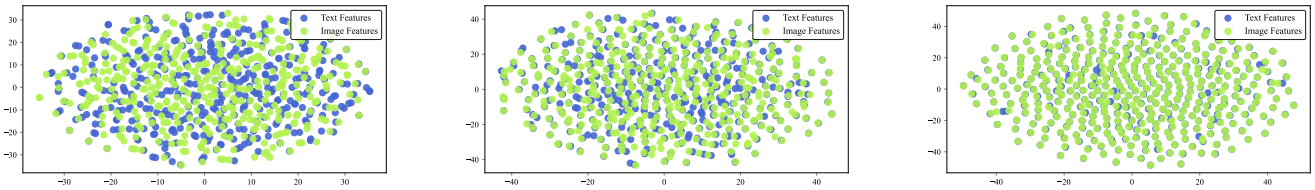
In this section, we provide additional explanations for experiment details not covered in the paper.
Image to caption conversion. Since current GraphLLM baselines do not support processing image features, we convert images into captions using VLMs in the text+vision experiments to enhance textual modality with auxiliary information. This facilitates multimodal graph reasoning. The model used for this purpose is Qwen-VL-Chat [1].
L(V)LMs-Based Baseline Experiment Execution. In the experiments, we frequently mention using LLaMA and LLaVA. All these experiments were conducted with the assistance of vLLM [5]. vLLM is a high-performance library for efficient LLM inference and serving. It provides state-of-the-art serving throughput with optimizations such as PagedAttention, continuous batching, CUDA acceleration, FlashAttention, and speculative decoding, ensuring low-latency execution. vLLM seamlessly integrates with Hugging Face models, supports various decoding strategies, and enables tensor/pipeline parallelism across diverse hardware platforms. In our experiments, we utilized vLLM to efficiently serve LLaMA and LLaVA, enabling scalable inference for text-based and MMG reasoning tasks while ensuring computational efficiency and high throughput.
Hyper-Parameter Settings. We provide a detailed discussion of the hyper-parameter settings used in our experiments. For Stage 1, we usually employ one layer (up to two) of GraphTransformer for structure-aware text-image alignment and we sample $\sim$**10** nodes ($\mathcal{V}_s$) to feed into the GVLM. For Stage 2, we typically select 10-15 neighbors to provide neighbor context and conduct 10 epochs of instruction tuning using LLaMA3.1-8B with early stop strategy. We use a four-layer MLP as the MAPR, and set λ = 0.01. For link prediction experiments, we only provide the neighbor context of the first node in the prompt, but these are common neighbors with the other node. The projection layer consists of two layers. For GraphPrompter [6], we use LLaMA3.1-8B as the final LLM for inference. For LLaGA [2], we follow the original paper and adopt the same setting, where Vicuna [3] serves as the primary foundational large language model. We truncate the final tokens input length to 512. All experiments involving LLM deployment were conducted on two A100-SXM4-80-GB GPUs. For GraphLLM-based baselines, we did not evaluate the vision-only setting. This is because such frameworks are inherently text-centric by design, and we followed their original modeling philosophy without extending them to vision-only scenarios. Additionally, we experimented with using image captions alone to support inference within these models, but the performance was significantly worse compared to text-only or image+text settings. Therefore, we did not include the vision-only results in the paper.

1.3. t-SNE Visualization of GVLM Alignment

To further illustrate the qualitative differences between the three models in Fig. 1, we visualize their aligned text and image features using t-SNE on two multimodal graphs, *Movies* and *Reddit*. For each dataset, we randomly sample a



(a) Movies – Frozen CLIP (b) Movies – Tuned CLIP (c) Movies – Mario's GVLM (ours)



(d) Reddit – Frozen CLIP (e) Reddit – Tuned CLIP (f) Reddit – Mario's GVLM (ours)

Figure 1. t-SNE visualizations of aligned multimodal features on *Movies* (top) and *Reddit* (bottom) for the three models in Fig. 1 in the paper. For each dataset, we project a randomly sampled subset of nodes from the full graph, using their aligned text and image representations as input to t-SNE. Comparing the six panels reveals how different alignment strategies affect the relative organization of text and image features in the shared latent space.

subset of nodes from the full graph and project their aligned text/image representations to 2D. This subsampling allows us to focus more closely on the structural differences between models while still capturing representative patterns. The six panels in Fig. 1 show the resulting distributions for the three models on *Movies* (top row) and *Reddit* (bottom row), respectively. Beyond the overall layout, we observe consistent qualitative trends across the six panels in Fig. 1 in the paper. On both *Movies* and *Reddit*, the frozen CLIP features form two loosely overlapping clouds, indicating a sizeable gap between text and image representations. Fine-tuning CLIP shrinks this gap and slightly tightens the clusters, but the two modalities still remain partially misaligned. In contrast, Mario produces a much more intertwined manifold where text and image features are almost co-located and organized along smoother global structures, suggesting that our graph-conditioned alignment achieved by Mario's GVLM substantially improves cross-modal consistency while preserving meaningful semantic variation.

1.4. Comparison with MMGCN and MGAT

In the main paper, MMGCN [8] and MGAT [7] are excluded, as they focus primarily on recommendation-style tasks and showed weak performance in our setting through initial experiments. For completeness, we provide here a small-scale comparison to substantiate this design choice. Table 2 reports their node classification accuracy on *Movies* and *Arts*, alongside representative "text+image" GNN baselines under the same experimental protocol.

Overall, MMGCN and MGAT do not show clear advantages over standard GNNs. On *Movies*, MMGCN is essentially on par with GCN and still below GATv2, while MGAT performs worse than all three GNN baselines. On *Arts*, MMGCN and MGAT slightly outperform some GNNs, but the gains are modest and all these methods remain far from the strong multimodal models and Mario reported in the main tables. Since SAGE, GATv2, and GCN are already treated as weak baselines in our core comparison, adding MMGCN and MGAT there would not change the conclusions; we therefore only include them in this appendix section for completeness. A similar conclusion can also be drawn from MLaGA [4].

Table 2. Node classification accuracy (%) on *Movies* and *Arts* for additional multimodal baselines (MMGCN, MGAT) and representative unimodal GNNs (text+image settings).

Model	Movies	Arts
SAGE	44.07	85.35
GATv2	49.29	81.19
GCN	46.96	76.76
MMGCN	46.79	86.63
MGAT	40.17	87.25

1.5. Additional GNNs Zero-Shot Results

In our zero-shot experiments in the paper, we assess the transferability of graph neural networks (GNNs) to new datasets, without re-training their core parameters. Specifically, when transitioning between datasets, we retain the trained GNN model, including its network architecture and learned parameters, and only replace the classifier layer corresponding to the new dataset. This approach ensures that the underlying graph feature extractor remains unchanged, allowing us to evaluate the generalization capacity of differ-

Table 3. Frozen Mario versus LoRA-Tuned Mario (Accuracy %).

Model	Trainable Params	Movies	Reddit	CDs	Arts
Node Classification (Trainable parameters are from Stage 2)					
Frozen Mario	18,886,656 (0.2346%)	50.85	93.60	60.45	89.69
Mario + LoRA	22,294,528 (0.2768%)	53.63	95.30	63.43	92.13
Link Prediction (Trainable parameters are from Stage 2)					
Frozen Mario	18,886,656 (0.2346%)	90.90	89.00	88.60	86.30
Mario + LoRA	22,294,528 (0.2768%)	93.90	91.30	92.70	89.96

ent models under domain shifts.

Table 4 presents the additional zero-shot transfer results across different models. This result serves as a supplement to Table 3 in the paper (where GraphLLMs adopt the text-only setting, and the other baselines adopt the text+vision setting). For the results below without explicit modality specification, the **text-only modality** is used (different from the setting in the paper). We evaluate the same two transfer settings: (1) Toys → Movies, where models trained on the Toys dataset are directly applied to the Movies dataset, and (2) Toys+Movies → CDs, where models trained on both the Toys and Movies datasets are tested on the CDs dataset. The evaluation is conducted under two tasks: NC (Node Classification Accuracy) and LP (Link Prediction Accuracy).

Across both transfer settings, our Mario significantly outperforms all baselines, demonstrating strong zero-shot adaptation capabilities. In contrast, traditional GNNs such as GCN, GATv2, and SAGE struggle to generalize, exhibiting considerably lower performance. For instance, in the Toys → Movies setting, GCN achieves an NC score of only 3.29, while Mario achieves 41.00, more than 10 times higher. A similar trend is observed in Toys+Movies → CDs, where Mario attains an NC score of 54.32, substantially outperforming all baselines.

Furthermore, while MLP-based models (both text-only and vision-only versions) show moderate performance in link prediction, they underperform in node classification due to their inability to leverage structural dependencies effectively. These results underscore the limitations of conventional GNNs in zero-shot scenarios and highlight the advantages of our Mario model in learning transferable multimodal representations.

1.6. Frozen vs. LoRA-Tuned Mario

We also find that LoRA-tuned Mario outperforms its frozen counterpart, and ***both exceed all baselines by a large margin***. As shown in Table 3, LoRA tuning yields consistent gains of about 1.7–3.0 points in node classification accuracy across all four datasets (e.g., from 50.85 to 53.63 on *Movies* and from 89.69 to 92.13 on *Arts*), and similarly improves link prediction by roughly 2–4 points. These improvements come with only a tiny increase in the number of trainable parameters, from 18.9M (0.2346%) to 22.3M (0.2768%) of the full LLM, indicating that Mario is already strong in a frozen-LLM regime while a lightweight LoRA adapter can further boost performance without sacrificing parameter efficiency.

Table 4. Zero-Shot Results (Accuracy %).

Model	Toys → Movies		Toys+Movies → CDs	
	NC	LP	NC	LP
MLP	6.12	52.60	7.04	50.20
GCN	3.29	62.13	10.01	64.17
GATv2	4.32	64.47	8.13	67.97
SAGE	3.11	55.83	6.14	59.63
MLP(Vision Only)	4.61	52.13	9.06	46.09
Mario-8B (Ours)	**41.00**	**86.60**	**54.32**	**82.50**

1.7. Ablation Study of LLM Backbone

To assess the robustness of Mario across different LLMs, we conduct an ablation study using a range of LLM backbones, including both LLaMA-based and non-LLaMA families. As summarized in Table 5, Mario consistently delivers strong performance regardless of the specific LLM used, highlighting the generalizability of our framework.

Within the LLaMA2 family, increasing model size from 7B to 13B results in negligible improvement: on Arts, accuracy rises slightly from 91.06% to 91.23%, while performance on Toys slightly drops from 81.20% to 80.93%. Similarly, when switching from LLaMA2 to Vicuna-v1.5 (also LLaMA2-based), results remain largely consistent, indicating that mere scaling or minor tuning of the base LLM does not significantly alter performance in our multimodal graph reasoning tasks.

More importantly, Mario remains effective even when paired with structurally different LLMs. Using FLAN-T5-XXL, a T5-style encoder-decoder model, Mario achieves 92.08% on Arts and 81.63% on Toys, outperforming all LLaMA2 variants. Furthermore, Mario-8B (LLaMA3), our best-performing configuration, achieves 92.13% and 82.58% on Arts and Toys respectively, demonstrating

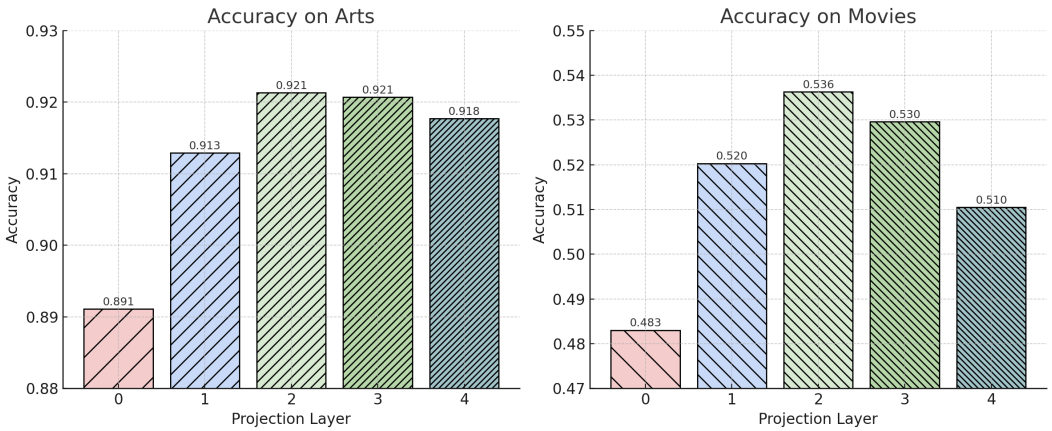


Figure 2. Sensitivity analysis of the projection layer in Arts and Movies

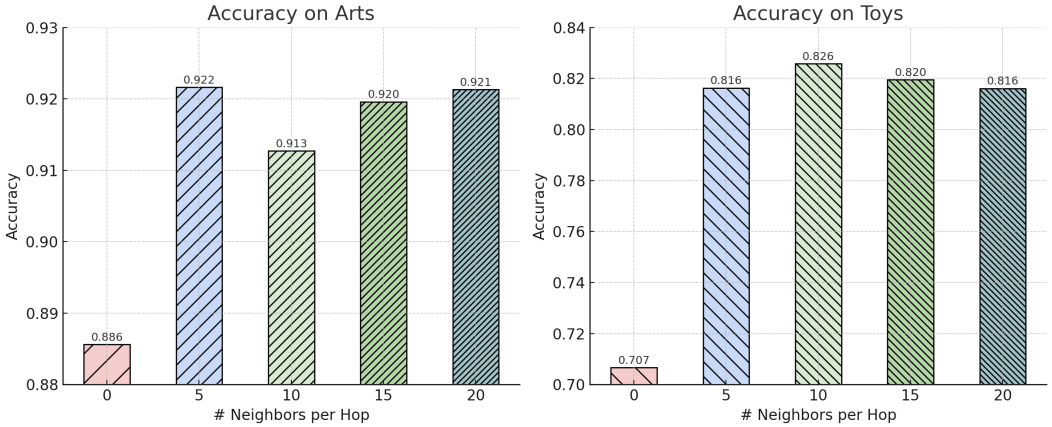


Figure 3. Sensitivity analysis of the number of neighbors per hop in Arts and Toys

stronger capability than its LLaMA2 predecessors.

These observations collectively indicate that Mario's architectural design—rather than the choice of LLM backbone—is the key contributor to its strong performance. Whether applied to decoder-only (LLaMA), instruction-tuned (Vicuna), or encoder-decoder (FLAN-T5) models, Mario exhibits consistent gains, underscoring its backbone-agnostic robustness in multimodal graph reasoning.

Table 5. Ablation Study of Different LLMs. (Accuracy %)

Different Size	**Arts**	**Toys**
Mario-7B (LLaMA2)	91.06	81.20
Mario-13B (LLaMA2)	91.23	80.93
Mario-7B (Vicuna-v1.5)	91.09	81.07
Mario (FLAN-T5-XXL)	92.08	81.63
Mario-8B (LLaMA3)	**92.13**	**82.58**

1.8. Sensitivity Analysis

To assess the effectiveness of our designed instruction templates that incorporate multimodal node features, we conducted a sensitivity analysis on two critical components: the number of projection layers and the length of the neighbor context. These factors directly influence how effectively multimodal information is aligned and delivered to the LLM for reasoning. As shown in Figure 2, introducing a projection layer consistently improves performance over the baseline without projection. Notably, employing two layers yields the best or near-best results across both Arts and Movies datasets. This suggests that a lightweight projection module facilitates better multimodal alignment without incurring excessive complexity, enhancing the model's ability to interpret visual-textual signals.

Similarly, Figure 3 illustrates that incorporating a limited number of neighbors per hop significantly boosts performance compared to the zero-neighbor setting. For instance, in the Toys dataset, adding neighbor context improves ac-

Table 6. Heterophily Ratios of Benchmark Datasets

Dataset	Movies	Toys	Arts	CDs	Goodreads	Reddit
Heterophily Ratio	0.53	0.26	0.34	0.69	0.33	0.04

curacy by over 10%. However, further increasing the number of neighbors yields marginal or unstable gains, indicating that a moderate amount of structural context is optimal. These results highlight the importance of integrating a controlled amount of structural information into the prompt, allowing the LLM to better contextualize the target node during reasoning.

1.9. Variance Analysis

Following prior GraphLLM studies, we initially omitted variance reporting. However, our experiments reveal that the variance of our method is relatively small—typically around ± 0.07 or ± 0.14 across three random runs. For reference, Table 7 summarizes partial variance scores on representative datasets and tasks (Metric: Accuracy).

Table 7. Partial variance results of Mario across datasets and tasks over 3 runs.

Method	**Movies (NC)**	**Arts (LP)**
Mario (Single Focus)	53.63 ± 0.07	89.96 ± 0.14
Mario (Mix Training)	50.98 ± 0.08	92.60 ± 0.12

1.10. Quantitative Analysis of Modality Preference

This subsection explains the statistics in the Venn diagram of Figure 1 in the paper. Specifically, the six numbers in Figure 1(b) can be grouped into three categories: (i) the proportion of nodes that can be correctly classified *only* by the template corresponding to a single modality; (ii) the proportion of nodes that can be correctly classified *only* when two templates are both correct (rather than counting a node as correct if either template is correct); and (iii) the proportion of nodes that can be correctly classified by all three template types. The proportions in the first category are 2.65%, 2.25%, and 2.05%; those in the second category are 7.71%, 7.40%, and 6.98%; and the third category accounts for 70.96%. These numbers sum to 100%. Therefore, the statement that "about 30% of nodes cannot be correctly classified by all templates jointly" is computed as $100\% - 70.96\% = 29.04\%$, which is approximately 30%. All percentages are normalized within the set of nodes correctly classified by at least one template.

1.11. Robustness against Varying Heterophily

To reduce overfitting to locally uniform neighborhoods and to expose the model to richer semantic context, we adopt a multi-hop neighbor selection strategy. Expanding the receptive field beyond immediate neighbors allows Mario to retrieve distant yet relevant nodes, so the router is not forced to rely solely on short-range label similarity when forming prompts. To quantify structural diversity in our benchmarks, we compute each graph's heterophily ratio, defined as the fraction of edges linking nodes with different labels.

Despite the heterophily ratios varying widely across datasets—from near-homophilic graphs such as Reddit (0.04) to strongly heterophilic ones like CDs (0.69) and Movies (0.53)—Mario consistently maintains strong performance. This suggests that our Stage-1 feature–based similar neighbor selection remains reliable across different structural regimes, confirming that it generalizes well even when local neighborhoods are not label-coherent. Importantly, while this finding is complementary to Observation 7 in the paper, it addresses a different question: here we show robustness of the selection strategy under varying heterophily, rather than characterizing the spatial pattern of modality preferences within the graph.

1.12. Training Compute Analysis

We compare Mario's Stage 1/2 with all baselines on identical data under a compute-matched budget (Table 8), which reports the training cost and resulting performance on our main datasets. To ensure that every method had sufficient opportunity to converge, we initially capped each Stage-1 run at 2 GPU-hours on A100 SXM4 80GB GPUs and terminated runs that remained unconverged at the cap. In practice, however, we observed that the GVLM and all graph-based baselines consistently converged within 1 GPU-hour, with no notable gap in training overhead across methods, indicating that our comparisons are conducted under a largely fair and compute-balanced setting. We further include a new Tuned CLIP baseline for a stronger reference; since its optimization is typically more compute-intensive than GVLM/GNN-style training, we set its compute cap to 1 GPU-hour as well to maintain fairness given that the other methods already converge within this budget. As shown in Table 8 (the lower part), MAPR converges in roughly half the epochs of the baselines (stage2), with an average total runtime only 0.25 (rather than 3×) GPU-hours higher. Finally, the resulting average accuracies under these compute caps closely match those reported in Table 4 and Figure 4 in the paper, with no noticeable discrepancies.

Table 8. Detailed Cost Breakdown (Stage 1 & 2).

Method	Arts	Reddit	Movies	#Epoch	Avg Acc(%)
Tuned CLIP/Other Baselines vs. GVLM (stage1) (Columns 2–4: GPU-hours)					
GCN	0.95	0.94	0.88	36.9	78.17 $_{\uparrow 2.79}$
SAGE	0.91	0.90	0.88	34.3	77.98 $_{\uparrow 3.04}$
GATv2	0.93	0.91	0.89	33.0	78.08 $_{\uparrow 2.92}$
MLP	0.90	0.87	0.85	37.0	77.78 $_{\uparrow 3.30}$
Tuned CLIP	0.99	0.98	0.99	22.3	78.01 $_{\uparrow 3.00}$
GVLM	0.92	0.91	0.88	23.0	80.35
Single-template Variants vs. MAPR (Stage2) (Columns 2–4: GPU-hours)					
Text-only	5.65	4.02	4.11	6.3	78.56 $_{\uparrow 2.28}$
Image-only	5.65	4.58	4.11	6.6	78.93 $_{\uparrow 1.80}$
Text+Image	5.77	4.09	4.23	6.3	78.50 $_{\uparrow 2.36}$
Mario (MAPR)	5.82	4.15	4.35	3.0	80.35

1.13. Prompt Template

The prompt templates used for adaptive multimodal graph instruction tuning in the two multimodal graph reasoning tasks, node classification and link prediction, are shown in Table 9. Since the templates for different modalities are broadly similar, differing mainly in which modality-specific features of the anchor node and its neighbors are embedded—we present the template for the text+image case as an illustrative example.

1.14. Case Study

To better understand how Mario behaves on multimodal graphs, we conduct a qualitative case study on both tasks. We compare Mario-8B against three strong closed-/API-based L(V)LMs—ChatGPT-5.1-Thinking, Gemini-3-Pro, and Qwen3-Max—on several representative nodes and node pairs drawn from the Movies, Toys, and CDs graphs (Figs. 4–15). Because these models are accessed only through high-level APIs, we cannot inject special feature tokens when prompting as we do for Mario. Instead, we adopt a uniform and conservative prompting protocol: for each case, we present the anchor node (or node pair) together with its neighbors using the same high-level templates as in Table 9, and we input each neighbor's raw text and image jointly to ensure a fair comparison.

For the node classification case shown in Figs. 6–7, the anchor node's text describes the content of a lecture series, whereas the associated image focuses almost entirely on after-sales information (lifetime warranty and replacement policy) and provides very little semantic signal about the lecture itself. Mario's MAPR, conditioned on both the anchor's multimodal features and its local subgraph, correctly infers that the image is not the preferred modality for this classification task and routes the node through a text-centric template. This decision matches the underlying graph semantics and illustrates that the router is able to down-weight visually salient but task-irrelevant information.

Across all the illustrated cases, Mario's behavior is consistently competitive with, and sometimes superior to, strong closed-source L(V)LMs. In several examples, all closed models converge to the same intuitive but graph-inconsistent label, while Mario is the only method that predicts the correct class—for instance, in the case of Figs. 4–5, where Mario is the only model that assigns the ground-truth category and all other systems fail. These qualitative results further corroborate that Mario is an effective and reliable framework for multimodal graph reasoning.

Table 9. Prompt Templates for Node Classification and Link Prediction Tasks. Note that this template is designed to include both text and image features of the node. If the input is text-only or image-only, simply retain the corresponding single modality feature.

Task	**Prompt Template**
Node Classification	I'm starting a node classification task in the <dataset>. Each node represents a <product> with text and image features, and edges indicate <relationship>. Given a target node, the raw text is ..., the text feature is <text feature> and the image feature is <image feature>. The neighbors are described in the following template: <text feature>, <image feature>, and <label>. It has the following neighbors at hop 1: N1: <1-hop neighbor 1 text feature>, <1-hop neighbor 1 image feature>, <1-hop neighbor 1 label> N2: <1-hop neighbor 2 text feature>, <1-hop neighbor 2 image feature>, <1-hop neighbor 2 label> N3: ..... It has the following neighbors at hop 2: N1: <2-hop neighbor 1 text feature>, <2-hop neighbor 1 image feature>, <2-hop neighbor 1 label> N2: <2-hop neighbor 2 text feature>, <2-hop neighbor 2 image feature>, <2-hop neighbor 2 label> ..... Based on the information provided, please classify the target node into one of the following categories: {all_categories}.
Link Prediction	I'm starting a link prediction task in the <dataset>. Each node represents a <product> with text and image features, and edges indicate <relationship>. Given the two nodes: Node 1: The raw text is ... the text feature is <text feature>, and the image feature is <image feature>. Node 2: The raw text is ... the text feature is <text feature>, and the image feature is <image feature>. The neighbors of node 1 (common neighbors with node 2) are described in the following template: <text feature>, <image feature>. It has the following neighbors at hop 1 (Directly connected): N1: <1-hop neighbor 1 text feature>, <1-hop neighbor 1 image feature> N2: <1-hop neighbor 2 text feature>, <1-hop neighbor 2 image feature> N3: ..... It has the following neighbors at hop 2 (Indirectly connected by shared neighbors): N1: <2-hop neighbor 1 text feature>, <2-hop neighbor 1 image feature> N2: <2-hop neighbor 2 text feature>, <2-hop neighbor 2 image feature> ..... Based on the information provided, please determine whether a link exists between the two nodes. Answer "yes" if a link exists or "no" if it does not.

Figure 4. A case from the Movies dataset in the NC task that Mario identifies as preferring Text+Image modality information.

Case Study: Node Classification-Text+Image-Movies

Anchor node raw text:

In a strange dark age based on Celtic myths, the Divine Empire's path of conquest seems unstoppable... until a savage priest makes a critical mistake while attempting to resurrect a Demon Lord! Now the scales of fate tip in the other direction.

Label list:

'A&E Home Video'
'Art House & International'
'BBC'
'Blu-ray'
'Boxed Sets'
'Classics'
'Criterion Collection'
'Fully Loaded DVDs'
'Genre for Featured Categories'
'HBO'
'Holidays & Seasonal'
'Independently Distributed'
'Movies'
'Music Artists'
'Musicals & Performing Arts'
'Paramount Home Entertainment'
'Science Fiction & Fantasy'
'Studio Specials'
'TV'
'Walt Disney Studios Home Entertainment'

ChatGPT-5.1-Thinking:

Science Fiction & Fantasy ✗

Gemini-3-Pro:

Science Fiction & Fantasy ✗

Qwen3-Max:

Science Fiction & Fantasy ✗

Mario-8B:

Genre for Featured Categories ✓

Figure 5. Anchor node's image in Figure 4.

Figure 6. A case from the Movies dataset in the NC task that Mario identifies as preferring Text-only modality information.

Case Study: Node Classification-Text-only-Movies

Anchor node raw text:

Roots of Human Behavior; Description: While human history is usually studied from the perspective of a few hundred years, anthropologists consider deeper causes for the ways we act. Now, in these 12 engrossing lectures, you'll join an expert anthropologist as she opens an enormous window of understanding for you into the thrilling legacy left by our primate past.

Label list:

'A&E Home Video'	'Art House & International'
'BBC'	'Blu-ray'
'Boxed Sets'	'Classics'
'Criterion Collection'	'Fully Loaded DVDs'
'Genre for Featured Categories'	'HBO'
'Holidays & Seasonal'	'Independently Distributed'
'Movies'	'Music Artists'
'Musicals & Performing Arts'	'Paramount Home Entertainment'
'Science Fiction & Fantasy'	'Studio Specials'
'TV'	'Walt Disney Studios Home Entertainment'

ChatGPT-5.1-Thinking:

TV ✗

Gemini-3-Pro:

Genre for Featured Categories ✓

Qwen3-Max:

Movies ✗

Mario-8B:

Genre for Featured Categories ✓

Figure 7. Anchor node's image in Figure 6

Figure 8. A case from the Movies dataset in the NC task that Mario identifies as preferring Image-only modality information.

Case Study: Node Classification-Image-only-Movies

Anchor node raw text:

Castle: Season 6; Description: After Castle's stunning romantic proposal to Beckett, what happens next? TV's magnetic crime-fighting couple faces a whole new set of challenges as they juggle wedding plans and their most intriguing cases yet in ABC's CASTLE: THE COMPLETE SIXTH SEASON. Beckett's new job with the Justice Department takes her away from the wisecracking love of her life. But Castle's devotion to his new fiancee -- and her fascinating line of work -- jeopardizes her career and creates a chain of events that might separate them forever. Back on the home front, Castle is none too pleased to discover his daughter has seemingly been captivated by, and now living with, her new, free-spirited boyfriend.

Label list:

'A&E Home Video'	'Art House & International'
'BBC'	'Blu-ray'
'Boxed Sets'	'Classics'
'Criterion Collection'	'Fully Loaded DVDs'
'Genre for Featured Categories'	'HBO'
'Holidays & Seasonal'	'Independently Distributed'
'Movies'	'Music Artists'
'Musicals & Performing Arts'	'Paramount Home Entertainment'
'Science Fiction & Fantasy'	'Studio Specials'
'TV'	'Walt Disney Studios Home Entertainment'

ChatGPT-5.1-Thinking:

TV ✗

Gemini-3-Pro:

TV ✗

Qwen3-Max:

TV ✗

Mario-8B:

Boxed Sets ✓

Figure 9. Anchor node's image in Figure 8

Figure 10. A case from the Toys dataset in the LP task that Mario identifies as preferring Text+Image modality information.

Case Study: Link Prediction-Text+Image-Toys

Node pair raw text:

Node 1:
The Crazy Scientist series, a collection of science tricks, was created by a joint venture of 2 crazy scientists and the Purple Cow. A combination bound to create an excellent and yet crazy experience! The Crazy Scientist Young Researches is a set of science tricks for kids to try out and discover 20 fun and fascinating facts about the world around you. Create excellent and crazy experiences that can be enjoyed by the entire family! Each science trick comes with a simple yet clever scientific explanation. A perfect STEAM gift! Challenge your brainpower and make intriguing discoveries about the world around us. Have fun experimenting and learning with the Young Researches amazing activities. Provide children the opportunity to become real researchers and follow easy instructions of science experiments that can be conducted using common household materials. Whats included? The box contains 20 activity cards with detailed instructions. Recommended for children ages 6 and up. Some science tricks may require adult supervision as indicated.

Node 2:
Learn the scientific principles behind optical illusions with the 4M Illusion Science kit. Experiment with 20 classic optical illusions included in this kit. The kit includes illusion trick cards, spinning top with illusion cards, 3D picture cards, markers, 3D glasses, and more. A 20-page instruction book is included, describing the science of optical illusions and how to create a wide range of illusory effects. Perfect for young scientists with an interest in optics. Recommended for ages 7 years and up.

Label list:
'yes' The two toys are co-purchased.
'no' The two toys are not co-purchased.

ChatGPT-5.1-Thinking:
No ✗

Gemini-3-Pro:
Yes ✓

Qwen3-Max:
Yes ✓

Mario-8B:
Yes ✓

(a) Node 1. (b) Node 2.

Figure 11. Node Pair's images in Figure 10.

Figure 12. A case from the CDs dataset in the LP task that Mario identifies as preferring Text-only modality information.

Case Study: Link Prediction-Text-only-CDs

Node pair raw text:

Node 1:
You Can Do It Yoga for MS Volume 2 DVD; Description: This DVD contains 2 complete classes. The first is a beginner/gentle yoga class. It includes some floor poses and some standing poses along with a guided meditation. Runtime: 54 minutes The second class is a beginner/intermediate yoga class. It includes some floor poses and some standing and balancing poses along with a guided meditation. Runtime: 50 minutes,This DVD contains 2 complete classes.

Node 2:
Thoughts Become Things; Description: You create your own reality and by changing your thoughts, words, and actions in the simplest of ways, you can create fantastic change. - Mike Dooley".

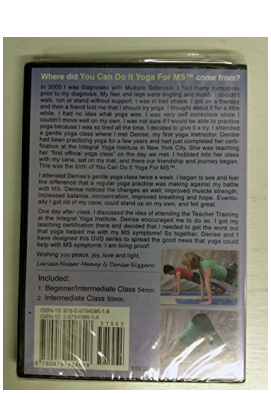
Label list:
'yes' The two CDs are co-purchased.
'no' The two CDs are not co-purchased.

ChatGPT-5.1-Thinking:
No ✗

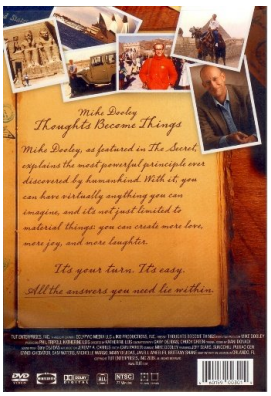
Gemini-3-Pro:
Yes ✓

Qwen3-Max:
No ✗

Mario-8B:
Yes ✓



(a) Node 1.



(b) Node 2.

Figure 13. Node Pair's images in Figure 12.

Figure 14. A case from the CDs dataset in the LP task that Mario identifies as preferring Image-only modality information.

Case Study: Link Prediction-Image-only-CDs

Node pair raw text:

Node 1:

Howard Lovecraft And The Frozen Kingdom; Description: After visiting his father in Arkham Sanitarium, young Howard Lovecraft ignores his father&#146s ominous warning and uses the legendary Necronomicon to open a portal to a strange, frozen world filled with horrifying creatures and grave danger. Alone and scared, Howard befriends a hideous creature he names Spot who becomes his companion throughout their treacherous journey across the Frozen Kingdom.

Node 2:

A Serbian Film (Uncut) by Srdjan Todorovic; Description: Milos, a retired adult film star, leads a normal family life with his wife Maria and six-year old son Petar in tumultuous Serbia, trying to make ends meet. A sudden call from his former colleague Layla will change everything. Aware of his financial problems, Layla introduces Milos to Vukmir – a mysterious, menacing and politically powerful figure in the adult film business. A leading role in Vukmir's production will provide financial support to Milos and his family for the rest of their lives.

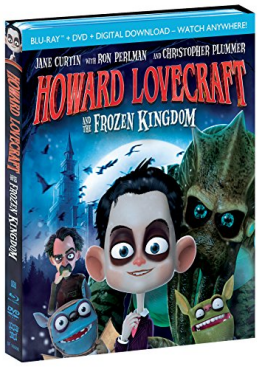
Label list:
'yes' The two CDs are co-purchased.
'no' The two CDs are not co-purchased.

ChatGPT-5.1-Thinking:
No ✓

Gemini-3-Pro:
No ✓

Qwen3-Max:
Yes ✗

Mario-8B:
No ✓



(a) Node 1.

(b) Node 2.

Figure 15. Node Pair's images in Figure 14

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization. *Text Reading, and Beyond*, 2(1):1, 2023. 1

[2] Runjin Chen, Tong Zhao, AJAY KUMAR JAISWAL, Neil Shah, and Zhangyang Wang. Llaga: Large language and graph assistant. In *Forty-first International Conference on Machine Learning*, 2024. 1

[3] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, march 2023. *URL https://lmsys. org/blog/2023-03-30-vicuna*, 3(5), 2023. 1

[4] Dongzhe Fan, Yi Fang, Jiajin Liu, Djellel Difallah, and Qiaoyu Tan. Mlaga: Multimodal large language and graph assistant. *arXiv preprint arXiv:2506.02568*, 2025. 2

[5] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626, 2023. 1

[6] Zheyuan Liu, Xiaoxin He, Yijun Tian, and Nitesh V Chawla. Can we soft prompt llms for graph learning tasks? In *Companion Proceedings of the ACM on Web Conference 2024*, pages 481–484, 2024. 1

[7] Zhulin Tao, Yinwei Wei, Xiang Wang, Xiangnan He, Xianglin Huang, and Tat-Seng Chua. Mgat: Multimodal graph attention network for recommendation. *Information Processing & Management*, 57(5):102277, 2020. 2

[8] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1437–1445, 2019. 2