GraphVLM: Benchmarking Vision Language Models for Multimodal Graph Learning

Supplementary Material

A. Dataset Statistics

Our experimental benchmark comprises six datasets spanning two distinct domains, with detailed statistical characteristics presented in Table 5.

Dataset	#Nodes	#Edges	Avg.#Degree	#Classes	Domain
Movies	16,672	218,390	26	19	E-commerce
Toys	20,695	126,886	12	18	E-commerce
Grocery	84,379	693,154	16	20	E-commerce
Arts	28,195	197,428	14	7	E-commerce
CDs	36,272	844,878	47	15	E-commerce
Reddit	99,638	1,167,188	23	50	Social Network

Table 5. Dataset statistics.

B. Baseline Implementations

We evaluate a diverse set of state-of-the-art methods spanning three major categories of graph learning paradigms.

- **GNN-based methods.** We include the conventional GCN [15], the widely used GraphSAGE [10], and a non-graph baseline MLP [26]. For multimodal graph learning, we evaluate MMGCN [35] and MGAT [32], which explicitly fuse heterogeneous modalities within graph structures. To ensure fair comparison, all models are trained and evaluated under consistent experimental settings, following the MM-Bench framework [42]. We additionally adopt the state-of-the-art UniGraph2 [12]. However, due to the $O(n^3)$ time complexity of shortest-path distance (SPD) computation in the original implementation, we omit the SPD module for efficiency in our evaluation.
  For the CLIP-F-S configuration, we randomly sample five hop-1 neighbors for each anchor node and employ CLIP as the visual encoder within a contrastive learning objective. The training hyperparameters are summarized in Table 6.

Learning Rate	#Neighbors	Epoch	Batch Size	Temperature	Optimizer
1×10^{-5}	5	1	16	0.5	Adam

Table 6. Training hyperparameters for the CLIP-F-S model.

- **LLM-based methods.** We assess multiple GraphLLMs with distinct architectures, including GraphPrompter [22], LLaGA [4], GraphGPT [31], GraphTranslator [40], and the recent MLaGA [5]. For all approaches requiring multimodal embeddings, we employ the pre-trained CLIP [25] encoder to extract unified visual–textual representations. All models are evaluated following their official implementations to ensure methodological fidelity.
- **VLM-based methods.** We further include the state-of-the-art multimodal large language models Qwen-VL-7B [2], Qwen2.5-VL-7B [3] and LLaVA-1.5-7B [20] for predictor experiments, with Qwen-VL-7B also serving as the aligner in RQ2. All models are evaluated under their official zero-shot and fine-tuning protocols. For structural information injection, we select the top-3 most similar nodes based on cosine similarity from the anchor node's hop-1 neighbors.

C. Prompt Templates

Image summary prompt. We use Qwen-VL-7B [2] to generate the image summaries as shown in Table 9.
Non-structure-aware aligner prompt. We synthesize the original textual information with generated image summaries using an LLM, following the prompts in Table 10.
Structure-aware aligner prompt. To incorporate structure-aware multimodal information, we design our prompts for modality synthesis as shown in Table 11.
VLM-as-Predictor prompt design. To directly enable VLM as the predictor, we design the prompt accordingly, as shown in Table 12 and Table 13. In fine-tuning strategies, the prompt includes *'Assistant: <true label>'*, while in in-context learning, it is excluded.

D. Efficiency–Effectiveness Trade-off

We analyze this trade-off from the perspective of cross-domain transferability, with results reported in Table 7. VLM-based models achieve substantially higher transfer accuracy than GNN baselines when evaluated on target datasets with or without supervised fine-tuning. This transfer advantage can help explain the growing interest in LLM/VLM-based graph methods, despite their higher computational cost. We collect the training and inference time on a specific dataset for reference in Table 8.

Model	Settings	Movies	CDs	Grocery	Arts
MLP	Text+Image (CLIP)	4.90	6.85	9.17	14.14
GCN	Text+Image (CLIP)	7.69	10.05	22.02	12.55
GraphSAGE	Text+Image (CLIP)	3.73	6.20	5.56	20.69
MLP	Text+Image (CLIP-F-S)	5.50	8.79	14.12	18.14
GCN	Text+Image (CLIP-F-S)	3.16	7.61	8.66	18.12
GraphSAGE	Text+Image (CLIP-F-S)	1.20	9.85	11.40	18.14
Qwen2.5-VL	No SFT (Zero-shot)	**14.96**	**34.94**	**68.79**	**69.34**
Qwen2.5-VL	Text+Image (Non-structure)	**13.16**	**37.02**	**71.95**	**68.10**
Qwen2.5-VL	Text+Image (Structure)	**12.17**	**37.17**	**71.20**	**72.26**

Table 7. Transfer ability comparison: Node classification accuracy (%) when trained on Toys and evaluated on other datasets.

Stage	GNN-based				LLM-based		VLM-based	
	GCN	SAGE	MMGCN	MGAT	LLaGA	GraphGPT	Qwen-VL	Qwen2.5-VL
Training	~2min	~2min	~3min	~3min	~17min	~60min	~5h	~5h
Inference	~10s	~10s	~15s	~20s	~10min	~30min	~20min	~20min

Table 8. Efficiency comparison on the Movies dataset.

Movies:	<image input> Given an image of a movie from the Amazon movies dataset , generate a concise and detailed summary. Focus on describing key visual concepts. Ensure the summary is informative and useful for understanding the product as described in user reviews, without losing critical details or introducing unnecessary information.
Toys:	<image input> Given an image of a toy from the Amazon toys dataset , generate a concise and detailed summary. Focus on describing key visual concepts. Ensure the summary is informative and useful for understanding the product as described in user reviews, without losing critical details or introducing unnecessary information.
Grocery:	<image input> Given an image of a grocery from the Amazon grocery dataset , generate a concise and detailed summary. Focus on describing key visual concepts. Ensure the summary is informative and useful for understanding the product as described in user reviews, without losing critical details or introducing unnecessary information.
CDs:	<image input> Given an image of a CD from the Amazon CD dataset , generate a concise and detailed summary. Focus on describing key visual concepts. Ensure the summary is informative and useful for understanding the product as described in user reviews, without losing critical details or introducing unnecessary information.
Arts:	<image input> Given an image of an artwork from the Amazon Art dataset , generate a concise and detailed summary. Focus on describing key visual concepts. Ensure the summary is informative and useful for understanding the product as described in user reviews, without losing critical details or introducing unnecessary information.
Reddit:	<image input> Given an image of a post from the Reddit dataset, generate a concise and detailed summary. Focus on describing key visual concepts. Ensure the summary is informative and useful for understanding the post as described in the caption, without losing critical details or introducing unnecessary information.

Table 9. Prompts used to generate a text description of the image by VLM

Movies:	Given the text information of a product from the Amazon Movies dataset: <text information>. Image summary: <image summary> Questions: Using the title, description, and image summary of the product provided above, create an informative and concise description that effectively highlights the product's key features.
Toys:	Given the text information of a product from the Amazon toys dataset: <text information>. Image summary: <image summary> Questions: Using the title, description, and image summary of the product provided above, create an informative and concise description that effectively highlights the product's key features.
Grocery:	Given the text information of a product from the Amazon grocery dataset: <text information>. Image summary: <image summary> Questions: Using the title, description, and image summary of the product provided above, create an informative and concise description that effectively highlights the product's key features.
CDs:	Given the text information of a product from the Amazon CD dataset: <text information>. Image summary: <image summary> Questions: Using the title, description, and image summary of the product provided above, create an informative and concise description that effectively highlights the product's key features.
Arts:	Given the text information of a product from the Amazon Art dataset: <text information>. Image summary: <image summary> Questions: Using the title, description, and image summary of the product provided above, create an informative and concise description that effectively highlights the product's key features.
Reddit:	Given the text information of a post from the Reddit dataset: <text information>. Image summary: <image summary> Questions: Using the caption and image summary of the post provided above, create an informative and concise description that effectively highlights the post's key features.

Table 10. Prompts for the non-structure-aware aligner cases.

Movies:	Given the text information of a product from the Amazon movies dataset: <text information>. Image summary: <image summary>. Also given the information of co-purchased or co-reviewed products: text information: <neighbor text information>, image summary: <neighbor image summary> (or, if unavailable: 'No co-purchased or co-reviewed product information is available.') Questions: Using the product's title, description, and image summary provided above, along with any co-purchase or co-review data, generate a concise yet informative description of the product.
Toys:	Given the text information of a product from the Amazon toys dataset: <text information>. Image summary: <image summary>. Also given the information of co-purchased or co-reviewed products: text information: <neighbor text information>, image summary: <neighbor image summary> (or, if unavailable: 'No co-purchased or co-reviewed product information is available.') Questions: Using the product's title, description, and image summary provided above, along with any co-purchase or co-review data, generate a concise yet informative description of the product.
Grocery:	Given the text information of a product from the Amazon grocery dataset: <text information>. Image summary: <image summary>. Also given the information of co-purchased or co-reviewed products: text information: <neighbor text information>, image summary: <neighbor image summary> (or, if unavailable: 'No co-purchased or co-reviewed product information is available.') Questions: Using the product's title, description, and image summary provided above, along with any co-purchase or co-review data, generate a concise yet informative description of the product.
CDs:	Given the text information of a product from the Amazon CD dataset: <text information>. Image summary: <image summary>. Also given the information of co-purchased or co-reviewed products: text information: <neighbor text information>, image summary: <neighbor image summary> (or, if unavailable: 'No co-purchased or co-reviewed product information is available.') Questions: Using the product's title, description, and image summary provided above, along with any co-purchase or co-review data, generate a concise yet informative description of the product.
Arts:	Given the text information of a product from the Amazon Art dataset: <text information>. Image summary: <image summary>. Also given the information of co-purchased or co-reviewed products: text information: <neighbor text information>, image summary: <neighbor image summary> (or, if unavailable: 'No co-purchased or co-reviewed product information is available.') Questions: Using the product's title, description, and image summary provided above, along with any co-purchase or co-review data, generate a concise yet informative description of the product.
Reddit:	Given the text information of a post from the Reddit dataset: <text information>. Image summary: <image summary>. Also given the information of co-commented posts: text information: <neighbor text information>, image summary: <neighbor image summary> (or, if unavailable: 'No co-commented post information is available.') Questions: Using the post's caption and image summary provided above, along with any co-commented data, generate a concise yet informative description of the post.

Table 11. Prompts for the structure-aware aligner cases.

Movies / Toys / Grocery / CDs / Arts:	Given the target product information on Amazon: Picture: <image input> Title and description: <text information>. Question: Based on the target product's picture, title, and description, which category does the target product belong to? Choose from the following options: <candidates set>. Assistant: <truth label>
Reddit:	Given the target post information on Reddit: Picture: <image input> Caption: <text information>. Question: Based on the target post's picture and caption, which category does the target post belong to? Choose from the following options: <candidates set>. Assistant: <truth label>

Table 12. Prompts for the non-structure-aware predictor case.

Movies / Toys / Grocery / CDs / Arts:	Given the target product information on Amazon: Picture: <image input> Title and description: <text information>. Co-purchased or co-reviewed products: Picture1: <image input>; Title1: <text information> ; Picture2: <image input>; Title2: <text information> ; Picture3: <image input>; Title3: <text information>. Question: Based on the target product's picture, title, description, and related products, which category does the target product belong to? Choose from the following options: <candidates set>. Assistant: <truth label>
Reddit:	Given the target post information on Reddit: Picture: <image input> Caption: <text information>. Co-commented posts: Picture1: <image input>; Caption1: <text information> ; Picture2: <image input>; Caption2: <text information> ; Picture3: <image input>; Caption3: <text information>. Question: Based on the target post's picture, caption, and related posts, which category does the target post belong to? Choose from the following options: <candidates set>. Assistant: <truth label>

Table 13. Prompts for the structure-aware predictor case.