GRAPHGPT-O: Synergistic Multimodal Comprehension and Generation on Graphs

Supplementary Material

7. Limitations

In our current approach, we treat the graph as homogeneous, simplifying all nodes and edges into a single type. However, real-world graphs often consist of multiple node and edge types, each with unique semantic meanings. Future research could address this limitation by extending GraphGPT-o to heterogeneous graphs, allowing for richer and more nuanced representations of complex structures.

8. Ethical Considerations

GraphGPT-o presents a new method for improving the structural understanding of MLLMs through graph-based alignment. This approach seeks to tackle current issues in MLLMs, such as the uncontrolled generation of unsuitable content and susceptibility to adversarial attacks. Although GraphGPT-o provides enhancements, it still depends on the MLLM foundation, making it subject to these inherent limitations. Ethical concerns, like the potential for misuse, unintended generation of inappropriate content, and exposure to adversarial manipulation, need careful attention when deploying GraphGPT-o in practical applications.

9. Experiment settings.

Datasets. We conduct experiments on three multimodal attributed graphs from distinct domains: ART500K, Amazon-Baby, and Amazon-Beauty. The ART500K dataset represents artworks, where nodes correspond to individual pieces, and edges indicate relationships such as shared authorship or genre. The Amazon datasets, comprising Amazon-Baby and Amazon-Beauty, represent product graphs. Here, nodes denote products, while edges capture co-view relationships. Each node in these graphs is enriched with a title and an image.

Metrics. To thoroughly assess the comprehension and generation capabilities of GRAPHGPT-O on multimodal attributed graphs, our evaluation focuses on two key aspects:

- The quality of the synthesized image and text, and how well they align.
- The text/image correspondence between synthesized nodes and the conditioned sub-graphs.

To evaluate the quality of the synthesized outputs, we use CLIP (CLIP-I2) scores to compare the synthesized images with the ground truth images, assessing image generation quality. We also measure the perplexity of the generated text to evaluate its coherence. Additionally, we calculate the CLIP (CLIP-IT) score of generated image-text pairs to assess image-text alignment.

To evaluate alignment with the conditioned sub-graph, we calculate the KL divergence (KL-DV) between the distributions of the neighbor nodes and generated node image-text CLIP scores.

Split. For training, we randomly sampled 40,000 nodes from each original dataset. For testing, we randomly selected 50 nodes and its related neighbors from the rest of the dataset.

Hyperparameters. In the implementation of GraphGPT-o, we utilize DreamLLM as the pre-trained backbone. Within the Graph Hierarchical Tokenization module, the learnable tokens, as well as all self-attention and cross-attention layers, are randomly initialized. We employ a pre-trained CLIP encoder as the fixed image and text encoder, with an additional MLP to resolve dimensional discrepancies.

Table 5. Hyper-parameter configuration for model training.

Parameter	ART500K	Beauty	Baby
learning rate	1e-5	1e-5	1e-5
Batch size per GPU	1	1	1
warmup ratio	3e-3	3e-3	3e-3
Epochs	1	1	1
loss weight of lm	1	1	1
loss weight of vm	5	5	5

10. More Experiment Results.

We demonstrate more cases generated by DreamLLM and GRAPHGPT-O with comparision with the ground truth.

Study on the impact of number of neighbors. Figure 5 shows that incorporating information from more neighbors can improve performance, but an excessive number may introduce noise, potentially hindering results.

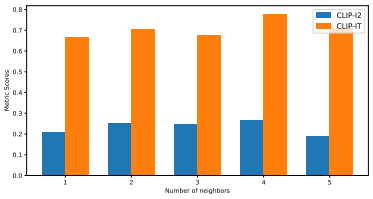

Figure 5. Study on the different number of neighbors on ART500K dataset.

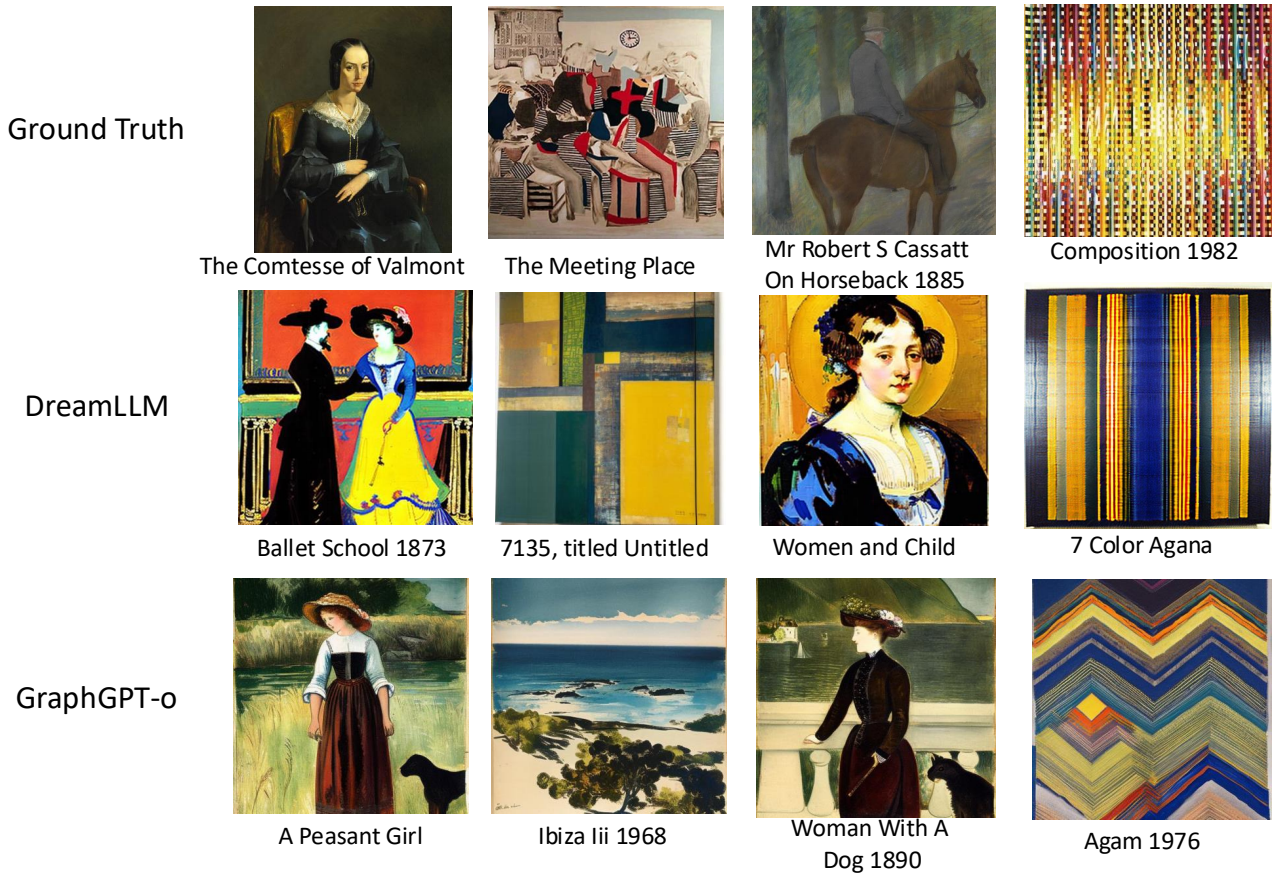

Figure 6. More cases for qualitative evaluation. Our method exhibits better consistency with the ground truth by better utilizing the graph information from neighboring nodes