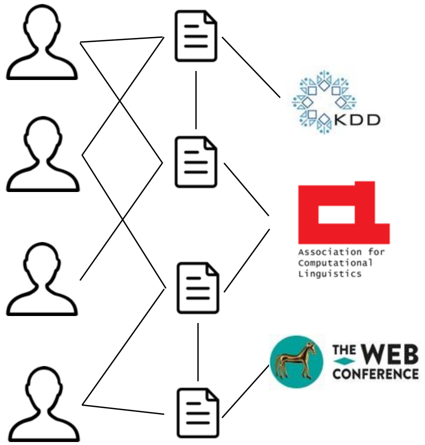
Multimodal Learning on Graphs

Bowen Jin

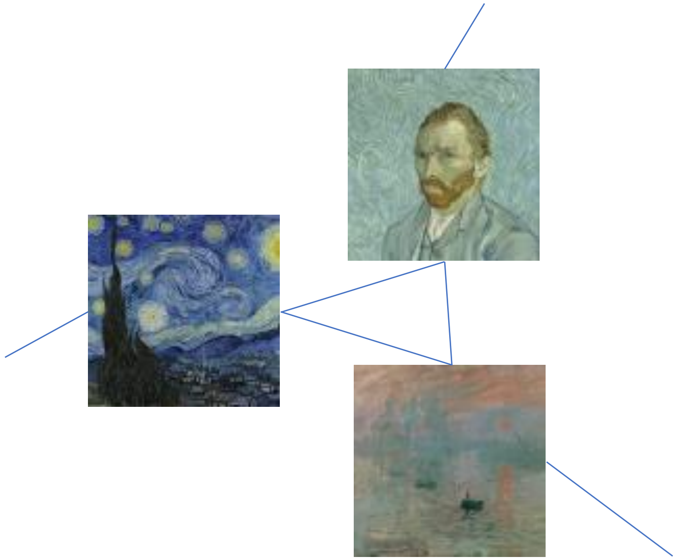
Oct 11, 2024

Introduction

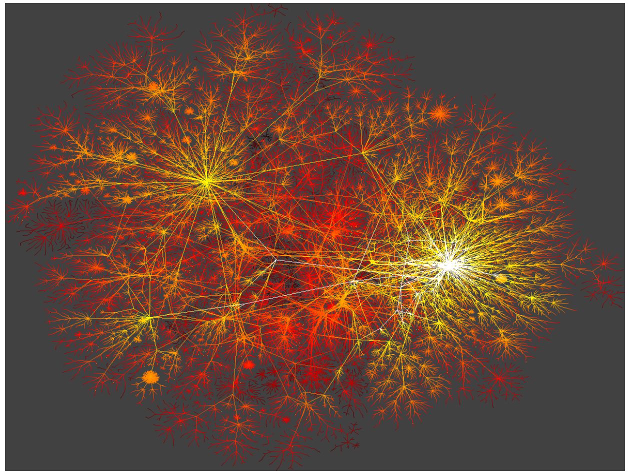
- Graphs in the real world are associated with multimodal attributes.
  - texts, images, videos, …



Academic Graphs



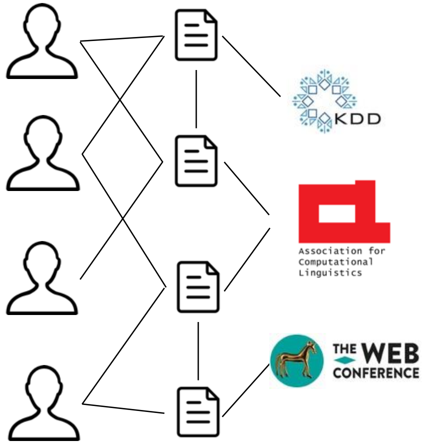
Artwork Graphs



World-Wide Web

Introduction

- Graphs in the real world are associated with multimodal attributes.
  - texts, images, videos, …



Text-attributed graphs

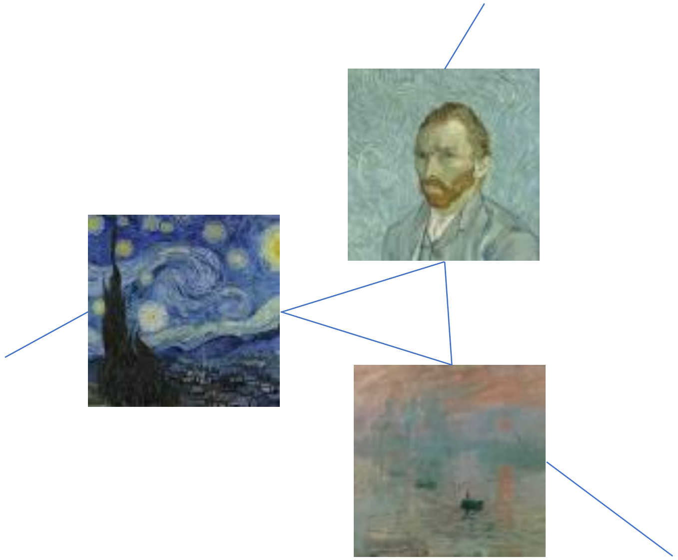
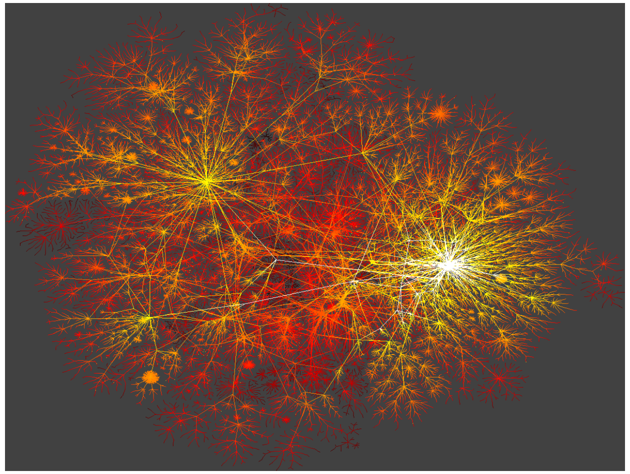


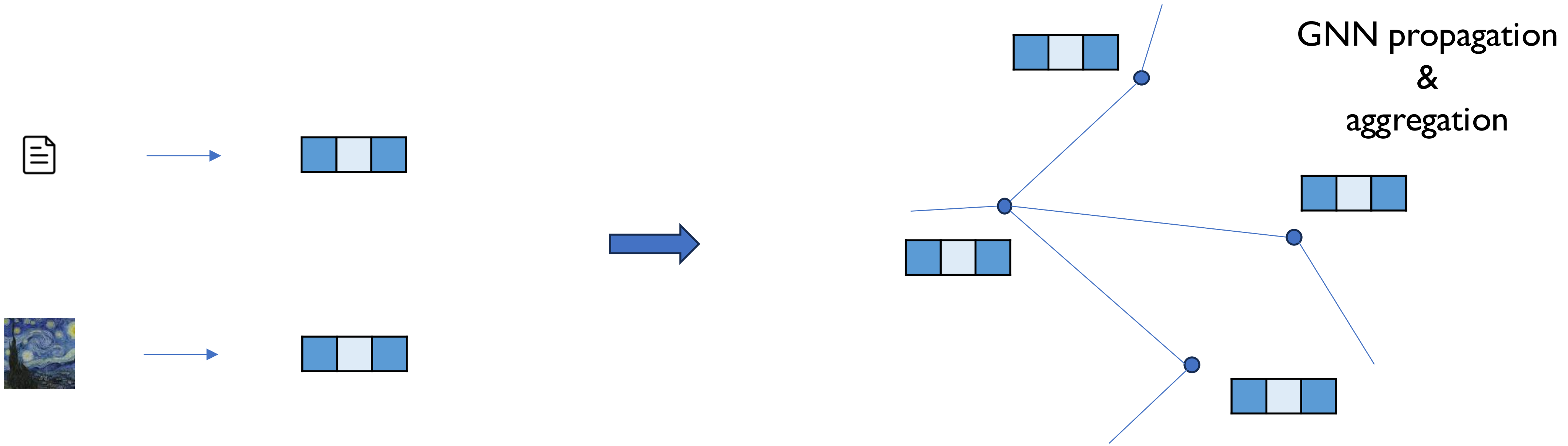
Image-attributed graphs



Multimodal-attributed graphs

Introduction

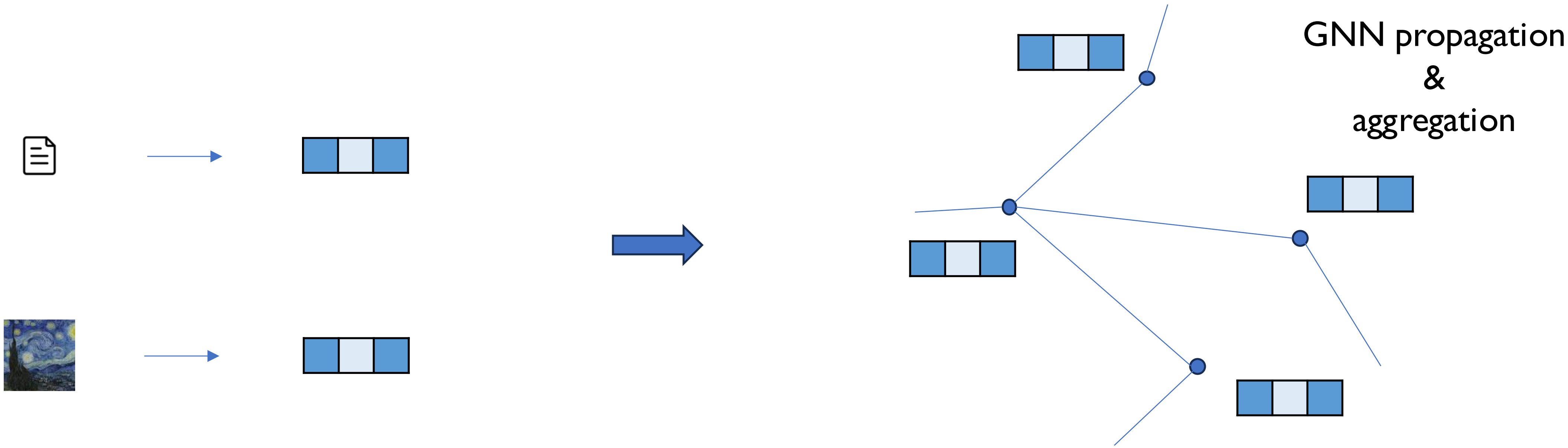
- Graph foundation model: GNN
  - It assumes that the node attributes can be represented as feature vectors.



Limitation 1: The rich node attributes (text, images, …) may not be well captured in a vector.

Introduction

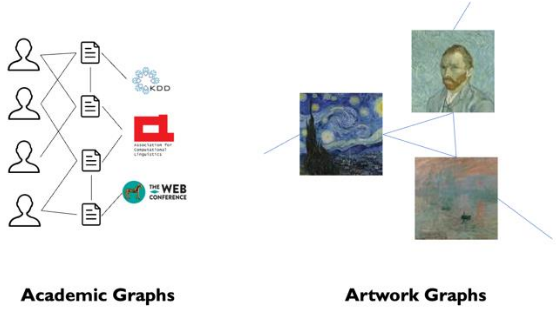
- Graph foundation model: GNN
  - It produces an embedding as output for each node.



Limitation 2: It mainly focus on representation learning tasks, while real-world scenarios can be more complex (text generation/image generation…).

Introduction

- Text / Image foundation model



Large Language Models

- Trained on a large text corpus.
- Expert in text understanding and generation.

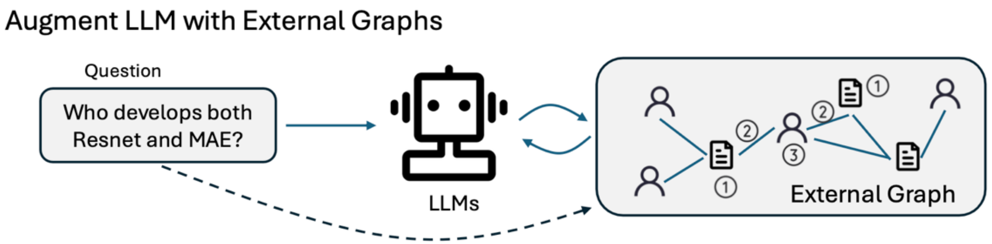
Stable Diffusion Models

- Trained on a large image corpus.
- Expert in image understanding and generation.

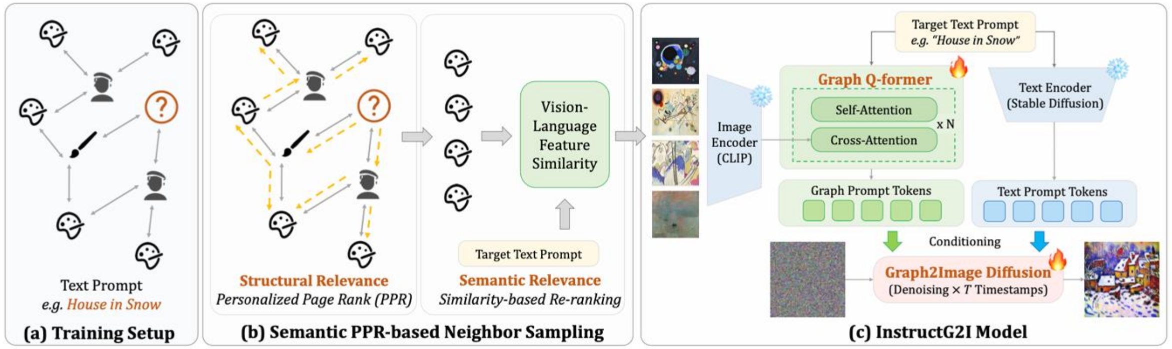
Limitation: They can not well encode the structure information associated with the texts and images.

Introduction

- Our works



Graph Chain-of-Thought



InstructG2I

- Augment the LLMs with an external graph corpus.
- LLMs interact with the graph via callable functions.

- Image generation with stable diffusion conditioned on a multimodal attributed graph.
- Graph context search with semantics-aware PPR, graph encoding with Graph Q-Former.

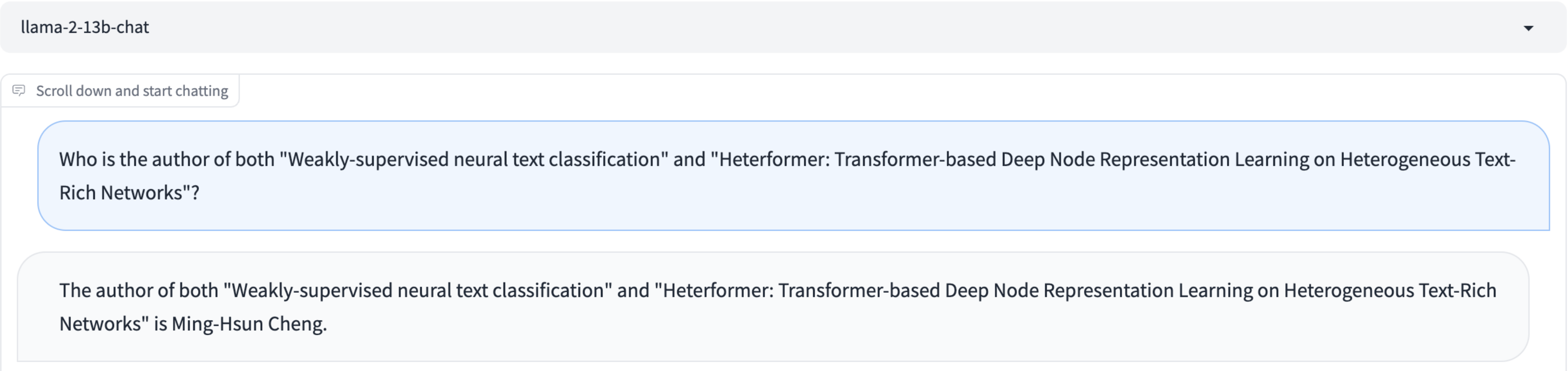
Graph Chain-of-Thought: Augmenting Large Language Models by Reasoning on Graphs

Bowen Jin, Chulin Xie, Jiawei Zhang, Kashob Roy, Yu Zhang, Zheng Li, Ruirui Li, Xianfeng Tang, Suhang Wang, Yu Meng, Jiawei Han

ACL 2024

Introduction

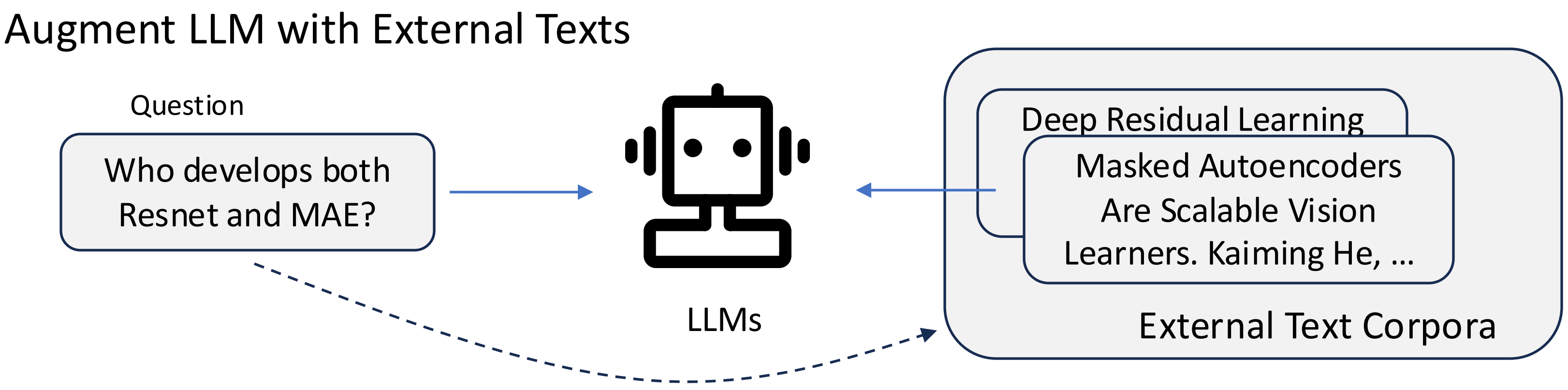
- **Motivation**
  - Large language models suffer from hallucination and misinformation.



Hallucinating!

Introduction

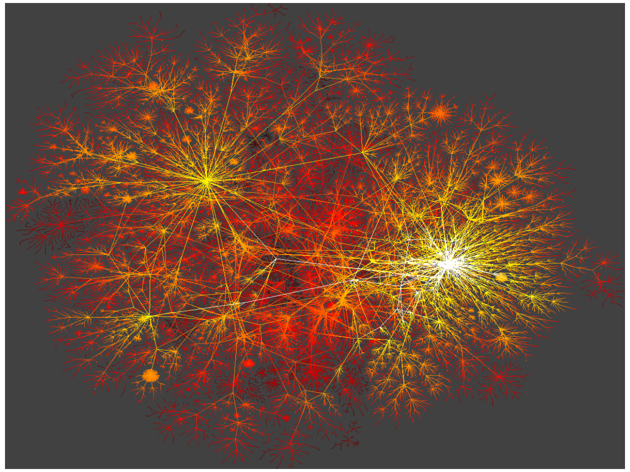
- **Motivation**
  - Existing works propose to augment LLMs with individual text units retrieved from external knowledge corpora to alleviate the issue (RAG).



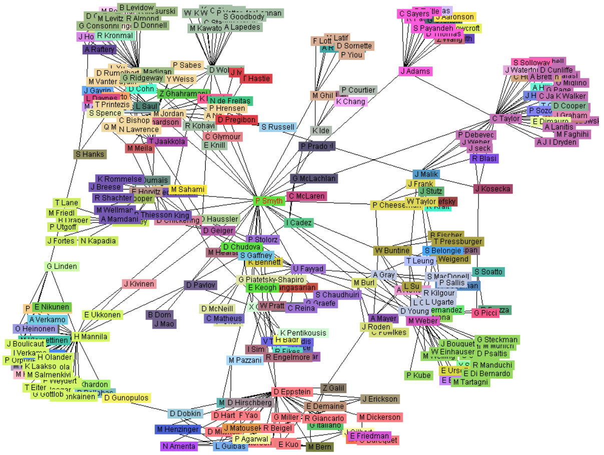
Retrieval-augmented generation (RAG)

Introduction

- **Motivation**
  - However, in many domains, texts are interconnected which form a (text-attributed) graph.
    - Legal case opinions are linked by citation relationships.
    - Web pages are connected by hyperlinks (Common Crawl).



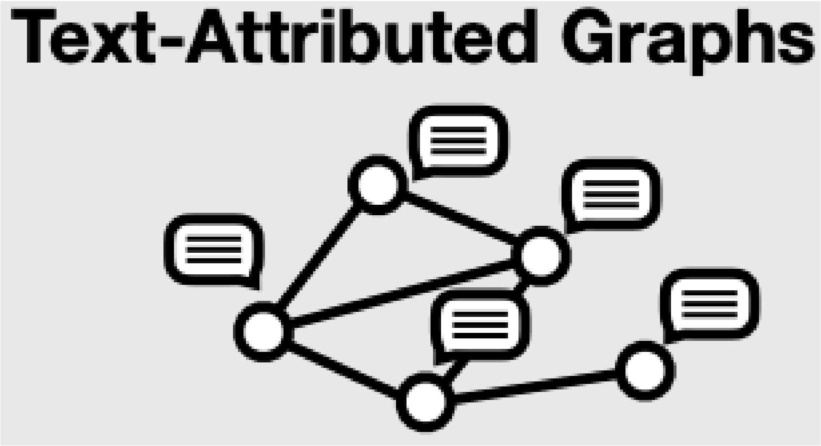
World-Wide Web



Co-author network

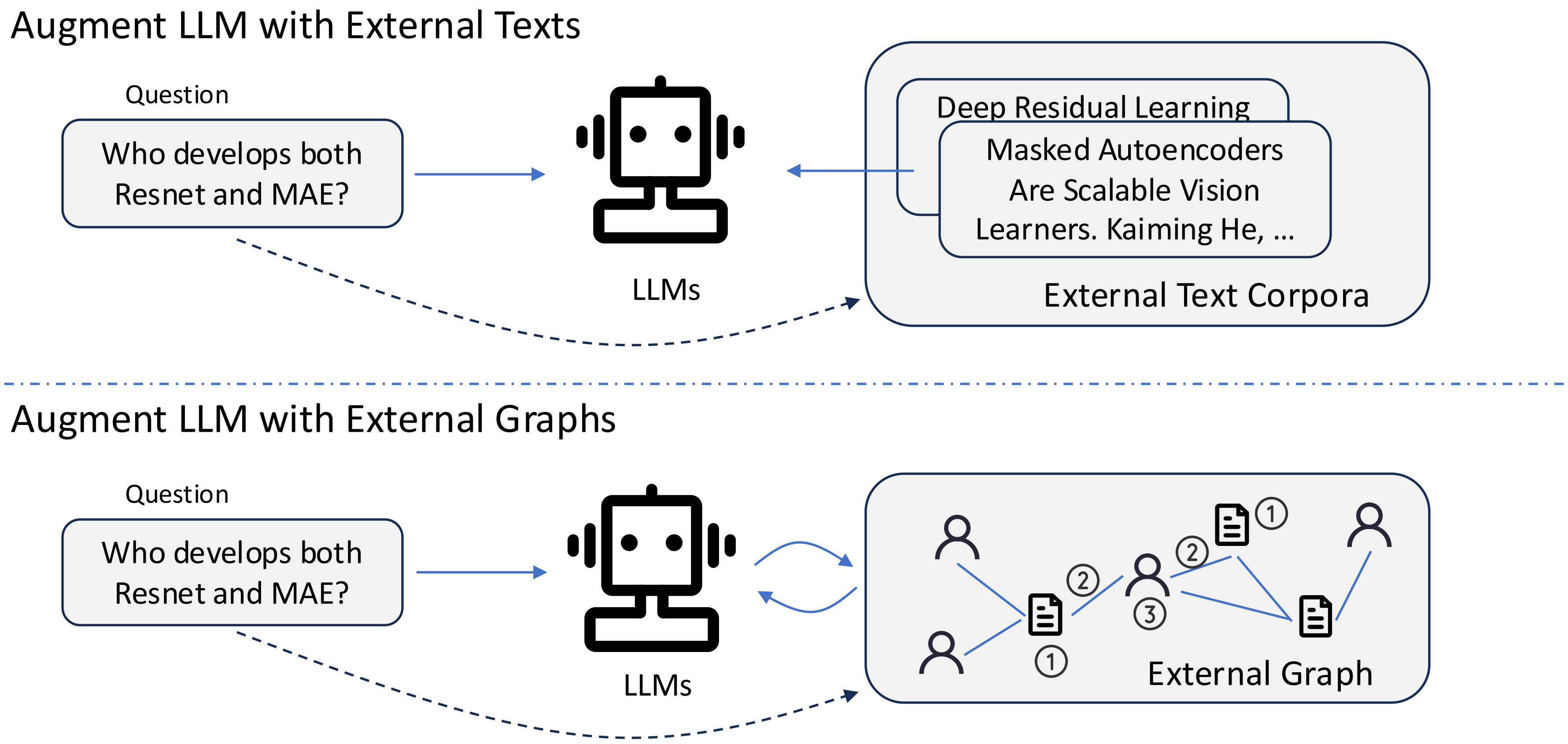
Introduction

- **Motivation**
  - However, in many domains, texts are interconnected which form a (text-attributed) graph.
    - Legal case opinions are linked by citation relationships.
    - Web pages are connected by hyperlinks (Common Crawl).
  - The knowledge in such graphs is encoded not only in single texts/nodes but also in their associated connections.



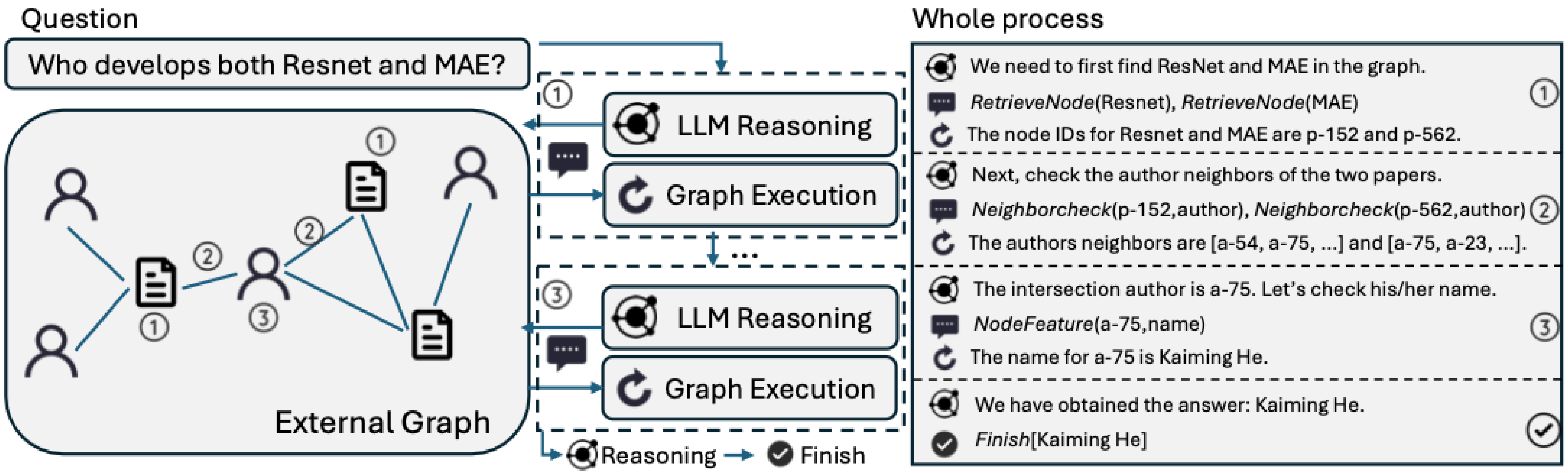
Introduction

- **Motivation**
  - This motivates us to explore the problem of augmenting LLMs with external graphs.



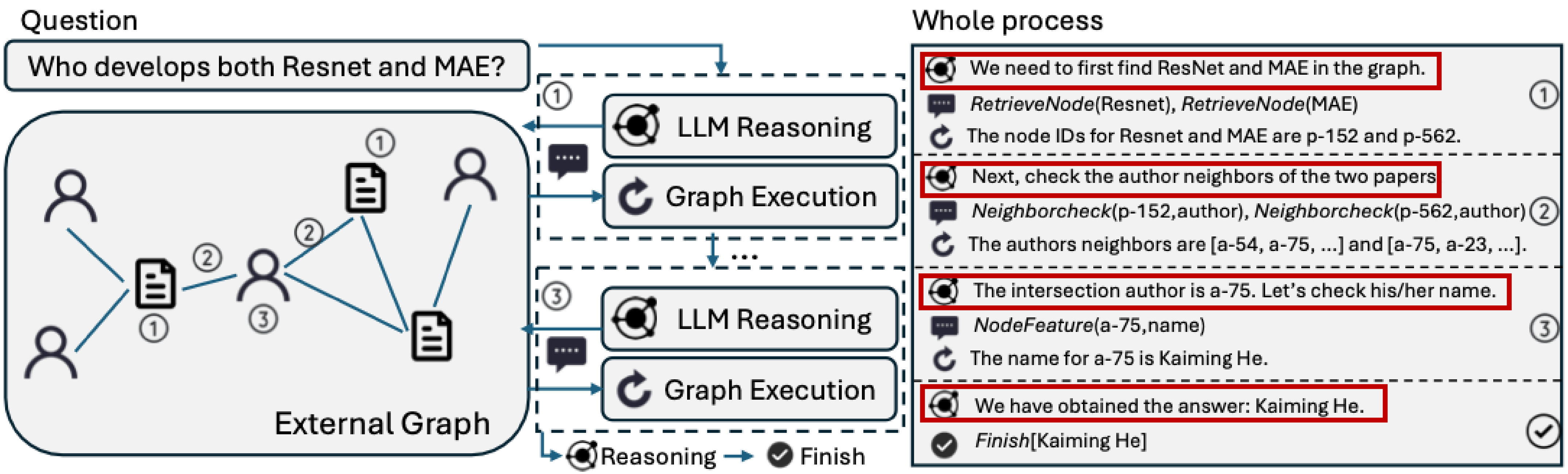
Graph Chain-of-Thought

- **Framework**
  - Iterative reasoning, interaction and execution.



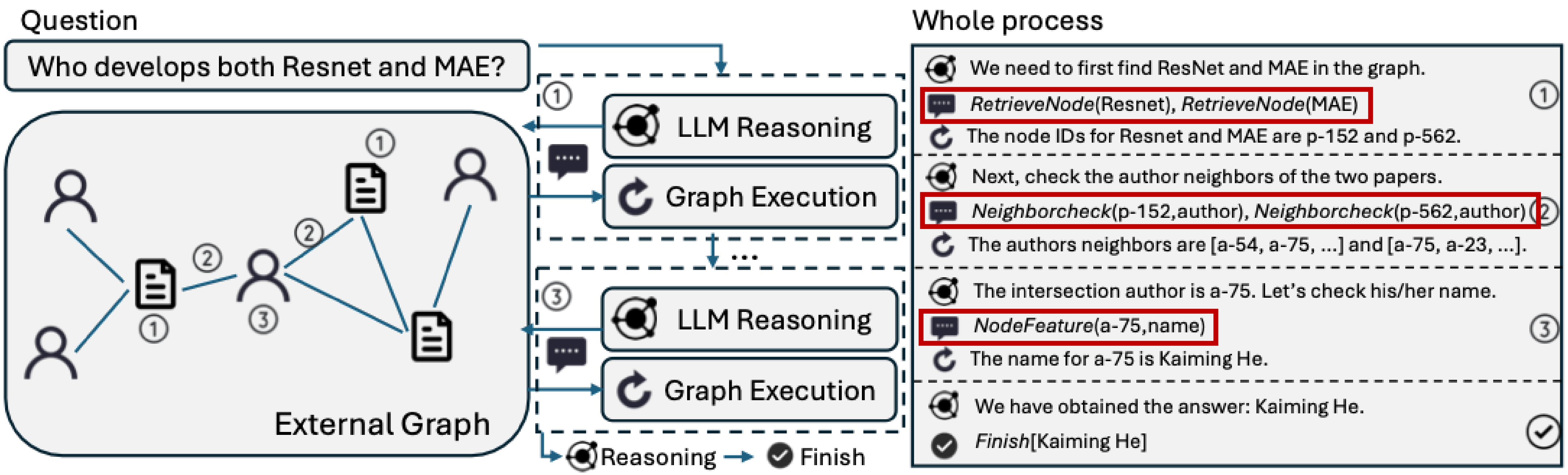
Graph Chain-of-Thought

- **LLM reasoning**
  - LLM conduct reasoning on what further external information from graph is needed.
  - If the question is answerable with the current contexts from graphs.



Graph Chain-of-Thought

- **Interaction between LLMs and graphs**
  - Let LLMs know how to interact with the graphs and fetch relevant information.

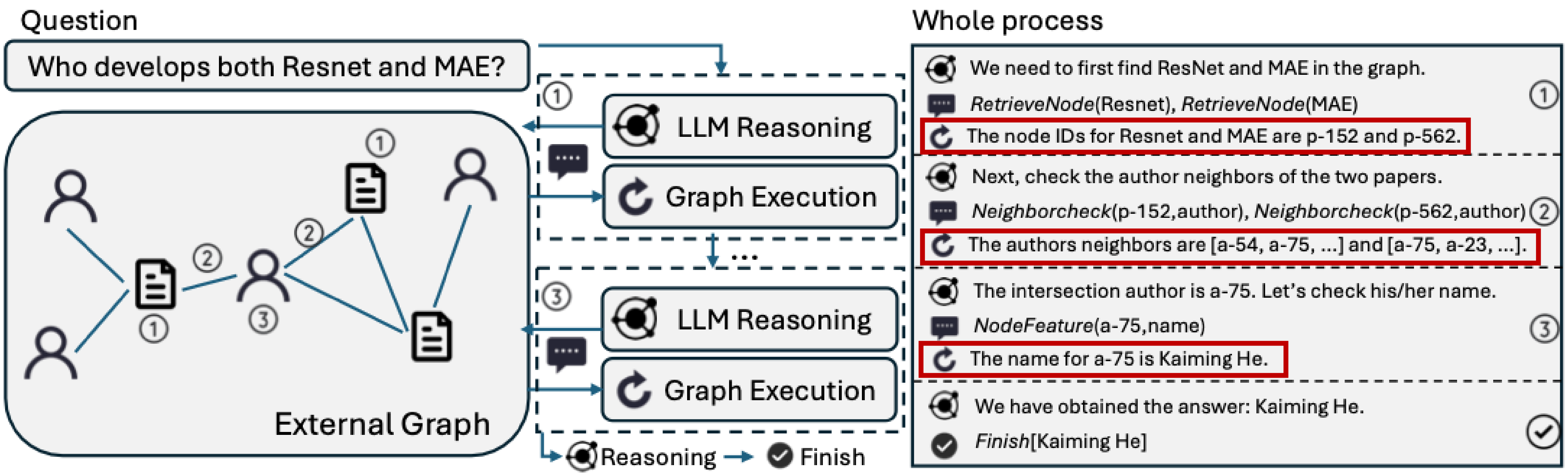


Graph Chain-of-Thought

- **Interaction between LLMs and graphs**
  - We pre-define four graph functions to cover both the semantic and structure information on graphs:
    - RetrieveNode(Text): Identify related nodes in the graph with semantic search.
    - NodeFeature(NodeID, FeatureName): Extract the textual feature information for a specific node.
    - NeighborCheck(NodeID, NeighborType): Return the neighboring information for a specific node.
    - NodeDregree(NodeID, NeighborType): Return the degree of a specific neighbor type for a node.

Graph Chain-of-Thought

- **Execution on graphs**
  - Call the functions and fetch relevant information from the graph.



Experiments

- **Overall performance**

	Model	Academic		E-commerce		Literature		Healthcare		Legal	
		R-L	GPT4score	R-L	GPT4score	R-L	GPT4score	R-L	GPT4score	R-L	GPT4score
Base	LLaMA-2-13b-chat	8.13	8.03	7.01	12.00	5.32	20.83	5.25	13.70	15.97	16.11
	Mixtral-8x7b	9.02	8.14	12.54	18.00	7.50	22.50	3.88	20.00	12.74	16.11
	GPT-3.5-turbo	6.05	12.80	9.18	23.50	10.43	26.67	5.83	14.44	10.51	20.00
Text RAG	LLaMA-2-13b-chat	8.69	8.52	9.23	12.50	7.61	20.00	1.44	5.93	15.37	16.67
	Mixtral-8x7b	8.44	8.02	23.14	29.50	13.35	27.92	3.22	16.67	19.69	25.00
	GPT-3.5-turbo	5.83	9.91	14.06	20.00	10.04	20.83	4.57	8.52	18.14	23.89
Graph RAG	LLaMA-2-13b	22.01	22.97	12.48	20.00	9.25	20.00	2.97	4.81	17.98	17.22
	Mixtral-8x7b	27.77	31.20	32.87	37.00	20.08	33.33	8.66	15.19	23.48	25.56
	GPT-3.5-turbo	18.45	26.98	17.52	28.00	14.94	24.17	8.69	14.07	18.66	22.22
	GRAPH-COT	**31.89**	**33.48**	**42.40**	**44.50**	**41.59**	**46.25**	**22.33**	**28.89**	**30.52**	**28.33**

- Graph-CoT outperforms all the baselines consistently and significantly.
- Base LLMs are exhibiting fairly poor performance, typically because the LLMs may not contain the knowledge needed to answer those questions.
- Graph RAG LLMs outperform text RAG LLMs in most cases since the former can provide more structure-aware context.

Experiments

- **How different LLMs perform in Graph-CoT?**

Model	**GPT4score**
GRAPH-COT	
w. LLaMA-2-13b-chat	16.04
w. Mixtral-8x7b	36.46
w. GPT-3.5-turbo	36.63
w. GPT-4	46.28

- An LLM with more advanced instruction following ability and reasoning ability (i.e., GPT-4) can contribute to better performance in Graph-CoT.

Experiments

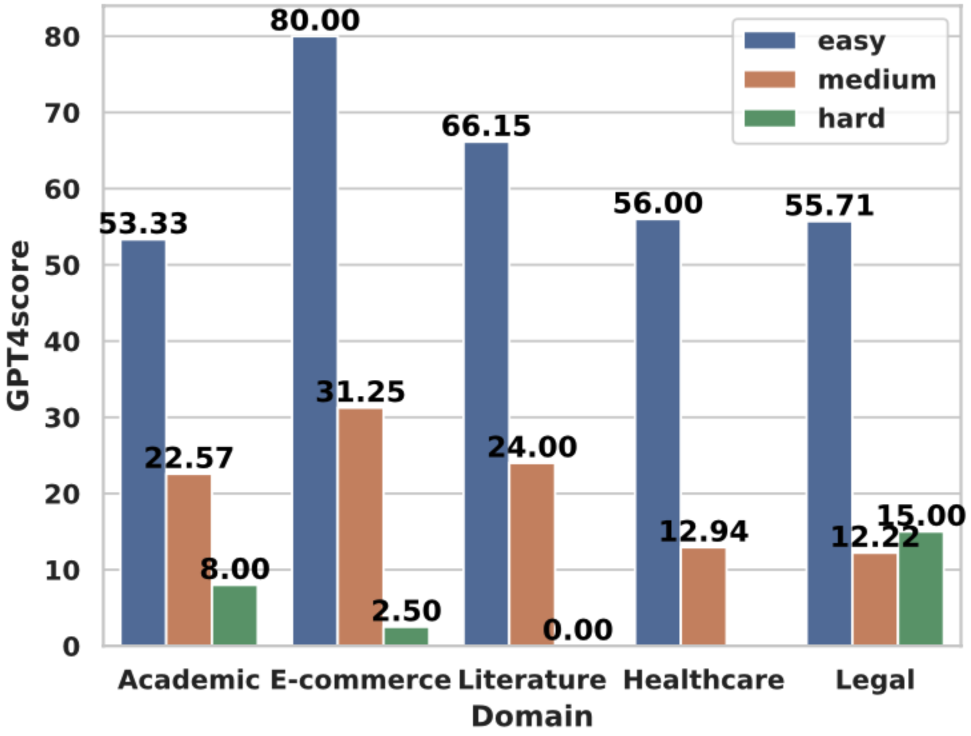
- **Graph RAG vs Graph-CoT**

Model	**GPT4score**
GPT-3.5-turbo	19.48
+ node retrieval	16.63
+ 1-hop subgraph retrieval	*23.09*
+ 2-hop subgraph retrieval	22.12
+ GRAPH-COT	**36.29**

- Retrieving 1-hop ego-graph performs the best, but still underperforms Graph-CoT.
- The number of nodes/texts grow exponentially as the hop number grows linearly.
- A large-hop ego-graph will lead to a super long context -> lost in the middle.

Experiments

- **Graph-CoT on questions of different difficulty levels**



- Graph-CoT performs relatively high on easy question (simple reasoning chain) while having worse performance on medium/hard questions (complex/inductive reasoning).

InstructG2I: Synthesizing Images from Multimodal Attributed Graphs

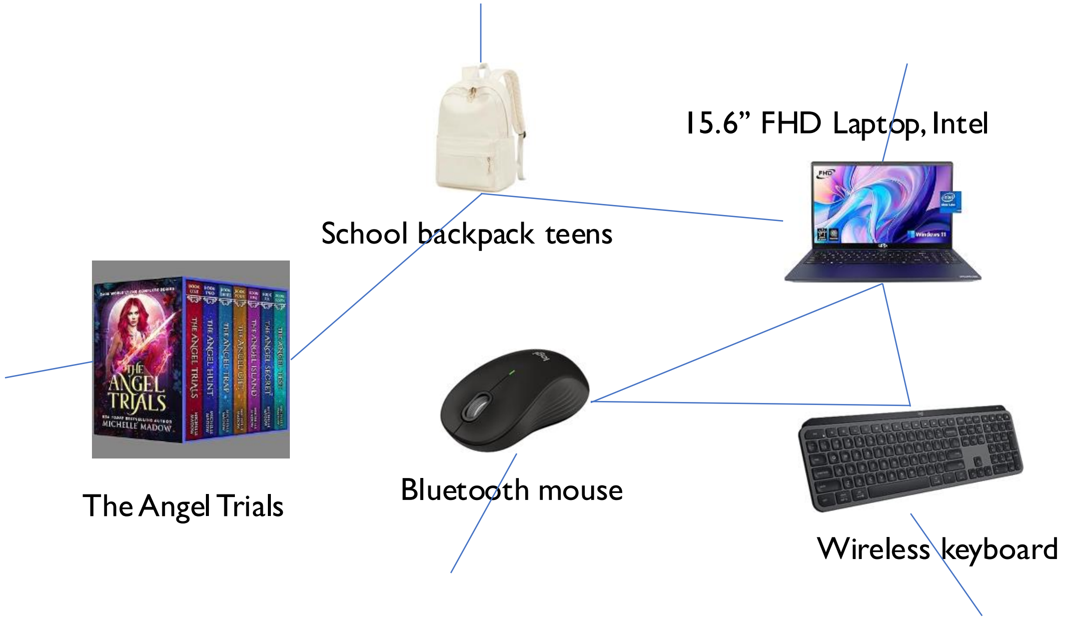
Bowen Jin, Ziqi Pang, Bingjun Guo, Yu-Xiong Wang, Jiaxuan You, Jiawei Han

NeurIPs 2024

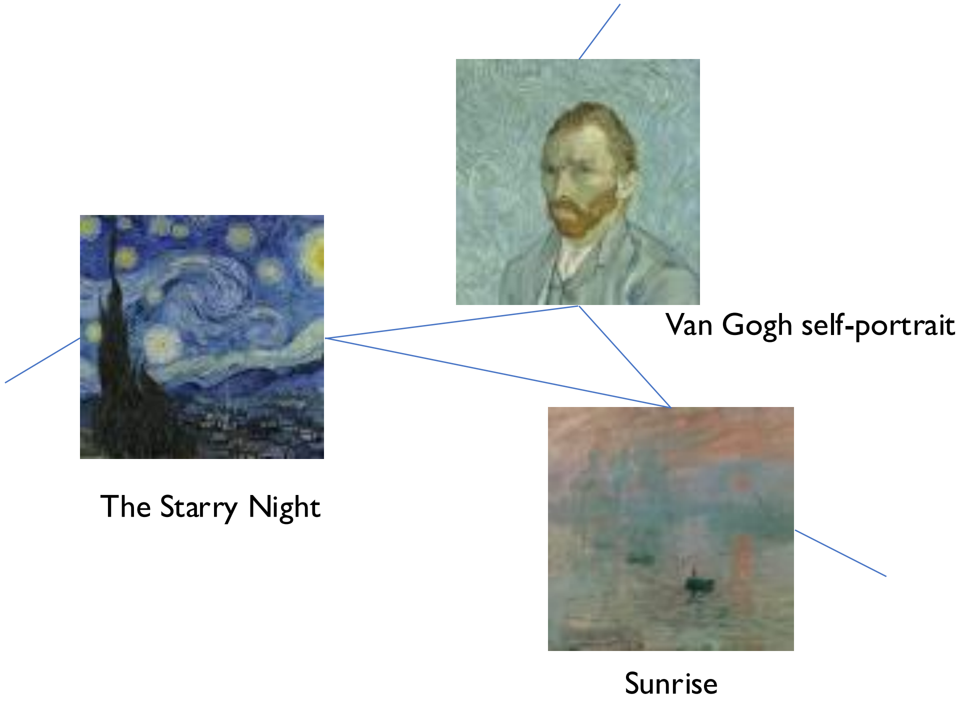
website: instructg2i.github.io

Introduction

- **Background**
  - In real world graphs, nodes are associated with text and image information ("multimodal attributed graphs").
  - E.g., product graphs in e-commerce, picture graphs in art domain.
  - Prev., we mainly focus on graphs with "text" ("text-attributed graph").



product graphs



art graphs

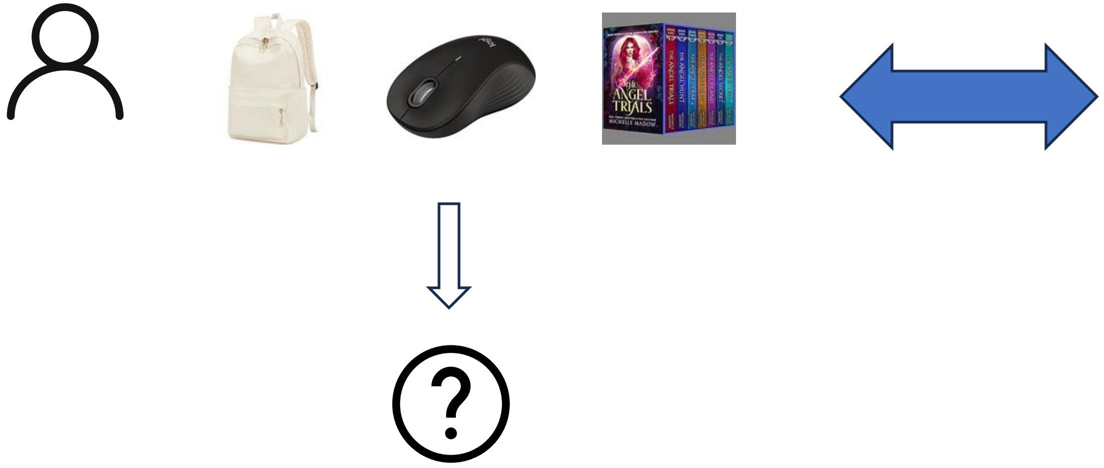
Problem

- **How we conduct node image generation on such graph?**
  - **Application on E-commerce**

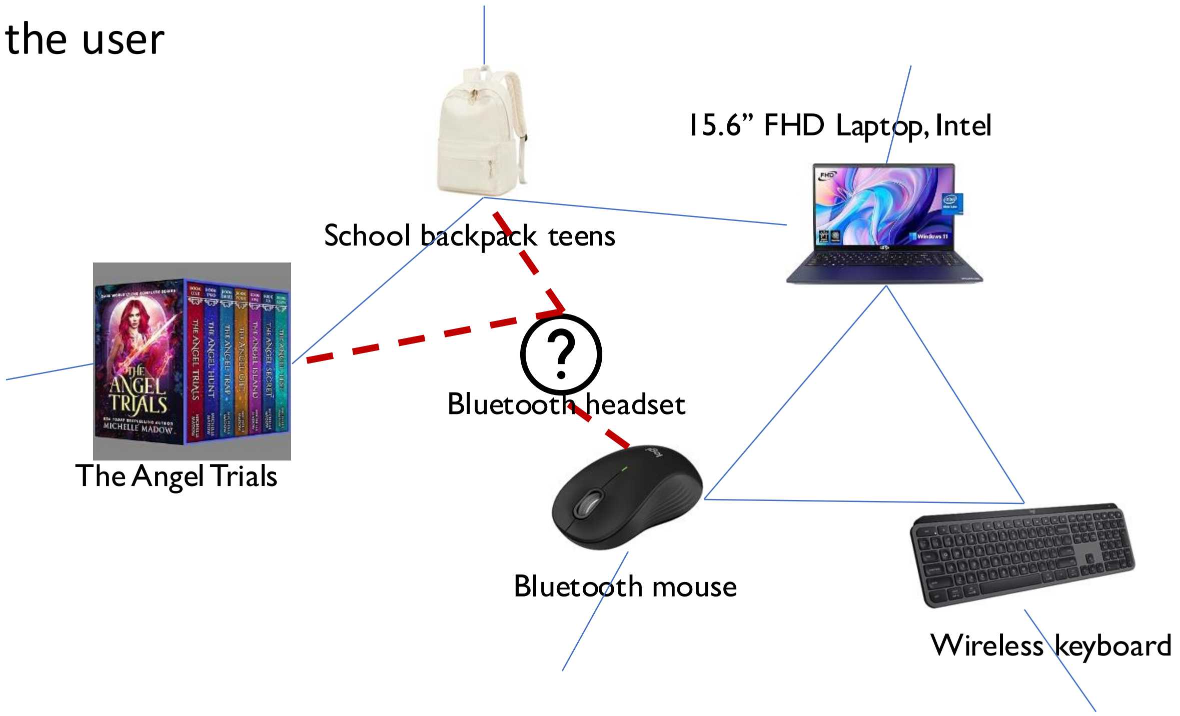
Generative recommendation

What the future item "Bluetooth headset" the user will be interested in looks like?

Purchase history



Bluetooth headset



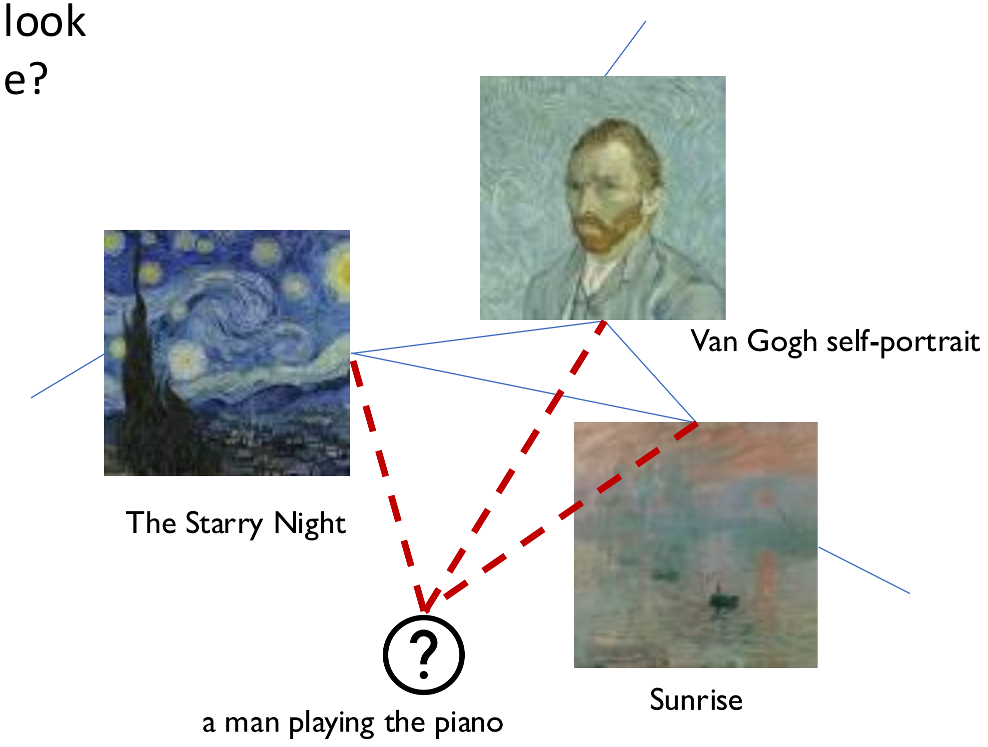
Problem

- **How we conduct node image generation on such graph?**
  - **Application on Art domain**

Virtual art creation
How will a picture titled "a man playing the piano" look like with 50% Monet style and 50% Van Gogh style?

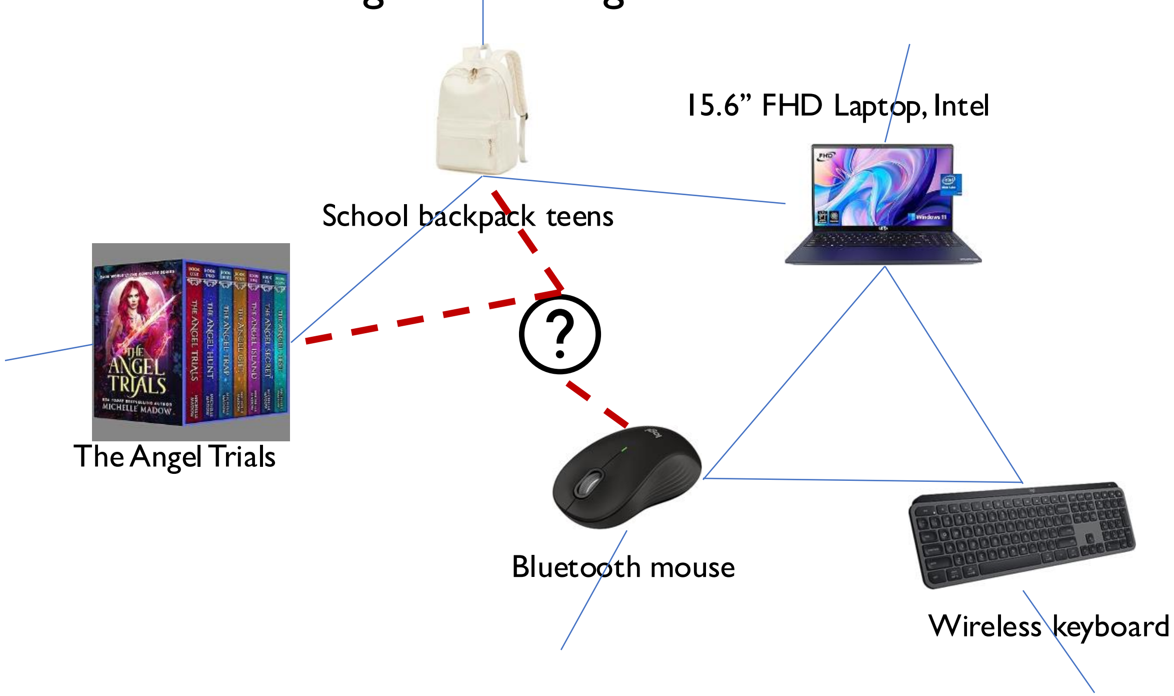
Monet arts & Van Gogh arts

a man playing the piano

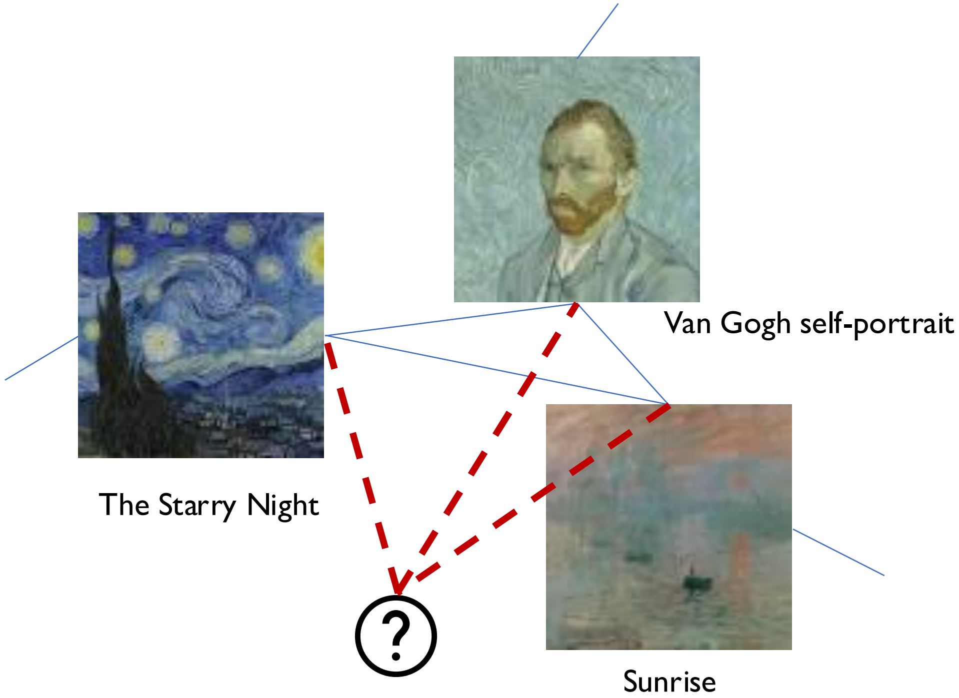


Problem

- **Task: Synthesizing Images from Multimodal Attributed Graphs**
  - Input:
    - A graph with multimodal attributes.
    - The neighbors of the target node on the graph.
    - Text description for the target node.
  - Output:
    - The image of the target node.



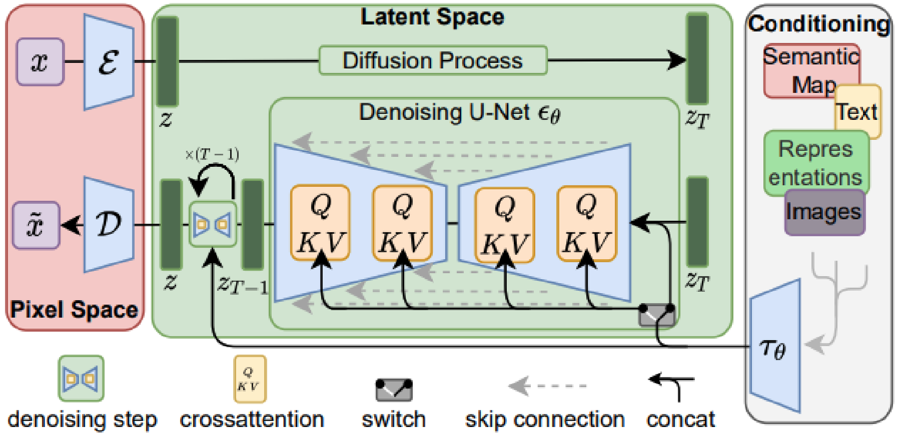
Generative recommendation



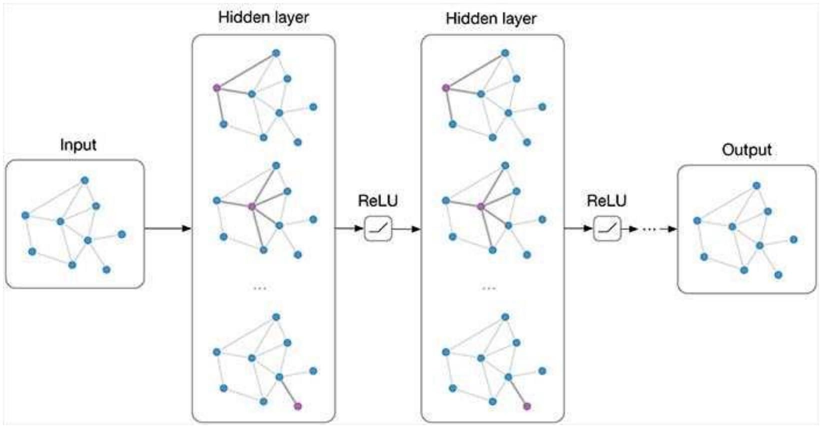
Virtual art creation

Problem

- **Existing works**
  - **Image generation with conditions**
    - Text-to-image generation: stable diffusions
    - Image-to-image generation: ControlNet, InstructPix2pix
    - No work on conditioning on graphs

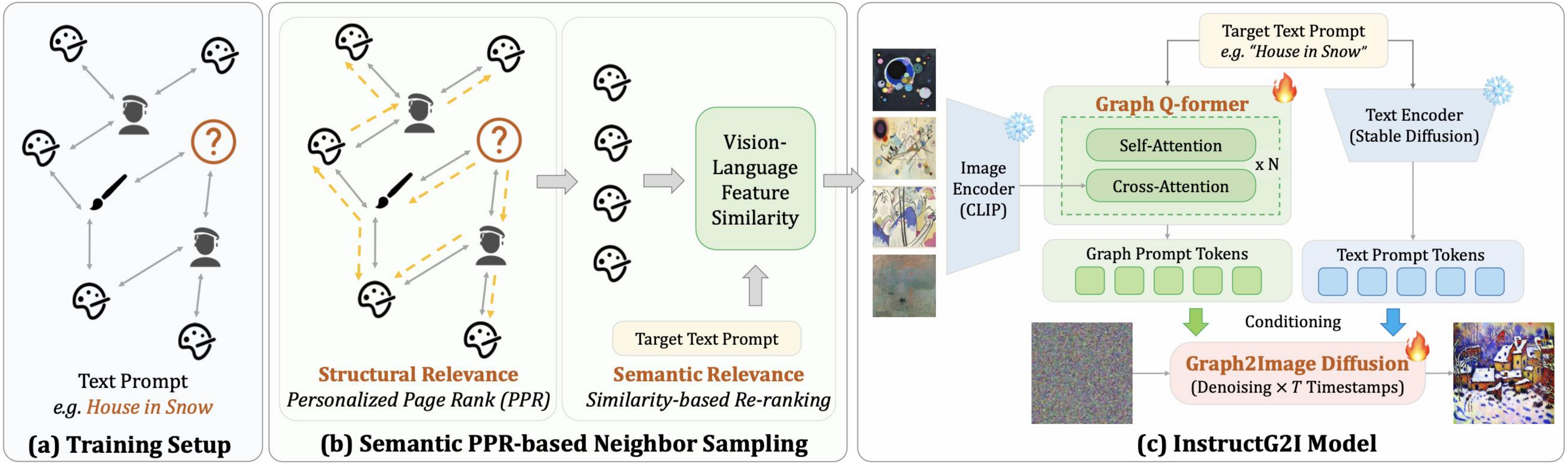


  - **Graph Neural Network**
    - GCN, GraphSAGE, …
    - They mainly focus on representation learning
    - Cannot handle generation tasks



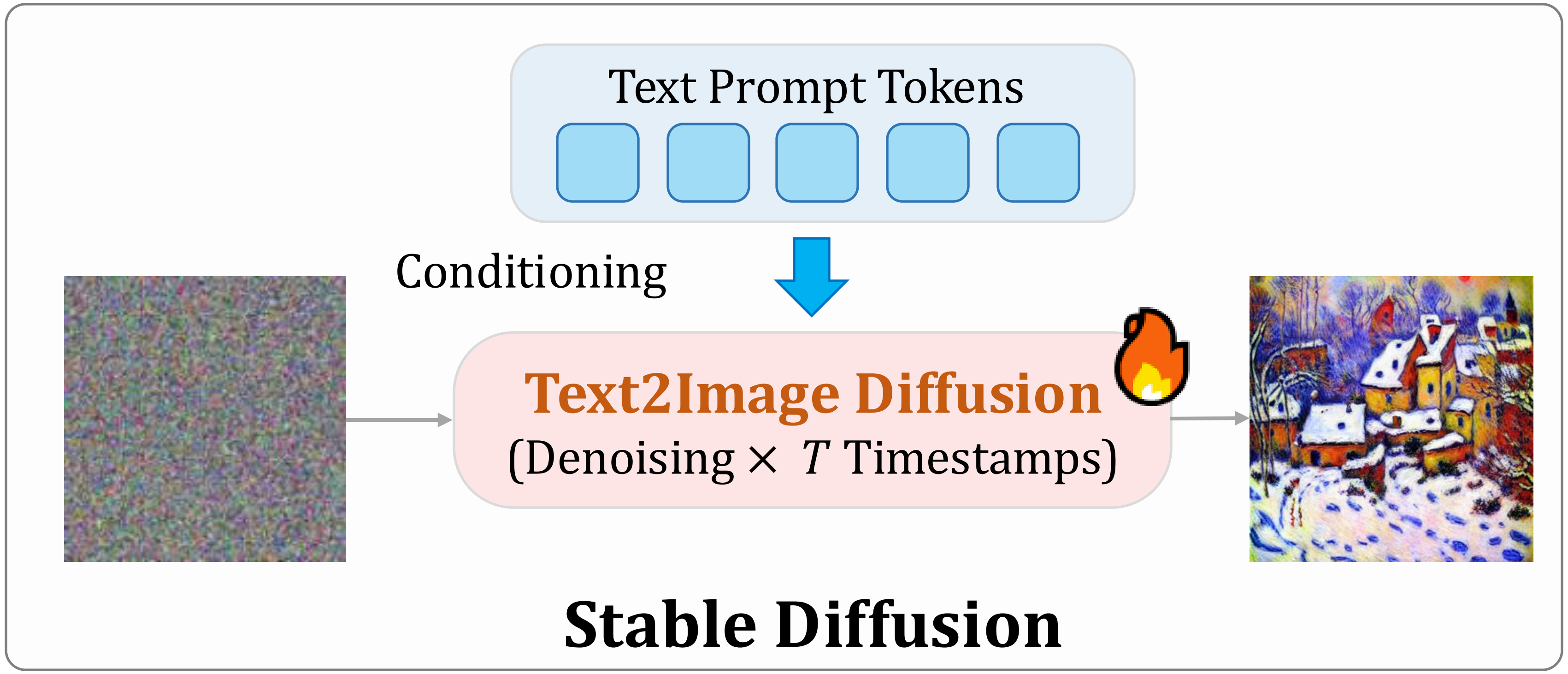
InstructG2I

- **Model Overview**



InstructG2I

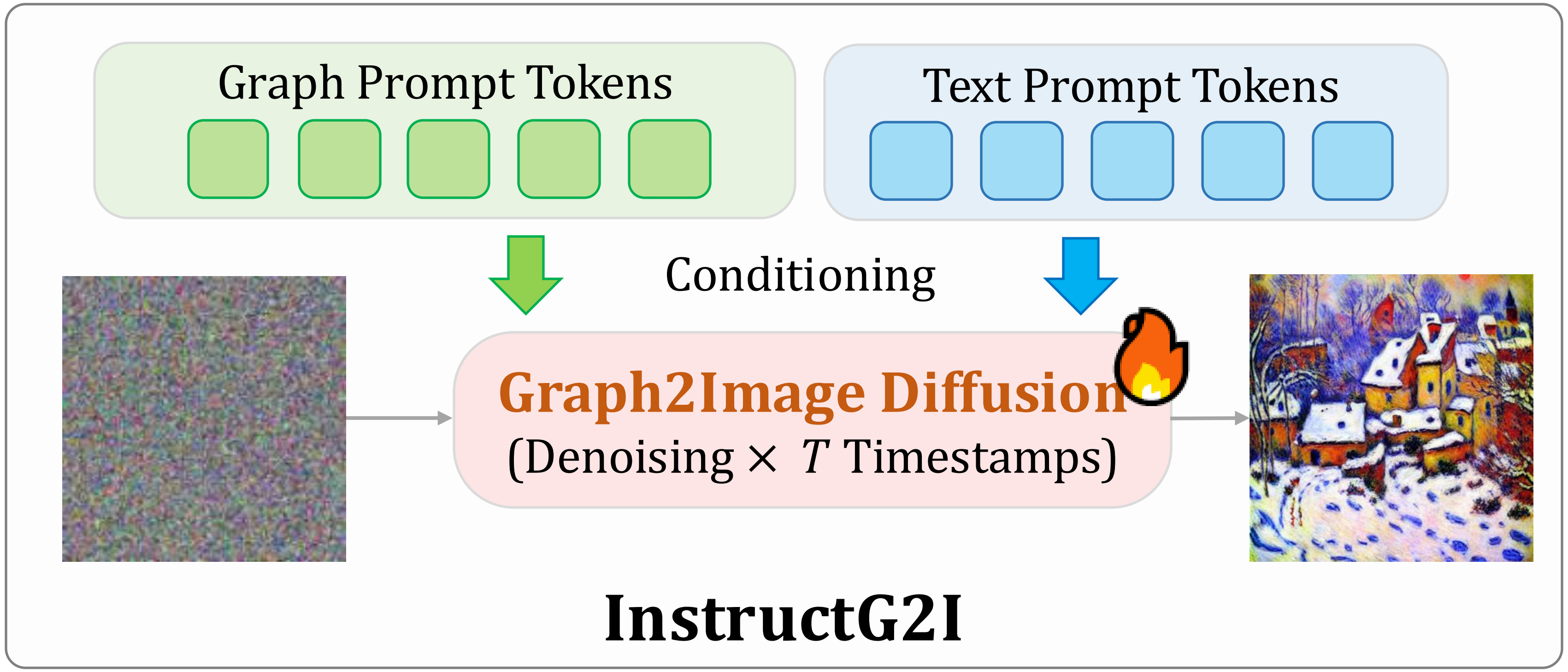
- **Stable diffusion (SD)**



$$\mathcal{L} = \mathbb{E}_{\mathbf{z}\sim \text{Enc}(x), c_T, \epsilon\sim\mathcal{N}(0,1), t}\left[\left\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, h(c_T))\right\|^2\right].$$

InstructG2I

- **Graph context-conditioned stable diffusion**

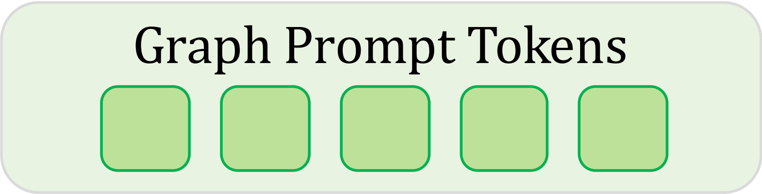


$$h(c_T, c_G) = [h_T(c_T), h_G(c_G)] \in \mathbf{R}^{d \times (l_{c_T} + l_{c_G})}$$

$$\mathcal{L} = \mathbb{E}_{\mathbf{z} \sim \mathrm{Enc}(x), c_T, c_G, \epsilon \sim \mathcal{N}(0,1), t} \left[\| \epsilon - \epsilon_\theta(\mathbf{z}_t, t, h(c_T, c_G)) \|^2 \right]$$

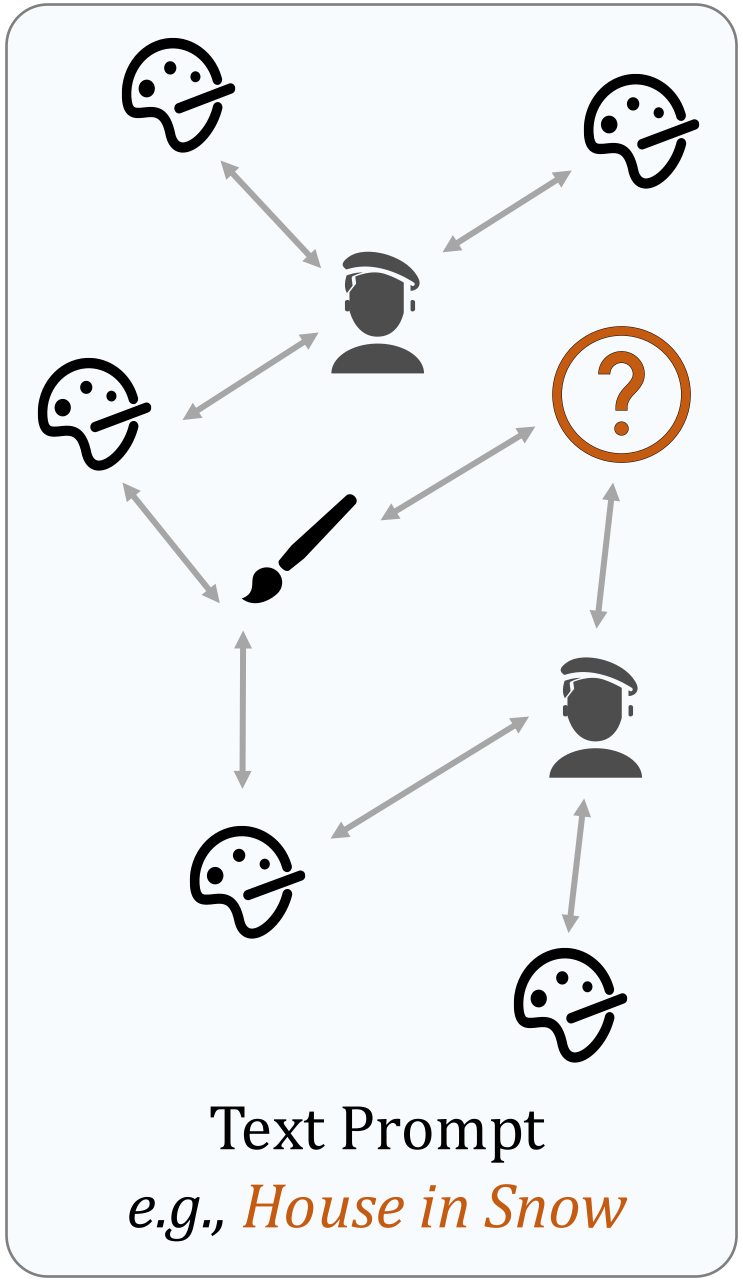
InstructG2I

- **How to get "Graph Prompt Tokens"?**



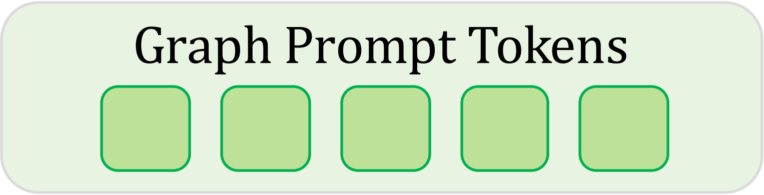
1. Find relevant context from the graph.
   -- **Semantic PPR-based Neighbor Sampling**

2. Compress graph context into tokens.
   -- **Graph Encoding with Text Conditions**



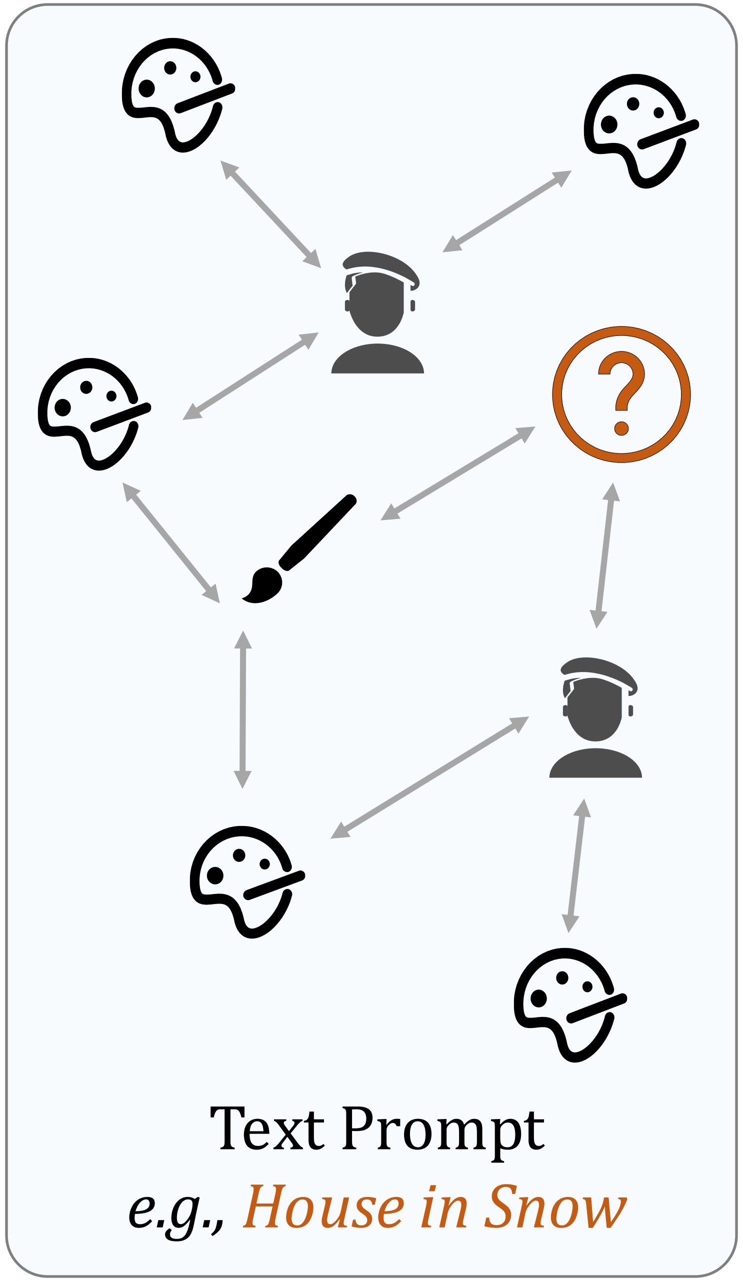
InstructG2I

- **How to get "Graph Prompt Tokens"?**



1. Find relevant context from the graph.
  -- **Semantic PPR-based Neighbor Sampling**

2. Compress graph context into tokens.
  -- **Graph Encoding with Text Conditions**



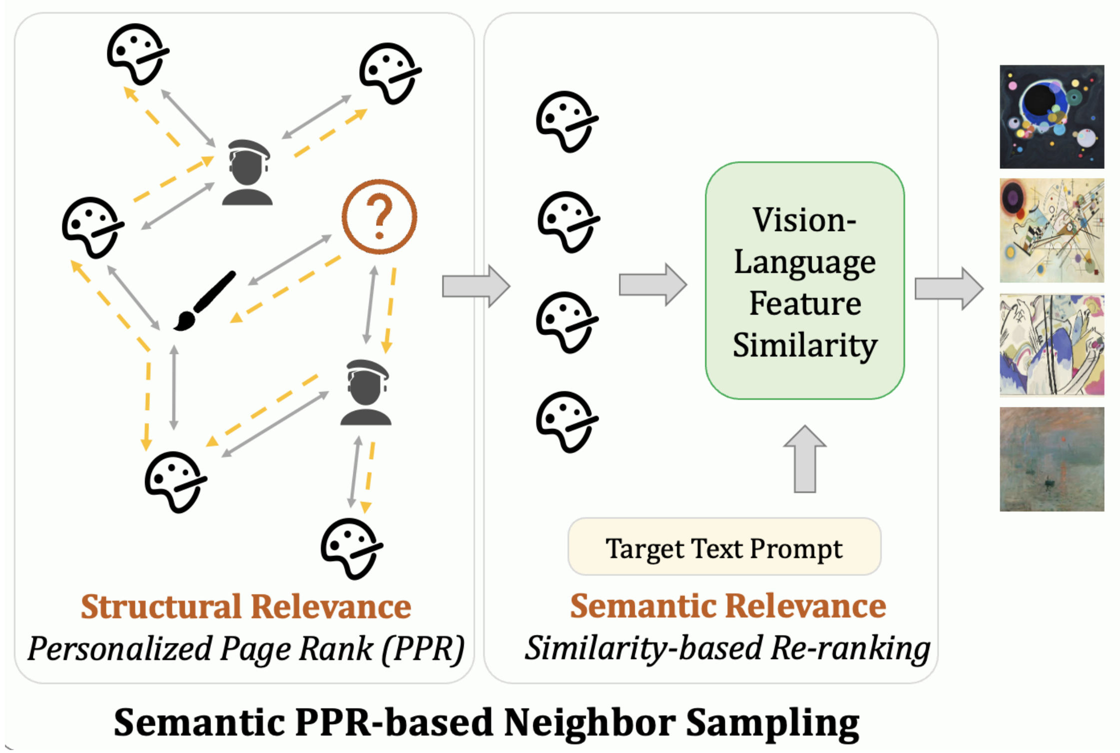
InstructG2I

- **Semantic PPR-based Neighbor Sampling**

Goal: Find relevant context from the graph for target node image generation.

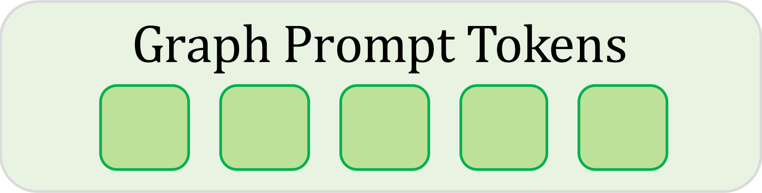
Step1: Structure relevance with Personalized Page Rank (PPR).

Step2: Semantic relevance with content similarity calculation.



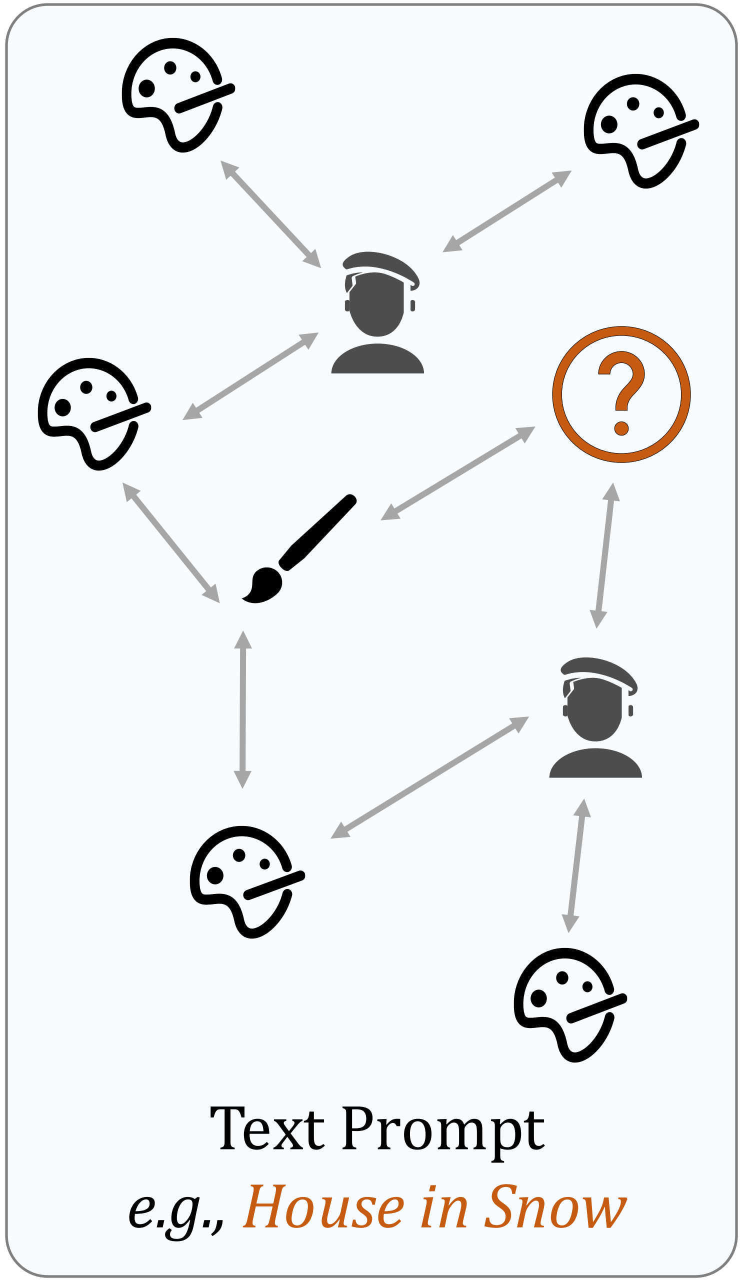
InstructG2I

- **How to get "Graph Prompt Tokens"?**



1. Find relevant context from the graph.
   -- **Semantic PPR-based Neighbor Sampling**

2. Compress graph context into tokens.
   -- **Graph Encoding with Text Conditions**



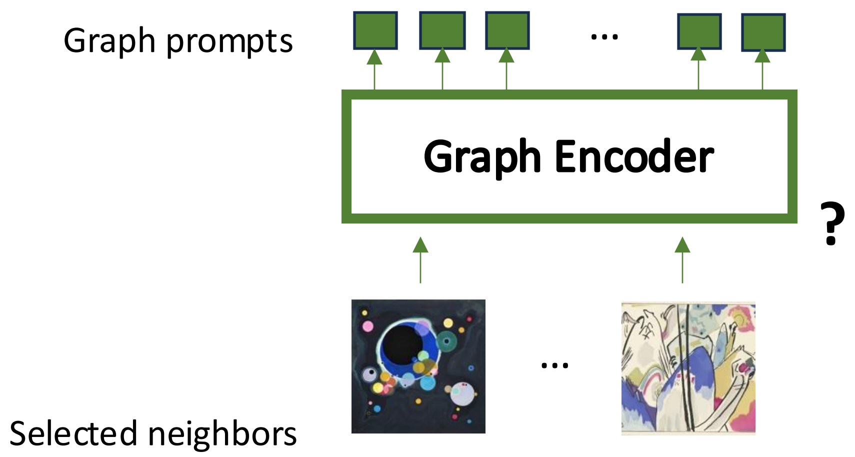
InstructG2I

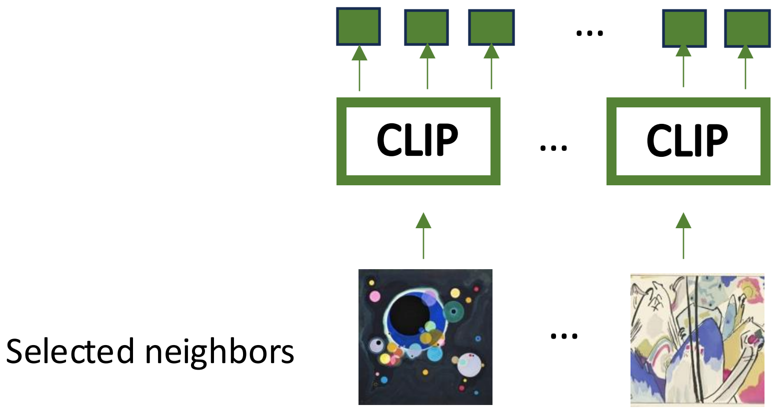
- **Graph Encoding: a simple baseline**

Goal: Compress graph context into tokens.

Cons:

- The neighbor feature extraction is isolated.
- The extracted features are general. They should be conditioned on our target goal (text prompt).

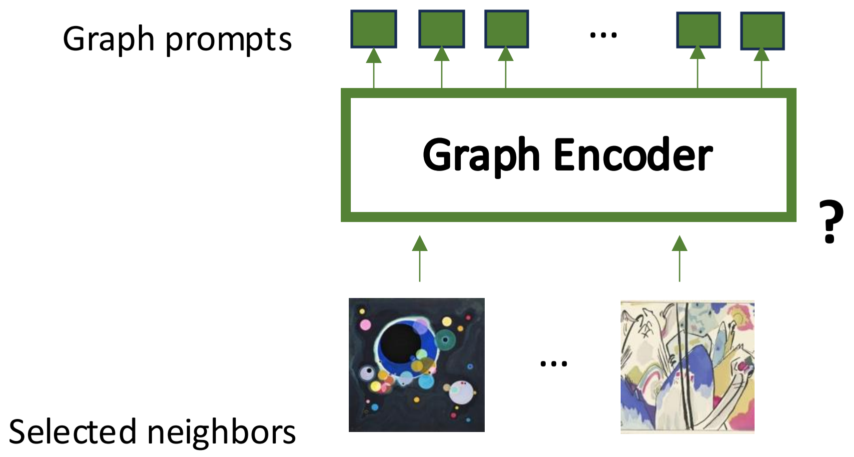


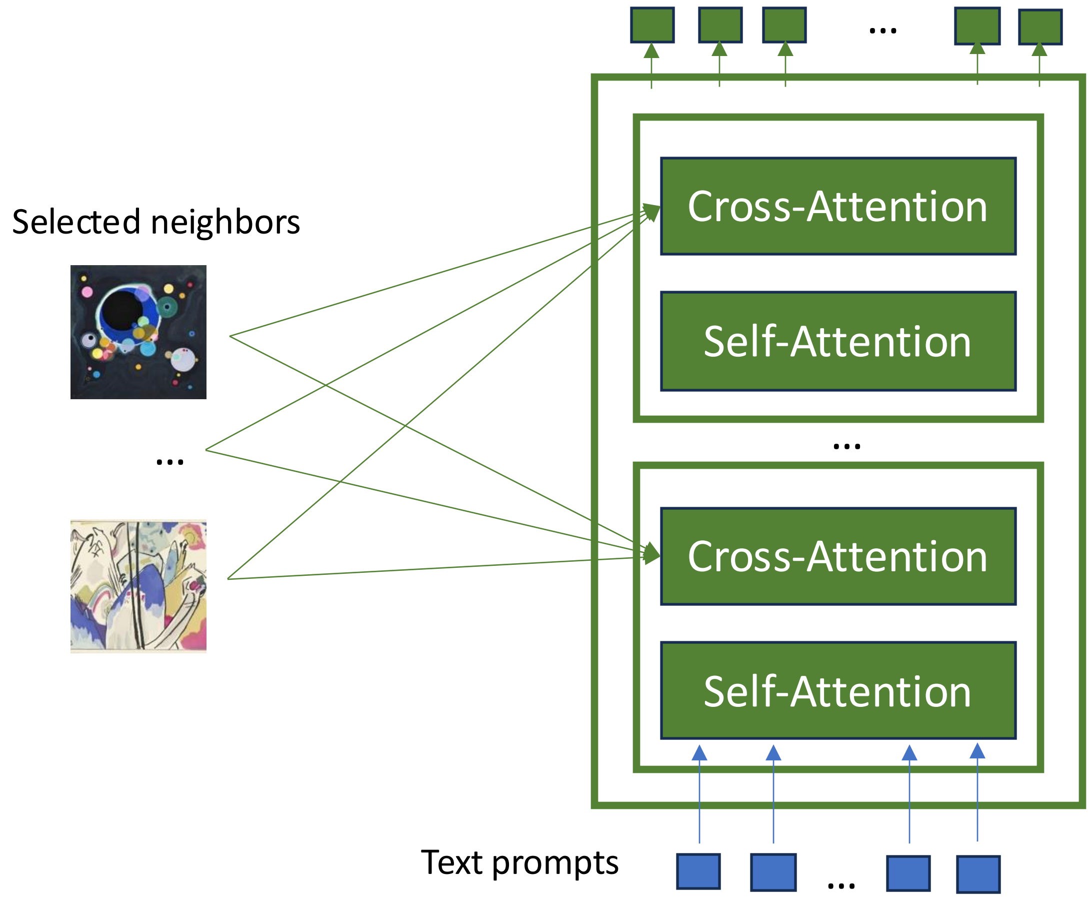


a simple baseline

InstructG2I

- **Graph Encoding with Text Conditions**

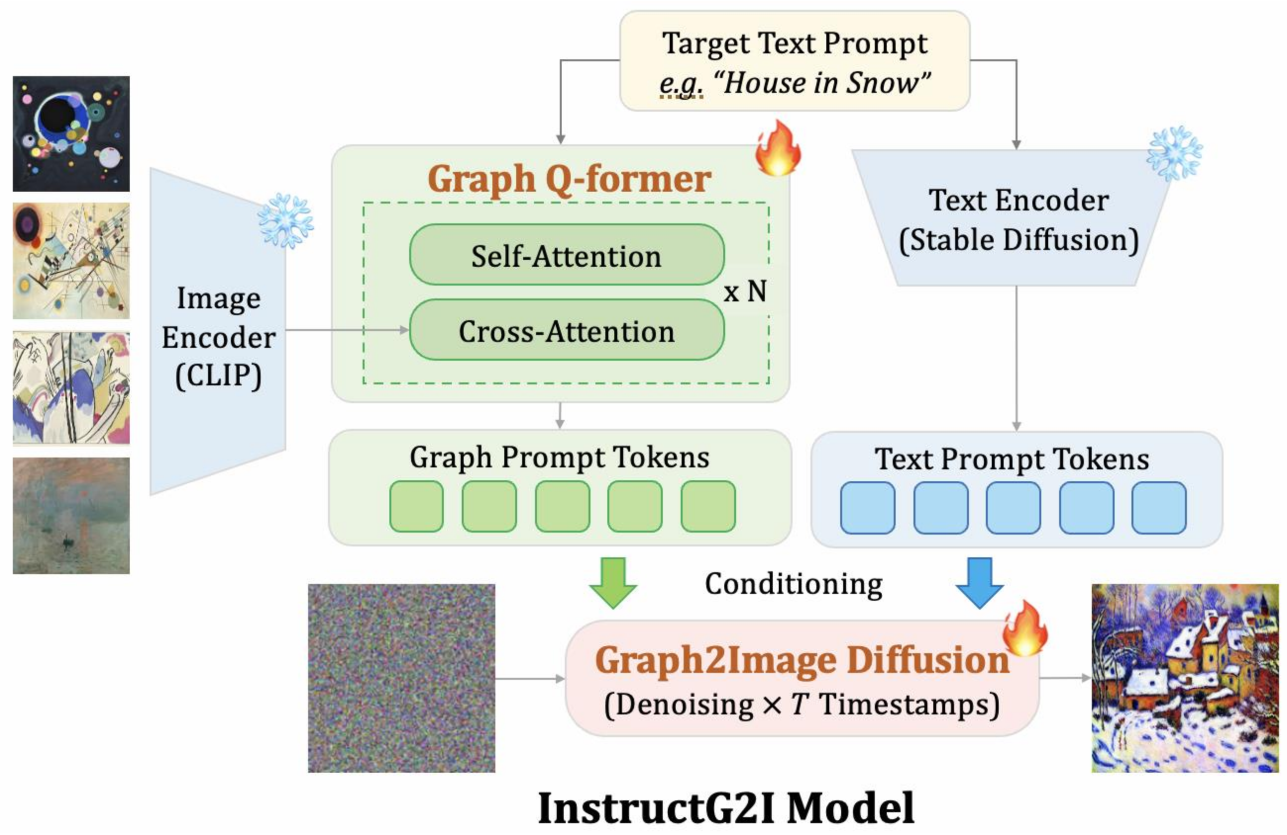




Ours: Graph Q-Former

InstructG2I

- **Graph Encoding with Text Conditions**



InstructG2I

- **How to make the image generation controllable?**

    - **Control the guidance weight between text and graph conditions.**

    - **Control multiple graph guidance.**

InstructG2I

- **Controllable Generation**

Goal: Balance the guidance weight from the text and graph.

Classifier-free guidance:

$$\hat{\epsilon}_\theta(\mathbf{z}_t, c) = \epsilon_\theta(\mathbf{z}_t, \varnothing) + s \cdot (\epsilon_\theta(\mathbf{z}_t, c) - \epsilon_\theta(\mathbf{z}_t, \varnothing))$$

Graph classifier-free guidance:

$$\begin{aligned}\hat{\epsilon}_\theta(\mathbf{z}_t, c_G, c_T) = \epsilon_\theta(\mathbf{z}_t, \varnothing, \varnothing) + s_T \cdot (\epsilon_\theta(\mathbf{z}_t, \varnothing, c_T) - \epsilon_\theta(\mathbf{z}_t, \varnothing, \varnothing)) \\ + s_G \cdot (\epsilon_\theta(\mathbf{z}_t, c_G, c_T) - \epsilon_\theta(\mathbf{z}_t, \varnothing, c_T)).\end{aligned}$$

InstructG2I

- **Controllable Generation**

Goal: Control from multiple graph conditions.

Graph classifier-free guidance:

$$\hat{\epsilon}_\theta(\mathbf{z}_t, c_G, c_T) = \epsilon_\theta(\mathbf{z}_t, \varnothing, \varnothing) + s_T \cdot (\epsilon_\theta(\mathbf{z}_t, \varnothing, c_T) - \epsilon_\theta(\mathbf{z}_t, \varnothing, \varnothing))$$
$$+ s_G \cdot (\epsilon_\theta(\mathbf{z}_t, c_G, c_T) - \epsilon_\theta(\mathbf{z}_t, \varnothing, c_T)).$$

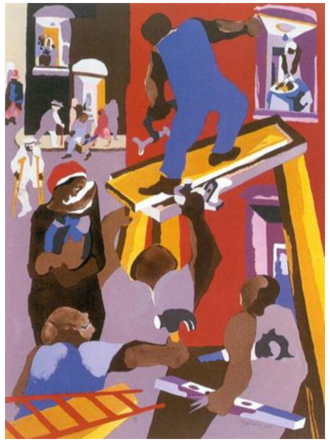
Multiple graph classifier-free guidance:

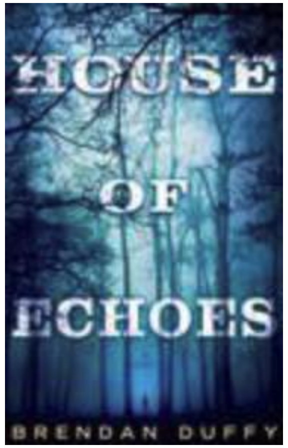
$$\hat{\epsilon}_\theta(\mathbf{z}_t, c_G, c_T) = \epsilon_\theta(\mathbf{z}_t, \varnothing, \varnothing) + s_T \cdot (\epsilon_\theta(\mathbf{z}_t, \varnothing, c_T) - \epsilon_\theta(\mathbf{z}_t, \varnothing, \varnothing))$$
$$+ \sum s_G^{(k)} \cdot (\epsilon_\theta(\mathbf{z}_t, c_G^{(k)}, c_T) - \epsilon_\theta(\mathbf{z}_t, \varnothing, c_T)),$$

Experiments

- **Datasets**
  - **ART500K**
    - nodes: artworks; edges: same-author, same-genre relationships.
    - text: title; image: picture.
  - **Amazon**
    - nodes: products; edges: co-view relationships.
    - text: title; image: picture.
  - **Goodreads**
    - nodes: books; edges: similar-book semantics.
    - text: title; image: cover image

Dataset	# Node	# Edge
ART500K	311,288	643,008,344
Amazon	178,890	3,131,949
Goodreads	93,475	637,210

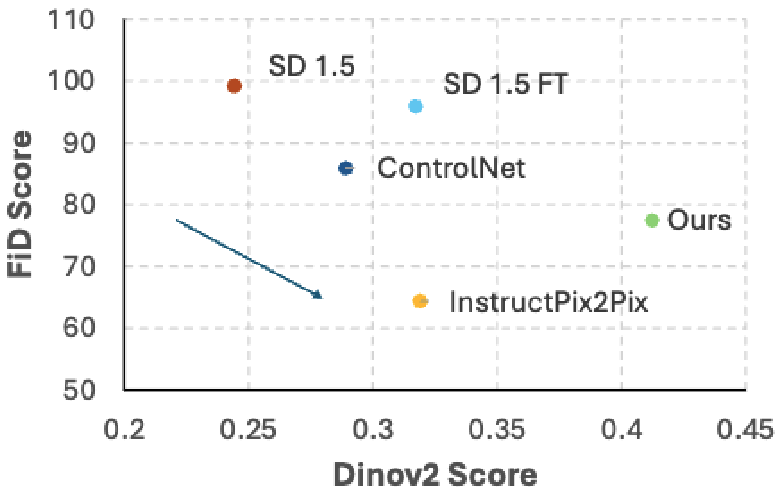




Experiments

- **Quantitative results**

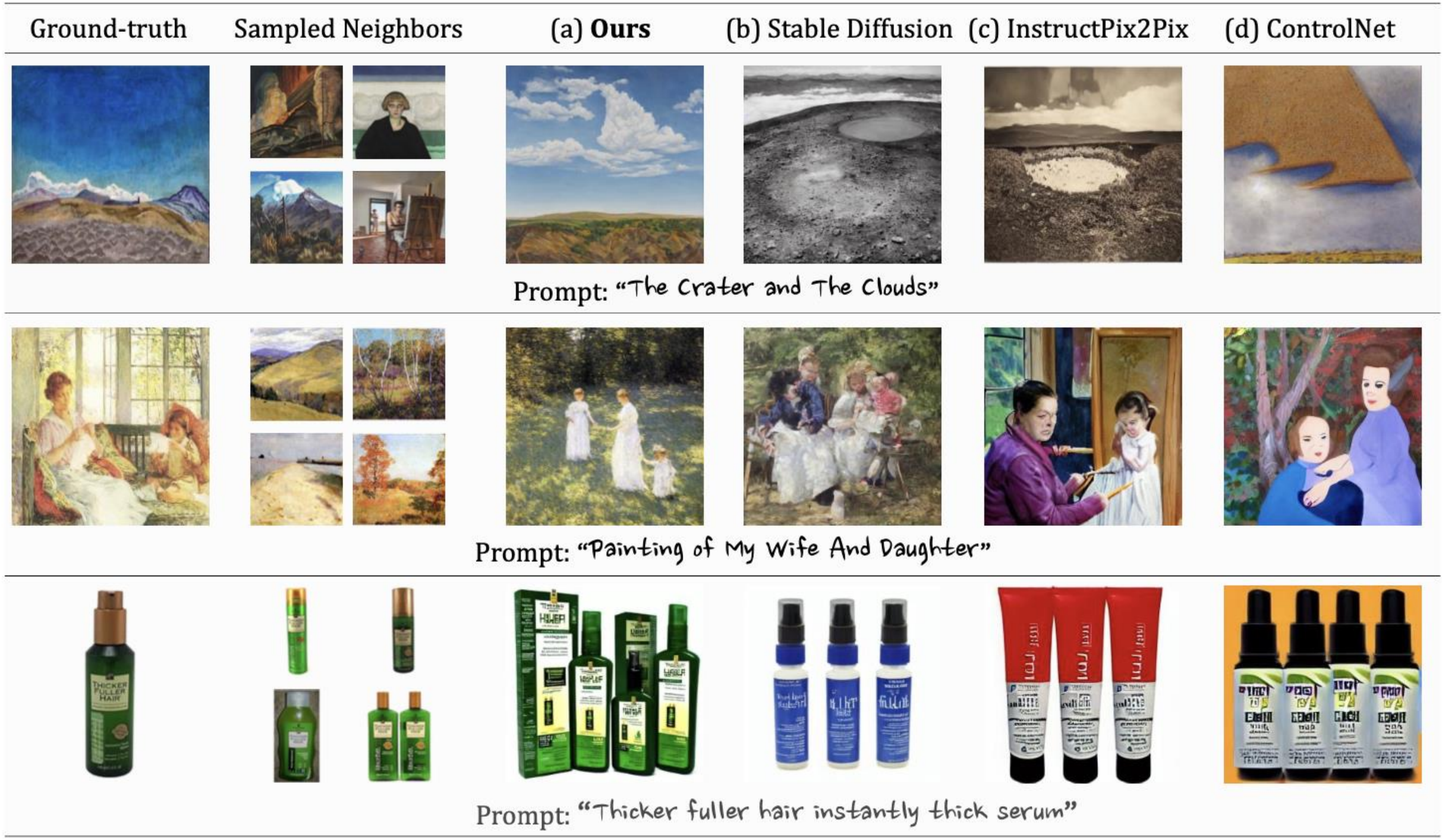
	ART500K		Amazon		Goodreads	
Model	CLIP score	DINOv2 score	CLIP score	DINOv2 score	CLIP score	DINOv2 score
SD-1.5	58.83	25.86	60.67	32.61	42.16	14.84
SD-1.5 FT	66.55	34.65	65.30	41.52	45.81	18.97
Instruct pix2pix	65.66	33.44	63.86	41.31	47.30	20.94
ControlNet	64.93	32.88	59.88	34.05	42.20	19.77
Ours	**73.73**	**46.45**	**68.34**	**51.70**	**50.37**	**25.54**



- Our model has consistently better performance than competitive baselines.

Experiments

- **Qualitative results**



- Our method exhibits better consistency with the ground truth.

Experiments

- **Same text prompts with different graph conditions**

Text: a man playing the piano

Pablo Picasso

Salvador Dali

Vincent van Gogh

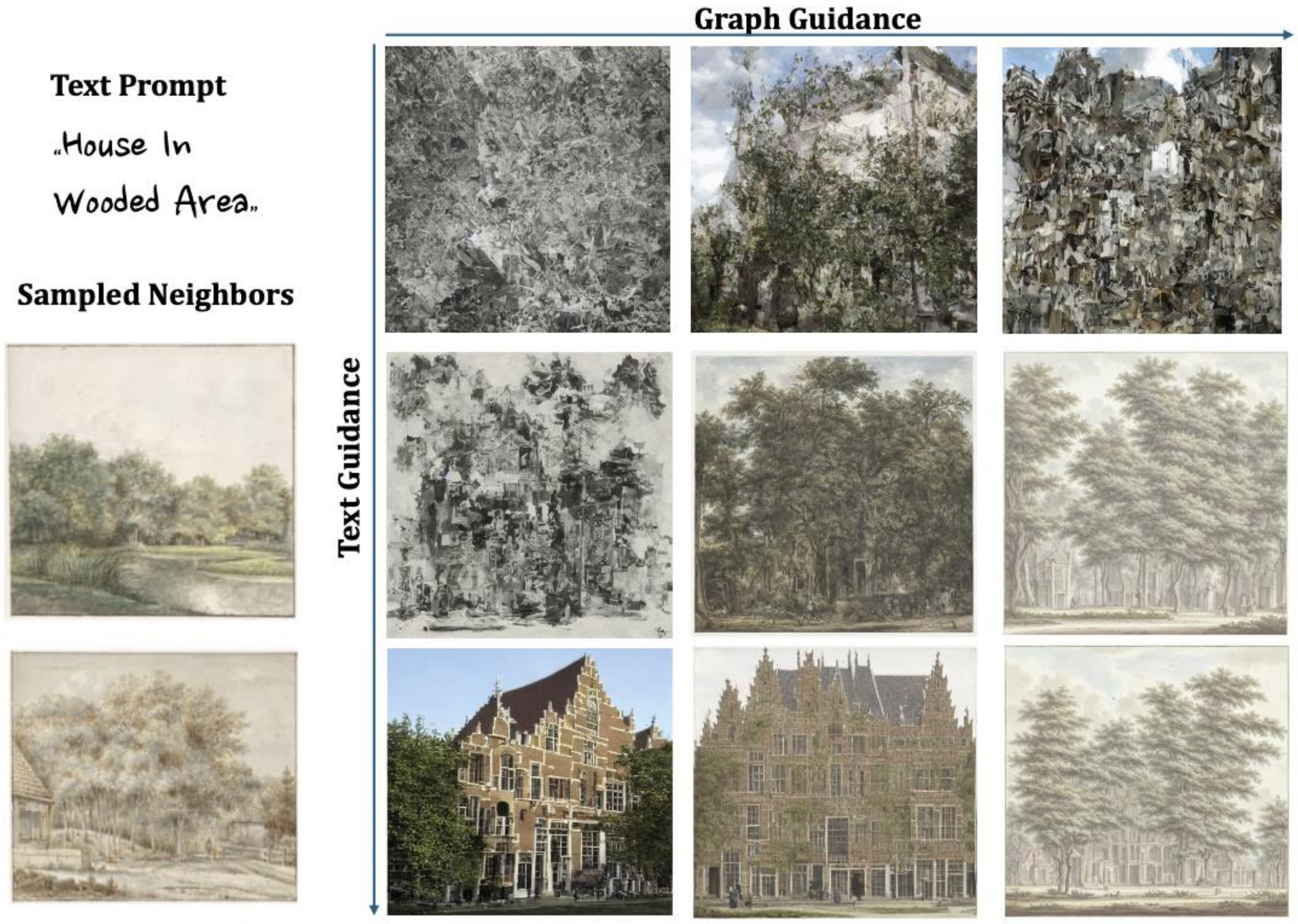
Gustave Courbet

Caravaggio

Max Beckmann

Experiments

- **Text and graph guidance study**



- As **text guidance** increases, the generated image incorporates more of the desired content.

- As **graph guidance** increases, the generated image adopts a more desired **style**.

Experiments

- **Single or multiple graph guidance**

Text: a man playing piano

- When **single** graph guidance is provided, the generated artwork aligns with that artist's style.
- As **additional graph guidance** is introduced, the **styles** of the two artists blend together.

Experiments

- **Single or multiple graph guidance**

Text: a house in the snow

Thank You !

Subscribe and learn more about our works!

Graph CoT

InstructG2I