InstructG2I: Synthesizing Images from Multimodal Attributed Graphs

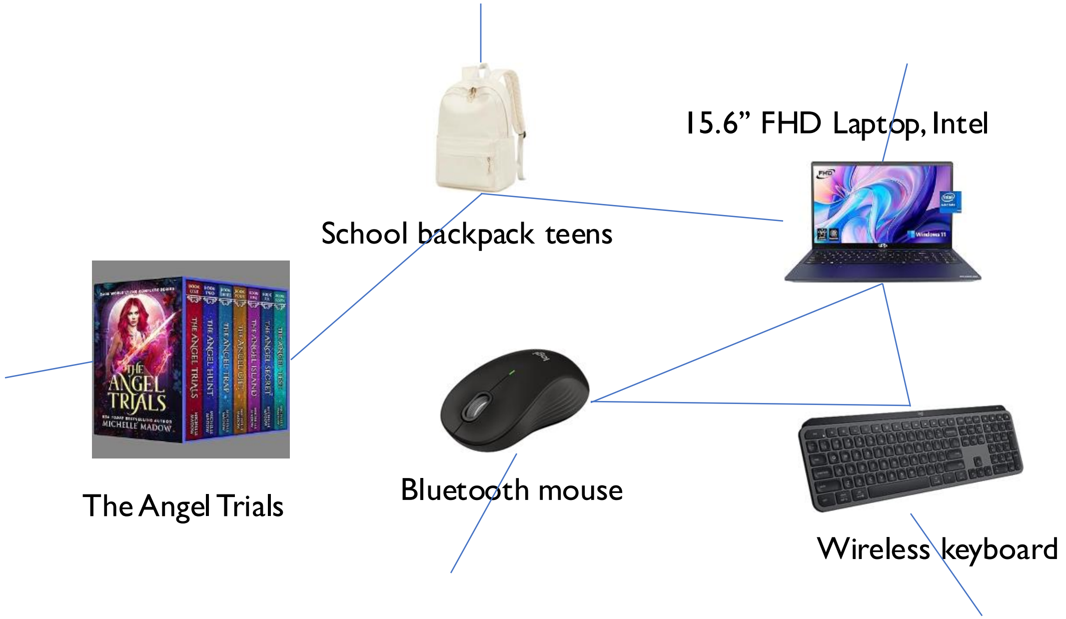
Bowen Jin, Ziqi Pang, Bingjun Guo, Yu-Xiong Wang, Jiaxuan You, Jiawei Han

NeurIPs 2024

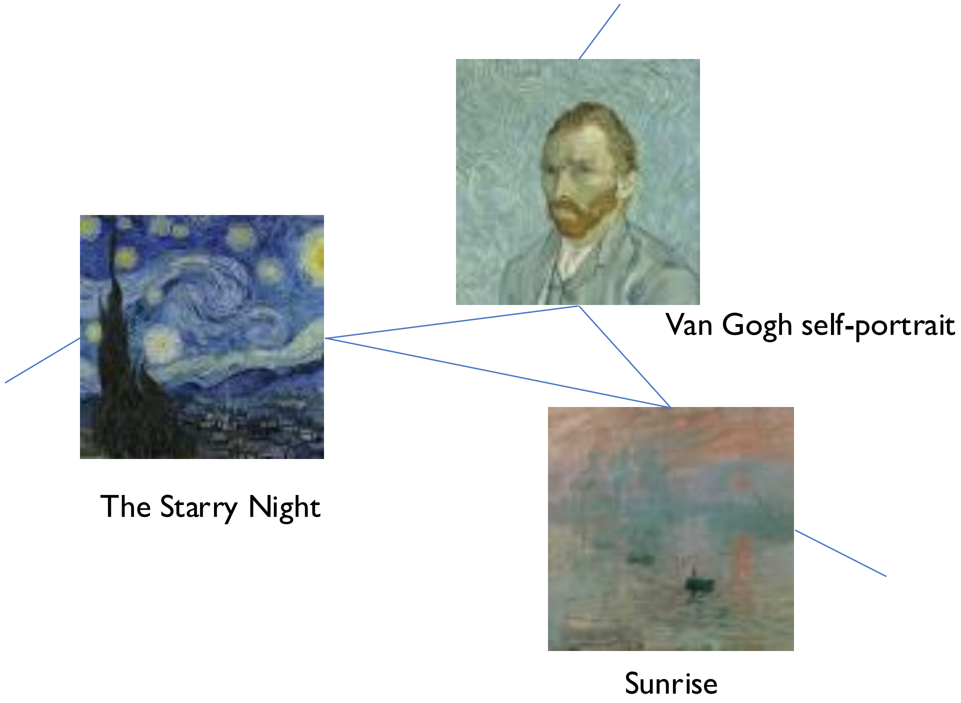
website: instructg2i.github.io

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

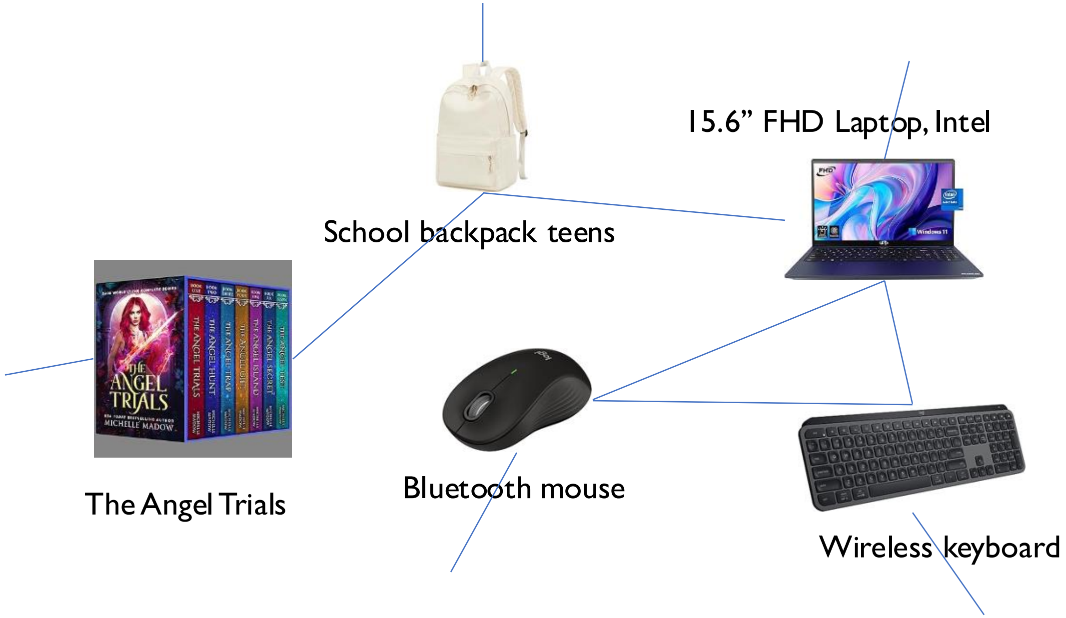
Introduction

- **Multimodal attributed graphs**
    - Text, Image and Graph

Text

Image

Graph Structure



product graphs

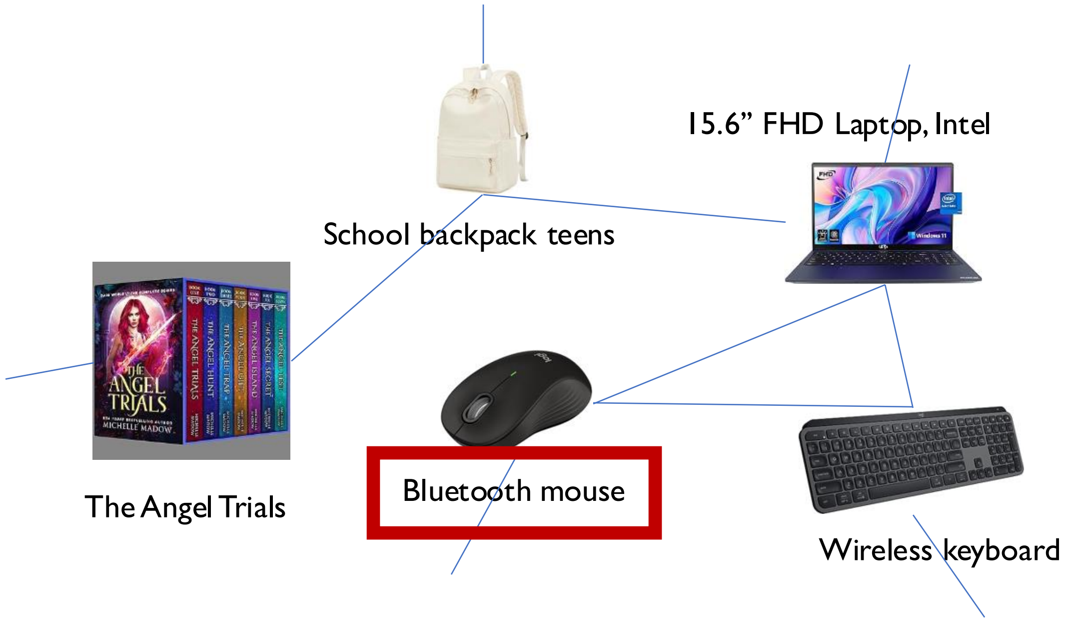
Introduction

- **Multimodal attributed graphs**
  - Text, Image and Graph

Text

Provide some features which is not conveyed by other modality.

E.g., we cannot know that this is a "**bluetooth**" mouse solely from the image or the graph structure.



product graphs

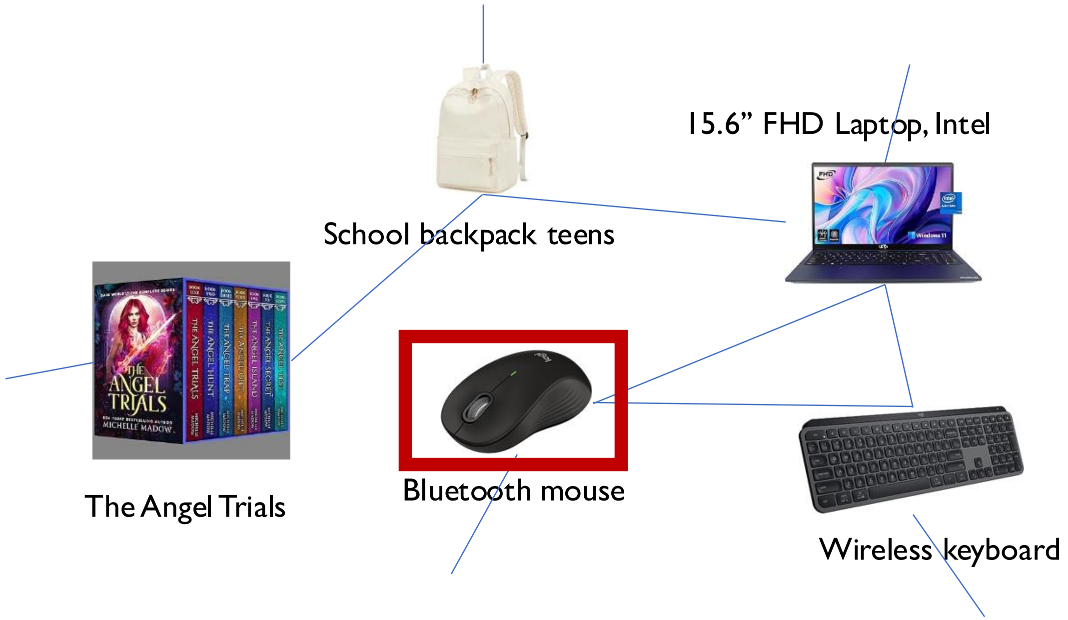
Introduction

- **Multimodal attributed graphs**
  - Text, Image and Graph

Image

Provide some features which is not conveyed by other modality.

E.g., we cannot know that this mouse is "**black**" solely from the text or the graph structure.



product graphs

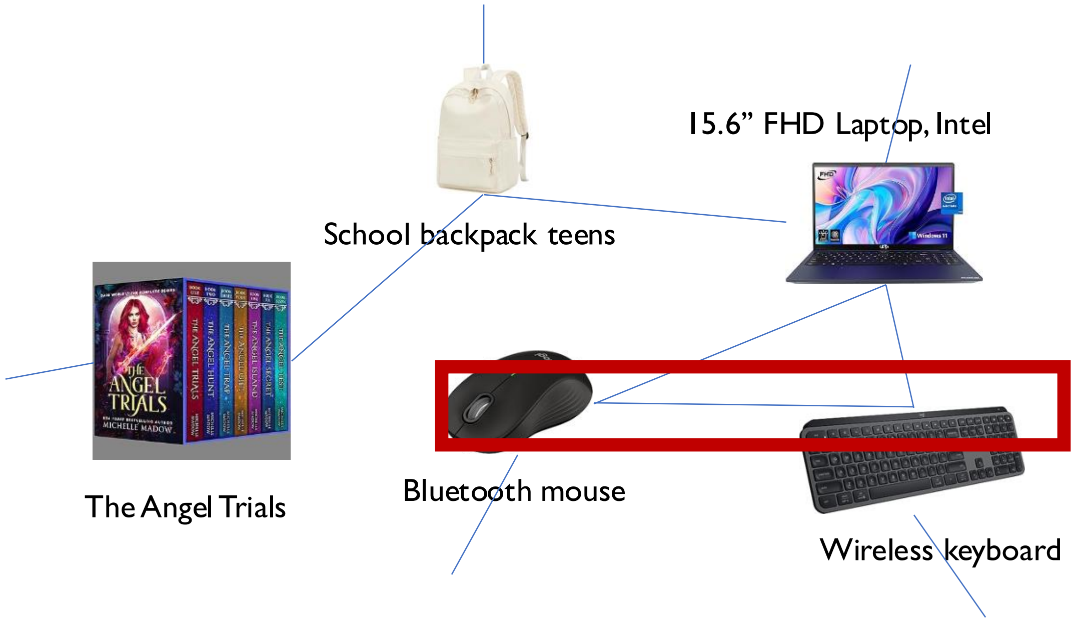
Introduction

- **Multimodal attributed graphs**
  - Text, Image and Graph

Graph Structure

Provide the positive semantic relation between nodes (i.e., their similarity).

E.g., we cannot know that this mouse and this keyboard are co-purchased by many users if we only have their texts and images.



product graphs

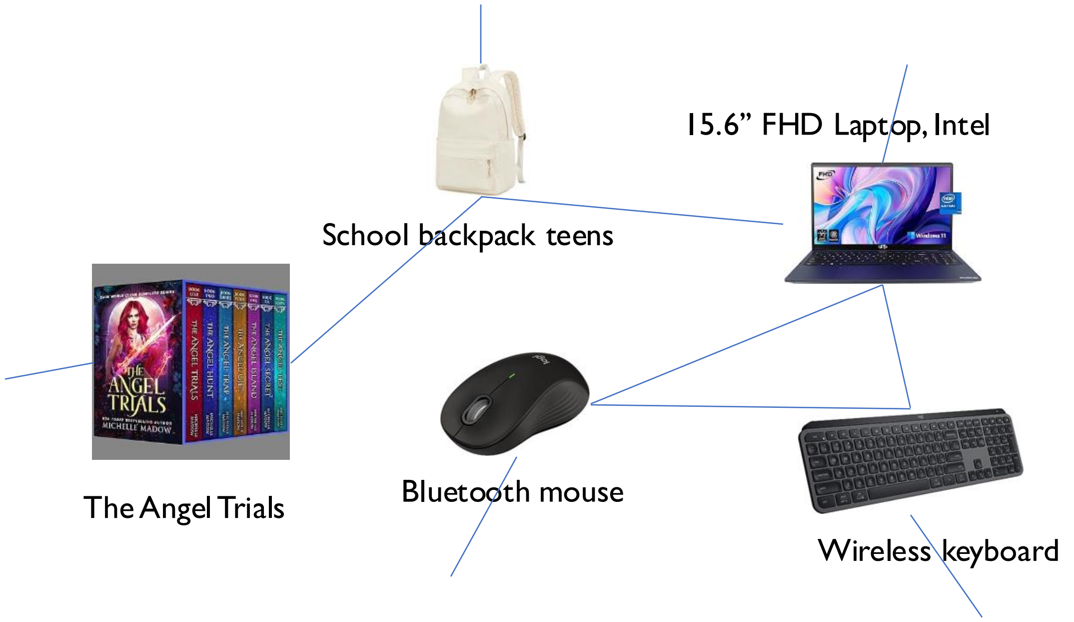
Introduction

- **Multimodal attributed graphs**
  - Text, Image and Graph

Text

Image

Graph Structure



product graphs

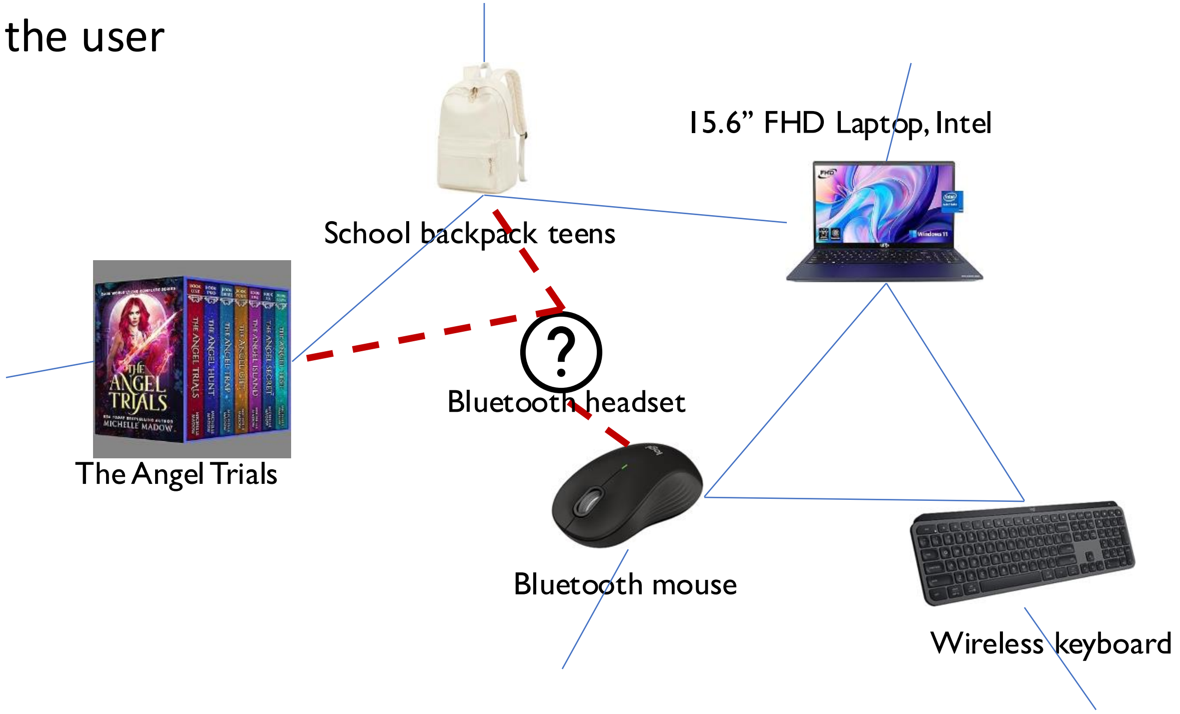
All three information are very important on learning on such graphs

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?



Problem

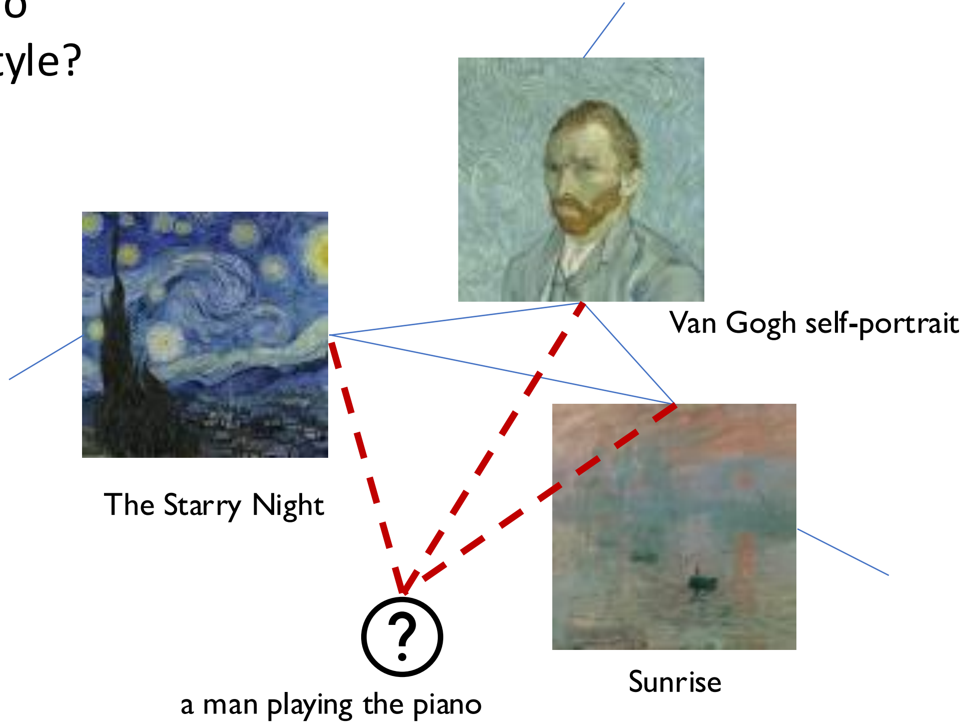
- **How we conduct node image generation on such graph?**
  - **Application on Art domain**

Virtual art creation

How will be a picture titled “a man playing the piano”
looks like with 50% Monet style and 50% Van Gogh style?

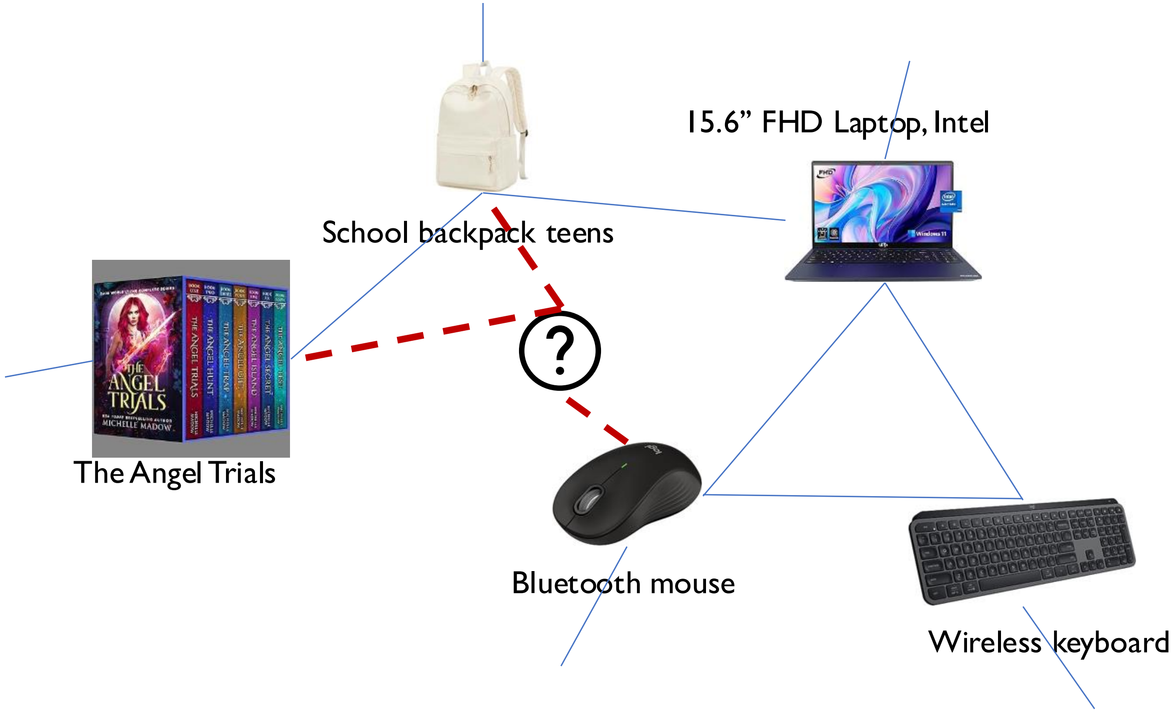
Monet arts & Van Gogh arts

a man playing the piano

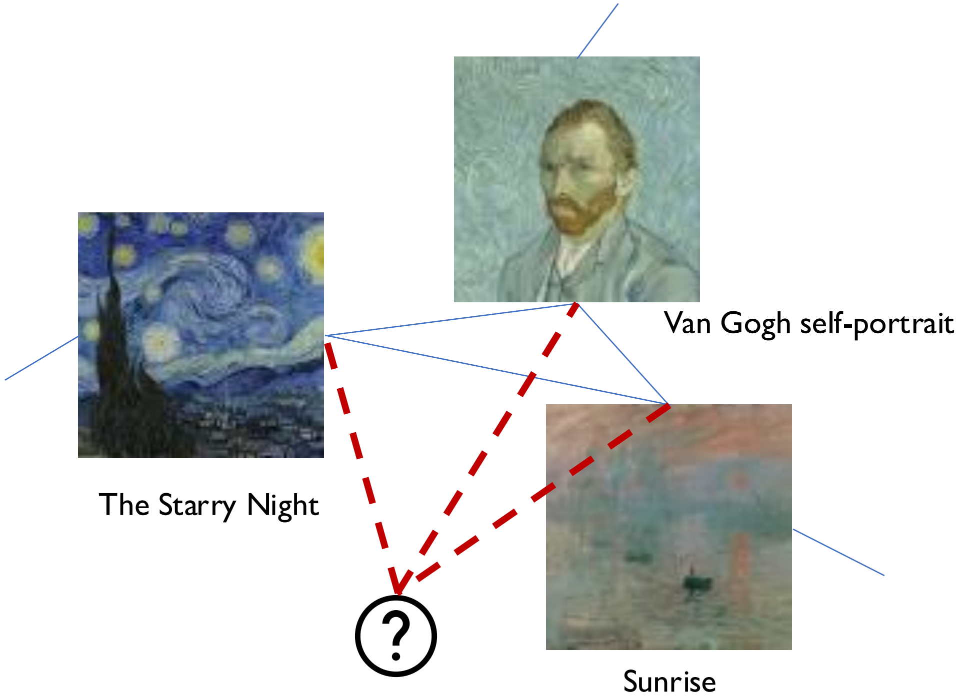


Problem

- **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.



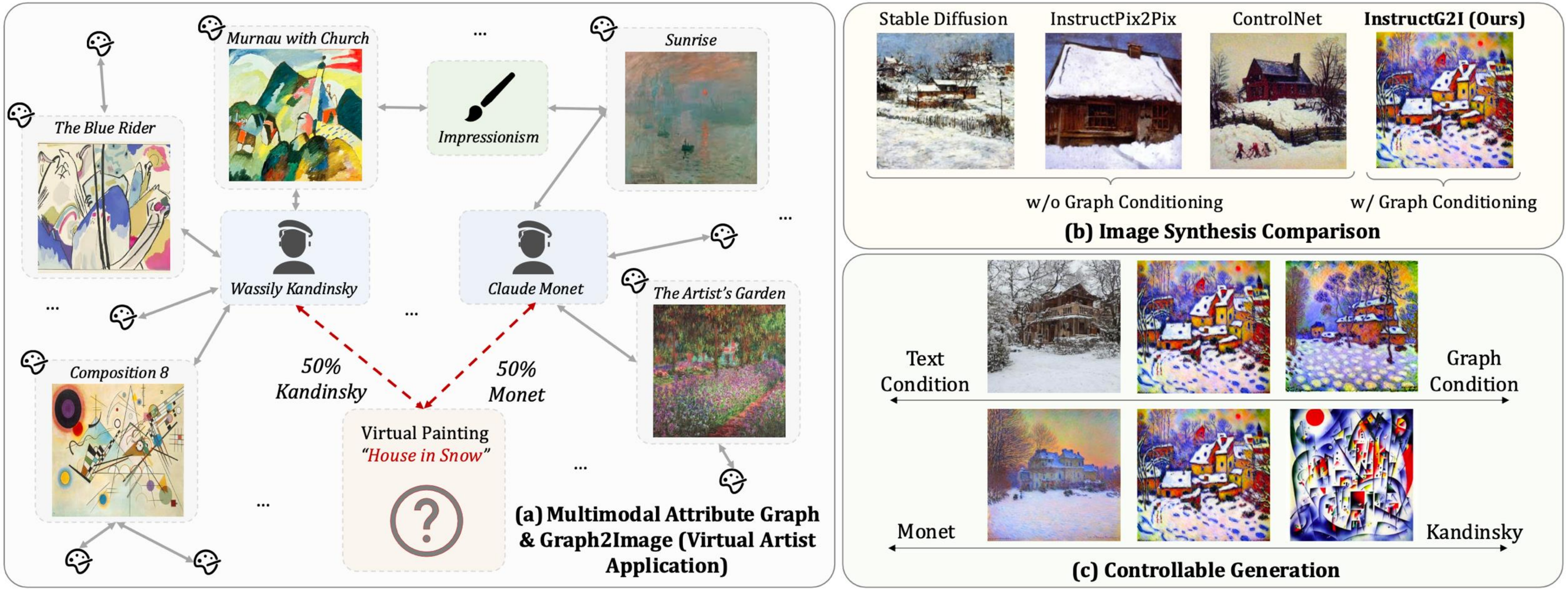
Generative recommendation



Virtual art creation

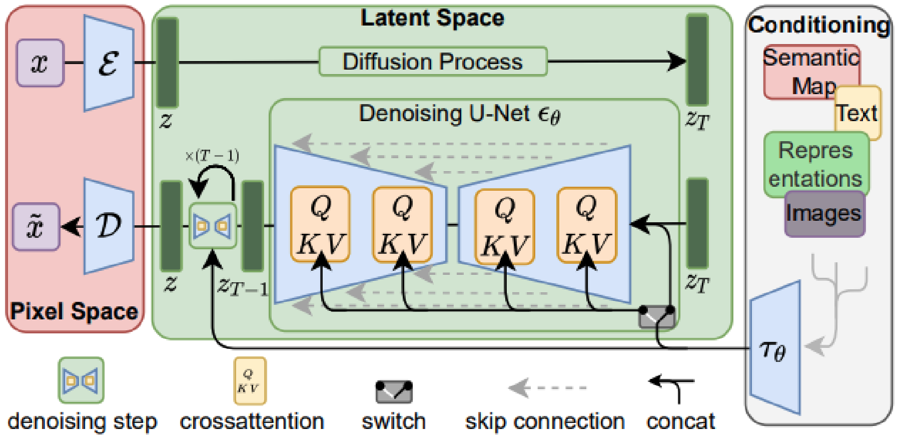
Problem

- **Task: Synthesizing Images from Multimodal Attributed Graphs**

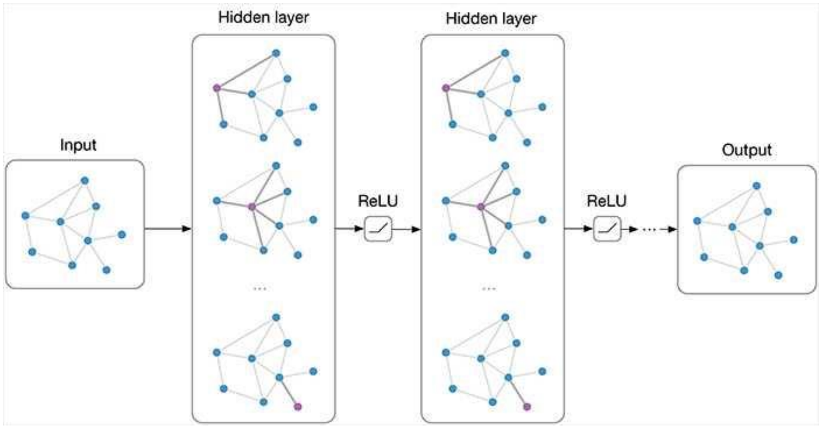


Problem

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

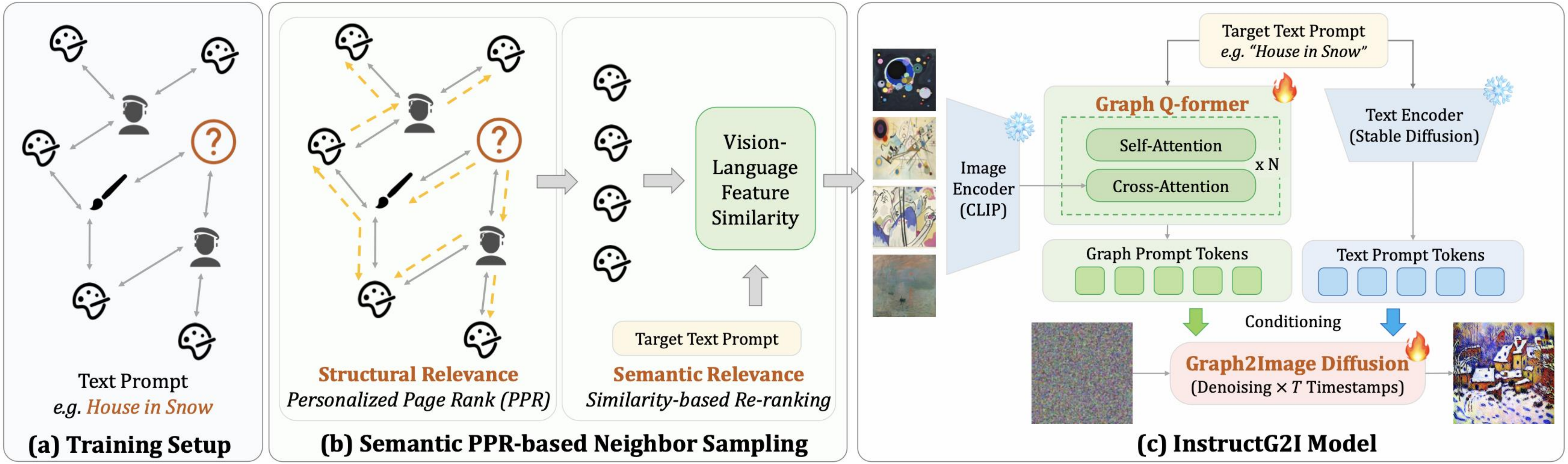


  - **Graph Neural Network**
    - GCN, GraphSAGE, …
    - They mainly focus on representation learning
    - Cannot handle generation tasks



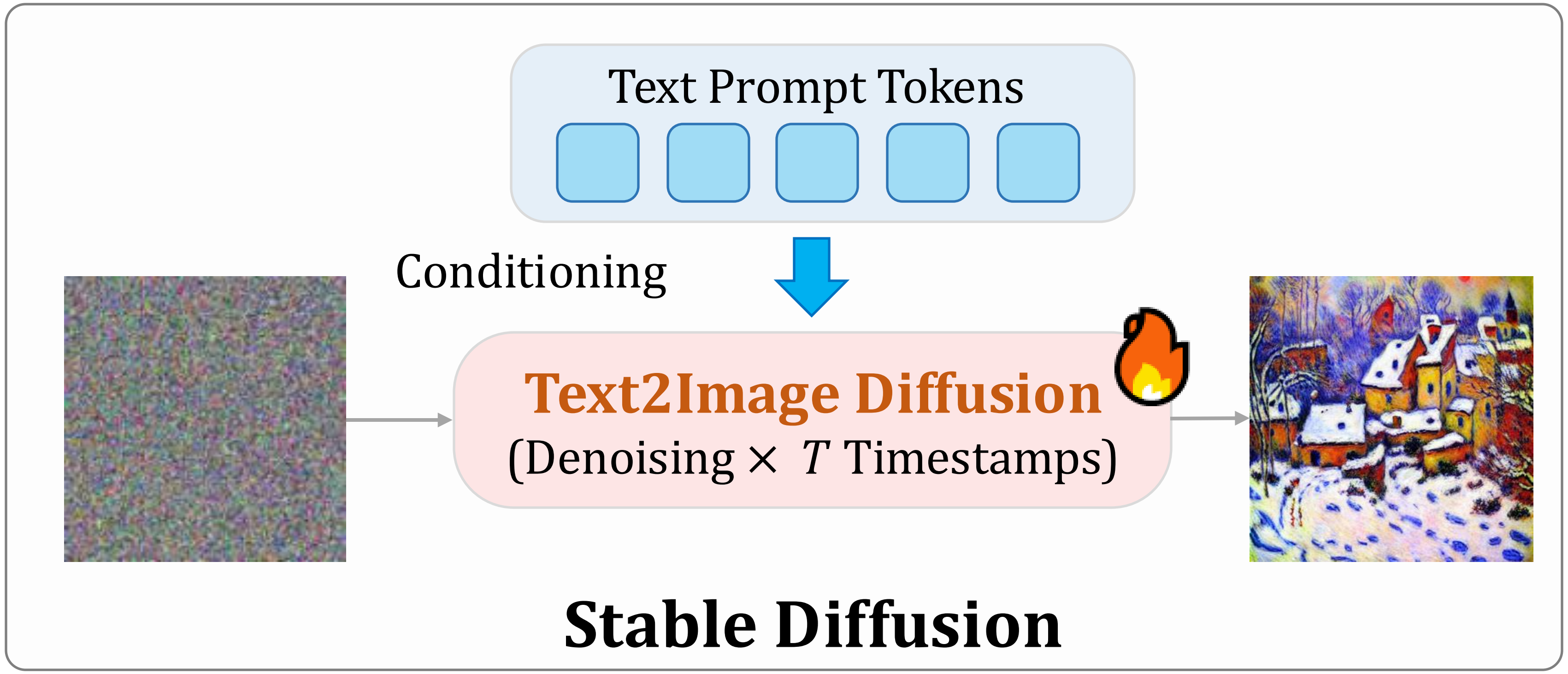
InstructG2I

- **Model Overview**



InstructG2I

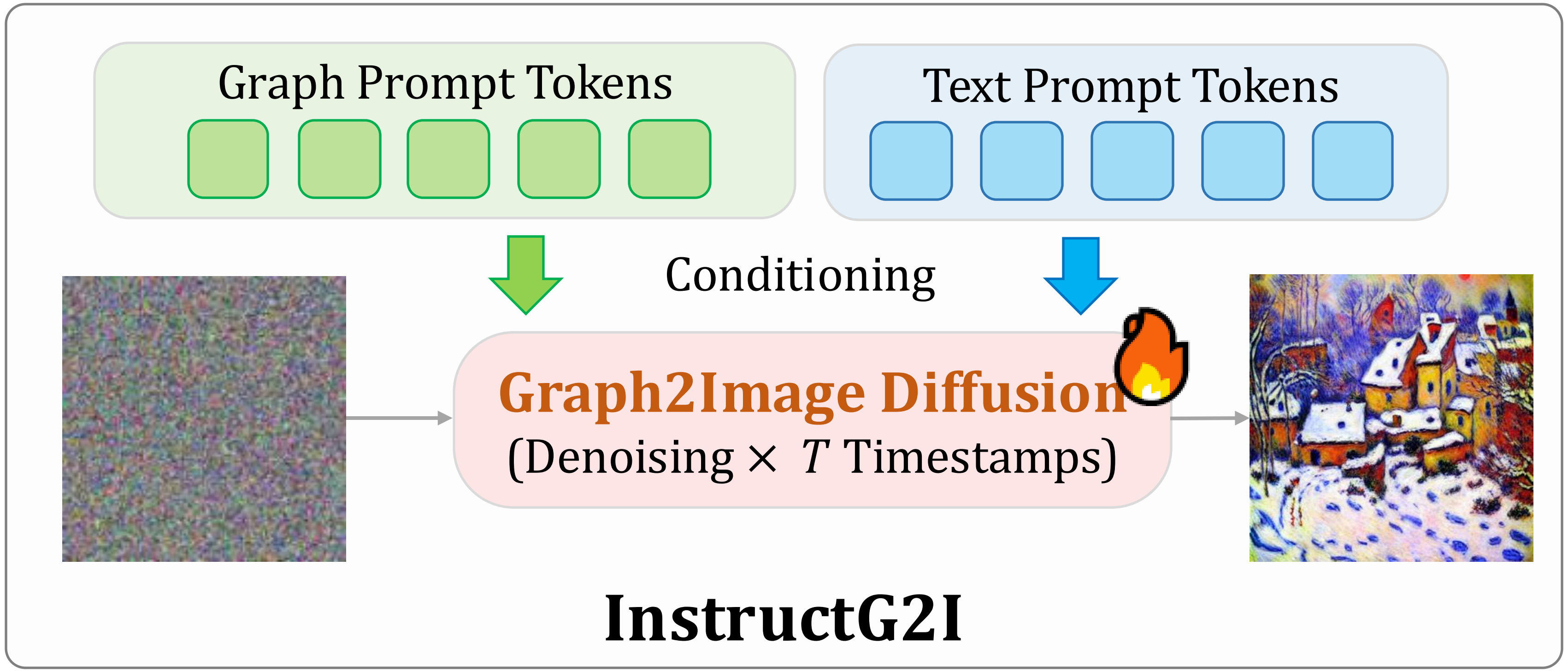
- **Stable diffusion (SD)**



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

InstructG2I

- **Graph context-conditioned stable diffusion**

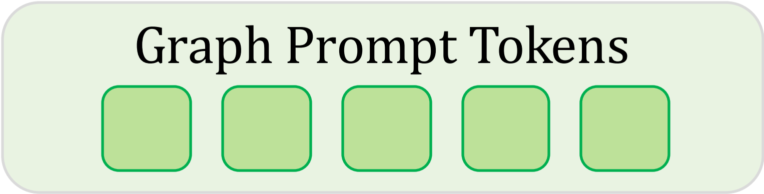


$$h(c_T, c_G) = [h_T(c_T), h_G(c_G)] \in \mathbf{R}^{d\times(l_{c_T}+l_{c_G})}$$

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

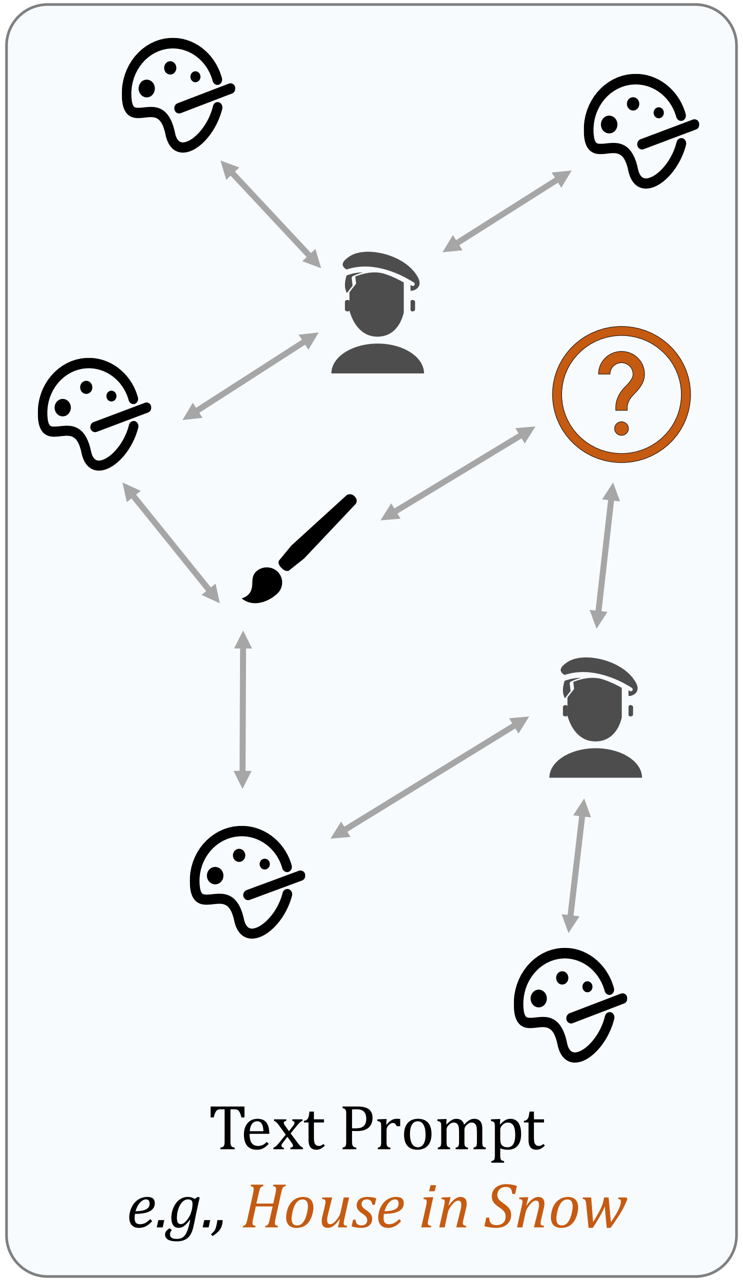
InstructG2I

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



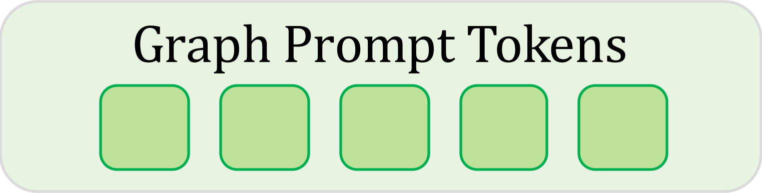
1. Find relevant context from the graph.
   -- **Semantic PPR-based Neighbor Sampling**

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



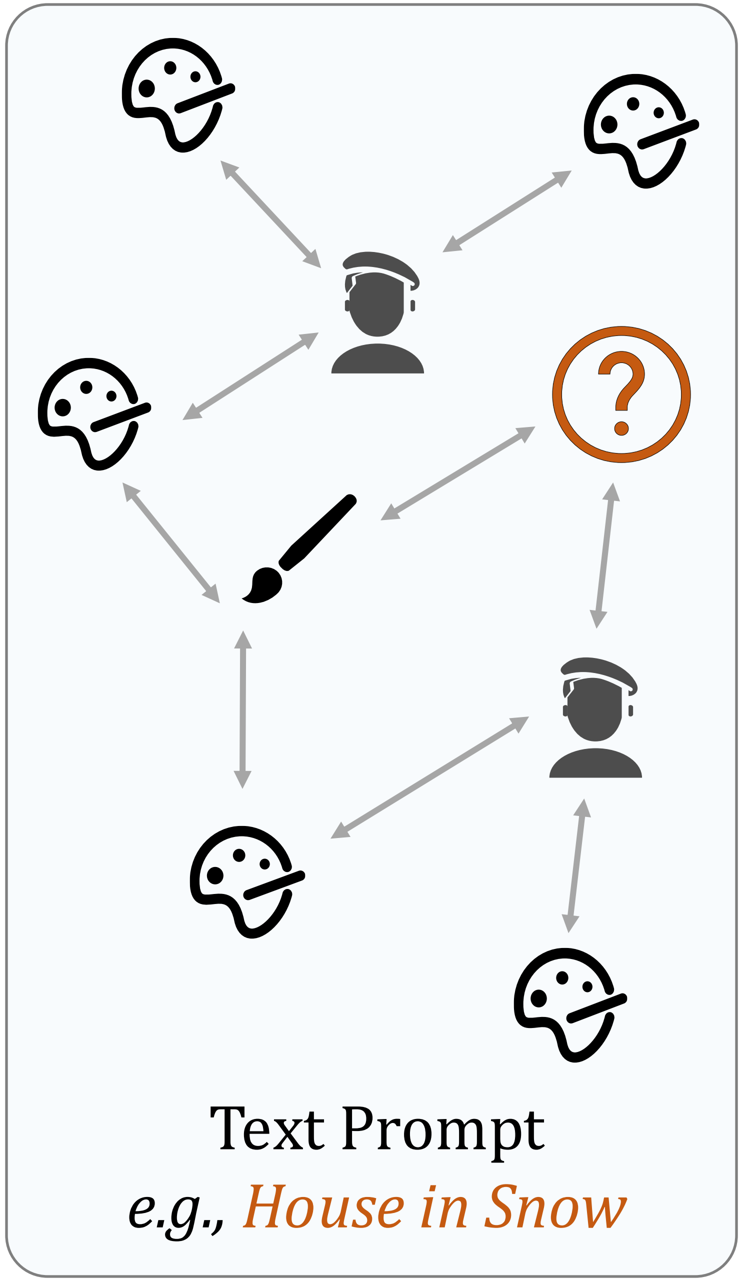
InstructG2I

- **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**



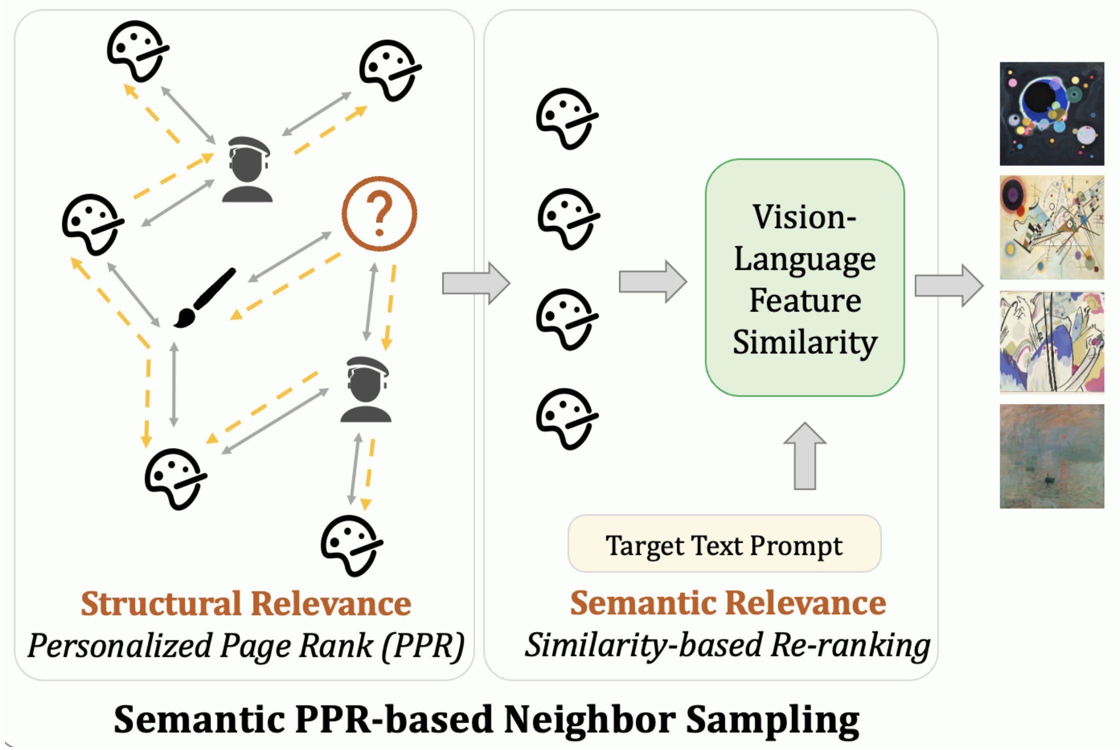
InstructG2I

- **Semantic PPR-based Neighbor Sampling**

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

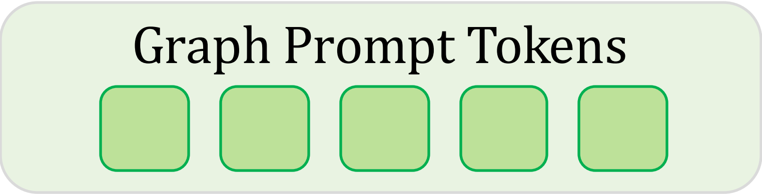
Step1: Structure relevance with Personalized Page Rank (PPR).

Step2: Semantic relevance with content similarity calculation.



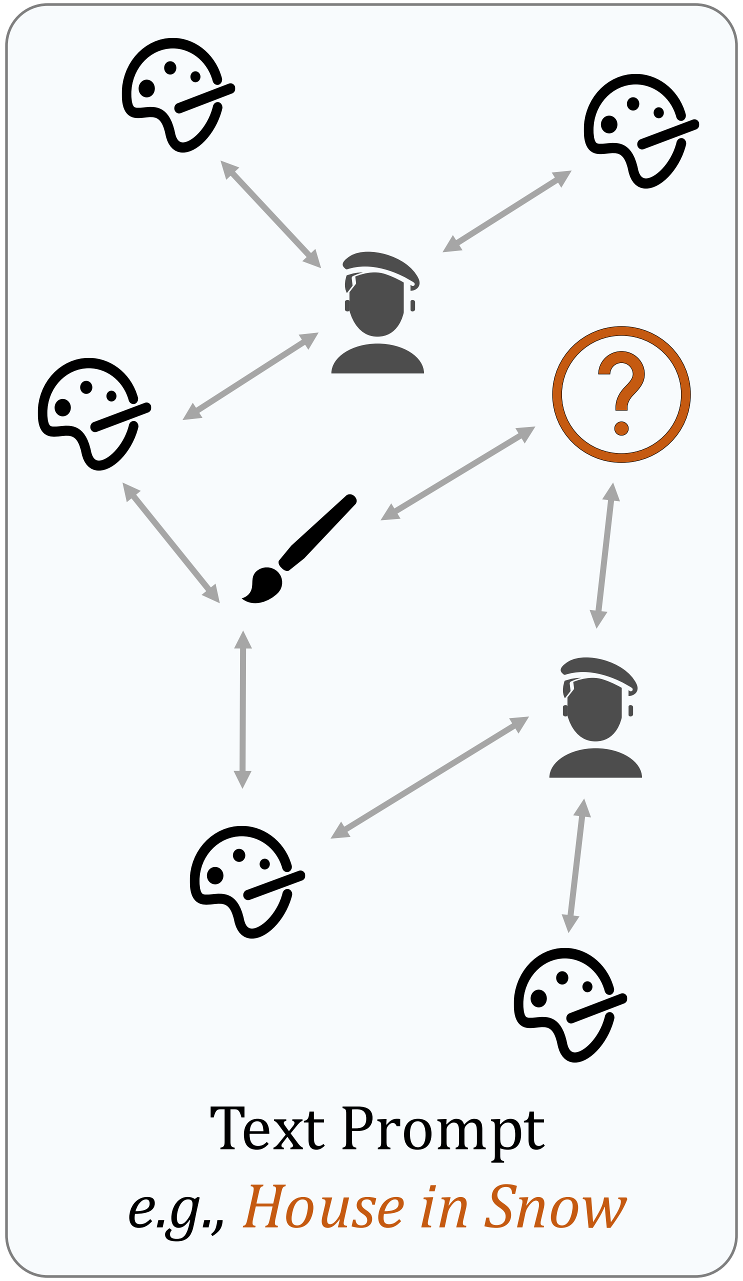
InstructG2I

- **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**



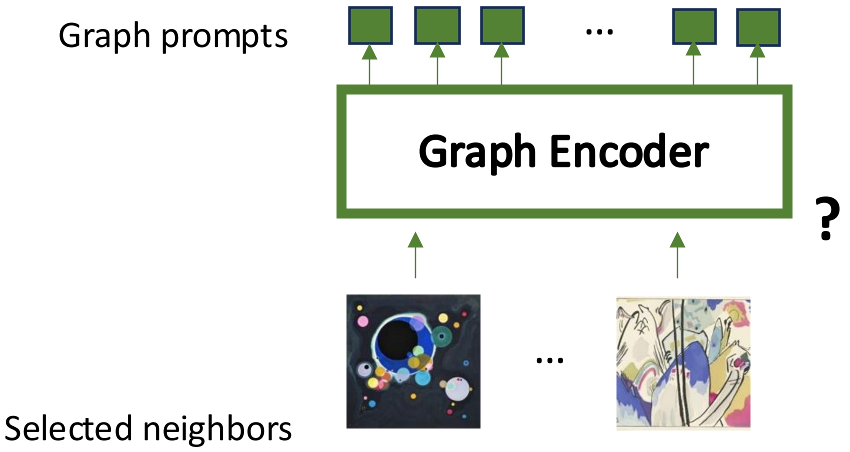
InstructG2I

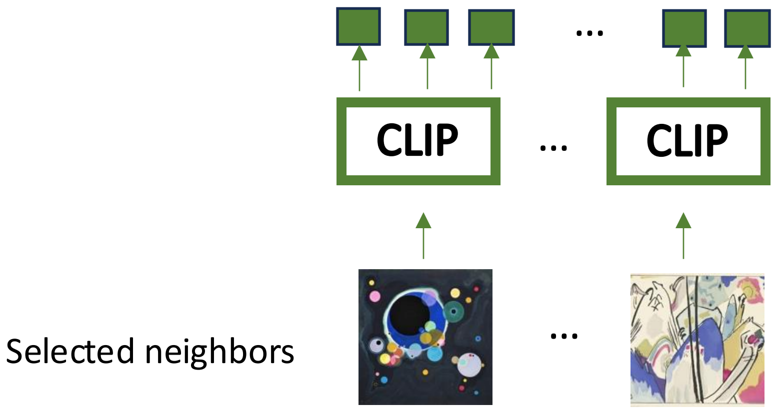
- **Graph Encoding: a simple baseline**

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).

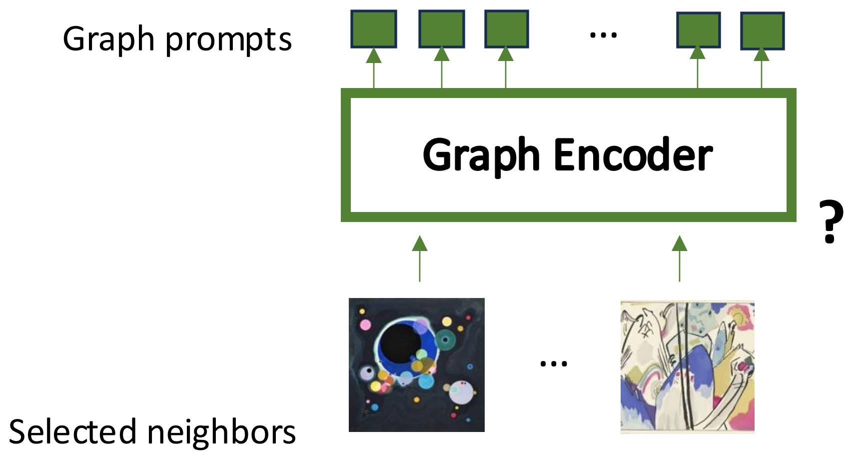


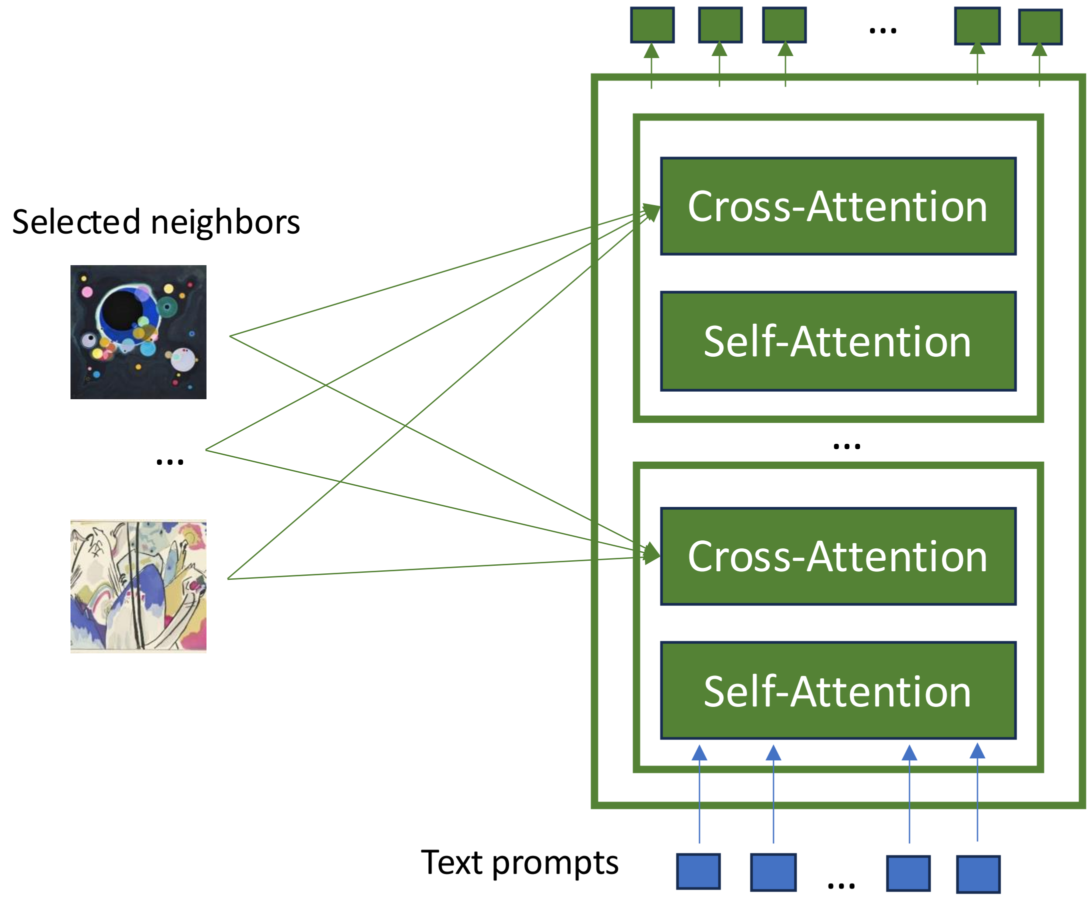


a simple baseline

InstructG2I

- **Graph Encoding with Text Conditions**

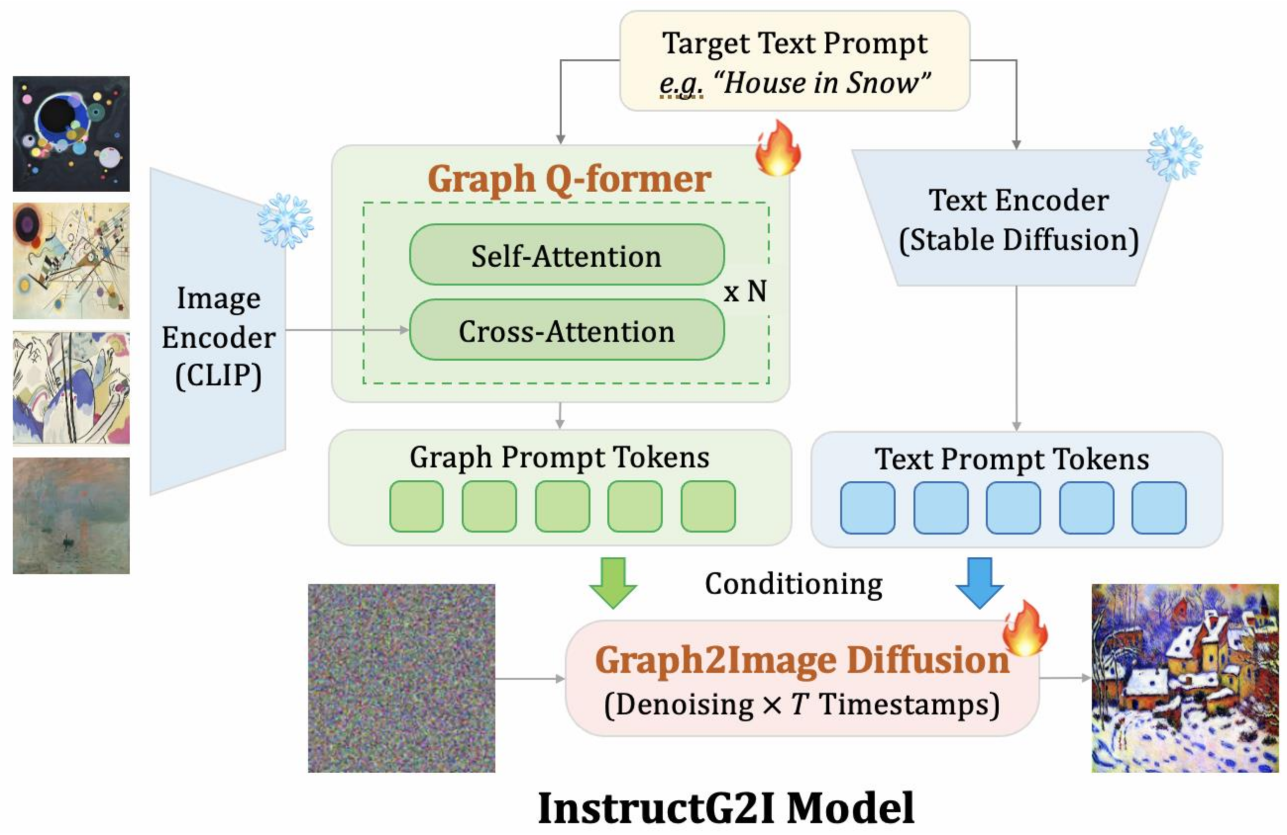




Ours: Graph Q-Former

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:

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

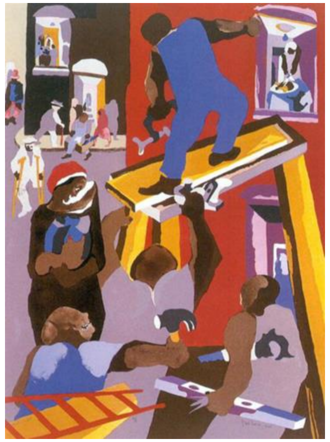
Multiple graph classifier-free guidance:

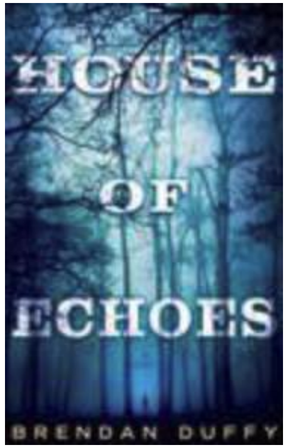
$$\hat{\epsilon}_\theta(\mathbf{z}_t, c_G, c_T) = \epsilon_\theta(\mathbf{z}_t, \varnothing, \varnothing) + s_T \cdot (\epsilon_\theta(\mathbf{z}_t, \varnothing, c_T) - \epsilon_\theta(\mathbf{z}_t, \varnothing, \varnothing))$$
$$+ \sum s_G^{(k)} \cdot (\epsilon_\theta(\mathbf{z}_t, c_G^{(k)}, c_T) -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

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

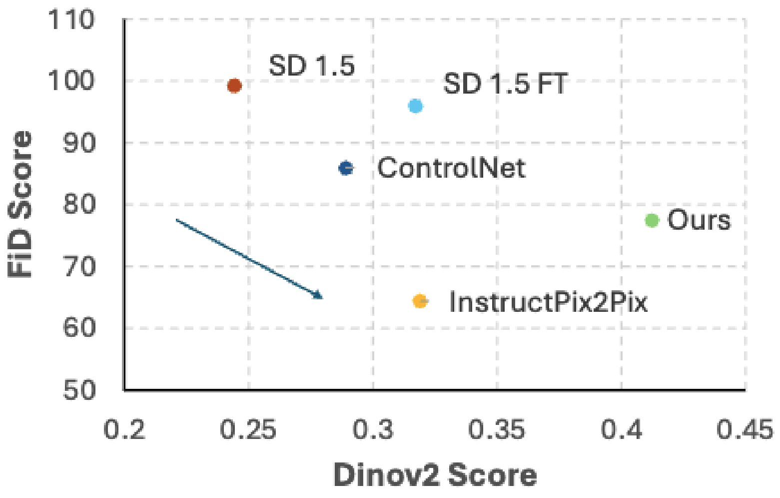




Experiments

- **Quantitative results**

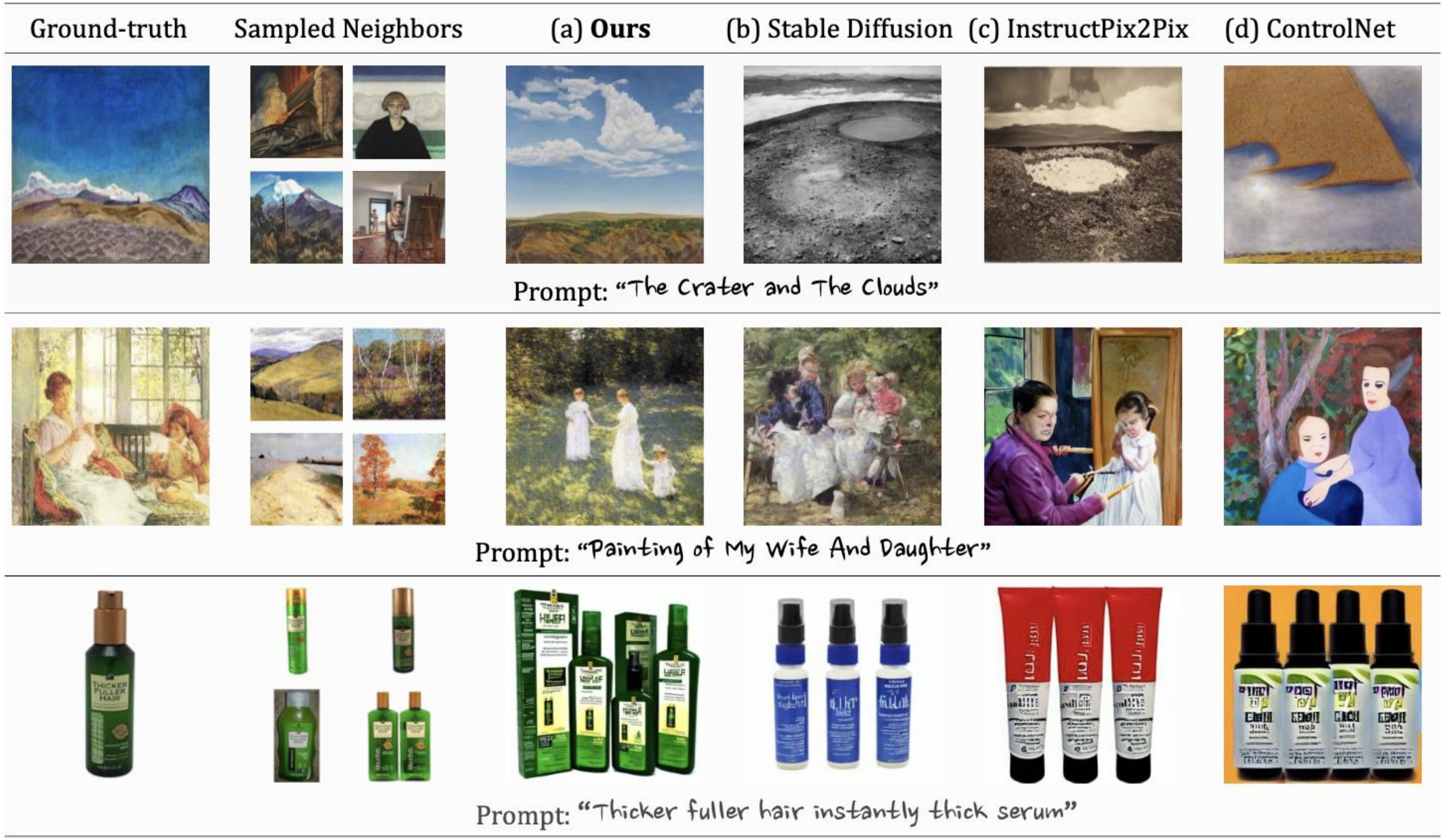
Model	ART500K		Amazon		Goodreads	
	CLIP score	DINOv2 score	CLIP score	DINOv2 score	CLIP score	DINOv2 score
SD-1.5	58.83	25.86	60.67	32.61	42.16	14.84
SD-1.5 FT	66.55	34.65	65.30	41.52	45.81	18.97
Instruct pix2pix	65.66	33.44	63.86	41.31	47.30	20.94
ControlNet	64.93	32.88	59.88	34.05	42.20	19.77
Ours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

Pablo Picasso

Salvador Dali

Vincent van Gogh

Gustave Courbet

Caravaggio

Max Beckmann

Experiments

- **Ablation study on graph condition variants**

	ART500K		Amazon		Goodreads	
Model	CLIP score	DINOv2 score	CLIP score	DINOv2 score	CLIP score	DINOv2 score
INSTRUCTG2I	**73.73**	**46.45**	**68.34**	**51.70**	**50.37**	**25.54**
- Graph-QFormer	72.53	44.16	66.97	48.18	47.91	24.74
+ GraphSAGE	72.26	43.06	66.07	43.40	46.68	21.91
+ GAT	72.60	43.32	66.73	46.58	46.57	21.45
IP2P w. neighbor images	65.89	33.90	63.19	40.32	47.21	21.55
SD FT w. neighbor texts	69.72	38.64	65.55	43.51	47.47	22.68

- InstructG2I consistently outperforms both variants.
- This demonstrates the advantage of leveraging image features on graphs and the effectiveness of our model design.

Experiments

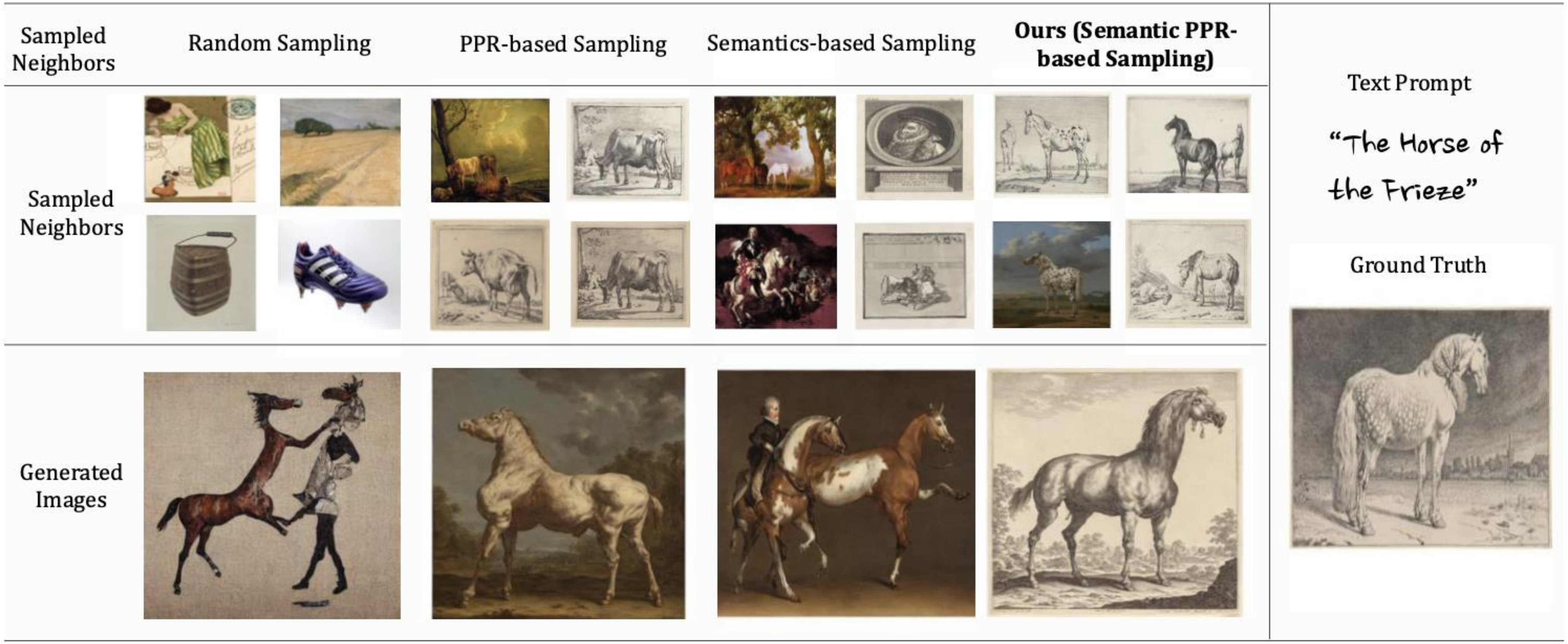
- **Ablation study on Graph-Qformer**

Model	ART500K		Amazon		Goodreads	
	CLIP score	DINOv2 score	CLIP score	DINOv2 score	CLIP score	DINOv2 score
INSTRUCTG2I	**73.73**	**46.45**	**68.34**	**51.70**	**50.37**	**25.54**
- Graph-QFormer	72.53	44.16	66.97	48.18	47.91	24.74
+ GraphSAGE	72.26	43.06	66.07	43.40	46.68	21.91
+ GAT	72.60	43.32	66.73	46.58	46.57	21.45
IP2P w. neighbor images	65.89	33.90	63.19	40.32	47.21	21.55
SD FT w. neighbor texts	69.72	38.64	65.55	43.51	47.47	22.68

- InstructG2I with Graph-QFormer consistently outperforms both the ablated version and GNN baselines.

Experiments

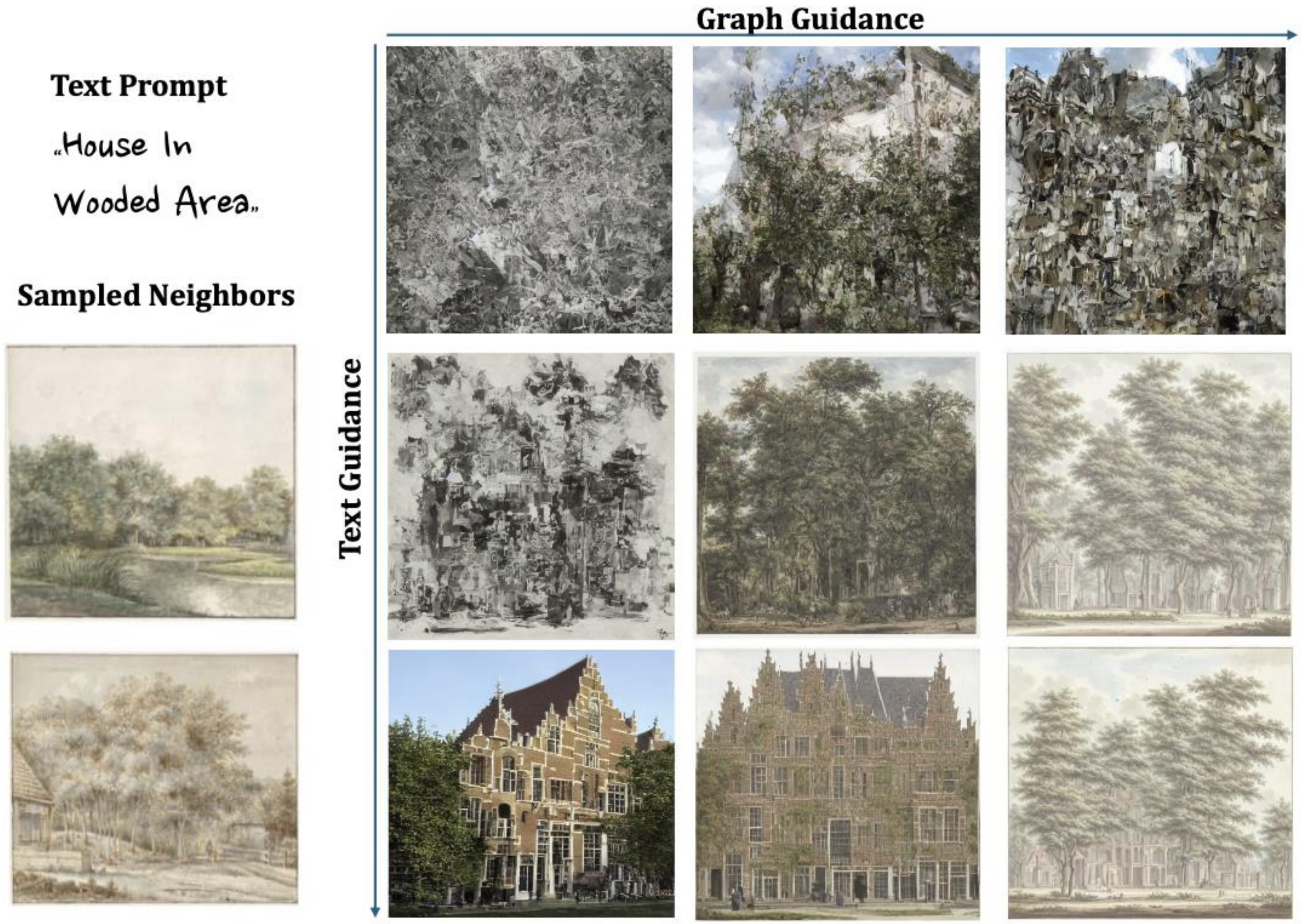
- **Ablation study of Semantic PPR-based Neighbor Sampling**



- Our sampling methods effectively identify neighbor images that contribute most significantly to the ground truth in both semantics and style.

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

Thanks