# DATS369 Machine Learning with Graphs Final Project: Graph Guided Latent Diffusion for Image Generation

**Yuanhe Guo**
NYU Shanghai
yg2709@nyu.edu

## Abstract

Recent advances in diffusion models have enabled high-quality image generation but remain limited in handling multi-image inputs for personalized outputs. This work introduces a novel framework that integrates graph convolutional networks (GCNs) with diffusion models, enabling coherent generation of child images from parent graphs. Inspired by Graph2Pix and Textual Inversion, we replace traditional text conditioning with graph embeddings to enhance control over the diffusion process. Experiments demonstrate that while our method exhibits a performance gap compared to existing approaches, it successfully captures salient visual features from multi-image inputs, opening up a wide range of future reseaches. Source code is available at https://github.com/RicercarG/DATS-SHU-369-ML-with-Graphs-SP25-Homework/tree/main/FinalProject

## 1 Introduction

Recent advances in generative models, such as GAN and Diffusion Models, have demonstrated the remarkable capability of machine learning to synthesize high-resolution images. While these models excel at text-guided generation, they also support high-quality image-to-image translation. However, their functionality is typically limited to single-input scenarios, restricting their potential for personalized output.

This limitation raises a critical challenge: Given a set of user-provided images, how can we generate novel images that effectively incorporate multiple input sources while preserving their semantic and structural relationships?

Existing approaches, such as ControlNet Zhang et al. (2023), enhance control by integrating multiple input modalities (e.g., depth maps). However, these methods often require explicit human intervention and struggle to scale efficiently when processing multiple images simultaneously.

In contrast, Graph2Pix Gokay et al. (2021) offers a promising solution by addressing multi-instance image-to-image translation through a graph-based framework. Based on Pix2PixHD Wang et al. (2018), Graph2Pix replaces the source image with an embedding generated by a Graph Convolutional Network (GCN) Kipf & Welling (2017). This design allows the model to aggregate information from an entire batch of input images.

Inspired by this approach, we propose a novel framework that integrates graph-based feature learning with diffusion models. By leveraging GCN embeddings as dynamic conditioning signals, our method enhances the diffusion process's ability to synthesize coherent outputs from multi-image inputs for personalized generation.

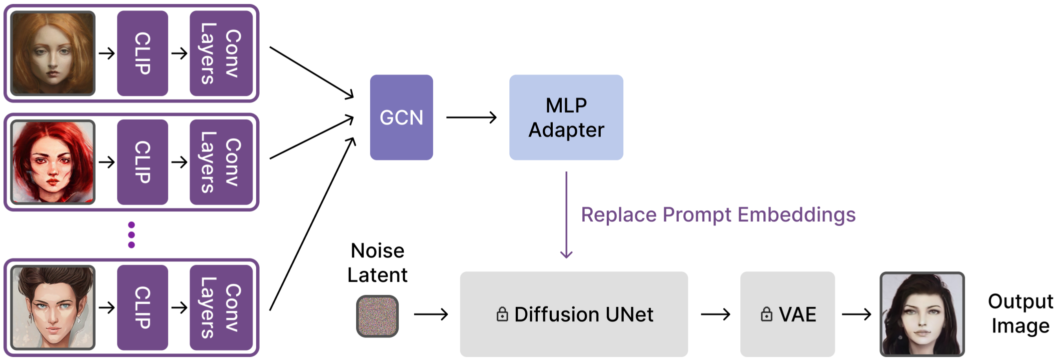

Figure 1: Pipeline of our proposed method

## 2 Methodology

Given the multi-step sampling nature of diffusion models, the image-to-image pipeline lacks control over the direction of the diffusion trajectory. To address this limitation, we draw inspiration from Textual Inversion Gal et al. (2022), a technique for personalizing text-to-image generation by associating prompt concepts with visual features through text encoder fine-tuning. While Textual Inversion only fine-tunes the tokenizer, our method replaces the entire text encoder with a graph-based encoder to better model hierarchical image relationships.

Fig. 1 illustrates the following. Given an input tree structure of parent images, we first process each image using CLIP Radford et al. (2021). Each CLIP embedding is initially a 1-channel tensor containing visual features. However, diffusion models typically require prompt embeddings as 2D tensors of shape (sequence length, feature dimension). To align with this requirement, we duplicate the CLIP embedding across all sequence positions (conventionally being 77 tokens) before applying graph convolutions. For child image generation, we follow the Graph2Pix Gokay et al. (2021) approach by using random noise as a placeholder for the missing child image.

The graph convolutional network (GCN) processes the hierarchical structure, extracting features from parent images. After processing, we extract the features corresponding to the child image and pass them through an MLP to enhance then representational capacity before feeding them into the diffusion model.

During training, we freeze all pretrained components of the diffusion pipeline (CLIP, UNet, and VAE) and only update the parameters of the GCN and MLP over the MSE loss. This strategy preserves the generative capabilities of the base model while enabling hierarchical image conditioning through graph-based feature learning.

## 3 Experiments

### 3.1 Dataset and Evaluation

Graph2Pix Gokay et al. (2021) introduces a dataset curated from Artbreeder [1], a platform enabling users to blend multiple images. For our experiments, we utilize their released subset containing $7,000$ images. Each graph represents a binary tree with a depth of two, generating seven images per graph. This configuration results in a total of $1,000$ graphs, comprising $850$ for training and $150$ for testing. To evaluate performance, we compute the Fréchet Inception Distance (FID) on the $150$ generated images from the test set.

### 3.2 Implementation Details

We did our experiments upon Stable Diffusion 1.5 Rombach et al. (2021). We implemented a 2 layer GCN of 77 channels, followed with a 2 layer MLP with ReLU activation. The total trainable

[1] https://www.artbreeder.com/

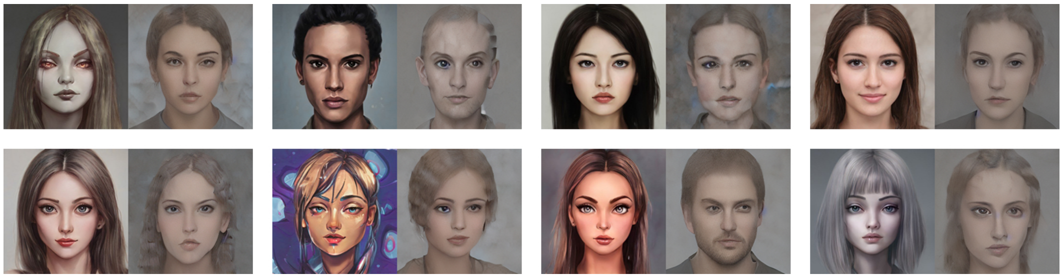

Figure 2: Qualitative results. Within each group, on the left is the ground truth, while on the right is our generated. All images are generated using DDIM Sampler, 20 steps with CFG scale 3.

| **Method** | CFG Scale | **FID ↓** |
|---|---|---|
| Graph2Pix Gokay et al. (2021) (Ours) | / | **62.88** |
| Ours | 7.5 | 423.62 |
| Ours | 3 | 282.51 |
| Ours | 1 | 286.98 |

Table 1: Comparison between our reproduced Graph2Pix model and our proposed method.

parameters are $1.64$ M. The model is trained of $60$ epochs with a constant learning rate of $2e-4$ with AdamW optimizer.

### 3.3 Main Results

Fig. 2 shows some qualitative results of the generated images. Compared with the ground truth image, our model does capture some parent features like eyes and nose. However, the color converges to an average state with low contrast grayish color. We suspect that our model is too small to capture sufficient information.

### 3.4 Ablation Study

As we use graph features as classifier-free guidance (CFG) during image generation, the CFG scale emerges as a critical hyperparameter during inference. If the CFG scale is set too high, the output tends to resemble random noise; conversely, an excessively low scale results in unstructured or uncontrolled outputs.

Table 1 presents quantitative results for different CFG scales, comparing our model's performance against a reproduced version of Graph2Pix on the same dataset. Results highlights the sensitivity of generative quality to CFG settings. However, our method with the optimal CFG scale still exhibits suboptimal results under identical conditions. This suggests room for improvement in our framework's conditioning mechanism.

## 4 Conclusions

In this work, we presented a novel approach to leveraging graph-structured input data as classifier-free guidance during diffusion model image generation. While our method exhibits a performance gap compared to the Graph2Pix framework, it demonstrates potential in capturing visual features from the input graph of images. This suggests that graph-based representations can serve as meaningful conditioning signals for diffusion models, even with limited training data or architectural constraints.

This work opens several promising directions for future research: (1) Investigate alternative feature extractors, such as VGG, to complement or replace CLIP for capturing both visual and semantic information from input images. (2) Explore the integration of full graph embeddings as conditional inputs, rather than relying only on the target child node's embedding. (3) Configure more trainable parameters for the model for better learning capacity. (4) Evaluate performance on diverse datasets beyond human faces to assess generalization across domains.

## References

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-Or, D. An image is worth one word: Personalizing text-to-image generation using textual inversion. *ArXiv*, abs/2208.01618, 2022. URL `https://api.semanticscholar.org/CorpusID:251253049`.

Gokay, D., Simsar, E., Atici, E., Ahmetoglu, A., Yuksel, A. E., and Yanardag, P. Graph2pix: A graph-based image to image translation framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 2001–2010, October 2021.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=SJU4ayYgl`.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models, 2021.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., and Catanzaro, B. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Zhang, L., Rao, A., and Agrawala, M. Adding conditional control to text-to-image diffusion models, 2023.