CS5340 Project Abstract

Ye Fangda, Bao Mingyang, Zhao Yilan, Yao Xiaolu, Hu Shiqi
School of Computing
National University of Singapore
{e1442448, e1351071, e0735460, e1503349, e1503369}@u.nus.edu

1 Introduction

Generating images that faithfully reflect the nuances and details of text prompts remains a paramount challenge in the field of text-to-image (T2I) diffusion models [1, 2, 3, 4, 5]. Extensive research [6, 7, 8, 9, 10] has underscored the pivotal role of text prompts in guiding the image synthesis process, with text embeddings being extensively explored to enhance image quality and fidelity.

Recent studies [11, 12, 13, 14, 15, 16] have further revealed that beyond text prompts, the input noise itself plays an equally critical role in shaping the generated imagery. Specifically, carefully selected or optimized noises, termed ’golden noises’, have been shown to facilitate the synthesis of images with improved semantic alignment to text prompts and enhanced overall quality. Despite the progress made by these methods [11, 13], several practical challenges remain, including difficulties in generalizing across diverse datasets and diffusion models, substantial time delays introduced by optimization processes, and a lack of adaptability to varying diffusion architectures without significant modifications.

Zhou et al.[16] addressed these challenges by proposing a novel noise prompt learning framework, termed NPNet, which efficiently transforms random noise into golden noise through a single forward pass of a compact neural network. Their work demonstrated significant improvements in image quality and generalization ability across various diffusion models, while introducing minimal computational overhead due to its plug-and-play nature.

However, their method implicitly assumes a consistent influence of text prompts on noise across all denoising timesteps. We posit that this assumption, while computationally efficient, may overlook the nuanced temporal interplay between text conditioning and optimal noise perturbation throughout the reverse diffusion process. Specifically, we hypothesize that the text prompt’s influence on the ideal noise transformation is not static, but rather evolves dynamically across different denoising stages. Therefore, in this exploratory course project, we aim to investigate the incorporation of temporal dependency into the noise prompt learning mechanism. We believe that explicitly modeling this temporal dimension could unlock further improvements in the quality and text alignment of generated images, by enabling a more fine-grained and stage-aware control over noise perturbation.

2 Related Work

2.1 Text-to-Image Synthesis with Diffusion Models

T2I synthesis using diffusion models has seen significant advancements in recent years. Diffusion models, initially introduced for generative tasks, have demonstrated remarkable success in generating high-quality images that align well with given text prompts. Ho et al. [17] proposed denoising diffusion probabilistic models (DDPMs), which laid the foundation for modern text-to-image generation. Ramesh et al. [18, 19] introduced the DALL·E models, leveraging transformer-based architectures for high-resolution image synthesis from text. More recent works, such as Stable Diffusion [5], have optimized diffusion processes to generate high-quality images in a more computationally efficient manner.

Preprint. Under review.

Classifier-free guidance (CFG) [20] has become a widely used technique for improving text alignment in T2I models, allowing models to generate images that better reflect the semantics of input text prompts. Additionally, models such as Imagen [21] and Parti [22] have further pushed the boundaries of photorealistic image synthesis by leveraging large-scale datasets and advanced architectural designs. Despite these advances, challenges remain in improving semantic faithfulness, compositional coherence, and efficiency in the image synthesis pipeline.

2.2 Noise Prompt Learning

Recent studies have identified the crucial role of noise in the diffusion process, beyond just text prompts. Optimizing noise input has been shown to significantly impact the quality and semantic alignment of generated images. Lugmayr et al. [14] proposed modifying the reverse diffusion process to refine generated images based on unmasked regions. Meng et al. [23] demonstrated that iterative denoising of the initial noise can lead to improved image quality compared to standard diffusion methods. Qi et al. [15] explored noise selection strategies to reduce truncation errors and enhance text-image consistency.

In an effort to systematically optimize noise, Chefer et al. [11] introduced Generative Semantic Nursing (GSN), which adjusts noise perturbations at each denoising step to better encode semantic information from text prompts. Similarly, InitNO [13] introduced an initial noise partitioning method to guide noise optimization, improving performance in compositional generalization tasks. However, these methods often suffer from limited generalization across datasets and diffusion models, high computational costs, and the need for manual fine-tuning.

To address these limitations, Zhou et al. [16] proposed NPNet, a lightweight neural network that learns to transform random noise into "golden noise" through a single forward pass. This approach enables more efficient and generalizable noise optimization without modifying the underlying diffusion model architecture. While NPNet improves text-image alignment and computational efficiency, it assumes a static relationship between noise and text prompts across all denoising timesteps. Our work builds on these findings by incorporating temporal dependency into the noise prompt learning framework, allowing for stage-aware noise transformations that better capture the evolving influence of text conditioning throughout the reverse diffusion process.

3 Methodology (Exploratory Stage)

3.1 Problem Formulation and Motivation

Current T2I diffusion models have demonstrated impressive capabilities, yet precise control over image generation based on text prompts remains a significant challenge (Figure 1, left). Recent advancements in Noise Prompt Learning, as exemplified by Zhou et al. [16], have highlighted the importance of optimizing the noise input in conjunction with text prompts (Figure 1, middle). These methods represent a notable step forward; however, they typically treat the noise prompt as a static entity applied uniformly across all denoising timesteps.

We hypothesize that this static noise prompt approach may overlook the inherent temporal dynamics of the diffusion process. Specifically, we believe that the optimal influence of text prompts on noise perturbation is not constant but rather varies throughout the reverse diffusion trajectory. Exploiting these temporal dependencies could potentially lead to more nuanced and higher-quality image generation. Therefore, we aim to investigate the incorporation of temporal awareness into noise prompt learning. Our overarching goal is to explore how to dynamically adapt noise prompts based on the evolving generation state to enhance text-image alignment and potentially address uncertainties inherent in the diffusion process.

3.2 Exploratory Ideas and Methodology

Given the exploratory nature of this course project and the emphasis on Uncertainty AI, our methodology is currently focused on investigating the following key ideas:

1. **Dataset Construction with Temporal Information:** We plan to explore modifications to the dataset construction process to explicitly incorporate temporal information relevant to the diffusion

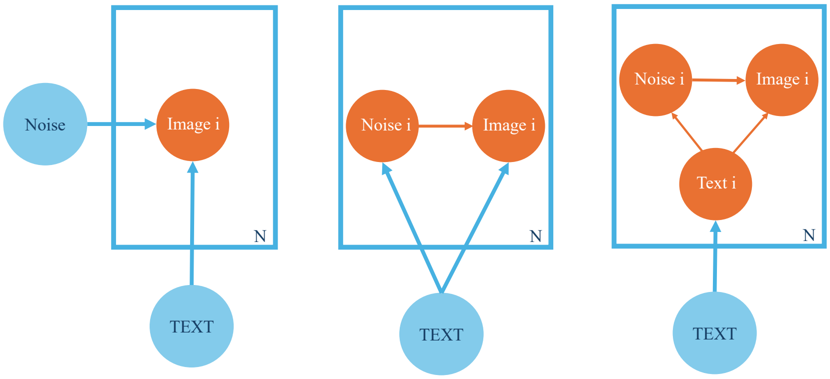

Figure 1: Evolution of Text-to-Image Model Paradigms. (Left) Traditional T2I models primarily condition on text prompts. (Middle) Current Noise Prompt frameworks consider both text and static noise prompts. (Right) Our exploratory approach: Temporally-Aware Noise Prompt Learning, incorporating time-varying noise manipulation.

process. This might involve analyzing and leveraging intermediate representations or timestep-specific features within existing diffusion model training datasets.

2. **Developing a Temporally-Aware Noise Prompt Model:** Our primary focus is on designing and exploring models capable of generating time-varying noise prompts. This entails investigating architectures that can adapt the noise prompt based on the current denoising timestep and potentially past states. Key questions we aim to address include:

   - *Markovian vs. Non-Markovian Modeling:* We will investigate the validity of the Markov assumption in noise evolution. Specifically, we will explore whether modeling noise transitions as a first-order Markov process is sufficient, or if incorporating information from states beyond the immediately preceding timestep is necessary for capturing richer temporal dependencies.
   - *Feedback Mechanisms (Image-Noise Feedback):* We will also tentatively explore the potential benefits of introducing feedback loops, where the characteristics of the partially generated image at each timestep could influence the noise prompt generation for subsequent steps. This would allow for a more interactive and potentially uncertainty-aware generation process.

3. **Computational Feasibility and Simplification:** Recognizing the computational constraints of a course project and the limited timeframe, we will prioritize approaches that are computationally feasible and can be implemented and evaluated within the given resources. This may involve simplifying the model architecture by focusing on adding temporal encodings or lightweight modules to existing diffusion models, rather than undertaking major architectural overhauls. This pragmatic approach will allow us to effectively explore the core concept of temporally-aware noise prompts within the project scope.

3.3 Uncertainty AI Theme Integration

Our exploration of temporally-aware noise prompts directly aligns with the theme of Uncertainty AI. By explicitly considering the temporal evolution of noise and its interaction with text prompts, we aim to better understand and potentially control the inherent stochasticity in diffusion models. Furthermore, by investigating feedback mechanisms and exploring different uncertainty quantification methods (as mentioned in Key Innovation 3, although not yet fully elaborated in our current exploratory approach), we seek to gain insights into how noise variations contribute to the uncertainty and robustness of text-to-image synthesis. This project, even in its preliminary stage, represents an initial step towards incorporating uncertainty awareness into the core mechanisms of generative diffusion models.

4 Ethical/Social Impact Statement

1. **Deepfakes and Misinformation Dissemination:** Diffusion models possess the capability to generate exceptionally realistic synthetic media, dramatically reducing the effort and resources

required to fabricate fake news, manipulate media content, or commit identity fraud. This poses substantial risks to public trust and societal stability. AI-generated images and videos could be readily weaponized for purposes of political manipulation, financial scams, and targeted malicious campaigns, eroding confidence in authentic information and potentially destabilizing social order.

2. **Intellectual Property and Copyright Challenges:** Diffusion models are trained on large-scale datasets, often containing copyrighted material. This leads to legal disputes over the authorship and ownership of AI-generated content. Artists and content creators may argue that AI-generated works infringe on their intellectual property rights. Therefore, establishing fair and ethical data usage policies is crucial for balancing innovation and copyright protection.

5 AI Tool Use

1. We use large language models (LLMs) to assist in searching for relevant papers and materials.
2. We utilize LLMs to assist us in understanding complex concepts and sentences in research papers.
3. We leverage LLMs to support our coding process.
4. We use LLMs to help us polish our paper.

References

[1] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

[2] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

[3] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[4] Pablo Pernias, Dominic Rampas, Mats L Richter, Christopher J Pal, and Marc Aubreville. Würstchen: An efficient architecture for large-scale text-to-image diffusion models. *arXiv preprint arXiv:2306.00637*, 2023.

[5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[6] K Team. Kolors: Effective training of diffusion model for photorealistic text-to-image synthesis. *arXiv preprint*, 2024.

[7] Buhua Liu, Shitong Shao, Bao Li, Lichen Bai, Zhiqiang Xu, Haoyi Xiong, James Kwok, Sumi Helal, and Zeke Xie. Alignment of diffusion models: Fundamentals, challenges, and future. *arXiv preprint arXiv:2409.07253*, 2024.

[8] Michael Toker, Hadas Orgad, Mor Ventura, Dana Arad, and Yonatan Belinkov. Diffusion lens: Interpreting text encoders in text-to-image pipelines. *arXiv preprint arXiv:2403.05846*, 2024.

[9] Qiucheng Wu, Yujian Liu, Handong Zhao, Ajinkya Kale, Trung Bui, Tong Yu, Zhe Lin, Yang Zhang, and Shiyu Chang. Uncovering the disentanglement capability in text-to-image diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1900–1910, 2023.

[10] Hu Yu, Hao Luo, Fan Wang, and Feng Zhao. Uncovering the text embedding in text-to-image diffusion models. *arXiv preprint arXiv:2404.01154*, 2024.

[11] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4):1–10, 2023.

[12] Sherry X Chen, Yaron Vaxman, Elad Ben Baruch, David Asulin, Aviad Moreshet, Kuo-Chin Lien, Misha Sra, and Pradeep Sen. Tino-edit: Timestep and noise optimization for robust diffusion-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6337–6346, 2024.

[13] Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. Initno: Boosting text-to-image diffusion models via initial noise optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9380–9389, 2024.

[14] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022.

[15] Zipeng Qi, Lichen Bai, Haoyi Xiong, and Zeke Xie. Not all noises are created equally: Diffusion noise selection and optimization. *arXiv preprint arXiv:2407.14041*, 2024.

[16] Zikai Zhou, Shitong Shao, Lichen Bai, Zhiqiang Xu, Bo Han, and Zeke Xie. Golden noise for diffusion models: A learning framework. *arXiv preprint arXiv:2411.09502*, 2024.

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[18] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[19] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[20] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[21] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[22] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

[23] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14297–14306, 2023.