Do It Yourself (DIY): Modifying Images for Poems in a Zero-Shot Setting Using Weighted Prompt Manipulation

Sofia Jamil[1] **Kotla Sai Charan**[1] **Sriparna Saha**[1] **Koustava Goswami**[2] **K J Joseph**[2]
[1] Department of Computer Science & Engineering, Indian Institute of Technology Patna, India
[2] Adobe Research
[1]{sofia_2321cs16, kotla_2101mc27, sriparna}@iitp.ac.in
[2]{koustavag, josephkj}@adobe.com

Abstract

Poetry is an expressive form of art that invites multiple interpretations, as readers often bring their own emotions, experiences, and cultural backgrounds into their understanding of a poem. Recognizing this, we aim to generate images for poems and improve these images in a zero-shot setting, enabling audiences to modify images as per their requirements. To achieve this, we introduce a novel *Weighted Prompt Manipulation (WPM)* technique, which systematically modifies attention weights and text embeddings within diffusion models. By dynamically adjusting the importance of specific words, *WPM* enhances or suppresses their influence in the final generated image, leading to semantically richer and more contextually accurate visualizations. Our approach exploits diffusion models and large language models (LLMs) such as GPT in conjunction with existing poetry datasets, ensuring a comprehensive and structured methodology for improved image generation in the literary domain. To the best of our knowledge, this is the first attempt at integrating weighted prompt manipulation for enhancing imagery in poetic language. Resources related to data and codes are available here: DIY

1 Introduction

Recent advancements in diffusion models have transformed the landscape of generative AI. These text to image generation models are pretrained on vast datasets of image-text pairs (Schuhmann et al., 2021, 2022), and leverage state-of-the-art techniques, including large-scale pre-trained language models (Devlin et al., 2019; Xia et al., 2021; Brown et al., 2020), variational autoencoders (Kingma et al., 2019), and diffusion-based architectures (Ramesh et al., 2021; Rombach et al., 2022). As a result, they excel in generating highly realistic and visually compelling images. However, current diffusion models often struggle to interpret metaphorical language, symbolism, and nuanced themes. Therefore, creative fields like poetry fail to directly generate relevant visuals and often lead to inconsistent or inaccurate visual outputs. To address this limitation, we propose **Weighted Prompt Manipulation**, a novel approach illustrated in Figure 1, designed to refine generated images in a real-time setting and adjust their alignment, especially for poetic content. Existing text-to-image editing techniques (Abdal et al., 2021; Bau et al., 2020; Lang et al., 2021) have demonstrated remarkable success in tasks such as image translation, style transfer, and appearance modification, all while preserving structural integrity and scene composition. Among these methods, attention layers play a pivotal role in regulating image layout and ensuring a coherent relationship between the generated image and its textual prompt. However, these techniques have not yet been applied in the domain of poems. Therefore, motivated by this, we specifically investigate the attribution of image generation in diffusion models, posing a fundamental question: *How do diffusion models generate images for poems?* To explore this, we employ prompt tuning and a systematic analysis of attention map generation, providing deeper insights into the underlying mechanisms of poem to image synthesis using diffusion-based models. Building on our findings, we introduce *Weighted Prompt Manipulation*, a technique designed to enhance image generation for poetic inputs by improving relevance and fidelity. Our key contributions include:

1. We introduce a new task of poem visualization, focusing on generating images that accurately capture the rich and intricate details conveyed in poetic text.

2. We propose a training-free **Weighted Prompt Manipulation** approach, which manipulates images by dynamically adjusting word importance in a real-time setting.

3. We provide a detailed analysis of text-to-image generation within diffusion models, leveraging heat

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 19657–19665
November 4-9, 2025

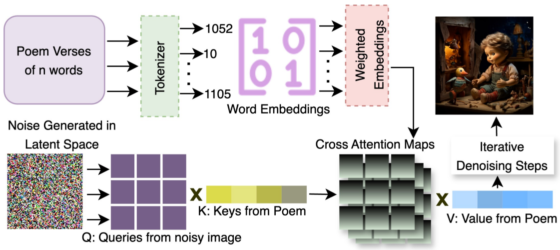

Figure 1: Architectural diagram of the poem-to-image generation process using our proposed *Weighted Prompt Manipulation* technique.

maps and attention maps to better understand how different parts of a poem influence image generation.

4. We conduct extensive quantitative and human evaluations to demonstrate that the diffusion model can be manipulated to enhance image generation by selectively reinforcing specific textual elements without significantly altering the existing visual composition.

2 Background and Related Works

Recent advancements in text-driven image manipulation have been significantly influenced by GANs (Brock et al., 2018; Karras et al., 2021, 2019) combined with image-text representations like CLIP (Radford et al., 2021). These approaches enable realistic image modifications using textual input (Gal et al., 2022b; Andonian et al., 2021; Patashnik et al., 2021; Goswami et al., 2024; Agarwal et al., 2023) However, while they perform well in structured domains (e.g., human face editing), they often struggle with diverse datasets where subjects vary significantly. To address this, fine-tuning methods (Houlsby et al., 2019; Ahn et al., 2024; Frenkel et al., 2024) allow models to learn novel styles from just a few images. However, these methods are prone to overfitting, leading to image degradation or content leakage. Alternative approaches, such as Textual Inversion (Gal et al., 2022a) and Hard Prompt Made Easy (PEZ) (Wen et al., 2023), aim to find optimal text representations (e.g., embeddings or tokens) that capture an object's characteristics without modifying the underlying text-to-image model parameters. Another line of research focuses on encoder-based methods (Chen et al., 2023; Gao et al., 2024; Wang et al., 2024; Wang et al.), which use visual encoders to extract image features and map them to text prompts. While these methods have set the standard in state-of-the-art performance, they remain limited by the capabilities of visual encoders, which often struggle with capturing fine-grained textures beyond abstract style information.

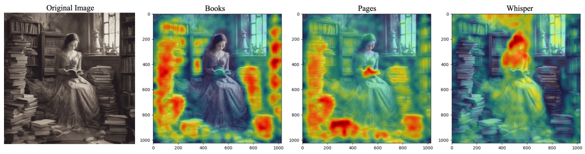

Figure 2: Heat maps for the generated image highlighting captured and missed words from the prompt. Readers are encouraged to zoom in for improved visibility.

3 Tasks Setups

Challenges in the Poem to Image generation: Building on the efficacy of the Playground (Liu et al., 2024) diffusion model in image generation, we conduct an in-depth analysis of how diffusion models process different words in poetry (Jamil et al., 2025a,c). As illustrated in Figure 2, diffusion models exhibit a strong bias toward visual elements, with the highest attention given to concrete objects ('books', 'page'). Diffusion models leverage CLIP embeddings, which are inherently designed to align textual descriptions with corresponding visual features. As a result, CLIP embeddings emphasize words containing visual objects, as they provide explicit semantic grounding for image synthesis. Additionally, the cross-attention mechanism in diffusion models determines how strongly each word contributes to the generated image. Certain words tend to have higher attention scores, guiding the model's output more effectively, whereas others, being more contextual than structural, receive lower attention weights and have less impact on the final image.

Proposed Solution:

To address the inherent bias of diffusion models toward certain words and their limited attention to others, we propose ***Weighted Prompt Manipulation (WPM)*** approach. As demonstrated in Figure 1, it is a systematic approach to dynamically adjust word influence during image generation. By assigning custom weight values to specific words in the prompt, we can enhance the model's focus on critical poetic elements, ensuring a more faithful and semantically rich visual representation. In our approach, words that naturally receive high attention are explicitly reinforced, while those that receive lower attention are strategically amplified to balance their contribution. Diffusion models use cross-attention mechanisms to determine the importance of each word in a text prompt. *WPM* modifies the default attention scores by explicitly assigning weights to different words, guiding the model to generate images that more accurately capture the semantic depth and poetic meaning. Each word in the prompt is assigned a scaling factor in parentheses. Words with higher weights are given greater prominence in the generated image, while those with lower weights are de-emphasized. As illustrated in Figure 4, the subsequent images are produced using *WPM*. To understand the weighting of certain words that are visually significant in poetry, we employed `GPT-4o-mini` for image instruction generation. Our method begins by providing GPT with an initial prompt as demonstrated in Figure 3:

> Refine the following poem into a weighted text prompt for a diffusion Model. Only apply weights to the most important visual words. Your response should only contain the weighted poem.

Figure 3: Initial Prompt for WPM

<Input Poem>
"Little girl, little girl, Where have you been?"
"Gathering roses, To give to the Queen."
"Little girl, little girl, What she gave you?
"She gave me diamond, As big as my shoe."
<GPT's Response (Weighted Prompt)>
Little girl, little girl, (girl:1.6) Where have you been?"
"Gathering (roses:1.7), To give to the (Queen:1.6)."
"Little girl, little girl, (girl:1.6) What she gave you?
"She gave me (diamond:1.8), As big as my (shoe:1.5)."

Figure 4: Examples of images generated using various weighted prompts, with corresponding weights displayed below each image.

The weighted output generated by GPT is then passed to the diffusion model. *WPM* processes input text by identifying attention markers, such as (word:1.5) to increase emphasis. The corresponding weights are applied to the text embeddings of their respective words and integrated into the model's dual text encoders. The final weighted embeddings are then used to condition image generation. *(Readers are encouraged to explore the implementation of WPM).*

4 Experiments

4.1 Implementation Details

We implemented our *WPM* approach using three text-to-image models, Playground V3 (Liu et al., 2024), Stable Diffusion XL (Podell et al., 2024), and Sana (Xie et al., 2024), selected for their training-free, pluggable design, enabling cross-architecture comparisons. To evaluate the alignment between the generated images and poems we employed BLIP (Li et al., 2022) to generate captions for the images and measure their similarity to the original poem. Similarly, we applied Long-CLIP (Zhang et al., 2024) to compute the cosine similarity between the poem and the generated image. Experiments were conducted on two benchmark datasets: *PoemSum* (Mahbub et al., 2023), containing 3,011 poems with curated English summaries from Poem Analysis, and *MiniPo* (Jamil et al., 2025b), comprising 1,001 nursery rhymes, both sourced from online platforms.

4.2 Results and Discussions

4.2.1 Quantitative Evaluation

To evaluate the effectiveness of our proposed methodology, we present the results in Table 1. Our *Weighted Prompt Manipulation* approach consistently outperforms direct poem as prompts. Given that the Long-CLIP score measures semantic consistency between text and image, the results demonstrate that incorporating weighted poems into diffusion models yields higher scores, particularly when using the optimal prompt refined through human feedback. Notably, our *WPM* technique is broadly applicable to all Stable Diffusion style models. Ex-

		Direct Poem	Prompt 1	Prompt 2	Prompt 3	Prompt 4
BLIP	Stable Diffusion	0.2243	0.2325	0.2340	0.2412	0.2352
	Playground V3	0.3296	0.3270	0.3408	0.3317	0.3272
	Sana	0.3148	0.3365	0.3380	0.3356	0.3354
LongClip	Stable Diffusion	0.2391	0.2245	0.2112	0.2309	0.2273
	Playground V3	0.2480	0.2449	0.2489	0.2507	0.2494
	Sana	0.2286	0.2388	0.2387	0.2384	0.2418

Table 1: Quantitative evaluation of generated images using different diffusion models on different prompts.

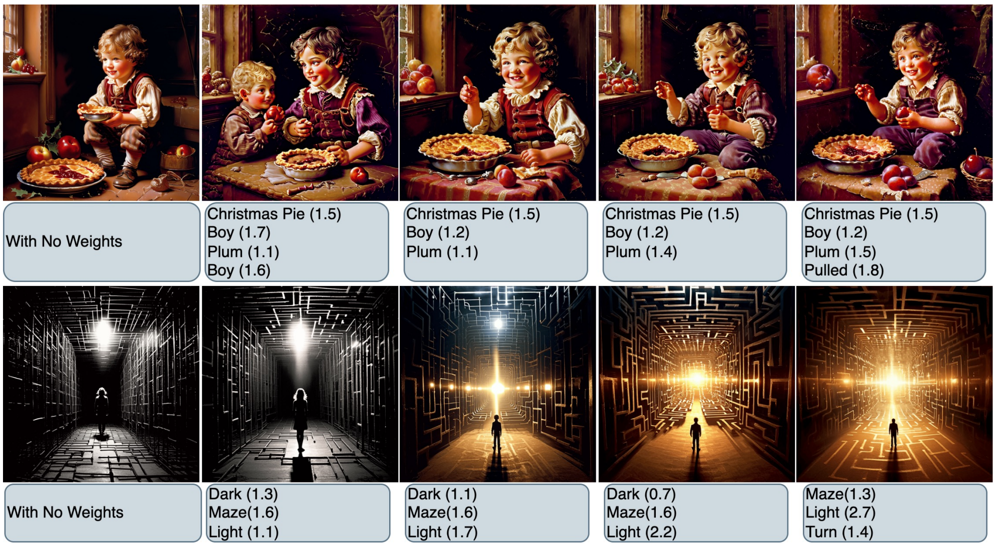

Figure 5: A comparison of generated images using different weights for various words in the same poem. All poems, along with their corresponding weighted prompts specified in the poem, are provided in the Appendix Table 3.

perimental results on SD 3.5 Medium and Playground v3 further validate the adaptability of our approach across various diffusion-based models.

Metrics	WPM	Without WPM(Only Poem)
Meaning	4.3	3.8
Visual Objects (Nouns)	4.2	4.35
Image Aesthetics	3.1	3
Action Depicted (Verbs)	3.9	3.2

Table 2: The results of Human Evaluation Scores in terms of expert ratings (1-5).

4.2.2 Human Evaluation

Given that existing automated metrics may not fully capture the quality of generated images and with no standardized metric available, we incorporated human evaluations. We selected 5% of the samples from the *PoemSum* dataset and had domain experts review images generated both with and without our *WPM* approach. Each image was evaluated based on four key criteria: interpretability of meaning, visual objects, image aesthetics, and action depicted. Participants rated each sample on a scale of 1 to 5, with higher scores indicating better quality. The final rating for each image was determined by averaging the scores provided by three experts. To ensure unbiased assessments, the evaluators were not informed of the model used to generate each image. As shown in Table 2, the results demonstrate that *WPM* significantly improves image generation in terms of semantic meaning and alignment. Moreover, we conducted qualitative evaluations to compare the results of *Weighted Prompt Manipulation* with those generated without it. Our observations indicate that images produced using weighted prompts are able to incorporate certain key elements that were otherwise missing when plain poems were used as prompts. As illustrated in Figure 5, when the diffusion model processes only the raw poem, the generated images tend to emphasize specific words *(pie, landscape, maze)* while completely ignoring others *(plum, smoke, light)*. However, by assigning greater importance to the previously ignored words, the updated images successfully incorporate those elements alongside the already emphasized ones.

5 Conclusion

In this work, we propose the task of poem-to-image manipulation based on the reader's interpretation in a zero-shot setting. Our novel *Weighted Prompt Manipulation* technique systematically modifies attention weights and text embeddings within diffusion models to add or remove certain elements in the poem-to-image generation. To evaluate the effectiveness of our method, we conduct extensive experiments on benchmark poetry visualization datasets. Our evaluation framework includes human assessments, qualitative analyses, and quantitative metrics, ensuring a comprehensive assessment of our approach. In future work, we aim to apply consistent weighted attention to phrases instead of individual words, making it a scalable poetry visualization tool that enables real-world applications in education, cultural preservation, and literary content creation.

6 Limitation

A key limitation of our *Weighted Prompt Manipulation (WPM)* approach is its effectiveness in handling poems that lack explicit visual elements or rely heavily on abstract concepts. Since our method primarily enhances image generation by adjusting prompt weights based on the presence of tangible objects and discernible themes, it struggles with highly conceptual or non-visual poetry. In such cases, where the essence of the poem cannot be easily translated into concrete imagery, *WPM* fails to introduce significant variations in the generated outputs. As a result, the images produced remain largely similar across different prompts, limiting the impact of our approach in capturing the deeper, non-representational meanings of such poems.

7 Ethical Consideration

A key ethical consideration involves the inherent biases present in diffusion models, which may reflect societal, cultural, or data-driven biases from the pre-trained models. These biases can potentially influence the generation of images related to poems on specific topics or forms, resulting in unfair or inappropriate outputs. To ensure compliance and ethical integrity, we also obtained formal approval from our institute's ethical review board (ERB) before utilizing the dataset and models for research purposes.

8 Acknowledgement

Sriparna Saha would like to acknowledge the funding from ADOBE Research for conducting this research.

References

Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. 2021. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (ToG)*, 40(3):1–21.

Aishwarya Agarwal, Srikrishna Karanam, K. J. Joseph, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. 2023. A-STAR: test-time attention segregation and retention for text-to-image synthesis. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 2283–2293. IEEE.

Namhyuk Ahn, Junsoo Lee, Chunggi Lee, Kunhee Kim, Daesik Kim, Seung-Hun Nam, and Kibeom Hong. 2024. Dreamstyler: Paint by style inversion with text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 674–681.

Alex Andonian, Sabrina Osmany, Audrey Cui, YeonHwan Park, Ali Jahanian, Antonio Torralba, and David Bau. 2021. Paint by word. *arXiv preprint arXiv:2103.10951*.

David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. 2020. Semantic photo manipulation with a generative image prior. *arXiv preprint arXiv:2005.07727*.

Andrew Brock, Jeff Donahue, and Karen Simonyan. 2018. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jingwen Chen, Yingwei Pan, Ting Yao, and Tao Mei. 2023. Controlstyle: Text-driven stylized image generation using diffusion priors. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7540–7548.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Yarden Frenkel, Yael Vinker, Ariel Shamir, and Daniel Cohen-Or. 2024. Implicit style-content separation using b-lora. In *European Conference on Computer Vision*, pages 181–198. Springer.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022a. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022b. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13.

Junyao Gao, Yanchen Liu, Yanan Sun, Yinhao Tang, Yanhong Zeng, Kai Chen, and Cairong Zhao. 2024. Styleshot: A snapshot on any style. *arXiv preprint arXiv:2407.01414*.

Koustava Goswami, Srikrishna Karanam, Prateksha Udhayanan, K. J. Joseph, and Balaji Vasan Srinivasan.

2024. Copl: Contextual prompt learning for vision-language understanding. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18090–18098. AAAI Press.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Sofia Jamil, Bollampalli Areen Reddy, Raghvendra Kumar, Sriparna Saha, Koustava Goswami, and K. J. Joseph. 2025a. Poemtale diffusion: Minimising information loss in poem to image generation with multi-stage prompt refinement. *Preprint*, arXiv:2507.13708.

Sofia Jamil, Bollampalli Areen Reddy, Raghvendra Kumar, Sriparna Saha, Joseph K. J, and Koustava Goswami. 2025b. Poetry in pixels: Prompt tuning for poem image generation via diffusion models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9224–9237, Abu Dhabi, UAE. Association for Computational Linguistics.

Sofia Jamil, Bollampalli Areen Reddy, Raghvendra Kumar, Sriparna Saha, K J Joseph, and Koustava Goswami. 2025c. Poetry in pixels: Prompt tuning for poem image generation via diffusion models. *Preprint*, arXiv:2501.05839.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2021. Alias-free generative adversarial networks. *Advances in neural information processing systems*, 34:852–863.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410.

Diederik P Kingma, Max Welling, et al. 2019. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392.

Oran Lang, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T Freeman, Phillip Isola, Amir Globerson, Michal Irani, et al. 2021. Explaining in style: training a gan to explain a classifier in stylespace. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 693–702.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *Preprint*, arXiv:2201.12086.

Bingchen Liu, Ehsan Akhgari, Alexander Visheratin, Aleks Kamko, Linmiao Xu, Shivam Shrirao, Chase Lambert, Joao Souza, Suhail Doshi, and Daiqing Li. 2024. Playground v3: Improving text-to-image alignment with deep-fusion large language models. *Preprint*, arXiv:2409.10695.

Ridwan Mahbub, Ifrad Khan, Samiha Anuva, Md Shihab Shahriar, Md Tahmid Rahman Laskar, and Sabbir Ahmed. 2023. Unveiling the essence of poetry: Introducing a comprehensive dataset and benchmark for poem summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14878–14886, Singapore. Association for Computational Linguistics.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2085–2094.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

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

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

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Haofan Wang, Matteo Spinelli, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. 2024. Instantstyle:

Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*.

Zhouxia Wang, Xintao Wang, Liangbin Xie, Zhongang Qi, Ying Shan, Wenping Wang, and Ping Luo. Styleadapter: A unified stylized image generation model without test-time fine-tuning.

Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2023. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *Advances in Neural Information Processing Systems*, 36:51008–51025.

Qiaolin Xia, Haoyang Huang, Nan Duan, Dongdong Zhang, Lei Ji, Zhifang Sui, Edward Cui, Taroon Bharti, and Ming Zhou. 2021. Xgpt: Cross-modal generative pre-training for image captioning. In *Natural Language Processing and Chinese Computing: 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13–17, 2021, Proceedings, Part I 10*, pages 786–797. Springer.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. 2024. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *Preprint*, arXiv:2410.10629.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. 2024. Long-clip: Unlocking the long-text capability of clip. *arXiv preprint arXiv:2403.15378*.

A List of Prompts

This section presents the complete set of prompts we provided to GPT for the purpose of generating weighted prompts tailored to each poem. These prompts are specifically designed to emphasize particular words or phrases within the poetic text, such as key metaphors, emotionally rich expressions, or visually evocative language to guide the image generation model (e.g., Stable Diffusion) in focusing more on those elements. By strategically assigning higher importance to selected terms, we ensure that the resulting visual outputs more accurately reflect the intended artistic and semantic essence of the original poem.

Poem	Weighted Prompt 1	Weighted Prompt 2	Weighted Prompt 3	Weighted Prompt 4
Little Jack Horner	Little Jack Horner (boy:1.5)	Little Jack Horner	(Little:0.9) (Jack:1.5) (Horner:1.5)	Little Jack (Horner:1.5)
Sat in a corner,	Sat in a (corner:1.5),	Sat in a (corner:1.2),	Sat in a (corner:1.7),	Sat in a (corner:1.4),
Eating his Christmas pie; Ce	Eating his (Christmas pie:1.6);	Eating his (Christmas pie:1.5);	(Eating:0.9) his (Christmas:1.6) (pie:1.6);	Eating his (Christmas:1.3) (pie:1.2);
He put in his thumb,	He put in his (thumb:1.5),	He put in his (thumb:1.2),	He put in his (thumb:1.5),	He put in his (thumb:1.1),
And he pulled out a plum,	And he pulled out a (plum:1.6),	And he pulled out a (plum:1.3),	And he pulled out a (plum:1.7),	And he pulled out a (plum:1.6),
And said, “What a good boy am I!"	And said, “What a good (boy:1.5) am I!”	And said, “What a good boy am I!”	And said, “What a good boy am I!”	And said, “What a (good:0.8) (boy:0.9) am I!”
	What (sound:1.6) was that?	What (sound:1.2) was that?	What (sound:1.7) was that?	What (sound:1.8) was that?
What sound was that?	I turn away, into the (shaking:1.7) room.	I turn away, into the (shaking:1.1) (room:1.3).	I turn away, into the (shaking:1.5) room.	I turn away, into the (shaking:1.5) (room:1.3).
I turn away, into the shaking room.	What was that (sound:1.6) that	What was that (sound:1.2) that	What was that (sound:1.7) that	What was that (sound:1.8)
What was that sound that came in on the dark?	came in on the (dark:1.5)?	came in on the (dark:1.1)?	came in on the (dark:1.5)?	that came in on the (dark:1.4)?
What is this maze of light it leaves us in?	What is this (maze:1.5) of	What is this (maze:1.2) of	What is this (maze:1.6) of	What is this (maze:1.6) of
What is this stance we take,	(light:1.6) it leaves us in?	(light:1.3) it leaves us in?	(light:1.6) it leaves us in?	(light:1.5) it leaves us in?
To turn away and then turn back?	What is this (stance:0.8) we take,	What is this (stance:1.1) we take,	What is this (stance:0.9) we take,	What is this (stance:1.2) we take,
What did we hear?	To turn away and then turn back?	To turn away and then turn back?	To turn away and then turn back?	To turn away and then turn back?
It was the breath	What did we (hear:1.5)?	What did we hear?	What did we hear?	What did we hear?
we took when we first met.	It was the (breath:1.6)	It was the (breath:1.3)	It was the (breath:1.5)	It was the (breath:1.9)
Listen. It is here.	we took when we first met.	we took when we first met.	we took when we first met.	we took when we first (met:1.4).
	Listen. It is (here:1.5).	Listen. It is (here:1.2).	Listen. It is here.	(Listen:1.1). It is here.

Table 3: These are the original poems that are passed as an input to the diffusion model for the results demonstrated in Figure 5.

Figure 6: A comparison of generated images using different weights for various words in the same poem. All poems, along with their corresponding weighted prompts are provided in the Grey Box below.

Prompt 1:	*Refine the following poem into a weighted text prompt for text-to-image models.* *Only apply weights to the most important visual words. Follow these strict rules:* *Identify and emphasize only the most critical visual elements. Avoid modifying too many words.* *Use weight (1.5-1.8) for words that should be prominent in the generated image.* *Use weight (0.7-0.9) for words that should appear less prominently.* *Do not modify auxiliary, abstract, or transition words.* *Maintain the structure and wording of the original poem.* *Your response should only contain the weighted poem.* *Example Input:* *'Underneath my outside face* *There's a face that none can see.* *A little less smiley,* *A little less sure,* *But a whole lot more like me.'* *Example Output:* *'Underneath my (outside:1.7) (face:1.7)* *There's a (face:1.7) that none can see.* *A little less (smiley:0.9),* *A little less sure,* *But a whole lot more like me.'* *Now apply these rules to the following poem:*
Prompt 2:	*Prompt: Transform the following poem into a weighted text prompt for text-to-image generation.* *Apply weights only to the most critical visual elements while preserving the poetic essence.* *Follow these strict rules:* *Weighting Guidelines:* *Incremental Weights (1.5 - 1.8) → Words that define the poem's core visual or emotional identity* *Apply to words that strongly shape the imagery, mood, or metaphor.* *Example: If the poem speaks of a storm, shadow, or teardrop, these evoke vivid visual elements and deserve higher weight.* *Prioritize nouns (objects, scenery, emotions with physical manifestations).* *Decremental Weights (0.7 - 0.9) → Words that modify or soften key visuals, but should not dominate* *Apply to words that exist only to describe or refine an image, rather than being the main focus.* *Example: If a poem describes a smiley face but the mood suggests hidden sorrow, "smiley" should be weighted lower to reduce its dominance.* *Use for adjectives or modifiers that subtly influence meaning but do not need strong emphasis.* *DO NOT modify auxiliary words, transition words, or abstract concepts that lack direct visual impact (e.g., "that," "none," "sure," "because").* *Output Format:* *Maintain the original poem's structure.* *Return only the transformed poem, with weights applied selectively and meaningfully.* *Do not add explanations, notes, or comments.* *Example Input:* *'Underneath my outside face* *There's a face that none can see.* *A little less smiley,* *A little less sure,* *But a whole lot more like me.'* *Example Output:* *'Underneath my (outside:1.7) (face:1.7)* *There's a (face:1.7) that none can see.* *A little less (smiley:0.9),* *A little less sure,* *But a whole lot more like me.'* *Now apply these rules to the following poem:*
Prompt 3:	*Prompt: Refine the following poem into a weighted text prompt for text-to-image models.* *Only apply weights to the most important visual words.* *Your response should only contain the weighted poem.*
Prompt 4:	*Prompt: Refine the following poem into a weighted text prompt for text-to-image models.* *Only apply weights to the most important visual words. Follow these strict rules:* *Identify and emphasize only the most critical visual elements. Avoid modifying too many words.* *Use weight (1.5-1.8) for words that should be prominent in the generated image.* *Use weight (0.7-0.9) for words that should appear less prominently.* *Do not modify auxiliary, abstract, or transition words.* *Maintain the structure and wording of the original poem.* *Your response should only contain the weighted poem.*

Table 4: List of prompts used in our study.