

What Do You See In a Poem? Image Generation from English Romantic Poetry

Md Ahsanur Rashid
Stanford University

mdrashid@stanford.edu

Shaoxiong Zhang
Stanford University

shaoxioz@stanford.edu

Funing Yang
Stanford University

funingy@stanford.edu

Abstract

The relationship between text and image has long intrigued scholars, and the task of generating images from poetry presents unique challenges due to the sophisticated literary connotations involved. This project aims to get better results by looking at multiple proposed solutions, implementing them and working on fine tuning. The motivation behind this work is to explore the deeper connections between poetic expressions and visual representations.

The methodology involves two different approaches - using Generative Adversarial Networks (GANs) based GALIP and Diffusion based SDXL models. The dataset used was the MINIPO dataset, which comprises pairs of nursery rhymes and corresponding images. We also created a hand-crafted English Romantic Poetry Dataset (RPD) featuring poems from prominent Romantic poets and images generated using Google’s Gemini agent. The planned evaluation method include both quantitative and qualitative approaches. Quantitative evaluations were to be performed by computing image-poem cosine similarity with the help of CLIP model, while qualitative evaluations were to involve independent reviews and ratings by literature students.

Key findings indicate that while the baseline GALIP model provided decent results, the more advanced SDXL, driven by LLM-generated poem summaries image prompts, performed significantly better. The advanced approach, which included poem summary, theme, emotion, and visual elements for generating image creation prompts, resulted in more accurate and coherent image generations. The report concludes with a discussion on the potential for future work, including the creation of more standardized datasets and the exploration of connected multiple imagery for longer poems.

1. Introduction

“A picture is worth a thousand words”. The relationship between text and image has been an intriguing topic for scholars across disciplines as people explore the interrelated nature between these, such as illustrations, graphic nov-

els, visual storytelling, etc. Among these tasks, poetry-to-image presents deeper challenges than typical text-to-image tasks. Compared to typical narrative input, poetry condenses sophisticated literary connotations, including symbolism, emotional nuance, and metaphor, making image generation more difficult than simply achieving factual accuracy and coherence. Though several attempts have been made to this space, there are still many open-ended challenges in creating coherent, relevant images that represent meaningful ideas synthesized from a poem with difficulties such as lack of quality training data of image-poem pairs, and lack of generalization, lack of quality evaluation metrics, etc. For specific domains, past scholarly efforts have targeted niche subsets of poetry-image generation such as Chinese classical poetry, children’s poetry, lyrics, etc., yet to our knowledge there have not existed novel models and methods targeted toward 19-century romantic English poetry, where the artistic expressions reached a new height with distinct characteristics. To this end, we propose an **end-to-end pipeline** with a comparative approach for several models based on Generative Adversarial Network (GANs) and Diffusion Models that take a **poem as input** and generate a **inspired image as output**. Where given a dataset, $D = \{P_1, P_2, P_3, \dots, P_i\}$ here P_i represents a poem, generate a corresponding dataset $= \{I_1, I_2, \dots, I_i\}$ such that I_i is the corresponding image that encapsulates the interpretation of the poem or parts of it.

2. Related Work

There have been numerous efforts in poetry-to-image generation and image-to-poetry generation, largely following the encoder-decoder model architecture that combines vision and language models in multi-component pipelines, from the traditional CNN-RNN models pre-2020 to recent emergence of Diffusion models [1], we have observed a great improvement for the quality and diversity of the images in the past couple of years. For the dataset, though overall the field remains in need of a larger volume of diverse labeled data for the purpose, there have been specific quality curated image-poem paired datasets featuring specific domains.

2.1. Poetry to Image

Lu et al. [8] proposed a semi-supervised framework for Chinese poem-to-painting translation using cycle-consistent adversarial networks, which has shown great promise in generating high-quality images specific to the Chinese painting domain using the novel joint learning approach from both paired and unpaired data. More recently, Jamil et al. Progress and insights can be found in related fields such as paragraph captioning, where hierarchical approaches [3] to decompose image and paragraphs into constituent parts greatly improved the coherence of the narrative and image as well as the details the narrative is able to capture. [2] proposed the "poemToPixel" pipeline with iterative prompt tuning and summarization, keyword extraction, and instruction generation with diffusion models, which contributes to more accurate and coherent generations with a more comprehensive approach of iterative, human-in-the-loop pipeline that bridged the gap between poetry embeddings and images.

2.2. Image to Poetry

Among earlier works, Reed et al. [11] first proposed using Generative Adversarial Networks (GANs) to synthesize text based on images, which has shown promising results from several benchmarks but suffers from challenges such as lack of image resolution and generalizability. Loller-Andersen and Gambäck [7] proposed an end-to-end, multi-module system for a poetry generation system satisfying rhythmical and rhyming requirements for English poetry from an input image using the Convolutional Neural Network (CNN) and the Long Short-Term Memory (LSTM) Neural Network. The system was able to generate 153 stanzas in a largely poetic and grammatically correct manner. However, the pipeline has shown inconsistent performance across comprehensive metrics. Building on the encoder-decoder architecture proposed above, [14] introduced memory networks to address the problems of topic drift and semantic inconsistency. A more recent work by Liu et al. [6] proposed the visual-poetic model that builds on top of an RNN-based generator for joint learning with two discriminators in multi-adversarial training.

In addition to supervised learning dependent on quality labeled datasets, [12] addresses the challenge of lack thereof with an unsupervised Image to Poetry Model (UI2P) with a generative adversarial network (GAN) and a contrastive learning algorithm that improved upon the diversity of the dataset.

2.3. Dataset

Although the datasets still remain scarce for the poetry to image domain, in recent years some notable datasets provided ground for the field focusing on specific niche areas.

Liu et al. [6] introduced Image2Poem, the first human-annotated image-poetry pairs with over 8000 pairs of English free verse and image pairs. In more diverse domains, Jamil et al. proposed the MiniPo [2] dataset, a novel multimodal dataset comprising 1001 children's poems and images, being the first to bridge the gap in the children's poetry domain. For datasets in other languages, Lu et al. recently introduced the Chinese Painting Description Dataset (CPDD) [8] which includes 3,217 Chinese poems and corresponding paintings, and Li et al. [4] also contributed a unique dataset focusing more than hand-picked works by Chinese artist Feng Zikai alongside 89000+ poem-painting pairs for larger model training. Overall, we observe newly emerged datasets that greatly help promote the research fields, but the overall volume is low, and few datasets are for image generation for a diverse range of style.

3. Methods

To generate images from poems, we leveraged the idea of CLIP feature to map the multi-modal inputs like text and image into the same latent space. Specifically, our methods integrate two existing cutting-edge models GALIP [13] (Generative Adversarial CLIP) and SDXL [9] (Stable Diffusion XL) and fine-tune using the MINIPO [2] dataset. In this section, we will detail the learning algorithms, model architectural design of GALIP and SDXL.

3.1. GALIP

GALIP is novel GAN model for text-to-image generation. The overall GALIP architecture is armed with the CLIP-based discriminator and CLIP-empowered generator, can synthesize more realistic complex images.

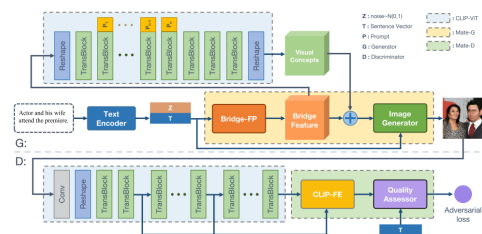


Figure 1. The architecture of the GALIP for text-to-image synthesis.

The model inputs contains poem and noise vector. The poem is first processed by a pre-trained CLIP text encoder to yield a global sentence vector, denoted as T . A noise vector Z is sampled from a Gaussian distribution which helps ensure the diversity of the synthesize high-fidelity images as model output.

3.1.1 CLIP-empowered Generator

The generator generates poem images utilizing the domain generalization ability of CLIP by inducing visual concepts from CLIP through bridge features and prompts.

The CLIP-empowered generator comprises a frozen CLIP-ViT for visual understanding, and a learnable mate-generator contains 3 key modules:

- Bridge Feature Predictor (Bridge-FP): takes the extracted poem summary sentence vector T and noise vector Z as input and translates them into implicit bridge features which are easier to generate and map to visual concepts through CLIP-ViT.
- Prompt Predictor: predicts text-conditioned prompts and append to CLIP-ViT.
- Image Generator: takes the visual concepts and bridge features as input through the neural network blocks and upsamples intermediate image features to produce image outputs.

The generator’s objective is to produce semantically consistent poem images that are indistinguishable from real ones by the discriminator. The generator loss is defined as:

$$L_G = -E_{G(Z,T) \sim P_g} [D(C(G(Z,T)), Z)] - \lambda E_{G(Z,T) \sim P_g} [S(G(Z,T), T)]$$

Where $S(G(Z,T), T)$ represents the cosine similarity between the encoded visual features of the generated image $G(Z,T)$ and the input text features T as computed by CLIP. λ is a coefficient for this text-image similarity term.

3.1.2 CLIP-based Discriminator

The discriminator assesses the quality of the generated images utilizing the visual understanding capability of CLIP. It consists of two main modules:

- CLIP-ViT image encoder: a pretrained module that converts the image into features.
- Mate-discriminator: a learnable module for adversarial training, collects image features from different layers of frozen CLIP-ViT and then further extracts informative visual features to form a quality assessor.

The discriminator’s objective is to detect unreal image parts more accurately to force the generator to produce more realistic images. The discriminator loss is defined as:

$$L_D = -E_{X \sim P_r} [\min(0, 1 + D(C(X), T))] - (1/2) E_{G(Z,T) \sim P_g} [\min(0, -1 - D(C(G(Z,T)), T))] - (1/2) E_{X \sim P_{mis}} [\min(0, -1 - D(C(X), T))] + k E_{X \sim P_r} [(\nabla_{C(X)} D(C(X), T) + \nabla_T D(C(X), T))^p]$$

Where D represents the Mate-D, C is the frozen CLIP-ViT within the discriminator, X is an image, T is the sentence vector, $G(Z,T)$ is a generated image, P_r , P_g , and P_{mis} denote distributions for real, synthetic, and mismatching data, respectively. k and p are hyperparameters for the gradient penalty.

3.2. SDXL

Stable Diffusion XL is a latent diffusion model for text-to-image tasks, featuring a large UNet backbone for rich representation capability and novel conditioning mechanism such as CLIP embedding for text encoding which highly useful for poem image generation.

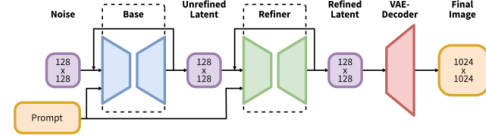


Figure 2. The architecture of the two-stage pipeline in SDXL.

3.2.1 Condition mechanism

SDXL introduces novel ways to condition the UNet backbone, provide more context and control including resolution transformations and cropping parameters besides text encoders, improved generation output quality. The conditioner embedding models will process the following information:

- Text Encoders: OpenCLIP ViT-bigG and OpenCLIP ViT-L takes the poem summary text input and output pooled text embedding after concatenation. This text encoder is trainable modules and contributes around one quarter of the overall parameter spaces.
- Original Image Resolution: additional timestep embedding using Fourier features. This prevents discarding smaller training images or introducing upscaling artifacts, and allowing users to control the resolution during inference.
- Image Cropping Parameters: additional embedding when sampled crop coordinates during training.

The embeddings above will be concatenated and condition the UNet backbone.

3.2.2 UNet Backbone

SDXL using a large UNet backbone with 2.6B parameters to include more attention blocks and cross-attention context. The UNet employs a heterogeneous distribution of transformer blockers, shifting computation to lower-level features.

During the latent diffusion training, SDXL learns to reverse a diffusion process that gradually adds noise to images by the learnable denoiser. When given a noisy data point, tries to predict the clean X_0 . The denoiser is trained via denoising score matching using the loss function:

$$E_{(x_0, c) \sim p_{data}(x_0, c), (\sigma, n) \sim p(\sigma, n)} [\lambda_\sigma |D_\theta(x_0 + n; \sigma, c) - x_0|_2^2]$$

Where $p(\sigma, n) = p(\sigma)\mathcal{N}(n; 0, \sigma^2)$, and $p(\sigma)$ is a discrete distribution over 1000 noise levels, and $\lambda_\sigma = \sigma^{-2}$. c represents the conditioning signal described above.

During sampling, SDXL employs classifier-free guidance to steer the iterative denoising process towards the conditional signal c .

3.2.3 Optional Refinement Stage

An optional separate refinement stage can be performed to further enhance the sample quality for detailed backgrounds and human faces. The refine stage model specialized in high-quality, high-resolution data and applies a noising-denoising process to the latent generated by the base SDXL model.

3.3. Method Limitation

Although both GALIP and SDXL are powerful poem to image solutions. We understand they all have limitations. GALIP may struggle with generating highly imaginary images, hard to train and extend to larger dataset. SDXL also faces challenges with very complex poetry summary prompts involving detailed spatial arrangements where different visual elements unintentionally merge. Our fine tune process with MINIPO dataset is aiming to mitigate these limitations and improve the poem image generation quality.

4. Dataset and Features

We have researched and selected existing pairs of poem-image data sets and constructed our own data set for the project. Ultimately, we decided to only pursue the existing dataset for the dataset but would still like to briefly discuss the approaches for our curated dataset.

4.1. Training Dataset: MINIPO

For the training dataset, we have chosen MINIPO [2], a new dataset that comprises 3011 pairs of nursery rhymes

Category	Statistics
Number of Poems	3011
Max Poem Length (in words)	346
Avg. Poem Length (in words)	50.47
Max Summary Length (in words)	491
Avg. Summary length (in words)	105.79

Table 1. Minipo Dataset Key Statistics

and corresponding images sourced from the internet. The main statistics of the dataset are outlined in Table 1.

While this dataset is not public, we got access by emailing the authors of the dataset. In the poem-image paired dataset, it was possible to work with 2565 pairs. These were the pairs that had proper images poems and we were able to use LLM to get the image generation prompt for. For the remaining ones, we could not get the proper image generation prompt from the LLM, Gemini in this case. The dataset also included generated poem summarization that went through a feedback loop that included experts in the field. There are a few malformed poems in the dataset.

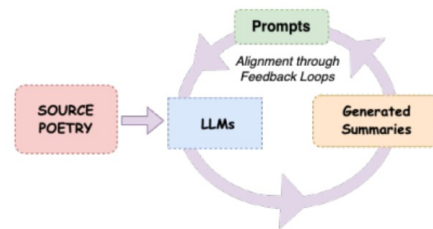


Figure 3. Poem summarization in MINIPO dataset.

We did a 80-10-10 split for Training-Validation-Test. For our more advanced approach, we followed the data extraction approach by Sofia Jamil et. al. [2]. In their paper, they extracted emotion, theme and visual elements from the poems, as well as the summary.

original_title	original_post	poem_text	poem_summary	extracted_emotion	extracted_theme	extracted_visual_elements	generated_image_prompt
Dear John, DeaMichael & Harper		'Dear John, Dea the poem explores	disgust	low		{rouins: 'challenges', 'protagonist', 'a shattered lens reflects a	
Parrot by Steven Steve Smith		'Parrot' 'Chits old parrot, sick	dislike	death		{rouins: {parrot, 'surroundings', 'a stenciled list parrot, feathers	
Dot of Snow in Robert Frost		The simplicity, The sight of a crow	joy	nature		{rouins: {poor, 'moor', 'snow', 'A joyful post, bathed in the	
Suburban Snow Green Harwood		'Suburban Snow A mother's precious	fear	childhood		{rouins: {mother, 'know', 'music', 'A heightened child hides an	

Figure 4. Minipo dataset sample.

	Unnamed: Title	Poet	text	ctext	Poem Link or _summary
0	0	Dear John, Michael S.	Dear John, Dear Coltrane" by Michael S. Harper is a powerful f a love		https://ws/The old parrot,
1	1	Parrot by SStevie Smi	ac "Parrot&" depicts the declining health of a wondrously be The old		https://rev/This poem explor
2	2	Parost of Sn Robert Ro	The simplicity, in the end, is the key element of the work as na The way a		https://ws/The sight of a c
3	3	Suburban / Owen Han	"Suburban Sonnet" by Owen Harwood is a poem about the diffi She		https://ger/ A mother practi
4	4	Unending / Rabindran	"Unending Love" by Rabindranath Tagore is a heartfelt poem all seem to		https://alls/ The speaker ex

Figure 5. Features extracted from the dataset - emotion, visual elements and theme.

4.2. English Romantic Poetry Dataset (RPD)

We have handcrafted the English Romantic Poetry Dataset (RPD), which consists of 150 poetry-image pairs

featuring poems from prominent English Romantic poets and an image generated with Google’s Gemini agent.

4.2.1 Collection Method

The poem corpus was compiled by selecting major poetry from several prominent English Romantic poets from the 17th to 20th century, including William Wordsworth, Samuel Taylor Coleridge, John Keats, Percy Bysshe Shelley, William Blake, Lord Byron, and Emily Dickinson. To maintain consistency and uniformity in input size, we include only the first 10 or so lines in the dataset if the poem is longer. All poems were in the public domain and had authoritative publications on the poetry foundation with open access.¹ The corresponding images for each poem were generated using the Gemini AI Assistant with the standard prompt “Generate an image for: {poem}”. The generated image was then evaluated for its congruence with the poem’s content in areas such as key themes, imagery, and sentiment consistency. We include the image as a paired data point if the generated image was assessed to be reasonable based on the above criteria, if not, a subsequent image was requested from Gemini, and the evaluation process was repeated until achieving a satisfactory match.

4.2.2 Validations

The validation of the image-poem pairings was carried out using a two-tier approach: independent reviews and ratings from students and graduates of English literature majors, and cosine similarity between the poetry and image embeddings for quantitative evaluations for the dataset validity. For independent validation, two literature students not directly involved in the image generation or primary evaluation process independently reviewed and rated a subset of randomly selected 20 image-poem pairs with the criteria of **relevance**, **thematic coherence**, and **overall impression** from 0-10. This independent review served as a crucial step in mitigating potential biases and enhancing the overall quality and relevance of the dataset by identifying and rectifying any significant mismatches. For cosine similarity evaluation, we have constructed visual-poetic embeddings in the same latent space using the CLIP model [10] and calculated the cosine similarity between the input poem and generated image. The cosine similarity results from the 150 images and poems have an average of 0.255 with the cosine similarity ranges between 0.201 and 0.320. In addition, the both human evaluations returns an average of 5 for all 3 criteria we have crafted out above. Hence, we eventually decided shortly after the project milestone that we should not use the dataset and pursued an existing, curated dataset

¹<https://www.poetryfoundation.org/>

instead as mentioned above. However, the evaluation methods that have already been setup for both qualitative and quantitative evaluations are adopted for the final model output. The dataset has potential to be leveraged in the future with some areas of enhancements: constraining the output image genre to be more consistent to focus on one or two genres for more coherence with the romantic poems (i.e., oil paintings), improving the prompting method where we can introduce multiple dimensions such as keywords related to the poem, theme, summary of the poem instead of simply querying the agent.

5. Experiments/Results/Discussion

5.1. Experiments

Our implementation of GALIP and SDXL are built on top of the existing official github repositories, <https://github.com/tobran/GALIP> and <https://github.com/Stability-AI/generative-models> respectively. We made multiple bug fixes and dataset preparation logic in order to get the inference path running and perform fine tuning with MINIPO dataset, we fully reproduce the paper [2] by implementing an algorithm to extract key elements in the poem as the prompts of pretrained diffusion model. Although the SDXL fine tuning didn’t go through due to the GPU resource limitation and time limits, we’ve done all the work needed for environment setup, dataset loading, training config tuning, and evaluation metrics implementation.

5.2. Poem Key Extraction

Poem text and summary were provided in the MINIPO dataset. Using these, we extracted emotion, theme and visual elements. We asked an LLM to generate image generation prompt based on these three key information.

```
Processing 'Crossed Land by Christine Keeler' by Christine Keeler...
Output sent to you:
Emotion: Sadness, Melancholy
Classified Theme: Nature
Extracted Visuals: ['rivers', 'shadows', 'slumber', 'tranquility']...
user query:
Generate an image prompt (under 50 words) for a poem with the following details:
Emotion: Sadness, Melancholy
Classified Theme: Nature
Key Visual Elements: rivers, shadows, slumber, tranquility
Generated Image Prompt: 'A tranquil river scene bathed in the soft glow of twilight. Long shadows stretch across the water, reflecting the slumbering trees...'

Processing 'Language Lesson 1076 by Heather Pitloch' by Heather Pitloch...
Output sent to you:
Emotion: Sadness, Melancholy
Classified Theme: Interpretation
Extracted Visuals: ['interpretation', 'boundaries', 'love', 'language', 'liberties']...
user query:
Generate an image prompt (under 50 words) for a poem with the following details:
Emotion: Sadness, Melancholy
Classified Theme: Interpretation
Key Visual Elements: interpretation, boundaries, love, americans, language, liberties
Generated Image Prompt: 'A surreal landscape depicting a diverse American family gathered around a table, engaged in lively conversation, with trees...'

```

Figure 6. Features extracted from the dataset - emotion, visual elements, theme and LLM generated image generation prompt.

5.3. GALIP Baseline Result

We implemented our methods on 1 Nvidia L4 GPU with 24 GB memory in GCP. We choose GALIP as our baseline and we’re using the pretrained model checkpoint with COCO dataset. Below are some of the baseline results using GALIP model:



Figure 7. This old parrot, sick and full of rage, longs for his jungle home but suffers in his new, urban surroundings. He is plagued by illness and waits for death to relieve him of his misery.



Figure 8. A young child, clutching a worn teddy bear, peers from behind a sheer curtain. Their wide eyes reflect the storm raging outside, blurring the words of a poem clutched in their tiny hand.

5.4. SDXL Fine Tuning

We successfully deployed the SDXL inference path using the pretrained model checkpoint [stable-diffusion-xl-base-1.0](#). And we use that to verify our poem key extraction result and we successfully reproduce the Poetry in pixels paper [2] and got decent results.

We then convert our dataset to [webdataset-format](#) in order to fine tune the SDXL model. We also created a torch lightning training yaml config that load the pretrained SDXL model checkpoint. The full training yaml config is included in the supplemental material. We listed the key configs:



Figure 9. This old parrot, sick and full of rage, longs for his jungle home but suffers in his new, urban surroundings. He is plagued by illness and waits for death to relieve him of his misery.

- Learning rate: 1e-4
- Epoch: 1000
- Scheduler: LambdaLinearScheduler with warm up steps 1000, f_start as 1e-6.
- Denoiser: DiscretDenoiser with eps scaling and legacy DDPM discretization.



Figure 10. A young child, clutching a worn teddy bear, peers from behind a sheer curtain. Their wide eyes reflect the storm raging outside, blurring the words of a poem clutched in their tiny hand.

- Loss: standard diffusion loss.
- Sampler: Euler EDM sampler.

However, we failed to train the model due to GPU resource constraints. The model itself needs around 11GB memory and the GPU OOM after few training epochs, we didn't get the chance to successfully fine tune with MINIPO dataset before the project due and thus we don't have the chance to do hyperparameter tuning as well.

5.5. Evaluations

There are several evaluations approaches being proposed for similar poetry-image generation tasks. For qualitative metrics, one can evaluate the relevance in content, emotions, and theme [6], and semantic consistency [8] between the image and poem pairs. For quantitative metrics, [6] proposed the visual-poetic embedding metric to evaluate the semantic similarity between the image and poetry pairs within the same latent space, and several existing, more generic evaluation metrics stand out from past literature: Image-Text Contrastive Learning Loss (ITC) [5], Image-Text Matching (ITM) Loss [5], Fréchet Inception Distance (FID) [8], and Distribution Consistency Error (DCE) metric [8]. For our evaluation process, similar to the own custom dataset evaluation, we are employing both quantitative and qualitative evaluations with the approaches and results below.

5.5.1 Quantitative Evaluations

Inspired by the CLIP [10] model's image-text multi-modal understanding by mapping them to the same latent space and maximize the agreement, as well as [6]'s approach on

semantic similarity along the similar lines, we are using the image-poem embedding cosine similarity using CLIP with a patch size of 32*32 pixels.

5.5.2 Qualitative Evaluations

For qualitative evaluations, we recruited two English literature major graduates not directly involved with the project to help independently review and rate our model output for a selection of poems. Each rater was asked to rate on a scale of 0-10 on:

- Relevance
- Thematic Coherence
- Overall Impression

We planned to do this for a selection of dataset poem-image pairs as well as a selection of our model’s output for the same set of poems. The intention was to not only check how the model did but also to compare the feedback against the provided dataset for a more even comparison.

However, we failed to train the model due to GPU resource constraint. As such, we skipped the qualitative evaluation.

6. Conclusion/Future Work

We worked on creating our own dataset. Since we didn’t get good evaluation, we switched gears to using the MINIPO dataset. Our baseline model was based on GAN and our improved model was based on Diffusion.

While our baseline GAN model gave somewhat decent results, our more advanced Diffusion model, driven by LLM generated image prompts fared much better. Since in the second approach, we took poem summary, theme, emotion and visual elements to generate image creation prompts, that basically made the image generation easier.

For future work, we would like to create a more standard dataset, given the dearth of good, public datasets in this area. Another area of focus would be generating connected multiple imagery, which can be thought of as storytelling for longer poems.

7. Appendices

7.1. More Results

Poem: Conrad Siever’ by Edgar Lee Masters presents the liveliness of the northern-spy apple tree and the impact of autumn on the surroundings. ‘Conrad Siever’ by Edgar Lee Masters is a poem based on the theme of autumn. Death and decay features from the beginning of the poem. The poet feels for the withering trees and rotting plants the farm

has turned into a graveyard. But, the poet isn’t sitting there. He sits under the northern-spy apple tree which still has its time. Winter is not far ahead. Then everything will be changed, even the poet’s favorite apple tree. However, at last, the poet feels satisfied with the red apples that he can still enjoy at that time of the year and imagines those apples to be the living epitaphs of the northern-spy.

Extracted summary with emotion, visual elements and themes: A solitary, gnarled apple tree in an overgrown garden, its branches reaching towards a fading sunset. Two souls, formed from soft light, embrace within the tree’s hollow, their love a last echo of life.



Figure 11. Poem to image generation.

7.2. Fine Tuning Experiment

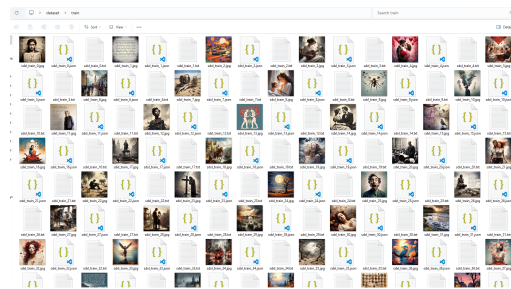


Figure 12. MINIPO in webdataset format for SDXL fine tuning.

8. Contributions & Acknowledgements

Ahsan contributed to the development of the custom dataset. He also worked on the SDXL model and wrote part of the final report. Shaoxiong was involved in creating the custom dataset as well. He contributed to the

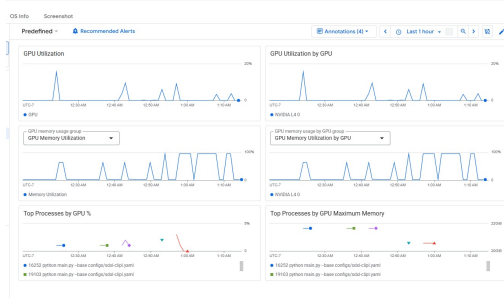


Figure 13. GPU OOM after few training epochs.

GALIP & SDXL models and also helped with the final report. Tuning focused on assessing the custom dataset (quantitative/qualitative), researching for a usable dataset, and writing a part of the final report.

9. References

References

- [1] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models, 2020. [1](#)
- [2] S. Jamil, B. A. Reddy, R. Kumar, S. Saha, J. K. J, and K. Goswami. Poetry in pixels: Prompt tuning for poem image generation via diffusion models. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9224–9237, Abu Dhabi, UAE, Jan. 2025. Association for Computational Linguistics. [2](#), [4](#), [5](#), [6](#)
- [3] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei. A hierarchical approach for generating descriptive image paragraphs, 2017. [2](#)
- [4] D. Li, S. Wang, J. Zou, C. Tian, E. Nieuwburg, F. Sun, and E. Kanoulas. Paint4poem: A dataset for artistic visualization of classical chinese poems, 2021. [2](#)
- [5] J. Li, R. R. Selvaraju, A. D. Gotmare, S. Joty, C. Xiong, and S. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021. [6](#)
- [6] B. Liu, J. Fu, M. P. Kato, and M. Yoshikawa. Beyond narrative description: Generating poetry from images by multi-adversarial training. In *Proceedings of the 26th ACM international conference on Multimedia*, MM '18, page 783–791. ACM, Oct. 2018. [2](#), [6](#)
- [7] M. Loller-Andersen and B. Gambäck. Deep learning-based poetry generation given visual input. In *International Conference on Innovative Computing and Cloud Computing*, 2018. [2](#)
- [8] Z. Lu, T. Guo, and F. Wang. Semi-supervised chinese poem-to-painting generation via cycle-consistent adversarial networks, 2024. [2](#), [6](#)
- [9] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. *CVPR*, 2023. [2](#)
- [10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision, 2021. [5](#), [6](#)
- [11] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis, 2016. [2](#)
- [12] X. Ren, X. Chai, and M. Mao. Generating chinese poetry from images based on deep learning. In *Proceedings of the 2023 3rd International Conference on Big Data, Artificial Intelligence and Risk Management, ICBAR '23*, page 134–138, New York, NY, USA, 2024. Association for Computing Machinery. [2](#)
- [13] M. Tao, B.-K. Bao, H.-Z. Huang, Z.-N. Liu, J. Lin, and C. Xu. Galip: Generative adversarial clips for text-to-image synthesis, 2023. [2](#)
- [14] L. Xu, L. Jiang, C. Qin, Z. Wang, and D. Du. How images inspire poems: Generating classical chinese poetry from images with memory networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32, 03 2018. [2](#)