2023 IEEE Sixth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE) | 979-8-3503-3128-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/AIKE59827.2023.00017

Prompt Recommendations for AI Art

Hyelim Yang, Kapil Wanaskar, Harshika Shrivastava, Shahbaz Mansahia, Shakshi Richhariya, Magdalini Eirinaki
Department of Computer Engineering
San Jose State University
San Jose, USA

***Abstract*—One of the main areas where generative AI models thrive is image synthesis or generation. This work highlights the importance of quality prompts in generating compelling artworks and delves into four principal methodologies for generating prompt recommendations: text embeddings, ensemble models, text with image embeddings and object detection for feature extraction. Multiple traditional and neural network-based models are explored for feature vector representation. Furthermore, the study explores the incorporation of image embeddings, the user's preferred art styles for tailored recommendations, and the inherent challenges in evaluating these systems. We also propose a novel methodology for evaluating such systems, in the absence of ratings or preference scores, using graph analysis and community detection algorithms. This work distinctly contributes to the prompt recommendation domain and complements previous works in the AI art generation landscape.**

***Index Terms*—AI Art Generation, Prompt Recommendation System, Text Embeddings, Text-Image Embeddings, Prompt Engineering, Community detection**

I. Introduction

In the dynamic world of digital art, users can create amazing artworks by just entering the "prompts" or text strings on the artificial intelligence (AI) platforms such as Dall-E [1] and Midjourney [2]. For instance, a prompt like "white cat sitting on a window in disney theme" can give beautiful and imaginary results to the user. Appropriate prompts yield more desired results and improve the customer experience. Our study aims to make the process of prompt selection easier by offering prompt recommendations to the user. However, a big challenge in such application domains is the absence of explicit ratings (or even implicit preference scores). Unlike typical recommendation systems the data available are limited to the user id, their prompt, and the respective image. Therefore, evaluating prompt recommendation algorithms cannot be done similarly to the case of traditional recommender systems (e.g. using scores such as RMSE, Precision@n etc). In this work, we propose a novel methodology, leveraging methods like graph analysis and community detection algorithms to assess the performance of the recommendation algorithms. While there's a lot of research on AI art creation, our work specially focuses on prompt recommendation, addressing a clear gap. In what follows, we review the related work, design details, algorithm insights, and conclude with the experimental evaluation and our conclusions.

II. Related Work

Bau et al. [3] delved deep into the use of generative adversarial networks (GANs) for semantic photo manipulation, which allowed them to recreate and modify real-life photos. Oppenlaender et al. [4] treated prompt engineering as a unique artistic technique for AI art production, investigating how participants could critically assess, craft, and refine prompts. Abdallah et al. [5] ventured into biomaterials design and employed text-to-image AI models, emphasizing the creation of detailed visual interpretations. Shen et al. [6] highlighted a concept at the crossroads of AI and security, introducing "prompt stealing attacks" on text-to-image models. Finally, Wang et al. [7] focused on refining text prompts to enhance the emotional undertones of AI-generated visuals.

Our work, on the other hand, concentrates on prompt recommendation for general AI art generation. Even though we are aware of GANs' potential for image modification, it is not our main goal. Instead, we have created a system that helps people improve the AI-based image generating process by providing pertinent prompt ideas. We use AI for art creation, drawing inspiration from [5], but our aim is to give consumers specific quick suggestions based on either the images they provide or textual data. Wang et al.'s attempt to change prompts for emotional depth [7] diverges from our approach, which is focused on the mechanics of providing the appropriate prompts to users. Finally, while Shen et al. [6] deal with security concerns , our research prioritizes user experience, aiming to suggest the most suitable prompts.

III. System Design and Implementation details

In this section, we discussed the architecture of our project. Designed for flexibility, our system is built to compare various methods side by side. We focused on the following algorithms to arrive to our conclusions:

- Algorithm-1. *Text embeddings and recommendations* provide an efficient way to capture and quantify the semantic relationships in textual prompts.
- Algorithm-2. *Ensemble models for text embeddings and image feature recommendations* offer a multi-modal approach that concurrently considers both textual and visual elements.
- Algorithm-3. *Text and image embedding recommendations followed by image to image recommendations* provide more relevant options to the user. This is also supplemented by *user cluster formation-based recommendations*, providing a comprehensive recommendation system that encapsulates the wide spectrum of user queries.
- Algorithm-4. *Text embeddings with image feature extraction* (using YOLOv4) that is blind to the prompts. This

979-8-3503-3128-8/23/$31.00 ©2023 IEEE
DOI 10.1109/AIKE59827.2023.00017

algorithm would take an image into consideration and try to detect the objects from the default common objects in context (COCO) classes to enrich our dataset and provide additional features for recommendation.

The decision to implement these four methods is driven by our aim to build a versatile, adaptive, and precise recommendation system. These approaches, in conjunction, cater to diverse user inputs and deliver a more satisfying user experience, ensuring the generation of desired and accurate artistic outcomes.

Each algorithm produces a similarity matrix between prompts, which allows for the graph representation of the output of each algorithm. This representation allows for a comparative analysis of the algorithms, achieved by examining graph characteristics and deploying the Louvain algorithm for community detection, an evaluation approach that can be applied in different domains in the absence of ratings. The reason for selecting each algorithm is further discussed in each algorithm's section.

IV. Materials and Methods

A. Dataset

The dataset used for the project is a subset of the DiffusionDB dataset[1] generated in August 2022. We started our experimentation with 5,000 prompts and utilized three features for our recommendation system within our dataset, namely "image" which has the AI art image, "prompt" which has the text string used to generate that image, and "user name" which has the string of the user name who has entered the prompt.

B. Data preprocessing

For preprocessing of prompt data, we removed common stopwords using NLTK library's stopwords module[2] and emojis using Emoji dictionary[3]. We also reduced words to their root form using a stemming algorithm using Porter stemming algorithm[4] from the NLTK library. For instance, the prompt: "the popes hard rock band, with instruments" is preprocessed into prompt: "pope hard rock band instrument".

C. Algorithm-1

This algorithm aims to prompt recommendations given a prompt as input. We utilized Term Frequency-Inverse Document Frequency (TF-IDF) and Count Vectorization to extract the features and create feature vectors from the text prompt. TF-IDF evaluates the importance of a word in a document relative to a collection of documents or a corpus. Count Vectorization represents text data based on the frequency of words, disregarding the order of the words (bag-of-words representation). Once the feature vectors are created, we can use any vector similarity metric to measure the similarity between pairs of prompts.

In our system, we use cosine similarity: $CosineSim(A, B) = \frac{A \cdot B}{||A|| \times ||B||}$ where A and B are prompt feature vectors of the same length. The result is a square and symmetrical matrix representing the similarity scores between prompts. To make a recommendation, we sort these similarity scores and identify the top-n neighboring prompts.

D. Algorithm-2

Algorithm-2 goes deeper into the world of embeddings to understand how text prompts connect with AI-generated images. We look at different ways to represent words and ideas: Word2Vec [8] is like a smart code that groups words that have similar meanings, Global Vectors for Word Representation (GloVe) [9] looks at which words often come together, Bidirectional Encoder Representations from Transformers (BERT) [10], gets the bigger picture of a sentence, and Contrastive Language–Image Pre-Training (CLIP) [11] can understand both pictures and words.

Digging into this, the way we represent words can make a big difference in suggesting art prompts. In our experimental evaluation, we observed that Word2Vec and GloVe are good at finding prompts that sound similar, which is helpful for simple and clear art ideas. BERT is great when the prompt is based on stories or famous sayings because it gets the whole idea. Meanwhile, CLIP is special because it can match prompts with art that is either very detailed or abstract. By using these different methods, Algorithm-2 helps us understand how different tools can give different suggestions, and it highlights the importance of picking the right tool for the kind of art suggestion we want to give.

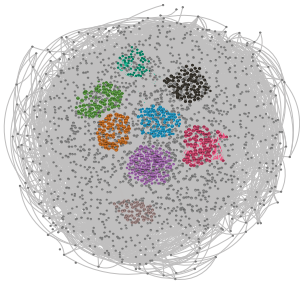
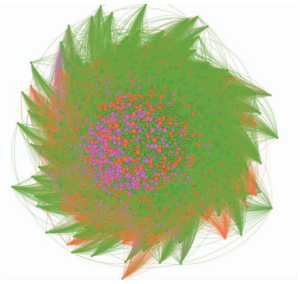
E. Algorithm-3

Algorithm-3 explores a hybrid approach, leveraging both text and image embeddings to enhance the prompt recommendation process. This model is designed to serve three objectives:

1) Implementing a prompt-to-prompt recommendation system incorporating image data alongside text. We generated text embeddings using TF-IDF vectorizer and image embeddings using the InceptionV3 model. These embeddings were merged to form a hybrid representation of the prompts. Our model constructs a similarity matrix based on cosine similarity (as defined previously), enabling it to recommend more relevant and precise prompts.
2) Implementing a prompt-to-prompt recommendation system by recommending prompts based on image input. The model matches the image input (expressed using embeddings) with the similar image embeddings, and suggests the prompts that are semantically similar to the input image.
3) Implementing user-based recommendations by finding user-user similarity based on their artwork interest. We found clusters of users who are interested in the same type of art forms by utilizing user and prompt features from the dataset. We build our recommendation system to recommend to user clusters.

[1] https://poloclub.github.io/diffusiondb/
[2] https://www.nltk.org/
[3] https://pypi.org/project/emoji/
[4] https://www.nltk.org/_modules/nltk/stem/porter.html

TABLE I
ALGORITHM COMPARISON

	Algorithm-1	Algorithm-2	Algorithm-3
Number of nodes	3587	4940	282
Number of edges	29020	7255901	391
Average weighted degree	13.53	2204.88	2.24
Graph density	0.005	0.595	0.01
Modularity	0.943	0.07	0.941
Number of communites	607	4	88
Graph after the community detection			

This approach not only caters to the semantic alignment of the prompts but also incorporates visual cues from images and the unique preferences of individual users, thereby making the recommendation system more comprehensive and accurate.

F. Algorithm-4

With Algorithm 4 we attempted to use the object detection model YOLO v4 [12] to explore an alternative way of leveraging the image properties in the prompt recommendation process. We applied feature extraction from images using YOLOv4 and the default COCO classes as references. We then combined them with the text to get hybrid embeddings. Latent diffusion models using text-to-image pipelines usually generate prompt-blind items and features in a stylized format for image-aesthetic purposes. The idea was to use the detected objects in specific styles based on their similarity to the user's generations and prompts for further prompt refinement and recommendation. Unfortunately, due to the lack of computational resources, lack of enough data to run and evaluate this algorithm, and the limitations of the default COCO classes' diversity, we could not perform a thorough experimental evaluation by the time this paper was written. However, we could think of this as an approach to enrich the dataset using the space projected within the images. We leave this as a task for future work.

V. EXPERIMENTAL EVALUATION

AI generated image prompt datasets such as the one we used in our experiments, are very different from datasets that are typically used in recommder systems in many ways. The most important different is the absence of explicit or ratings or implicit preference scores, since each user uses different prompts each time. Therefore, evaluating the proposed algorithms using scores such as RMSE, MAE, Precision@N, NDCG, etc. was impossible. We therefore propose an alternative methodology, using graph theory and community detection, to compare the three different recommendation algorithms based on their ability to identify and recommend semantically similar prompts.

A. Graph Theory and Community Detection

All previously presented approaches generate a vector representation of the prompt (text-only or text-and-image), allowing us to calculate the pair-wise similarities between prompts. We represent each algorithm by a graph where nodes correspond to prompts, and edges connect nodes if the similarity between prompts exceeds a threshold of 0.7. Gephi[5] was employed to visualize graphs, apply the Louvain algorithm [13] for community detection, and compute various parameters of graphs. We then recorded the parameters, including the number of nodes and edges, average weighted degree, graph density, and average clustering coefficient. After the Louvain algorithm for community detection within each graph, we compute modularity as a measure to evaluate how well a network is partitioned into communities or clusters. Essentially, a network with high modularity demonstrates dense connections within clusters and sparse connections between different clusters.

We summarize our findings in Table I. Our observations reveal that Algorithm-2 yields the highest number of nodes and edges. It finds 4940 nodes out of 5000 entire data, showing 99% of prompts exhibit similarities within the data. This implies its capability to identify a larger number of similar prompts. Among all three algorithms, Algorithm-1 showcases the highest modularity, suggesting a stronger community structure. In contrast, Algorithm 2 exhibits the lowest

[5]https://gephi.org

modularity, indicating a less distinct community structure. This might be due to Algorithm-2's proficiency in grasping subtle nuances within prompts and establishing connections with other prompts based on these intricate understandings. For the graphs after the community detection, the community that includes more than 5% of all nodes was colored. The rest of the communities were colored gray. Through this unconventional evaluation approach, we can glean insights about each algorithm's performance and make an informed choice about which to implement depending on the desired characteristics of the recommendation system.

B. Comparison of embedding-based approaches

Out of all the algorithms evaluated, Algorithm-2 performs well according to our observations. In this section, we provide a comparison based on graphs representing three different text embedding algorithms: Word2Vec, Glove, and BERT.

TABLE II
COMPARISON OF EMBEDDING-BASED APPROACHES

	Word2Vec	GloVe	BERT
Number of nodes	4628	4758	4939
Number of edges	518883	3011295	6831524
Average weighted degree	0.66	0.785	0.851
Graph density	0.048	0.266	0.56
Modularity	0.374	0.133	0.053
Number of communites	118	58	11

From Table II, we observe a pattern concerning the relationship between the complexity of the embedding and the resultant graph's properties. As the complexity of the embedding increases, there is an upward trend in the number of nodes and edges, the average weighted degree, and graph density. This is particularly true for the BERT embedding, which manages to find at least one similar prompt for 99 percent of the prompts provided the similarity threshold is set at 0.7. However, a contrasting trend is observed in the results of community detection. With increased embedding complexity, both modularity and the number of detected communities decrease. The BERT embedding, due to its inherent intricacy, can discern nuanced meanings in each prompt and consequently establish connections that other embeddings might overlook. This suggests that the BERT embedding is able to bridge the semantic gap between different communities, leading to a reduction in the number of communities and lower modularity.

VI. CONCLUSION

Our project aims to recommend prompts for generating AI art. We proposed three algorithms and compared them qualitatively with graph analysis and community detection (we also proposed a fourth algorithm but left the evaluation as part of our future work). We found that text embedding approaches produced good results, allowing for the generation of similar prompts efficiently. However, the performance is subjective to the desired outcome. When we consider a single outcome to be our goal, Algorithm-2 is the best approach in terms of serendipity and novelty. Due to the absence of conventional evaluation metrics, we propose an innovative approach utilizing graph structures and community detection for qualitative analysis of three recommendation algorithms. Algorithm-2 demonstrates a nuanced connection-building ability. A comparative examination of Word2Vec, Glove, and BERT text embeddings demonstrates how embedding complexity relates to graph properties, with the BERT embedding excelling in discerning prompt nuances. Algorithm-3 showcases a more cohesive community structure in our evaluation. Algorithm-3 is desirable in a use case where the desired outcome is higher relevance in the recommender system. This can be attributed to the high modularity metric and small number of nodes of the graph associated with this algorithm.

We leave the choice to the user as a prompt recommender's use would be conventionally targeted at creative users and therefore, the desired goal would vary with the workflow of the user. We also suggest further experimentation with a hybrid approach where Algorithm-3 is utilized in downstream tasks while Algorithm-2 is used for exploration. Surprisingly, increased complexity leads to reduced modularity and community count due to the BERT embedding's inter-community bridging capability in Algorithm-2, thus making an option worth exploring. Our research offers insights into recommendation algorithms, text embeddings, and their impacts within a novel evaluation framework.

REFERENCES

[1] Dall-E, https://openai.com/product/dall-e-2
[2] Midjourney, https://www.midjourney.com
[3] D. Bau, H. Strobelt, W. Peebles, J. Wulff, B. Zhou, J. Zhu, and A. Torralba, "Semantic photo manipulation with a generative image prior," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1-11, Aug. 2019, doi: https://doi.org/10.1145/3306346.3323023.
[4] J. Oppenlaender, R. Linder, and J. Silvennoinen, "Prompting AI Art: An Investigation into the Creative Skill of Prompt Engineering," 2023. [Online]. Available:arXiv:2303.13534.
[5] Y. K. Abdallah, and A. T. Estévez, "Biomaterials Research-Driven Design Visualized by AI Text-Prompt-Generated Images," *Designs*, vol. 7, no. 2, pp. 1-27, March 2023, doi:https://doi.org/10.3390/designs7020048.
[6] X. Shen, Y. Qu, M. Backes, and Y. Zhang, "Prompt Stealing Attacks Against Text-to-Image Generation Models," 2023. [Online]. Available:arXiv:2302.09923.
[7] Y. Wang, S. Shen, and B. Y. Lim, "RePrompt: Automatic Prompt Editing to Refine AI-Generative Art Towards Precise Expressions," *In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–29, Apr. 2023, doi:https://doi.org/10.1145/3544548.3581402.
[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," 2013. [Online]. Available:arXiv:1301.3781v3.
[9] J. Pennington, R. Socher, and C. D. Manning (2014) "GloVe: Global Vectors for Word Representation," 2014. [Online]. Available:https://nlp.stanford.edu/projects/glove/
[10] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019. [Online]. Available:arXiv:1810.04805.
[11] A. Radford, *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, pp. 8748-8763, July 2021.
[12] A. Bochkovskiy, C. Wang, and H. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," 2020. [Online]. Available:arXiv:2004.10934v1.
[13] V.D. Blondel *et al.*, "Fast unfolding of communities in large networks," *J. Stat. Mech.*, vol. 10008, pp. 1-12, 2008.