Ideal Prompt Generation for Text-Conditional Image Synthesis from Large Text Corpora

Sumuk Shashidhar, Deparment of Computer Science, University of Illinois at Urbana-Champaign

***Abstract*—This research paper proposes a novel method for generating effective text prompts for image generation from large bodies of natural language text from descriptive literature, eliminating the need for manual prompt engineering. The proposed method utilizes transformer-based architectures and a separate sub-model to extract meaningful text and transform it into a coherent prompt. The effectiveness of the method is evaluated through a comparative analysis with human-generated prompts using widely recognized metrics such as SPICE and CIDEr. The experimental results show that the proposed method can significantly reduce the time and effort required to generate a large number of images while maintaining semantic similarity with human-generated prompts**

I. Introduction

In recent times, the emergence of text-conditonal diffusion-based image synthesizers, such as DALLE-2, Midjourney, and Stable Diffusion, has sparked a notable surge in interest in AI-generated art and images. These synthesizers leverage a diffusion process that is conditioned on text prompts provided by the user, which must be meticulously crafted to invoke the intended visual output. Due to the vast number of hyperparameters that can be tuned to completely alter the image, AI artists engage in extensive experimentation to generate the desired result. The creators of these diffusion models acknowledge these limitations and provide manuals and guides to educate users on prompt creation and engineering, as evidenced by the creators of Midjourney.[1]

However, this approach is not scalable for generating a large number of images, such as in scenarios where numerous images are required to be generated (book illustration, movie scene generation, game prototyping etc.). Each image requires the prompt to be engineered from scratch and tuned, which can be a time-consuming and prohibitive process.

In this paper, we propose a novel method for generating effective text prompts from large bodies of natural language text. By eliminating the need for manual prompt engineering, our method can significantly reduce the time and effort required to generate a large number of images. It can also be used to generate varied prompts for a single body of text, based on temperature variations, enabling artistic creativity.

II. Task Description

The extraction of meaningful text from a large corpus of natural language entails the partitioning of the input into smaller, semantically coherent chunks. These chunks are subsequently transformed into prompts that are suitable for image generation. This process can be achieved in a modular fashion by utilizing transformer-based architectures that are optimized to capture the salient sensory information required for image generation. Additionally, a separate sub-model can be utilized to convert the extracted information into a coherent prompt by learning to emphasize the most significant aspects of the input.

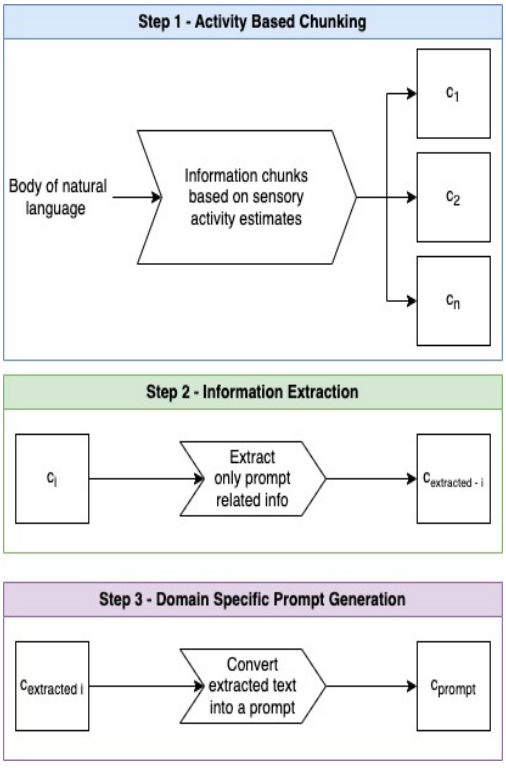

Figure 1. Model Flow

The assessment of the effectiveness of our proposed approach can be achieved through a comparative analysis between the generated prompts by our machine learning model and those that are manually crafted by human

experts to depict the same set of visual elements. This evaluation methodology is particularly useful since our model can capture not only the overall human intention but also emphasize the critical aspects that artists prioritize when representing a given piece of information. To conduct this comparison, we employ widely recognized metrics, such as SPICE [2] and CIDEr [3] that are commonly utilized in the natural language processing domain to measure the semantic similarity between two texts.

III. Background and Related Work

This section presents an overview of prior research in diverse natural language processing domains that can be utilized to address the challenge of generating efficient prompts for image generation. By examining existing literature, we can gain insights into the various techniques and methodologies that have been explored in the past to tackle similar problems, which can aid in designing effective solutions.

A. Activity based chunking

In this section, we present a strategy to quantify the relevant text in our input that pertains to the visual output, which involves leveraging concepts from information theory. However, determining the amount of information contained in natural language is a challenging task, as it is not a discrete quantity and can have subjective interpretations. Prior research in the field of summarization has proposed soft computing approaches to address the challenge of quantifying information in text, and transformers, which were introduced in the seminal paper by Vaswani et al. [4] in 2017, has greatly advanced the state-of-the-art in this domain.

Numerous groups, including Facebook, Microsoft, and Google Research, have developed various extractive and abstractive summarization models, such as BART [5], T5 [6], and BERTSum [7], using large datasets like CNN/Daily Mail [8], Gigaword [9], and XSum [10]. Among these models, Pegasus [11] stands out, as it achieves high scores on popular benchmarks for summarization, such as ROUGE [12], as it uses a pre-training objective that involves predicting masked sentences in multi-sentence tasks, which closely aligns with the task of summarization.

Leveraging these summarization models can aid us in determining the optimal chunk size by evaluating how the summary produced changes when varying amounts of the source prompt are provided. This strategy can be useful in quantifying the relevant text for the visual output and improving the overall effectiveness of activity based chunking.

B. Information Extraction

The task of extracting visual information is challenging due to the novelty of image generators based on latent diffusion [13], and requires an innovative approach. We propose to treat this as a variation of two well - researched tasks in natural language processing: machine, conditional summarization, and text-to-text translation. To achieve this, we propose utilizing transfer learning.

1) Transfer learning

To effectively approach the task of visual information extraction, it is crucial to equip our initial model with a strong understanding of human language. However, due to the limited dataset available for the narrow domain, it is not feasible to solely train the model on this data. Instead, a preferable approach is to train the model on a large and diverse dataset across domains, which can later be fine-tuned for our specific dataset. This approach is known as transfer learning and has been extensively researched in recent years.

One notable example of a successful transfer learning approach is the T5 transformer developed by Raffel et al. [6]. The T5 model was trained on a data-rich task, enabling it to acquire a comprehensive understanding of language that can be effectively applied to downstream tasks. This approach offers several advantages, such as reducing the computational requirements and enabling us to work with a smaller dataset for the same task. Furthermore, T5 is highly versatile and can apply its acquired knowledge in innovative ways.

We can fine-tune T5 to transform a given chunk of natural language text to an image prompt by training it on a dataset of image prompts and their corresponding text. This approach can be useful in generating a prompt that is semantically similar to the given text, which can be used to generate conceptually accurate images.

C. Assessment Parameters

In order to ascertain the efficacy of our information extraction methodology, it is essential to assess it against human-generated prompts, specifically designed to produce the same visual output. Numerous metrics have been proposed over the years to evaluate the caliber of machine-generated text. By considering the given problem as one pertaining to machine summarization, we can draw upon an extensive array of literature on the evaluation of machine-generated language quality. The ROUGE score, an early effort to gauge machine-generated text quality, demonstrated human agreeability scores of up to 80% across various single-document summarization tasks [14]. Subsequently, METEOR [15] was developed, boasting a significantly higher human agreeability score of up to 96.4%. However, these met-

rics are extremely sensitive to n-gram overlap, which is damaging to our current task at hand. As evidenced in [16], n-gram overlap is neither sufficient nor necessary for two sentences to convey identical meanings.

Fortunately, in recent times, substantial research has been conducted within the domain of automatic image captioning to evaluate the quality of machine-generated captions across various datasets. The CIDEr score [3] emerged as one of the pioneering metrics in the image captioning realm, tailored to the MS-COCO dataset to achieve elevated levels of human agreeability. The SPICE score [2] surpassed previous metrics by incorporating the notion that *semantic propositional content plays a crucial role in human caption evaluation*, and by subsequently constructing a scene graph for assessment purposes. Although this metric is the most comprehensive to date, certain limitations with specific datasets have been identified by [17].

IV. Method

We approach this problem in three stages:

A. Chunking

In order to facilitate the generation of meaningful images from the input text prompt T_{source} , it is necessary to first segment it into logical chunks or segments based on some threshold value t. The threshold t is determined by the level of detail required in the generated images, with higher values of t corresponding to longer and more complex prompts, which may be less stable.

We define a chunker function C, that takes as input the prompt T_{source} and the number of desired chunks n and produces a set of labeled chunks $\{c_1, c_2, \ldots, c_n\}$ where i ranges $[1 \ldots n]$. The value of n is determined by the length of the input text, measured in terms of the number of tokens l, and the chosen threshold value t. Specifically, we can compute n as follows:

$$n = \lceil \frac{l}{t} \rceil \tag{1}$$

The chunker, C, is defined as follows:

$$C(T_{\text{source}}, n) = \{c_1, c_2, \ldots, c_n\} \tag{2}$$

B. Information Extraction and Prompt Generation

We aim to generate an array of potential prompts for each chunk, denoted as c_i. To achieve this, we introduce the prompt generator function, G, which takes the chunk as input and outputs an array of n prompts. We denote each prompt as p_j, where j varies from 1 to n.

$$G(c_i, n) = \{p_1, p_2, \ldots p_n\} \tag{3}$$

1) Evaluation Criteria

In order to assess the effectiveness of the prompt generator G, we employ an evaluation metric denoted by E, which involves a comparison of the generated prompts against various human-generated counterparts. The metric is defined as a function that accepts both the generated prompts and the human-generated prompts as inputs and outputs the upper third quartile of the variability score distribution v.

$$E(\{p_1, p_2, \ldots p_n\}, \{h_1, h_2, \ldots h_k\}) = \mathcal{Q}_3(v) \tag{4}$$

where:

$$\mathcal{Q}_3(v) = \text{upper quartile boundary of } v$$

This variability score is computed on an individual basis for each pairing of generated and human-generated prompts using the function V.

$$V(p_i, h_j) = v \qquad \forall i \in [1 \ldots n], j \in [1 \ldots k] \tag{5}$$

V is established as a weighted aggregation of several semantic and n-gram similarity metrics M_x that gauge the resemblance between the two input texts $\mathcal{T}_1$ and $\mathcal{T}_2$. Each metric is normalized to the range $[0, 1]$ and the weights are input w_α, which is the metric for M_α, reflecting its relative significance in the computation of the overall variability score. The weights are determined by the user based on the desired trade-off between the metrics.

$$V(\mathcal{T}_1, \mathcal{T}_2) = \sum_{\alpha}^{x} w_\alpha \cdot M_\alpha(\mathcal{T}_1, \mathcal{T}_2) \tag{6}$$

C. Hyperparameter Tuning

We introduce a language-specific hyperparameter set, S, which consists of a set of tokens that do not modify the semantic meaning of prompt, but rather control the visual style or apply additional constraints to the output image through references to external factors such as artistic styles or dimensions.

We define a multi-label logistic regression model, $\mathcal{H}$, specifically trained on the hyperparameter set for the language S, that takes in an input prompt and a relevance threshold r. The model outputs the best hyperparameters, that are relevant to the input prompt, based on the relevance threshold r.

$$\mathcal{H}(p_j, r) = (h_x, h_y, h_z) \tag{7}$$

V. Experimenal Results

A. Dataset

The dataset $\mathcal{D}_p$ utilized in this study encompasses a collection of 300 samples of natural language text, paired with a corresponding list of human-created image prompts and representative images generated by the Midjourney Image Generation Platform. An additional dataset of hyperparameters, $\mathcal{D}_h$, is used in order to aid in the generation of tuned images. The majority of the natural language text originates from classical and contemporary English fiction literature, while the image prompts, hyperparameters and generated images are created by the paper author. However, the dataset's representation is imbalanced, as the pre-selected text samples contain a disproportionate amount of sensory and visual narrative information compared to generic public domain text.

Note: The current dataset includes human prompts for only one artist. Further advancements can be made in the dataset's effectiveness by augmenting the dataset with additional human prompts sourced from various artists and expanding the dataset's natural language text samples. This will models to generalize better to the prompt engineering process employed by various artists, as opposed to a single artist. The author will release this dataset as open-source under the GNU GPL-v3 license, thereby enabling future research in this area.

1) Dataset Entry Example

Here, we can look at one of the sample rows in the dataset $\mathcal{D}_p$, as well as some hyperparameters present in the dataset $\mathcal{D}_h$.

Original Text: John had always felt drawn to the great outdoors. He loved the feel of the wind on his face and the sun on his skin. He loved the sense of freedom and adventure that came with exploring the natural world. And today, as he stood at the base of the towering cliff, he felt that familiar rush of excitement. As he looked up at the jagged rock face, he noticed another man already scaling the cliff. This man was muscular and strong, with bulging biceps and powerful legs that propelled him upwards. John couldn't help but feel a twinge of envy as he watched the man climb, effortlessly finding handholds and footholds that seemed invisible to everyone else.

Human Created Prompt: A muscular man climbing a cliff.

Hyperparameter Set: DSLR, Realistic, Grunge, Dark, Bokeh

Generated Image:

Figure 2. Image generated by Sumuk Shashidhar, via Midjourney

B. Experimental Setup

We use Pegasus [11] as the backend for the chunking model C. We use the T5X model [18], provided by Google Research as our primary prompt generator G, and train it using our main dataset, $\mathcal{D}_p$, using our variability metric v as a loss function. We use a basic multi-class logistic regression model as our hyperparameter tuner $\mathcal{H}$.

For our initial test, we define $t = 0.6$, an acceptable score of $v = 0.5$, and classification accuracy of 0.8. We use WMD, SPICE, CIDEr and BLEU as our evaluation metrics. The dataset is partitioned into training and testing sets using an 80/20 ratio.

C. Results

We see that our model learns to extract words that are generally used in visual contexts. It also learns to keep semantic context intact, at the cost of allowing words that are not related to visual contexts into the final generated prompt. Due to a large threshold in the current run of our experiment, we also notice prompts with a significant amount of visual information. The following are the weights, $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ for the metrics M_x that we use. These are calculated by our logistic regression tuner, based on the dataset.:

Weights: [[0.00131709] [0.51856935] [0.17440857] [0.30570499]]

1) Example Result

We use the same source text as in our dataset entry example in V-A1. We get the following prompts as an output.

- a towering cliff with a jagged rock face and a muscular man climbing it with bulging biceps and powerful legs, while another man stands at the base of the cliff, feeling the wind on his face and sun on his skin. *V: 0.44681*

- standing in front of a towering cliff with the sun shining down on him, feeling drawn to the great outdoors and the sense of freedom and adventure that comes with exploring the natural world. *V: .93758*
- a man watching another man climb a towering cliff with ease, feeling a twinge of envy as he notices the other man's strength and ability to effortlessly find handholds and footholds that are invisible to everyone else. *V: 0.95377*

Figure 3. Resultant image of first machine generated generated prompt

Note: It is worth noting that the output images are not identical in nature. However, by appropriately tuning the threshold hyperparameter, it is possible to generate more focused prompts that lead to more visually consistent images.

2) Baseline Performance

We establish a baseline for evaluating the effectiveness of our proposed process by considering the performance of the image generation task without utilizing our proposed method, i.e., by simply utilizing the source text as the prompt.

Figure 4. Image generated by text source

Upon comparison with the baseline image, we observe that although it is visually coherent, it lacks semantic context and fails to maintain the scenario depicted in the source text. In contrast, the generated prompts using our proposed method preserve the situation and the context of the source text. While there is room for further improvements in the generation process, it is evident that our proposed method is capable of producing more visually consistent and accurate images.

Table I
BASELINE VS. MODEL PERFORMANCE - LOWER IS BETTER (PART 1)

Prompt	**Performance Score (v)**
Baseline	0.99889
Generated Prompt 1	0.44681
Generated Prompt 2	0.63758
Generated Prompt 3	0.65377
Human Prompt	0.0

Table II
BREAKDOWN OF MODEL PERFORMANCE - LOWER IS BETTER

Prompt	M_1	M_2	M_3	M_4
Baseline	0.32855	0.99996	0.99963	0.99957
Generated Prompt 1	0.32057	0.35594	0.42849	0.60868
Generated Prompt 2	0.24839	0.73540	0.62211	0.48215
Generated Prompt 3	0.18621	0.82237	0.65796	0.36740
Human Prompt	0.0	0.0	0.0	0.0

Table III
WEIGHTS

Alpha	**Weight**
α_1	0.00131709
α_2	0.51856935
α_3	0.17440857
α_4	0.30570499

VI. DISCUSSION

Our study has successfully shown that the proposed methodology for generating text prompts is highly effective in producing visually consistent and accurate images, in comparison to the baseline. By exploiting the capabilities of T5X for prompt generation and building upon it, we have achieved a high-quality extraction of visual information from the source text.

We assert that this proof of concept for the translation of the same text into different formats, without losing its essence, holds significant implications for the machine learning community. This technique can be effectively employed in other scenarios, such as transcribing emails, so that they match the preferences of the particular recipient. Similarly, it can be employed in the writing process to ensure that the same content is conveyed differently

to different audiences, thereby increasing its impact. The potential applications of domain-local finetuned machine learning models are indeed remarkable.

A. Improvements and Future Work

The current study presents an innovative approach for generating high-quality images using text prompts. However, there are several areas for improvement and future work. One of the major limitations of the current dataset is that it is limited to human prompts from a single artist, thereby restricting the model's generalization to different artists' prompt engineering processes. To address this, it is recommended to expand the dataset with prompts from multiple artists and a wider range of natural language text samples. This can be achieved by collecting data from the providers of the diffusion models or collaborating on an open source dataset. Additionally, it is suggested to avoid using a service like Amazon's Mechanical Turk program for this task, as prompt curation requires specialized skills.

Incorporating a more robust hyperparameter tuning process could potentially improve the performance of the generated images. This can be achieved by exploring more sophisticated models for hyperparameter tuning or using optimization techniques such as Bayesian optimization or genetic algorithms. Furthermore, improving the evaluation metrics can lead to a more accurate assessment of the model's performance. While current metrics such as SPICE and CIDEr are effective in measuring semantic similarity between texts, they may not fully capture the quality of generated prompts in the context of image synthesis. Developing new evaluation metrics tailored to the task of text-prompt generation for image synthesis could provide a more accurate assessment of the model's performance.

Lastly, given the fast-paced advancements in AI-generated art and images, it is essential to adapt the proposed method to work effectively with newer models and architectures. It is suggested to incorporate the proposed technique directly into a text-conditional diffusion model to negate the need for an external model. This can ensure the continued applicability of the proposed method in the face of evolving AI technologies.

VII. Conclusion

This paper proposes a new approach for generating text prompts that can be used with text-conditional diffusion-based image synthesizers. The method eliminates the need for manual prompt engineering and reduces the time and effort required to generate a large number of images. Specifically, our model operates as a text-to-text translator, converting conventional, literature-based text into visually oriented text that is more readily interpreted by the image generation model.

Our model outperforms the baseline approach, where the entire prompt is input into the generator. The primary distinction observed is an improvement of up to 50% with average scores of (0.44681, 0.63758, 0.65377) and higher. This can be attributed to the prompts generated by our model closely resembling the training data used for text-conditional image generators, such as Midjourney, Stable Diffusion, and DALLE-2. Additionally, the chunker component of the model ensures the appropriate amount of visual information is provided, avoiding excessively large prompts that result in images focusing solely on a single overarching concept.

Furthermore, our approach effectively balances the level of visual activity within a given prompt, enabling the generation of more focused images. While the results show promise, there are several avenues for future work, including expanding the dataset, improving hyperparameter tuning, refining evaluation metrics, and adapting the method to new image synthesis models.

The proposed method has significant potential for a variety of applications, such as book illustration, movie scene generation, and game prototyping. By addressing the challenges and building on the foundation laid by this research, we can contribute to the growth of AI-generated art and images and enable more scalable and efficient generation processes.

References

[1] *Prompt Explorations*. 2023. URL: https://docs.midjourney.com/docs/explore-prompting.

[2] Peter Anderson et al. "SPICE: Semantic Propositional Image Caption Evaluation". In: *ECCV*. 2016.

[3] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. *CIDEr: Consensus-based Image Description Evaluation*. 2015. arXiv: 1411.5726 [cs.CV].

[4] Ashish Vaswani et al. *Attention Is All You Need*. 2017. arXiv: 1706.03762 [cs.CL].

[5] Mike Lewis et al. *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*. 2019. arXiv: 1910.13461 [cs.CL].

[6] Colin Raffel et al. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. 2020. arXiv: 1910.10683 [cs.LG].

[7] Yang Liu. *Fine-tune BERT for Extractive Summarization*. 2019. arXiv: 1903.10318 [cs.CL].

[8] Karl Moritz Hermann et al. *Teaching Machines to Read and Comprehend*. 2015. arXiv: 1506.03340 [cs.CL].

[9] David Graff et al. "English gigaword". In: *Linguistic Data Consortium, Philadelphia* 4.1 (2003), p. 34.

[10] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. "Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium, 2018.

[11] Jingqing Zhang et al. *PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization*. 2020. arXiv: 1912 . 08777 [cs.CL].

[12] Chin-Yew Lin. "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013.

[13] Robin Rombach et al. *High-Resolution Image Synthesis with Latent Diffusion Models*. 2022. arXiv: 2112.10752 [cs.CV].

[14] Chin-Yew Lin. *ROUGE: A Package for Automatic Evaluation of Summaries*. URL: https://aclanthology.org/W04-1013.pdf.

[15] Satanjeev Banerjee and Alon Lavie. "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments". In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics, June 2005, pp. 65–72. URL: https://aclanthology.org/W05-0909.

[16] Jesus Gimenez and Lluis Marquez. "Linguistic Features for Automatic Evaluation of Heterogenous MT Systems". In: *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 256–264. URL: https://aclanthology.org/W07-0738.

[17] Mert Kilickaya et al. *Re-evaluating Automatic Metrics for Image Captioning*. 2016. arXiv: 1612.07600 [cs.CL].

[18] Adam Roberts et al. *Scaling Up Models and Data with `t5x` and `seqio`*. 2022. arXiv: 2203.17189 [cs.LG].