GEM: Generating Engaging Multimodal Content

Chongyang Gao\textsuperscript{1}, Yiren Jian\textsuperscript{2}, Natalia Denisenko\textsuperscript{1}, Soroush Vosoughi\textsuperscript{2} and V.S. Subrahmanian\textsuperscript{1}

\textsuperscript{1}Northwestern University
\textsuperscript{2}Dartmouth College
chongyanggao2026@u.northwestern.edu, yiren.jian.gr@dartmouth.edu, nataliadenisenko2027@u.northwestern.edu, soroush.vosoughi@dartmouth.edu, vss@northwestern.edu

Abstract

Generating engaging multimodal content is a key objective in numerous applications, such as the creation of online advertisements that captivate user attention through a synergy of images and text. In this paper, we introduce GEM, a novel framework engineered for the generation of engaging multimodal image-text posts. The GEM framework operates in two primary phases. Initially, GEM integrates a pre-trained engagement discriminator with a technique for deriving an effective continuous prompt tailored for the stable diffusion model. Subsequently, GEM unveils an iterative algorithm dedicated to producing coherent and compelling image-sentence pairs centered around a specified topic of interest. Through a combination of experimental analysis and human evaluations, we establish that the image-sentence pairs generated by GEM not only surpass several established baselines in terms of engagement but also in achieving superior alignment.

1 Introduction

The advent of social media platforms has significantly elevated the importance of multimodal content, spanning online advertising, image-text posts, and educational materials. In these digital arenas, the ability to generate engaging image-text posts is crucial for capturing user interest and maintaining relevance to current topics. To this end, exploring varied contexts in which advertisements can thrive presents a groundbreaking strategy:

- Figure 1(a) depicts a proposed social media campaign for the Crazy Leopard Lodge, a hotel located near a safari park, aiming to attract tourists with captivating image-text posts created by our GEM system.
- Figure 1(b) demonstrates the capability of our GEM system in producing an appealing advertisement for Pizza Amore, emphasizing the authenticity of its Neapolitan thin-crust pizza with an enticing image.
- Utilizing the GEM system, Figure 1(c) unveils a creative social media campaign for the upcoming drone-themed video game, “Drone Wars,” designed to engage potential gamers with an engaging slogan and visual appeal.
- As illustrated in Figure 1(d), Pacific Airways utilizes the GEM system to promote its affordable yet luxurious business class fares with engaging image-text posts, combining striking visuals with concise, compelling text.

The examples highlighted in Figure 1 not only comply with the character limits of Twitter but also showcase the adaptability of our content across diverse social media platforms such as Facebook.

The field of generative models has achieved significant progress, enabling the creation of varied uni-modal content, from poetic compositions [Agarwal and Kann, 2020] to advanced code snippets [Rozière et al., 2023] and describing visual content [Jian et al., 2023; Yang et al., 2021]. Additionally, diffusion models have emerged as a powerful method for producing highly realistic images [Rombach et al., 2021; Ho et al., 2020].

However, despite these advancements, the synthesis of multimodal content, especially the coherent integration of image and text components, remains an area less ventured into. Current models often focus on narrow domains, constrained by the limitations of their training datasets, which hampers the diversity and engagement of the generated outputs [Lee et al., 2022; Hu et al., 2022; Qiao et al., 2019; Wah et al., 2011].

To bridge this gap, we propose GEM, a comprehensive framework tailored for generating engaging multimodal image-text posts, aiming to boost user interaction while ensuring harmony between image and text elements. GEM sets itself apart by prioritizing content engagement within specific constraints, like the 280-character limit on Twitter. The framework employs a dual-phase approach: initially training an engagement discriminator that works in tandem with a diffusion model to derive continuous prompts for creating contextually relevant images. It then leverages an iterative algorithm, which combines the engagement image generator with a pre-trained text paraphrase model, to craft coherent and captivating image-text posts on chosen subjects.

Through a combination of empirical analysis and human assessments, we validate the GEM framework’s superior performance in generating image-text posts that outperform var-
ious benchmarks in engagement metrics. GEM demonstrates exceptional capability in producing aligned and engaging multimodal content in an open-vocabulary scenario, effectively addressing existing shortcomings in this research area.

In the following sections, we present a comprehensive review of existing research in the fields of content generation, engagement metrics, and the development of adaptable prompts that enable controlled generation. We delve into the architecture and operational methodology of GEM, highlighting its key components such as the engagement classifier for directed generation, the prompt-based image generation system, and the iterative methodology for creating image-text posts. The discussion concludes with an analysis of empirical findings and the potential applications of our framework. The primary contributions of this research include:

- Introducing a comprehensive framework for the simultaneous generation of engaging image-text posts, with a particular focus on enhancing advertising initiatives.
- Developing an engagement discriminator trained on Instagram data, which aids in learning engaging continuous prompts for the diffusion model. This is complemented by an iterative refinement process aimed at producing highly coherent and engaging image-text posts.
- Providing quantitative and qualitative evidence to showcase our framework’s superiority in generating image-text posts that are both more engaging and better aligned.

2 Related Work

2.1 Social Media-Centric Generation

The digital age has ushered in a dramatic augmentation in the development and proliferation of large vision-and-language models specializing in the generation of visually appealing and semantically consistent content prevalent on social media platforms. Generative Adversarial Networks (GANs) [Goodfellow et al., 2020] have emerged as a popular choice, proficient at synthesizing images that resonate with textual descriptions, a fundamental component in the creation of social media posts. Noteworthy advancements include Conditional GANs [Isola et al., 2017; Gao et al., 2024], mitigating bias generation [Liu et al., 2021], and others that have pushed the boundaries of realistic generation, promising a more immersive experience for social media users. The quest for enhanced multimodal data handling has also shifted towards transformer-based architectures, combining the capabilities of GANs and transformers [Lee et al., 2021; Zhang et al., 2022] to foster a more harmonized generation process. In a bid to foster content that engages users on a deeper level, a new generation of models such as DALL-E-2 [Ramesh et al., 2022] have been conceptualized, which excel at generating high-quality images guided by textual prompts, a critical feature for enhancing user engagement on social media platforms.

2.2 Engagement in Social Media Contexts

With social media platforms burgeoning as hubs of user interaction and content sharing, the prerogative to craft engaging content has intensified. Despite the focus of generative models traditionally centered around the fluency of text, recent research endeavors have striven to infuse a level of engagement in the generated content conducive to fostering user interaction on social media platforms. Emphasizing the emotional resonance of the content through sentiment analysis [Mathews et al., 2016] has been identified as a potent strategy to amplify user engagement. Furthermore, integrating elements of humor [Yoshida et al., 2018] or puns [Chandrasekaran et al., 2017] can potentially catalyze stronger responses from the audience, fostering a vibrant social media environment. This research delineates a novel pathway, introducing an innovative classifier geared towards the generation of image-caption pairs characterized by high engagement scores, a vital metric in assessing content efficacy on social media platforms.
2.3 Leveraging Learnable Prompts for Controllable Generation in Social Media Contexts

As computational models burgeon in complexity and size, encapsulating billions of parameters, the traditional approach of end-to-end fine-tuning has become increasingly impracticable. Consequently, the paradigm has shifted towards the utilization of learnable “prompts” to guide the generative processes more effectively [Brown et al., 2020]. These prompts, acting as textual precursors or learnable vector embeddings, effectively navigate the model outputs, fostering a more controlled and flexible generation process crucial for crafting targeted content on social media platforms. In the realm of language modeling, various strategies have been developed, including AutoPrompt [Shin et al., 2020], Prefix-Tuning [Li and Liang, 2021], and others, to optimize the utilization of prompts, fostering more nuanced and controlled outputs. The applicability of this technique has also transcended to computer vision [Jia et al., 2022; Zhou et al., 2022] and multimodal pre-training [Jian et al., 2024; Li et al., 2023], establishing its critical role in fine-tuning content generation strategies, especially in the context of social media platforms where the user engagement is paramount. The comparative analysis of these models provides a comprehensive perspective on the advancements in the field and demonstrates that GEM can generate engaging image-text pairs in an open-vocabulary setting.

3 The GEM Model

Our GEM model establishes a novel approach to generating multimodal content that is highly engaging, utilizing a fusion of text and image components based on GPT [Brown et al., 2020] and the recent text-to-image Stable Diffusion (SD) model, respectively. Figure 2 shows the training process within the GEM framework that involves a twofold procedure: an engagement classifier is trained to provide the engaging score, and then continuous prompts for diffusion model are learned with the supervision of both engaging and reconstruction losses. During inference, we proposed an iterative algorithm, as illustrated in Alg. 2, which enhances the alignment between text and generated images.

### 3.1 Preliminaries

The cornerstone of the GEM method is the deployment of fusion models, establishing a Markov process that iteratively infuses Gaussian noise with a variance denoted by $\beta_t$ into the data, which is subsequently utilized to reverse the diffusion process and reconstruct the data from the induced noise. The foundational equation for forward diffusion, originating from the initial image data $x_0$ produced during the first stage of GEM, is represented as:

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t} \epsilon \quad (1)$$

where $\alpha_t = 1 - \beta_t$, and $\epsilon$ is sampled from a normal distribution. Stable diffusion consists of an autoencoder and a UNet denoiser. The autoencoder first converts the image $x_0$ into the latent space $z_0$ and then reconstructs it. A modified UNet [Ronneberger et al., 2015] denoiser is used to estimate the noise $\epsilon_\theta(z_t, t, c)$ in the latent space, where $\theta$ refers to the parameters of the UNet denoiser. $z_t$ is the latent map in the time step $t$, which can be calculated using Equation 1. $c$ is the conditional information, which is the text in GEM.

In our setting, we have access to a pre-trained diffusion model [Rombach et al., 2021] $f_d$ and an engaging multimodal content dataset [Kim et al., 2020] with image-text-score triplets ($Img, Txt, Score$).

### 3.2 Engagement Classifier for Generation Guidance

We first train an “engagement” classifier to provide engagement guidance for training continuous prompts to generate images with high engagement scores. The engagement classifier $f_c$ contains a pre-trained vision-language model, namely ViLT [Kim et al., 2021], to extract multimodal features from the text and image pairs. Next, we attach a fully connected module on top of ViLT to assess the engagement level of the pairs. The fully connected module consists of two fully connected layers and one ReLU [Agarap, 2018] activation layer between them.

The engagement classifier, denoted by $f_{cls}$, is trained on the engagement dataset [Kim et al., 2020] using image-text-score triplets ($Img, Txt, Score$) until convergence, as illustrated in Alg. 1. The engagement scores are binarized to allow a classification task and we use the cross-entropy loss to
perform the supervised training.

3.3 Prompt-based Image Generator

We adopt Stable Diffusion ($f_d$) \cite{rombach2021high} as our base image generation model for text-to-image generation. To further control the generation, we concatenate a set of learnable vectors ($cp$) as continuous prompts [Li and Liang, 2021] and text embeddings. The output of the diffusion model $I_e$ can be denoted as follows:

$$I_e = f_d(z, t, \text{Concat}(cp, \text{Emb}(T xt))),$$  \hspace{1cm} (2)

where $t$ is the time step uniformly sampled from $1, ..., T$, and $T$ is the length of time steps. $z \sim N(0, I)$ is the latent map, which is the Gaussian noise. $\text{Emb}$() is the text embedding layer of diffusion model $f_d$, and $\text{Concat}$ denotes the concatenation operation. The dimension of $cp$ is $l \times d$. $p \in \mathbb{R}^{l \times d}$, where $l$ is the length of the continuous prompt and $d$ is the hidden dimension size of the text embedding. Stable diffusion uses a pre-trained CLIP text encoder \cite{radford2021learning}, and the hidden dimension size is 768. To encourage the generation of engaging images, we forward the diffusion output $I_e$ and input $T xt$ to the pre-trained engagement classifier $f_{cls}$ introduced in Section 3.2. By increasing the positive confidence of $f_{cls}$ in predicting a generated pair $(I_e, T xt)$, the classifier $f_{cls}$ back-propagates the gradients into the diffusion model $f_d$ to produce more engaging images. To be precise,

$$L_{\text{engaging}} = \text{CrossEntropy}(f_{cls}(I_e, T xt), 1),$$  \hspace{1cm} (3)

where 1 in Equation 3 refers to the fact that the label of $(I_e, T xt)$ is an engaging pair. In addition to the $L_{\text{engaging}}$ loss, we keep the standard reconstruction loss \cite{rombach2021high} in the training process of continuous prompts. In particular, the reconstruction loss computes the difference between the predicted noise and the original noise, i.e.:

$$L_{\text{rec}} = \mathbb{E}_{x, e, t} \left[ \epsilon - f_d(z_i, t, e) \right]^2,$$  \hspace{1cm} (4)

where $z_i$ is acquired by encoding the input image and $x$ into the latent space using stable diffusion’s autoencoder, which is the noisy version of $x$ and can be calculated using $\epsilon \sim N(0, I)$ in Equation 1. $\epsilon$ is the original noise, and $e$ is $\text{Concat}(cp, \text{Emb}(T xt))$. Thus, the total loss for learning our continuous prompts is

$$L_{\text{total}} = L_{\text{engaging}} + w \cdot L_{\text{rec}}$$  \hspace{1cm} (5)

, where $w$ is the weight, which we set to 1 in our experiments because the scales of these two losses are comparable. During training, as illustrated in Alg. 1, we freeze the parameters of diffusion model $f_d$ in order to maintain the model’s ability to generate high-quality images while retaining control over the generation’s engagement score. Therefore, the only learnable parameters in our framework are in $cp$. Similar strategies have been found to be both efficient and effective in fine-tuning Transformer-based language models in NLP tasks \cite{liester2021text-image, li-and-liang-2021, liu-etal-2022}.

3.4 Iterative Image-Text Post Generation

A well-aligned image and text pair requires the image and text to share substantial similarities. Based on this intuition, we devise an iterative procedure for generating the image and text pairs with the help of CLIP-based similarity scores. Figure 3 and Algorithm 2 illustrate our iterative algorithm for image and text alignment.

Given the text, we first use the image generator described in Section 3.3 to generate an engaging image, $I_e$, and then calculate the similarity between the original text and $I_e$ as $\text{Sim}_{O}$, as shown in Lines 5-6 of Alg.2. We then generate the new text, $T xt'$, using the GPT-3.5 as text paraphraser, as shown in Line 7 of Alg. 2. A pre-trained CLIP model is used to measure the similarity between the generated image $I_e$ and all the text, $T xt, T xt'$. In particular, we can generate any number of $T xt'$. In our experiments, we generate 10 $T xt'$ instances. If $(T xt, I_e)$ has the highest similarity score, then we return $(T xt, I_e)$ as the generated image and text pair, as illustrated in Lines 15-16.

If there exists a $T xt'$ such that the similarity score of $(T xt', I_e)$ is greater than the similarity score of $(T xt, I_e)$, the $T xt'$ with the highest similarity score is used as the new $T xt$, as shown in Line 10. Next, we compare the similarity between $(T xt, I_e)$ and $\text{Sim}_{O} + S$, where $S$ is the threshold used to constrain the similarity value to be at least $S$ greater than the similarity of the original text and generated image. In our experiments, $S$ is set to 1. If the similarity of $(T xt, I_e)$ is greater than $\text{Sim}_{O} + S$, the algorithm returns $(T xt, I_e)$, as illustrated in Lines 11-12. If not, we use the image generator to generate a new image given the new $T xt$ and generate text $T xt'$ based on the new $T xt$, as shown in Lines 13-14. In order to make the algorithm more efficient and avoid the problem of the endpoint not being achieved, we limit the maximum number of iterations (it is set to 50 in our experiments). The iterative image and text pair generation process enables our method to generate more well-aligned and engaging image and text pairs, as illustrated in Section 4.5. We can address any additional constraints via the prompts for both text paraphrase and prompt-based image generator. For example, we use the prompt, "within 280 words," to satisfy the word limits of Twitter.
4 Experiments

This section details our experimental setup to evaluate the quality of generated image-text posts. All models are assessed through a twofold evaluation approach: quantitatively using CLIP similarity scores (detailed in Section 4.4) and qualitatively through human evaluation that analyzes engagement levels of the generated posts (described in Section 4.5).

4.1 Dataset

We utilized the Instagram Influencer Dataset [Kim et al., 2020] to train the engagement classifier and the continuous prompt mechanism of the stable diffusion model for image generation. The dataset comprises 10,180,500 Instagram posts from 33,935 Instagram influencers. There are nine categories: beauty, family, fashion, fitness, food, interior, pet, travel, and others. We find the text in the posts is too short, sometimes empty, and does not describe the image sufficiently. Hence, we use the BLIP-2 [Li et al., 2023] model to provide the caption for the Instagram image instead of using the original text in the post. Moreover, we generate engaging captions with the prompt, “Write something engaging and interesting for the image”. We defined the engagement scores by the mean value of the likes. Specifically, we divided the likes of each post by the mean value of the likes of all posts as the engagement scores. If the likes of the post are greater than the mean value, the engagement score of that post is 1. We sampled the datasets into two sub-datasets, each containing 20,000 samples. One sub-dataset is used to train the engagement classifier, and the other is used to train the continuous prompt for stable diffusion.

For evaluation samples, we prepare the initial text using five distinct themes - crowds, vehicles, nature, architecture, and notable individuals- and each topic encompasses 25 sentences. Initial textual prompts were generated using GPT-3.5 without imposing restrictions on text length. These prompts are the initial inputs for baselines and our framework.

4.2 Baselines

We benchmark the results against three established baseline models: 1) the stable diffusion model utilizing original sentences as input, referred to as D; 2) the stable diffusion model with paraphrased sentences as inputs, referred to as D + P; and 3) the stable diffusion model that incorporates a combination of continuous prompts and paraphrased sentence embeddings as input, termed as D + P + CP. Notably, depending on the types of captions we used for training continuous prompts, as described in Section 4.1, we have two types of continuous prompts. One is CP (Caption) that is trained with caption generated by BLIP-2 directly, and the other one is CP (Engaging) that is trained with engaging captions generated by BLIP-2 using specific prompts. Our GEM has two types based on the type of continuous prompts, which are GEM(C) and GEM(E).

4.3 Implementation Details

We employed the pre-trained VILT [Kim et al., 2021] model to extract features from the image-text pairs. This procedure is followed by the integration of a fully connected module to downscale the VILT dimension from 768 to 120, accompanied by ReLU activation and another fully connected layer tasked with predicting engagement labels. The model underwent a training phase spanning 10 epochs, with a learning rate of 1e-3 and a batch size of 64, optimized through stochastic gradient descent with a momentum of 0.9.

A grid search was executed to find the optimal learning rate from the set 1e-2, 1e-3, 1e-4, since the training of the model is not sensitive to the learning rate for predicting engagement scores. We use the pre-trained stable diffusion v1.5 1 and set the dimension to (prefix_length, 768) when training the learnable continuous prompt. In our setting, the prefix_length = 2. The learning rate is 1e−5 with gradient accumulation steps of 4. The batch size is set to 16, and we use the Adam optimizer, following the default hyper-parameters and settings provided by HuggingFace. For the iterative creation of image-text posts, we leveraged the GPT 3.5 as the text paraphrase, with the prompt, “Please rewrite the following sentences and make them more engaging and interesting.” Mostly, the image-text pairs can be generated

1https://huggingface.co/runwayml/stable-diffusion-v1-5

Algorithm 2: Iterative Generation Process

<table>
<thead>
<tr>
<th>Line</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TP: text paraphraser</td>
</tr>
<tr>
<td>2</td>
<td>TXT: Input text</td>
</tr>
<tr>
<td>3</td>
<td>Sim: the similarity score provided by CLIP</td>
</tr>
<tr>
<td>4</td>
<td>S: threshold of the termination</td>
</tr>
<tr>
<td>5</td>
<td>( I_e = f_d(\text{Concate}(cp, \text{Emb}(TXT))) )</td>
</tr>
<tr>
<td>6</td>
<td>( \text{Sim}_O = \text{Sim}(TXT, I_e) )</td>
</tr>
<tr>
<td>7</td>
<td>( TXT' = TP(TXT) )</td>
</tr>
<tr>
<td>8</td>
<td>for Number of iteration do</td>
</tr>
<tr>
<td>9</td>
<td>if ( \text{Sim}(TXT', I_e) &gt; \text{Sim}(TXT, I_e) ) then</td>
</tr>
<tr>
<td>10</td>
<td>( TXT = TXT' )</td>
</tr>
<tr>
<td>11</td>
<td>if ( \text{Sim}(TXT, I_e) &gt; \text{Sim}_O + S ) then</td>
</tr>
<tr>
<td>12</td>
<td>( \text{return } (TXT, I_e) )</td>
</tr>
<tr>
<td>13</td>
<td>( I_e = f_{diff}(\text{Concate}(cp, \text{Emb}(TXT))) )</td>
</tr>
<tr>
<td>14</td>
<td>( TXT' = TP(TXT) )</td>
</tr>
<tr>
<td>15</td>
<td>else</td>
</tr>
<tr>
<td>16</td>
<td>( \text{return } (TXT, IMG) )</td>
</tr>
<tr>
<td>17</td>
<td>( \text{return } (TXT, I_e) )</td>
</tr>
</tbody>
</table>

Figure 3: Iterative image-text generation process for each iteration.
D + P + CP (Caption)
D + P + CP (Engaging)
GEM (Caption)
GEM (Engaging)

Table 1: Average similarity scores.

4.4 Similarity Score Analysis

To gauge the alignment between image and text pairs, we utilized the pre-trained CLIP model to compute similarity scores for each of the 125 image-text pairs, subsequently calculating the mean scores for each method. According to Table 1, a discernible decline in similarity scores was noted with the introduction of paraphrased text and continuous prompt (Caption/Engaging) inputs, as evidenced by a decrease of 0.44, 1.07, and 0.49 respectively.

Our iterative generation method mitigates the drop and elevates the similarity score, surpassing Diffusion + CP + P (C/E) by 2.63 and 1.54, respectively. Moreover, the GEM(C) model achieved the highest similarity score of 26.36, which demonstrated the promising ability of the iterative generation method to align the image and text. Above all, incorporating paraphrasing and continuous prompt methodologies led to a drop in similarity scores. Our iterative algorithm mitigates the problem and facilitates the generation of more aligned image-text pairs with the highest similarity scores.

4.5 Human Evaluation

To assess the effectiveness of our method relative to standard baselines, we conducted human evaluations using Amazon Mechanical Turk (MTurk). We paid participants $0.03 per task, which equates to an approximate hourly wage of $15, based on the average task completion time of 7 seconds. To ensure high-quality data, we only employed workers who had an approval rating of 99% or higher and had completed at least 10,000 tasks previously.

These evaluations were conducted in a pairwise manner.
As depicted in Figure 4, the image-text posts are sequen-
tally generated by Diffusion. As demonstrated in Figure 1, initially, we deploy GPT 3.5 to con-
coc-text advertisements, incorporating a “within 280 words”
clause in the prompts to comply with Twitter’s character limi-
tation. Subsequent to this preparation, our GEM(E) takes
charge, generating tweets that harmoniously pair images and
text, as evident from the well-aligned outputs in Figure. 1.

### 5 Limitations

The advent of social media has underscored the importance
creating engaging multimodal posts for wide-ranging dis-
semination across sectors such as education, healthcare, and
environmental advocacy. The GEM framework offers an effi-
cient pathway for entities aiming to bolster their online foot-
print through vibrant, dynamic content. Yet, the current it-
eration of GEM is designed to process exclusively textual
inputs, triggering a cascade that produces images and their
Corresponding text sequences. While beneficial for scenarios
where only textual outlines are at hand, this singular focus on
text as the seed for content generation may fall short of de-
ivering optimal image-text congruence. The challenge lies
in the inherent limitation of text-only inputs to fully convey
the nuanced visual details that might be critical for creating
perfectly aligned multimodal content.

Advancements in GEM should consider the integration of
a more versatile input mechanism, such as the capability to
interpret sketches alongside the text, offering a dual-modality
of inputs. This enhancement could significantly improve the
system’s flexibility and ensure a higher degree of harmony be-
tween the visual and textual components of generated posts,
addressing a critical limitation and paving the way for a more
refined and adaptable content generation platform.

### 6 Conclusion

In this work, we introduced the GEM framework, designed
to generate engaging and coherently aligned image-text posts
tailored for social media platforms. This framework em-
ploys a continuous prompt learning mechanism guided by an
engagement classifier alongside an iterative algorithm. The
outcomes of our experiments underscore the efficacy of the
GEM framework in creating more aligned image-text con-
tent that significantly boosts user engagement. These results
have profound implications for organizations looking to im-
prove their social media presence through deliberate content
creation, contributing to the advancement of multimodal con-
tent generation and social media marketing strategies. Future
research should consider the incorporation of diverse inputs
and further refinement of the generative process to promote a
more responsible and engaging social media environment.

<table>
<thead>
<tr>
<th>Methods</th>
<th>More Engaging</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diffusion v. D+P</td>
<td>D+P (55%)</td>
</tr>
<tr>
<td>Diffusion v. D+P+CP(C)</td>
<td>Diffusion (54%)</td>
</tr>
<tr>
<td>Diffusion v. D+P+CP(E)</td>
<td>D+P+CP(E) (60%)</td>
</tr>
<tr>
<td>Diffusion v. GEM(C)</td>
<td>GEM (60%)</td>
</tr>
<tr>
<td>Diffusion v. GEM(E)</td>
<td>GEM (63%)</td>
</tr>
<tr>
<td>D+P v. D+P+CP(C)</td>
<td>D+P+CP(C) (52%)</td>
</tr>
<tr>
<td>D+P v. D+P+CP(E)</td>
<td>D+P+CP(E) (52%)</td>
</tr>
<tr>
<td>D+P+CP(C) v. GEM(C)</td>
<td>GEM (58%)</td>
</tr>
<tr>
<td>D+P+CP(E) v. GEM(E)</td>
<td>GEM (53%)</td>
</tr>
<tr>
<td>GEM(C) v. GEM(E)</td>
<td>GEM (58%)</td>
</tr>
</tbody>
</table>

Table 2: Results of the human evaluation. The selected percentage
of more engaging methods is illustrated in parentheses.

Specifically, for image-text pairs that are generated by two
methods that are compared with the same initial text input,
the workers need to decide which method generates a more
engaging image-text pair. For each pair of compared meth-
ods, this comparison is performed for all 125 test samples
as described in Section 4.1, and the works are kept unaware
of the mechanisms underlying the generation to prevent any
preconceived bias. The order of the data presentation to the
workers was randomized to minimize any potential bias.

As demonstrated in Table 2, the GEM model is profi-
cient at generating cohesive and engaging image-text pairs,
and image-text pairs generated by our GEM(E) framework
markedly surpassed the engagement levels of all baseline
models, registering higher selection rates by margins of 13%,
3%, and 8% compared to D, D + P + CP (E), and GEM(C)
respectively. A comparative analysis between D and D + P
shows that text paraphrasing could notably enhance engage-
ment. Furthermore, the D + P + CP (E) model demonstrated
a 2% increase in selection frequency over D + P, thereby
highlighting the beneficial implications of incorporating en-
gaging continuous prompts for augmenting engagement.
Signi-
cantly, despite not explicitly steering the engagement gen-
eration, the deployment of the proposed iterative generation
algorithm amplified the engagement rate, recording an 8% and
3% increase in preference for GEM(C/E) compared to
D + P + CP (C/E), respectively. This trend insinuates that
fostering coherence between text and image elements can po-
tentially make them more engaging.

### 4.6 Case Studies

This subsection illustrates the proficiency of various meth-
ologies in crafting persuasive and engaging image-text posts.
As depicted in Figure 4, the image-text posts are sequen-
tially generated by Diffusion, D + P, D + CP + P (C/E),
and GEM(C/E), in order from left to right. The original input
sentence for each row corresponds to the input utilized by D
in the initial column.

Our analysis reveals that the generated text of GEM has
more details and is more attractive. The first example curated
by our GEM framework exhibits natural scenery poised to be
a potent tool in tourism promotion campaigns. In the second
row, our GEM(C) generates an attractive, colorful portrait,
and the image generated by GEM(E) is more interesting. The
subsequent example curated by our GEM framework exhibits
natural scenery poised to be a potent tool in tourism promo-
tion campaigns. We propose leveraging the third row’s ex-
amples to bolster architecture tour attendance. Interestingly,
the image instantiated by GEM(C) shows the interior of the
building with a pool, which makes it more engaging.

Venturing further, we showcase the adaptability of our
proposed pipeline by employing GEM to generate four distinct
image-text posts, apt for Twitter promotions by companies
in the hospitality, culinary, gaming, and aviation sectors, as
demonstrated in Figure 1. Initially, we deploy GPT 3.5 to con-
coca-t text advertisements, incorporating a “within 280 words”
clause in the prompts to comply with Twitter’s character limi-
tation. Subsequent to this preparation, our GEM(E) takes
charge, generating tweets that harmoniously pair images and
text, as evident from the well-aligned outputs in Figure. 1.
Ethical Statement

While the GEM project, to our knowledge, does not inherently raise specific ethical issues, it is essential to recognize the broader ethical landscape it operates within. Our framework relies on a pre-trained diffusion model, which, as highlighted in existing literature [Perera and Patel, 2023; Luccioni et al., 2023], may inherit and perpetuate biases present in its training data. Users and developers must remain vigilant about these biases and consider the ramifications of deploying such models in real-world scenarios.

The ability to generate engaging multimodal content introduces the risk of its exploitation for spreading misinformation or creating deceptive and harmful content. It is paramount to develop and integrate mechanisms that mitigate these risks, such as advanced filters and classifiers designed to identify and prevent the dissemination of inappropriate or unethical content. Additionally, the generation of content based on unimodal text inputs presents unique challenges in ensuring the alignment and relevance of the generated images, potentially leading to misrepresentations or misunderstandings. Efforts should be made to enhance the model’s understanding and handling of diverse inputs, promoting more accurate and contextually appropriate content generation.

Beyond these considerations, it is crucial to address ethical concerns related to privacy, consent, and the representation of individuals and communities. Ensuring that generated content does not infringe on privacy rights or contribute to the marginalization or stereotyping of any group is fundamental. As we advance in creating more sophisticated generative models, the development of ethical guidelines and the incorporation of ethical considerations into the design and deployment of these technologies become increasingly important to ensure they serve the public good and contribute to a responsible digital ecosystem.

Acknowledgments

We gratefully acknowledge support from ONR grants N00014-18-1-2670 and N00014-20-1-2407.

References


