GEM: Generating Engaging Multimodal Content

Chongyang Gao[1], **Yiren Jian**[2], **Natalia Denisenko**[1], **Soroush Vosoughi**[2] and **V.S. Subrahmanian**[1]
[1]Northwestern University
[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-

Figure 1: Sample image-text posts created by the GEM framework for various companies, demonstrating versatility and alignment with the specified criteria.

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 DALLE-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 GEMcan 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 diffusion models, establishing a Markov process that iteratively infuses Gaussian noise with a variance denoted by β_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:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon \tag{1}$$

where $\alpha_t = 1 - \beta_t$, and ϵ 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

Algorithm 1: GEM Training

```
# Step:  Number of training steps
# CLS: Engagement classifier
# E:Text embedding layer (FC)
# Diffuser:  Diffusion model.
# D_CLS: Dataloader for training C
# CP: Learnable continuous prompts
# D_CP: Dataloader for training CP

### Pre-training a engagment
 discriminator (ViLT based)
for i in Step:
    I, T, S = next(D_CLS)
    L_engaging = CE(CLS(I,T), S)
    L_engaging.backward()
    CLS.update()

### Learning a Diffuser (coninuous
 prompt learning)
for i in Max_Step:
    I, T, _ = next(D_CP)
    embed = concatenate(CP, E(T))
    I_gen = Diffuser(I, T)
    L_rec = MSE(I_gen, I)
    L_engaging = CE(CLS(I_gen, T), 1)
    L_total = L_rec + L_engaging
    L_total.backward()
    prompt.update()
return prompt
```

UNet [Ronneberger *et al.*, 2015] denoiser is used to estimate the noise $\epsilon_\theta(z_t, t, c)$ in the latent space, where θ 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

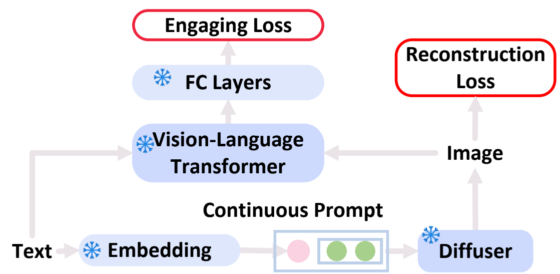

Figure 2: The training of continuous prompts. Given the text, its embedding is acquired using the embedding layer and is concatenated with continuous prompts. The image is generated by feeding the concatenated embedding into a stable diffusion model. Both the engagement loss and the reconstruction loss are used.

perform the supervised training.

3.3 Prompt-based Image Generator

We adopt Stable Diffusion (f_d) [Rombach *et al.*, 2021] 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:

$$I_e = f_d(z, t, Concate(cp, Emb(Txt))), \tag{2}$$

where t is the time step uniformly sampled from $1, ..., T$, and T is the length of time steps. $z \in \mathcal{N}(0, I)$ is the latent map, which is the Gaussian noise. $Emb()$ is the text embedding layer of diffusion model f_d, and $Concate$ 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 [Radford *et al.*, 2021], and the hidden dimension size is 768. To encourage the generation of engaging images, we forward the diffusion output I_e and input Txt 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, Txt), the classifier f_{cls} back-propagates the gradients into the diffusion model f_d to produce more engaging images. To be precise,

$$\mathcal{L}_{\text{engaging}} = CrossEntropy(f_{cls}(I_e, Txt), 1), \tag{3}$$

where 1 in Equation 3 refers to the fact that the label of (I_e, Txt) is an engaging pair. In addition to the $\mathcal{L}_{\text{engaging}}$ loss, we keep the standard reconstruction loss [Rombach *et al.*, 2021] 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.:

$$\mathcal{L}_{\text{rec}} = \mathbb{E}x, \epsilon, t\left[|\epsilon - f_d(z_t, t, e)|_2^2\right], \tag{4}$$

where z_t 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 $\epsilon \sim \mathcal{N}(0, I)$ in Equation1. ϵ is the original noise, and e is $Concate(cp, Emb(Txt))$. Thus, the total loss for learning our continuous prompts is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{engaging}} + w \cdot \mathcal{L}_{\text{rec}} \tag{5}$$

, 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 [Lester *et al.*, 2021; Li and Liang, 2021; Liu *et al.*, 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 Sim_O, as shown in Lines 5-6 of Alg.2. We then generate the new text, Txt', 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, Txt, Txt'. In particular, we can generate any number of Txt'. In our experiments, we generate 10 Txt' instances. If (Txt, I_e) has the highest similarity score, then we return (Txt, I_e) as the generated image and text pair, as illustrated in Lines 15-16.

If there exists a Txt' such that the similarity score of (Txt', I_e) is greater than the similarity score of (Txt, I_e), the Txt' with the highest similarity score is used as the new Txt, as shown in Line 10. Next, we compare the similarity between (Txt, I_e) and $Sim_O + S$, 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 (Txt, I_e) is greater than $Sim_O + S$, the algorithm returns (Txt, I_e), as illustrated in Lines 11-12. If not, we use the image generator to generate a new image given the new Txt and generate text Txt' based on the new Txt, 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.

Algorithm 2: Iterative Generation Process

TP: text paraphraser
Txt: Input text
Sim: the similarity score provided by CLIP
S: threshold of the termination
$I_e = f_d(Concate(cp, Emb(Txt)))$
$Sim\_O = Sim(Txt, I_e)$
$Txt' = TP(Txt)$
for *Number of iteration* **do**
    if $Sim(Txt', I_e) > Sim(Txt, I_e)$ **then**
        $Txt = Txt'$
        if $Sim(Txt, I_e) > Sim_O + S$ **then**
            $return\ (Txt, I_e)$
        $I_e = f_{diff}(Concate(cp, Emb(Txt)))$
        $Txt' = TP(Txt)$
    else
        $return\ (Txt, Img)$
$return\ (Txt, I_e)$

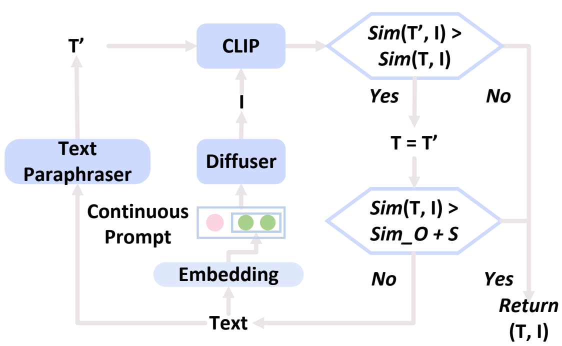

Figure 3: Iterative image-text generation process for each iteration.

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 continues 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 GEMhas 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 hyperparameters 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." Mostlty, the image-text pairs can be generated

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

Figure 4: Multimodal (image-text) pairs generated using various techniques. Columns are arranged from left to right to show pairs produced by different methods. Each row uses the text in the first column as the initial input.

Method	Similarity Score
Diffusion	25.24
Diffusion + P	24.8
Diffusion + P + CP (Caption)	23.73
Diffusion + P + CP (Engaging)	24.31
GEM (Caption)	**26.36**
GEM (Engaging)	25.85

Table 1: Average similarity scores.

within 10 iterations. Experiments were conducted using three RTX A6000 GPUs.

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.

Methods	More Engaging
Diffusion v. D+P	D+P (55%)
Diffusion v. D+P+CP(C)	Diffusion (54%)
Diffusion v. D+P+CP(E)	D+P+CP(E) (60%)
Diffusion v. GEM(C)	GEM(60%)
Diffusion v. GEM(E)	GEM(E) (63%)
D+P v. D+P+CP(C)	D+P+CP(C) (52%)
D+P v. D+P+CP(E)	D+P+CP(E) (52%)
D+P+CP(C) v. GEM(C)	GEM(C) (58%)
D+P+CP(E) v. GEM(E)	GEM(E) (53%)
GEM(C) v. GEM(E)	GEM(E) (58%)

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 methods, 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 biases. 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 proficient 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 engagement. Furthermore, the **D + P + CP (E)** model demonstrated a 2% increase in selection frequency over **D + P**, thereby highlighting the beneficial implications of incorporating engaging continuous prompts for augmenting engagement. Significantly, despite not explicitly steering the engagement generation, 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 potentially make them more engaging.

4.6 Case Studies

This subsection illustrates the proficiency of various methodologies in crafting persuasive and engaging image-text posts. As depicted in Figure 4, the image-text posts are sequentially 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 promotion campaigns. We propose leveraging the third row's examples 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 Figure1. Initially, we deploy GPT 3.5 to concoct text advertisements, incorporating a "within 280 words" clause in the prompts to comply with Twitter's character limitation. 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 of creating engaging multimodal posts for wide-ranging dissemination across sectors such as education, healthcare, and environmental advocacy. The GEM framework offers an efficient pathway for entities aiming to bolster their online footprint through vibrant, dynamic content. Yet, the current iteration 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 delivering 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 between 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 employs 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 content that significantly boosts user engagement. These results have profound implications for organizations looking to improve their social media presence through deliberate content creation, contributing to the advancement of multimodal content 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.

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

[Agarap, 2018] Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018.

[Agarwal and Kann, 2020] Rajat Agarwal and Katharina Kann. Acrostic poem generation. *arXiv preprint arXiv:2010.02239*, 2020.

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

[Chandrasekaran *et al.*, 2017] Arjun Chandrasekaran, Devi Parikh, and Mohit Bansal. Punny captions: Witty wordplay in image descriptions. *arXiv preprint arXiv:1704.08224*, 2017.

[Gao *et al.*, 2024] Chongyang Gao, Sushil Jajodia, Andrea Pugliese, and VS Subrahmanian. Fakedb: Generating fake synthetic databases. *IEEE Transactions on Dependable and Secure Computing*, 2024.

[Goodfellow *et al.*, 2020] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[Hu *et al.*, 2022] Minghui Hu, Chuanxia Zheng, Heliang Zheng, Tat-Jen Cham, Chaoyue Wang, Zuopeng Yang, Dacheng Tao, and Ponnuthurai N Suganthan. Unified discrete diffusion for simultaneous vision-language generation. *arXiv preprint arXiv:2211.14842*, 2022.

[Isola *et al.*, 2017] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[Jia *et al.*, 2022] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pages 709–727. Springer, 2022.

[Jian *et al.*, 2023] Yiren Jian, Tingkai Liu, Yunzhe Tao, Soroush Vosoughi, and Hongxia Yang. Simvlg: Simple and efficient pretraining of visual language generative models. *arXiv preprint arXiv:2310.03291*, 2023.

[Jian *et al.*, 2024] Yiren Jian, Chongyang Gao, and Soroush Vosoughi. Bootstrapping vision-language learning with decoupled language pre-training. *Advances in Neural Information Processing Systems*, 36, 2024.

[Kim *et al.*, 2020] Seungbae Kim, Jyun-Yu Jiang, Masaki Nakada, Jinyoung Han, and Wei Wang. Multimodal post attentive profiling for influencer marketing. In *Proceedings of The Web Conference 2020*, pages 2878–2884, 2020.

[Kim *et al.*, 2021] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.

[Lee *et al.*, 2021] Kwonjoon Lee, Huiwen Chang, Lu Jiang, Han Zhang, Zhuowen Tu, and Ce Liu. Vitgan: Training gans with vision transformers. *arXiv preprint arXiv:2107.04589*, 2021.

[Lee *et al.*, 2022] Hyungyung Lee, Sungjin Park, Joonseok Lee, and Edward Choi. Unconditional image-text pair generation with multimodal cross quantizer. *arXiv preprint arXiv:2204.07537*, 2022.

[Lester *et al.*, 2021] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[Li and Liang, 2021] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, 2021.

[Li *et al.*, 2023] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[Liu *et al.*, 2021] Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, Lili Wang, and Soroush Vosoughi. Mitigating political bias in language models through reinforced calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14857–14866, 2021.

[Liu *et al.*, 2022] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[Luccioni *et al.*, 2023] Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Analyzing societal representations in diffusion models. *arXiv preprint arXiv:2303.11408*, 2023.

[Mathews *et al.*, 2016] Alexander Mathews, Lexing Xie, and Xuming He. Senticap: Generating image descriptions with sentiments. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

[Perera and Patel, 2023] Malsha V Perera and Vishal M Patel. Analyzing bias in diffusion-based face generation models. *arXiv preprint arXiv:2305.06402*, 2023.

[Qiao *et al.*, 2019] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1505–1514, 2019.

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

[Ramesh *et al.*, 2022] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[Rombach *et al.*, 2021] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[Rozière *et al.*, 2023] Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023.

[Shin *et al.*, 2020] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, 2020.

[Wah *et al.*, 2011] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[Yang *et al.*, 2021] Xu Yang, Chongyang Gao, Hanwang Zhang, and Jianfei Cai. Auto-parsing network for image captioning and visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2197–2207, 2021.

[Yoshida *et al.*, 2018] Kota Yoshida, Munetaka Minoguchi, Kenichiro Wani, Akio Nakamura, and Hirokatsu Kataoka. Neural joking machine: Humorous image captioning. *arXiv preprint arXiv:1805.11850*, 2018.

[Zhang *et al.*, 2022] Bowen Zhang, Shuyang Gu, Bo Zhang, Jianmin Bao, Dong Chen, Fang Wen, Yong Wang, and Baining Guo. Styleswin: Transformer-based gan for high-resolution image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11304–11314, 2022.

[Zhou *et al.*, 2022] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.