Unveiling the Heart of AI-Generated Imagery

Ram Guttikonda

***Abstract*—Text-to-image and text-to-video generation models have emerged as transformative tools for producing stunning visual content from textual prompts, finding applications in industries like filmmaking, marketing, and personalized storytelling. However, their ability to capture and reflect the intricate emotions embedded in prompts remains largely uncharted territory. Emotional understanding is pivotal for tasks such as generating visuals from movie scripts, where depicting nuanced sentiments can define the narrative's emotional depth and contextual accuracy. Despite these demands, existing generative AI models face a critical limitation: over 70% of their training data is devoid of emotional intensity, leaving them ill-equipped to process sentiment-rich inputs effectively. This study pioneers the exploration of emotional attention in generative AI, specifically focusing on Stable Diffusion, marking a departure from earlier works that exclusively examined attributes, actions, and descriptive adjectives. Through meticulous cross-attention analysis, we evaluate how well the model attends to emotional words in prompts and adapts its outputs accordingly. Additionally, we investigate the model's behavior when emotional words are replaced with their opposites, revealing insights into its interpretative dynamics. The findings are compelling yet revealing: while the model generates images that align well with the content of prompts with high sentiment scores, the sentiment of the generated images often fails to reflect the intended emotional tone, achieving an accuracy of less than 10%. This exposes a fundamental gap in current generative AI systems, underscoring their inability to faithfully encode and attend to emotional cues. These results highlight the urgent need for sentiment-aware architectures and emotion-specific training paradigms. By uncovering these limitations, this research lays the groundwork for advancing generative AI to better understand and attend to human emotions. Future efforts could leverage these insights to develop emotion-sensitive frameworks, enabling applications in storytelling, therapeutic tools, and emotionally resonant content creation.**

***Index Terms*—Computer Vision, GenAI.**

I. Introduction

The fusion of language and vision, embodied in the task of text-to-image and text-to-video generation, represents one of the most captivating frontiers of artificial intelligence. These models have captivated researchers and practitioners alike, as they hold the promise of transforming creative industries, enabling machines to bring human imagination to life through the seamless translation of textual descriptions into visually compelling content. Applications span diverse domains—from crafting intricate concept art and visualizing cinematic scripts to revolutionizing marketing with personalized visuals tailored to specific narratives. Amidst this wave of innovation, much of the foundational groundwork for generative modeling owes its origins to Generative Adversarial Networks (GANs) [1], which pioneered the art of synthesizing realistic images.

GANs, with their adversarial framework of competing networks—a generator creating synthetic images and a discriminator evaluating their realism—ushered in a new era of image generation. Their iterative refinement process empowered groundbreaking models such as StyleGAN [2], which introduced style-based architectures, enabling unparalleled control over image features like texture, color, and structure. StyleGAN set a gold standard for image-to-image generation, elevating the photorealism and flexibility of generated visuals. However, these advancements, while monumental, were fundamentally constrained by their exclusive focus on image-to-image tasks, lacking the ability to interpret and condition outputs on textual descriptions.

This gap gave rise to text-to-image GANs [3], which sought to bridge language and vision by extending the adversarial framework to incorporate linguistic context. Models like StackGAN [4] adopted a stacked architecture to synthesize high-resolution images conditioned on textual descriptions, marking a pivotal step forward. Despite these strides, text-to-image GANs remain encumbered by notable limitations. First, the stacked architectures in text-to-image GANs frequently lead to entanglements between generators operating at different image scales, complicating the training process and undermining stability. Second, many existing models depend heavily on auxiliary networks to ensure semantic consistency between text and images, which inherently limits their flexibility and supervision capabilities. Third, the computational demands of cross-modal attention mechanisms, commonly employed for text-to-image fusion, restrict their effectiveness to specific image scales, thereby reducing their adaptability to diverse scenarios [5]. These inherent challenges have spurred the exploration of alternative methodologies, paving the way for the development of more robust and semantically coherent generative models.

With the limitations of GANs prompting the search for alternative approaches, diffusion models [6] have emerged as a ground breaking solution in the field of generative modeling. These models leverage a probabilistic framework to iteratively denoise random noise into coherent data, producing high-fidelity and diverse outputs. Among the most notable diffusion models are Stable Diffusion [7] and Latent Diffusion [8], which have redefined the landscape of generative AI. Stable Diffusion operates in a high-dimensional space, refining images with exceptional detail and quality, while Latent Diffusion introduces computational efficiency by conducting the denoising process in a lower-dimensional latent space, significantly reducing resource requirements without compromising output quality.

Diffusion models offer several advantages over GANs, making them a preferred choice for many applications. Firstly, their training process is inherently stable, eliminating adversarial dynamics that often lead to instability and mode collapse in GANs. Secondly, diffusion models excel in capturing the full data distribution, enabling them to generate diverse and coher-

ent outputs even for complex or abstract inputs. Lastly, their iterative refinement process ensures that each stage contributes to the quality of the final output, resulting in visuals that rival or surpass those generated by GANs, especially in terms of fine-grained detail and realism.

Despite their strengths, diffusion models are not without limitations. Compared to attention-based autoregressive models like DALL·E [9], which use transformers to directly model the relationship between textual prompts and visual outputs, diffusion models can be computationally intensive due to their iterative generation process. Additionally, while diffusion models are excellent at producing high-quality visuals, they may lack the contextual depth and nuanced alignment with text that transformer-based models achieve through attention mechanisms. This trade-off between fidelity and semantic alignment underscores the need for continued innovation to bridge these gaps in generative AI.

In parallel to the rise of diffusion models, Transformer architectures, first introduced by Vaswani et al. (2017) [10], have revolutionized the processing of textual and visual information. These models leverage attention mechanisms to capture complex relationships between elements within a sequence, enabling a deep understanding of context and semantics. This innovation has proven transformative for generative tasks, including text-to-image generation, by effectively aligning language with visual data.

A pivotal development in this domain was the introduction of CLIP [11] (Contrastive Language–Image Pretraining) by OpenAI, which serves as a bridge between textual descriptions and visual features. CLIP employs a contrastive learning framework to jointly train text and image encoders, enabling it to map textual inputs to their corresponding visual counterparts in a shared latent space. This capability has significantly enhanced models' ability to interpret nuanced textual prompts and generate outputs that align closely with semantic intent.

Building on CLIP's success, autoregressive transformer-based models like DALL·E and its successor DALL·E 2 have pushed the boundaries of text-to-image synthesis. These models combine the power of transformer architectures with the semantic understanding provided by CLIP, allowing them to generate intricate and contextually rich visuals from detailed textual descriptions. By sequentially predicting visual tokens in a manner akin to language modeling, DALL·E achieves a high degree of semantic alignment and creative diversity in its outputs.

While diffusion models like Stable Diffusion and Imagen [12] excel in photorealistic rendering and fidelity, autoregressive models such as DALL·E bring a unique strength in handling the contextual richness and subtle nuances of complex text prompts. These complementary approaches illustrate the diverse pathways being explored in generative AI, with transformers and attention mechanisms continuing to play a crucial role in advancing the alignment of language and vision.

Despite the remarkable advancements in generative AI, a significant challenge persists: the explainability and interpretability of these models. Most current systems function as opaque black boxes, providing little insight into how textual prompts are transformed into visual outputs. This opacity becomes especially problematic when evaluating how these models capture and reflect the emotional tone of input text. Explainability is essential not only for fostering trust and transparency but also for ensuring reliability in applications ranging from creative industries to sensitive domains like healthcare. Without a clear understanding of the internal mechanisms, users and developers face difficulties in aligning outputs with expectations or identifying biases in emotional interpretations. Bridging this gap is crucial for advancing both the usability and ethical adoption of generative AI systems.

To address the critical challenge of understanding how text-to-image models process emotional cues embedded in textual inputs, my research adopts a focused approach using advanced interpretability techniques. Central to this study is the application of DAAM (Diffusion Attention Attribution Map) [13] to trace how specific emotional words in a prompt influence different regions of the generated image. This method provides unique insights into the cross-modal relationships between textual inputs and visual outputs, enabling a detailed examination of the model's behavior.

Traditional techniques like CAM (Class Activation Mapping) and Grad-CAM (Gradient-weighted Class Activation Mapping) [14] have been widely used to interpret classification models by identifying which parts of an input are most responsible for a particular decision. However, these methods are primarily designed for classification tasks and are not directly applicable to generative models, which involve a complex interplay between text and image generation. DAAM, on the other hand, leverages cross-attention mechanisms to analyze the alignment between words in a prompt and the generated image, offering a much-needed capability to explore the influence of individual words on the generation process.

By utilizing DAAM, this research investigates how much attention the model assigns to emotional words in a text prompt and how this attention translates to specific regions of the generated image. This approach not only reveals which parts of the image are influenced by emotional words but also provides a clear mapping of the interaction between text and visual output. Through this analysis, the study aims to shed light on the underlying processes of text-to-image models, contributing to a deeper understanding of how these systems interpret and reflect emotional cues in their outputs.

To evaluate the generated images and their corresponding attention maps, sentiment measures are employed alongside annotations provided by a large language model (LLM). Human annotation, while ideal for assessing nuanced sentiment and emotional alignment, is often costly and time-intensive. Leveraging an LLM as an annotator offers a practical alternative, allowing for scalable and efficient evaluation. Although LLMs cannot fully replace human annotators due to potential biases or limitations in understanding context, they can be relied upon partially to provide consistent and objective insights, making them a valuable tool in the analysis pipeline. This approach strikes a balance between accuracy and feasibility, enabling a robust assessment of the model's performance.

Ultimately, this research aims to bridge the gap in explainability by providing insights into how text-to-image and text-to-video generation models process and translate emotional

(a) Original Image

(b) Generated Image

Fig. 1: Comparision of the original and the generated images for the caption ”An exhilarating moment of freedom and love as Rose spreads her arms like wings on the ship’s bow, her spirit soaring in Jack’s steady embrace from behind, against the endless horizon of the open sea.”

aspects of textual inputs into visual outputs. By evaluating the extent to which these models capture emotional nuances and how accurately they generate corresponding images, this study seeks to contribute valuable knowledge to the development of more transparent, interpretable, and emotionally aware generative models.

II. Related Works

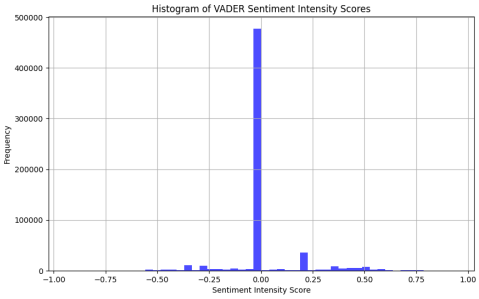
The integration of emotional understanding in text-to-image generation has been relatively underexplored, with most existing models focusing on capturing concrete attributes, actions, and adjectives from textual descriptions. However, recent research has begun to address this gap by investigating how generative models can represent and convey emotions in visual outputs.

Yang et al. introduced EmoGen [15], a framework designed to generate images that faithfully represent specified emotional categories. EmoGen constructs an emotion space and aligns it with the Contrastive Language-Image Pre-training (CLIP) space, providing a concrete interpretation of abstract emotions. The model employs attribute loss and emotion confidence mechanisms to ensure semantic diversity and emotional fidelity in the generated images. This approach outperforms state-of-the-art text-to-image methods in emotion accuracy, semantic clarity, and diversity.

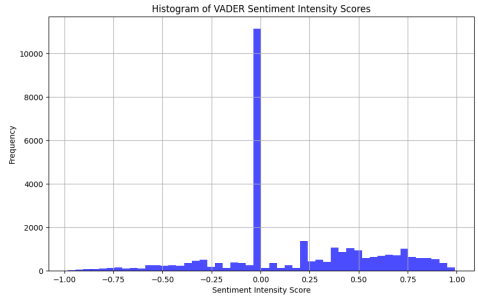
Weng et al. proposed the Affective Image Filter [16], a method that reflects emotions from text to images. This approach focuses on adjusting image attributes such as color and style to convey the desired emotional tone, addressing the limitations of fixed image content in effectively expressing emotions.

Paskaleva et al. developed a unified and interpretable emotion representation and expression generation framework [17]. Their method introduces a 3D numerical representation of emotions, allowing for fine-grained control over facial expressions in generated images. This model accommodates a broader range of emotions, including compound emotions, enhancing the expressiveness of generated visuals.

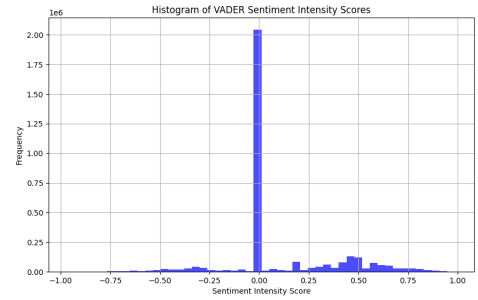
Despite these advancements, the field still lacks comprehensive studies focusing on how generative models attend to and represent emotional content in text-to-image generation. Most



(a) MS-COCO



(b) Flickr



(c) Conceptual Captions

Fig. 2: VADER Intensity Scores histograms for different Datasets

existing works have concentrated on attributes, actions, and adjectives, leaving the emotional dimension underexplored. This study aims to fill this gap by employing interpretability techniques to analyze how emotional language is translated into visual outputs, thereby contributing to a deeper understanding of how these models handle emotions.

III. Experimentation

A. Datasets

MS COCO [18], Flickr30k [19], Conceptual Captions [20], and LAION [21] are among the most widely used datasets for training text-to-image generation models. These datasets serve as benchmarks in the field, offering extensive collections of text-image pairs to train and evaluate the ability of models to align textual descriptions with visual outputs. Each dataset brings unique strengths but also shares limitations, particularly in their ability to capture emotional or sentiment-driven content.

	MS-COCO	Flickr	Conceptual Captions
No of datapoints chosen	331	331	667

TABLE I: Dataset Stats

The MS COCO dataset, widely used in computer vision and natural language processing, contains over 330,000 images annotated with five human-generated captions per image, resulting in approximately 1.5 million captions. These captions focus on objects, actions, and scene contexts, making MS COCO an essential resource for evaluating descriptive alignment in text-to-image models. However, its primary focus is on factual attributes and lacks annotations for emotions or sentiment, limiting its use for affective analysis.

The Flickr30k dataset comprises 31,783 images, each annotated with five human-generated captions, yielding approximately 158,915 captions. Compared to MS COCO, Flickr30k captions exhibit more linguistic richness and creativity, often capturing nuanced language that reflects subtle contexts. This makes Flickr30k valuable for analyzing text-to-image models' ability to handle varied language constructs. However, like MS COCO, Flickr30k does not include explicit emotional or sentiment annotations.

The Conceptual Captions dataset contains over 3 million image-caption pairs, making it one of the largest publicly available datasets for text-to-image tasks. Unlike MS COCO and Flickr30k, where captions are curated by humans, Conceptual Captions relies on automated pipelines to generate textual descriptions of web-sourced images. This large-scale approach provides diverse and generalizable data but often sacrifices precision and emotional depth in the captions, further underscoring the need for sentiment-rich datasets.

The LAION dataset is a groundbreaking resource for large-scale text-to-image generation, containing over 5 billion image-text pairs. This dataset is constructed using automated web scraping methods and refined using CLIP-based similarity measures to ensure alignment between text and images. Its sheer size and diversity make LAION the backbone for training state-of-the-art text-to-image models, including Stable Diffusion and DALL·E. The dataset spans a vast range of objects, scenes, and concepts, enabling models to generalize across varied prompts. However, LAION, like Conceptual Captions, primarily focuses on factual and descriptive alignment, offering little insight into emotional or sentiment-driven contexts. Moreover, its automated construction introduces potential noise, reducing the dataset's suitability for tasks requiring fine-grained emotional understanding.

Due to time and resource constraints, this research primarily focuses on the MS COCO, Flickr30k, and Conceptual Captions datasets. These datasets provide a foundational understanding of how generative models handle descriptive elements, such as attributes, actions, and contextual details. However, they are not designed to explore complex emotional or sentiment-driven inputs. Future work aims to address this limitation by incorporating datasets explicitly tailored for sentiment analysis, such as the ARTEMIS and Movie Description datasets.

The ARTEMIS [22] dataset (Affective Reasoning and Theory of Mind in Image Synthesis) focuses on emotional annotations for visual content, particularly in artworks. It contains approximately 80,000 images sourced from WikiArt, each labeled with emotional categories such as "joy," "sadness," or "anger," along with free-text explanations of the perceived emotion. ARTEMIS bridges the gap between visual aesthetics and emotional semantics, offering a unique resource for studying sentiment-rich text-to-image generation.

The Movie Description dataset [23] offers detailed textual descriptions of cinematic scenes, often imbued with narrative and emotional depth. These annotations frequently reflect sentiments such as tension, joy, or melancholy, tied to specific scenes. This dataset is particularly valuable for analyzing how generative models interpret and translate complex emotional narratives into visual outputs. Its rich contextual and emotional cues provide a robust foundation for studying sentiment-driven generative tasks.

By starting with MS COCO, Flickr30k, and Conceptual Captions, this research builds a foundation for understanding how generative models handle descriptive text. The future incorporation of sentiment-rich datasets like ARTEMIS and the Movie Description dataset will enable a deeper exploration of emotional representation, paving the way for the development of models that are both semantically aligned and emotionally intelligent.

To prepare the datasets for this study, captions from MS COCO, Flickr30k, and Conceptual Captions were filtered based on their sentiment intensity. This preprocessing step aimed to focus the analysis on captions with strong emotional tones, ensuring that the study targeted text-to-image models' ability to process and represent sentiment-rich textual inputs. The filtering process employed the VADER (Valence Aware Dictionary and sEntiment Reasoner) sentiment analysis tool, a lexicon and rule-based method widely recognized for its effectiveness in evaluating sentiment in short textual data, such as captions or tweets. By incorporating both lexical sentiment scores and contextual rules, VADER [24] provides nuanced evaluations of emotional content, making it suitable for this task.

VADER calculates a compound sentiment score for each text, ranging from -1 (most negative) to +1 (most positive). This compound score aggregates the positive, negative, and neutral sentiment components using weighted averages. It also applies contextual rules to adjust for nuances, such as the impact of negations (e.g., "not happy"), intensifiers (e.g., "very joyful"), or punctuation emphasis (e.g., "great!!!"). The resulting score offers a concise representation of the overall emotional tone and intensity of the caption.

For this study, captions were filtered based on the modulus of their compound sentiment score, denoted as $\|\mathrm{X}\|$, where X is the compound score of a caption. The modulus ensured that both highly positive and highly negative captions were prioritized, focusing on captions with pronounced emotional content regardless of polarity. Captions meeting the condition $\|\mathrm{X}\| > T$, where T is a predefined threshold, were included in the subset. This threshold was chosen empirically to ensure that captions with meaningful emotional intensity were

retained, while neutral or weakly emotional captions were excluded.

For example, captions like "A child laughing joyfully in a park" with a compound score of 0.85 ($\|\mathrm{X}\| = 0.85$) or "A bleak and abandoned building in the rain" with a compound score of -0.76 ($\|\mathrm{X}\| = 0.76$) were included due to their high emotional intensity. Conversely, neutral captions such as "A car parked on the street" with a compound score of 0.05 ($\|\mathrm{X}\| = 0.05$) did not meet the threshold and were excluded from the analysis.

This filtering process significantly reduced the size of the datasets but ensured that the remaining captions were enriched with sentiment-rich content. For MS COCO and Flickr30k, 331 captions each were selected that met the threshold criteria. In Conceptual Captions, 667 captions were chosen due to the larger initial dataset size and diversity of captions with pronounced emotional tones. For instance, in MS COCO, many neutral descriptions of objects or actions, such as "A red apple on a table," were excluded. On the other hand, captions conveying stronger emotions, such as "A heartwarming family gathering around a festive table" or "A somber, lonely figure walking through a foggy forest," were retained.

By employing this sentiment-based filtering, the study emphasizes captions with strong emotional tones, enabling a targeted evaluation of generative models' ability to attend to and reflect sentiment in their visual outputs. This preprocessing step was critical in addressing the often-overlooked emotional dimension of text-to-image generation, setting the foundation for subsequent experiments and analysis. The curated subsets, consisting of 331 captions from MS COCO, 331 captions from Flickr30k, and 667 captions from Conceptual Captions, form the core of the study and provide a robust basis for exploring sentiment representation in generative AI systems.

B. Method

The primary objective of this study is to analyze how text-to-image models attend to emotional cues embedded in textual inputs and how these cues influence the generated visuals. To achieve this, captions from the curated datasets were processed using the Stable Diffusion model, a state-of-the-art text-to-image generation framework. For each caption, the corresponding image was generated, and attention maps for specific emotional words in the captions were extracted using DAAM (Diffusion Attention Attribution Map). This approach enabled a detailed examination of how individual words, particularly those carrying emotional weight, influenced specific regions of the generated images.

Stable Diffusion is a diffusion-based generative model that iteratively refines random noise into coherent images conditioned on textual prompts. Unlike traditional GANs, which rely on adversarial frameworks, Stable Diffusion leverages a probabilistic framework that progressively denoises a latent representation of the image. During this iterative process, cross-attention layers play a crucial role in aligning textual descriptions with corresponding visual elements. These layers provide the foundation for DAAM, which interprets the attention dynamics between text and image.

DAAM (Diffusion Attention Attribution Map) is an interpretability tool specifically designed for diffusion-based models like Stable Diffusion. Unlike traditional interpretability techniques such as CAM or Grad-CAM, which work on static feature maps in classification tasks, DAAM leverages the iterative and cross-modal nature of diffusion models. The mathematical formulation of DAAM is as follows.

At each denoising step t, the cross-attention mechanism aligns textual tokens with latent image representations. Let the textual input be tokenized into N tokens, represented as $\mathbf{T} = \{t_1, t_2, \ldots, t_N\}$, and let the latent image representation at step t be $\mathbf{Z}_t$. The cross-attention weights are computed as:

$$\mathbf{A}_t = \text{softmax}\left(\frac{\mathbf{Q}_t \mathbf{K}_t^\top}{\sqrt{d_k}}\right),$$

where: - $\mathbf{Q}_t \in R^{H \times W \times d_k}$ is the query matrix derived from the latent image features. - $\mathbf{K}_t \in R^{N \times d_k}$ is the key matrix derived from the textual tokens. - d_k is the dimensionality of the key/query vectors. - $\mathbf{A}_t \in R^{H \times W \times N}$ is the cross-attention map that relates each token to spatial regions of the latent image.

For a specific word t_i, the attention scores across all spatial locations are extracted from $\mathbf{A}_t$:

$$\mathbf{A}_t^{(i)} = \mathbf{A}_t[:, :, i],$$

where $\mathbf{A}_t^{(i)} \in R^{H \times W}$ represents the attention map for the token t_i at step t.

To aggregate the influence of t_i across all diffusion steps t, the attention maps are averaged:

$$\mathbf{A}^{(i)} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{A}_t^{(i)},$$

where T is the total number of diffusion steps.

Finally, the attention map $\mathbf{A}^{(i)}$ is normalized to produce the Diffusion Attention Attribution Map (DAAM):

$$\mathbf{M}^{(i)} = \frac{\mathbf{A}^{(i)}}{\sum_{h=1}^{H} \sum_{w=1}^{W} \mathbf{A}^{(i)}[h, w]}.$$

Here, $\mathbf{M}^{(i)} \in R^{H \times W}$ represents the normalized influence of the token t_i on the generated image's spatial regions, ensuring that the map reflects the proportional contribution of t_i.

Using DAAM, attention maps were generated for key emotional words in the captions. For example, in the caption "A child laughing joyfully in a park," the word "joyfully" was identified as the primary emotional cue. DAAM produced an attention map highlighting the regions of the generated image most influenced by "joyfully," providing a visual representation of the model's interpretive process.

This mathematical framework ensures a rigorous and interpretable mapping of textual tokens to image regions, offering insights into how emotional language is processed by Stable Diffusion during the text-to-image generation process.

To identify the most significant emotional word in a caption, the study employs a masking-based approach. This method systematically evaluates the contribution of each word to the overall sentiment classification by temporarily replacing

The images and heatmaps generated for the captions ”A happy, beautiful bride getting out of a luxury car.” and ”A unhappy, beautiful bride getting out of a luxury car.” . Here the word happy is replaced by unhappy by the algorithm.

The images and heatmaps generated for the captions ”An older South American woman is knitting colorful hand warmers and enjoying herself.” and ”An older South American woman is knitting colorful hand warmers and suffer herself”. here the word enjoying is replaced by suffering.

Fig. 3: Working of the evaluation algorithm

individual words with a special ‘[MASK]‘ token and observing the impact on the classifier’s predictions. The algorithm first tokenizes the input sentence and computes a baseline confidence score for the predicted label using a pre-trained sentiment classification model.

Each word in the sentence is then iteratively replaced with the ‘[MASK]‘ token, and the model’s output probabilities for the predicted label are recalculated. The drop in confidence for the predicted label is recorded, and the word causing the maximum drop is identified as the most significant emotional word. For example, in the sentence ”A child laughing joyfully in a park,” masking ”joyfully” is likely to cause the largest drop, identifying it as the emotional focus.

This technique ensures precise identification of the word most influential to the sentiment of the caption, which is subsequently used for attention map extraction. The implementation manages computational efficiency by explicitly releasing memory during processing, enabling effective handling of large datasets.

After identifying the most significant emotional word in a given caption, the next step involves using this word and the caption as input to DAAM (Diffusion Attention Attribution Map) to generate an image and the corresponding heatmap. The heatmap reveals the specific regions of the generated image influenced by the emotional word, providing insights into how the text-to-image model attends to and integrates emotional cues during the generation process. To evaluate whether the identified word is being attended to appropriately, the caption is modified by replacing the most significant word with its antonym, and the process is repeated to analyze the model’s behavior.

For instance, consider the caption ”A calm and serene forest covered in snow.” Here, the word ”serene” might be identified as the most significant emotional word, as it contributes to the peaceful and tranquil atmosphere described in the caption. This caption is first processed by the Stable Diffusion model via DAAM to generate an image of the forest and an attention map for the word ”serene.” The heatmap would likely highlight regions such as the soft snow, gently falling flakes, or still trees, emphasizing the elements of the image that align with serenity.

Next, the word ”serene” is replaced with its antonym using the NLTK library, producing the modified caption ”A calm and chaotic forest covered in snow.” This introduces a contrasting emotional tone, which should result in corresponding changes in the generated image. The modified caption is processed to generate a new image and its heatmap for the word ”chaotic.” Ideally, this new image would reflect a more dynamic or

	LLM Annotation	Sent Classification	Heatmap Similarity
MS-COCO	70.63	3.6	71.6
Flickr	73.11	8.1	72.1
Conceptual Captions	72.19	4.4	75.8

TABLE II: Accuracy results of various evaluation methods

unsettling scene, such as windswept snow, broken branches, or a more cluttered forest landscape. The heatmap for "chaotic" should similarly focus on elements in the image that convey the sense of disorder introduced by the antonym.

By comparing the images and heatmaps generated for the original and modified captions, this method evaluates whether the model captures and represents the semantic and emotional differences between the two words. If the model effectively attends to the emotional words, the heatmaps should focus on similar regions of the image (e.g., snow and trees) but reflect changes in how these regions are depicted to align with the emotional context. For example, "serene" might emphasize soft and tranquil aspects, while "chaotic" highlights more dynamic or disruptive features.

The expectation is that replacing an emotional word with its antonym should alter the sentiment of the generated image and lead to subtle yet meaningful variations in the visual representation of entities within the scene. For instance, a serene forest might appear calm and harmonious, while a chaotic forest might appear disorganized and tumultuous. The results of this analysis, presented in the subsequent sections, demonstrate whether the model attends to and reflects these emotional shifts in its outputs, providing valuable insights into the capabilities and limitations of its attention mechanisms and its understanding of emotional language.

Algorithm 1 Generating Images and Heat map using DAAM for analysis

Caption $\mathcal{C}$, Pre-trained Text-to-Image Model $\mathcal{M}$, DAAM Framework, Sentiment Classifier $\mathcal{S}$, Tokenizer $\mathcal{T}$, Threshold T Generated Images $\mathcal{I}_1, \mathcal{I}_2$, Heatmaps $\mathcal{H}_1, \mathcal{H}_2$

Step 1: Identify the Most Significant Emotional Word.

Tokenize $\mathcal{C}$ using $\mathcal{T}$ to obtain tokens $\{t_1, t_2, \ldots, t_n\}$.

Compute the original classification score p_{orig} for $\mathcal{C}$ using $\mathcal{S}$.

for each token t_i in $\{t_1, t_2, \ldots, t_n\}$ **do** Replace t_i with `[MASK]` in $\mathcal{C}$.

Compute new classification score $p_{\text{mask},i}$ for the masked caption.

Calculate the drop in confidence $\Delta p_i = p_{\text{orig}} - p_{\text{mask},i}$.

Identify $t_k = \arg\max_i \Delta p_i$, the most significant emotional word.

Step 2: Generate Image and Heatmap for the Original Caption.

Input $\mathcal{C}$ and t_k to $\mathcal{M}$ using DAAM.

Generate the image $\mathcal{I}_1$ and heatmap $\mathcal{H}_1$ for t_k.

Step 3: Replace the Emotional Word with its Antonym.

Use the NLTK library to find the antonym t_k^{antonym} of t_k.

Replace t_k with t_k^{antonym} in $\mathcal{C}$ to form the modified caption $\mathcal{C}_{\text{antonym}}$.

Step 4: Generate Image and Heatmap for the Modified Caption.

Input $\mathcal{C}_{\text{antonym}}$ and t_k^{antonym} to $\mathcal{M}$ using DAAM.

Generate the image $\mathcal{I}_2$ and heatmap $\mathcal{H}_2$ for t_k^{antonym}.

Step 5: Analyze and Compare Results.

Compare $\mathcal{I}_1$ and $\mathcal{I}_2$ to assess sentiment change in the generated images.

Compare $\mathcal{H}_1$ and $\mathcal{H}_2$ to evaluate attention shifts for t_k and t_k^{antonym}.

return $\mathcal{I}_1, \mathcal{I}_2, \mathcal{H}_1, \mathcal{H}_2$.

IV. Evaluation

To evaluate whether text-to-image models effectively attend to and represent emotional tones embedded in textual prompts, one of the methods employed involved using a Large Language Model (LLM) as an annotator. This approach leveraged the interpretive and reasoning capabilities of the LLM to assess the alignment between the textual input and the corresponding generated image. Specifically, the LLM was provided with the text prompt and the generated image, and it was tasked with determining whether the image matched the given textual description.

The evaluation process was structured to elicit a clear and concise response from the LLM. For each input pair, the LLM was required to output a simple "Yes" or "No" answer, accompanied by a brief explanation of its reasoning. This reasoning typically described whether the elements depicted in the image corresponded to those described in the text, including attributes, actions, and other contextual details. For example, an image corresponding to the caption "A child smiling joyfully in a park" might be evaluated based on the depiction of a child, the act of smiling, and the park environment, along with any indication of joyfulness in the image.

While this method provides an automated and scalable approach to evaluating the generated images, it has significant limitations when used to assess emotional alignment specifically. The LLM may focus on overall alignment with the prompt, including descriptive elements like objects and actions, rather than exclusively considering the emotional tone conveyed by the most significant word in the caption. Furthermore, the LLM was not explicitly instructed to prioritize or isolate its evaluation of the emotional content, which could lead to responses that weigh attributes and contextual alignment more heavily than emotional accuracy.

Despite these limitations, using the LLM as an annotator offers a valuable initial metric for evaluating alignment between text prompts and generated images. It provides a broader perspective on whether the model captures the essence of the text but must be supplemented with additional evaluation methods to specifically address the representation of emotional tones in the generated outputs. This ensures a more comprehensive assessment of the model's ability to attend to emotional cues in the input text.

To further evaluate whether text-to-image models effectively represent emotional tones, the second method employs emo-

tion classification as a metric. This approach focuses on determining whether the emotional content of the caption aligns with the emotional tone conveyed by the generated image. By using emotion classification models, this method provides a quantitative measure of emotional consistency between the text and the corresponding visual output.

The evaluation begins by passing the caption through a pre-trained emotion classification model designed to analyze textual data. This model assigns an emotional label to the caption, such as ”joy,” ”sadness,” ”anger,” or ”neutral,” based on the emotional tone conveyed by the text. Simultaneously, the generated image is given as input to a similar emotion classification model trained on visual data, which predicts an emotional label for the image based on its visual features.

The core task is to compare the emotional labels assigned to the caption and the generated image. If both the caption and the image are classified with the same emotional label, the alignment is considered successful, indicating that the text-to-image model has effectively captured and represented the emotional tone of the caption in the generated image. For example, if the caption ”A joyful child playing in a sunny park” is assigned the label ”joy,” and the corresponding image of a child smiling in a bright park is also classified as ”joy,” the model successfully aligns emotionally.

This method provides a structured and automated way to evaluate emotional consistency across modalities. However, its reliability depends on the performance of the emotion classification models used. These models may have limitations in accurately predicting nuanced emotions or interpreting complex emotional cues, especially in generated images with subtle or ambiguous emotional content. Additionally, discrepancies in the way emotions are encoded in text and images could lead to misalignment even when the text-to-image model performs correctly.

Despite these challenges, emotion classification offers a robust framework for assessing emotional alignment. By focusing on directly measurable emotional labels, this method complements other evaluation metrics and contributes to a more comprehensive understanding of the model’s ability to attend to and reflect emotional tones in the text.

The third method evaluates the alignment of emotional attention in text-to-image models by analyzing the heatmap similarity between an original caption and its counterpart where the emotional word is replaced with its opposite. This approach provides insights into whether the model attends to appropriate regions of the image when generating visual representations of contrasting emotions while maintaining subtle contextual consistency.

In this process, the heatmaps for the original caption and the caption with the opposite emotional word are first generated using the DAAM (Diffusion Attention Attribution Map) method. These heatmaps highlight the regions of the image that the model attends to most strongly in response to the specified emotional word. The contours of the highest attention regions from both heatmaps are then extracted, and corresponding patches from the generated images are identified based on these contours.

Next, the similarity between the patches from the original and opposite-caption heatmaps is computed. This similarity is evaluated using a predefined range: if the similarity falls between 0.7 and 0.9, the model is considered to have performed well. The underlying hypothesis is that, when an emotional word is replaced with its opposite, the model will attend to similar regions of the image—indicating it understands the context—but will make subtle adjustments to reflect the change in emotion. Thus, the images and their attended regions are expected to exhibit a degree of similarity that is neither overly identical nor excessively dissimilar.

For instance, if the caption ”A serene lake at sunset” produces a heatmap focusing on the calm water and sky, replacing ”serene” with ”chaotic” might generate a heatmap that still focuses on the water and sky but highlights turbulent waves or stormy clouds. The patches extracted from these regions would reflect the intended emotional shift while preserving the broader scene context. The similarity score of the patches in this case would ideally fall within the defined range, demonstrating that the model has appropriately balanced its attention.

This method provides a quantitative way to assess the nuanced adjustments made by the model in response to emotional shifts in text. By focusing on heatmap similarity, it evaluates the model’s ability to attend to relevant regions consistently while capturing the intended emotional transformation. This approach not only measures the accuracy of emotional representation but also sheds light on the interpretability and robustness of the model’s attention mechanisms.

As observed from Table 2, text-to-image generation models demonstrate a moderate ability to attend to emotional words, yet definitive conclusions cannot be drawn due to inconsistencies across evaluation metrics. While LLM annotation shows an accuracy of 70%, this metric alone is insufficient as it relies on the LLM’s interpretive capabilities, which may not specifically focus on the emotional tone of the caption. LLMs often prioritize overall prompt alignment, such as attributes or actions, rather than isolating the significance of emotional words. This limitation underscores the need for human evaluation to provide a more nuanced and reliable assessment of the emotional alignment in generated images.

For the sentiment classification task, the accuracy is considerably lower, indicating potential misalignment between the emotional labels assigned to the caption and the image. This further emphasizes the inadequacy of relying solely on LLM-based annotation for evaluating emotional representation in text-to-image models.

The heatmap similarity metric provides more promising results, as it reveals that attention is often appropriately directed to relevant regions of the image during the generation process. However, while this method captures minute adjustments in attention when an emotional word is replaced with its opposite, it lacks a robust mechanism to confirm that the heatmap is attending to the correct regions in an interpretable manner. A more refined methodology is needed to ensure that the heatmaps not only demonstrate consistency but also accurately represent the intended emotional focus of the text prompt.

V. Conclusions and Future Direction

This study investigates the interpretability of text-to-image generation models, focusing on their ability to attend to emotional words in textual prompts. The methodology, comprising LLM annotation, sentiment classification, and heatmap similarity analysis, provides a structured framework for evaluation. While LLM annotation achieved 70% accuracy in identifying alignment, it has limitations in isolating emotional tones, often conflating them with attributes or actions. Sentiment classification further highlighted this challenge, showing low accuracy in matching emotional labels of captions and images. Heatmap similarity analysis offered promising results, suggesting that models often attend to appropriate regions, but it lacks a robust mechanism to confirm that the emotional focus is accurately represented.

Future work should expand the evaluation to broader datasets, including those with rich emotional contexts, such as ArtEmis and Movie Description datasets, to better assess emotional alignment. Incorporating human evaluation is crucial to provide nuanced and reliable evidence of the model's performance. Additionally, developing advanced evaluation techniques to isolate emotional aspects and refine the analysis of attention mechanisms will be essential. These steps can enhance our understanding of how text-to-image models process and represent emotions, improving their interpretability and reliability.

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[2] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[3] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *International conference on machine learning*. PMLR, 2016, pp. 1060–1069.

[4] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5907–5915.

[5] M. Tao, H. Tang, F. Wu, X.-Y. Jing, B.-K. Bao, and C. Xu, "Df-gan: A simple and effective baseline for text-to-image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 515–16 525.

[6] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[7] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[8] ——, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[9] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[10] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[12] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36 479–36 494, 2022.

[13] R. Tang, L. Liu, A. Pandey, Z. Jiang, G. Yang, K. Kumar, P. Stenetorp, J. Lin, and F. Ture, "What the daam: Interpreting stable diffusion using cross attention," *arXiv preprint arXiv:2210.04885*, 2022.

[14] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[15] J. Yang, J. Feng, and H. Huang, "Emogen: Emotional image content generation with text-to-image diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6358–6368.

[16] S. Weng, P. Zhang, Z. Chang, X. Wang, S. Li, and B. Shi, "Affective image filter: Reflecting emotions from text to images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 10 810–10 819.

[17] R. Paskaleva, M. Holubakha, A. Ilic, S. Motamed, L. Van Gool, and D. Paudel, "A unified and interpretable emotion representation and expression generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2447–2456.

[18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[19] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2641–2649.

[20] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of ACL*, 2018.

[21] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 278–25 294, 2022.

[22] P. Achlioptas, M. Ovsjanikov, K. Haydarov, M. Elhoseiny, and L. J. Guibas, "Artemis: Affective language for visual art," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 569–11 579.

[23] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele, "A dataset for movie description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3202–3212.

[24] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the international AAAI conference on web and social media*, vol. 8, no. 1, 2014, pp. 216–225.