Journal of Artificial Intelligence and Data Science (JAIDA)
e-ISSN: 2791-8335
https://dergipark.org.tr/pub/jaida

Vol. 5, No. 1, pp. 53-62, 2025
Research Article
Received: 31/05/2025 Accepted: 23/06/2025

Exploring Semantic Consistency in Generative Artificial Intelligence via Text-to-Image and Image-to-Text Transformation

Onur DOĞAN [1], Almila ALTINTAŞ [1], Buse YÜCETÜRK [1], Doğa AYDIN [1],
Fatih SOYGAZİ [1*], Yılmaz KILIÇASLAN [2]

Abstract
Recent advancements in artificial intelligence (AI) have brought Generative AI models dealing with Text-to-Image and Image-to-Text transformation to the forefront. While these models offer significant potential, their effectiveness hinges on the proper utilization of prompts – user-provided instructions guiding the generation process. It is crucial to ascertain the success of images generated through prompt input. In this context, following the text-to-image generation process, the creation of descriptive text for the produced image is also of significant importance. Thus, by comparing the input prompt and the output text, it becomes possible to determine the degree of success of the generatively produced image. This study discusses the semantic consistency between natural language used by humans and prompt language used by Generative AI models. We propose a novel approach: a natural language text-to-image model generates an image, which is then described in text by an image-to-text model, and this text is subsequently used as a prompt. A comparison module then identifies the prompt and corresponding image pair from a pre-built database that has the highest similarity to a human-generated description. This approach aims to maximize the benefit from existing AI models and promote explainability – a crucial principle in AI. This solution addresses the problem of improving the AI model's ability to generate human-like descriptions and enhancing the process of evaluating the accuracy of these descriptions.
Keywords: Generative artificial intelligence; image-to-text models; natural language processing; prompt engineering; text-to-image models.

1. Introduction

The emergence of Generative AI models capable of creating images from text represents a significant advancement in computer vision and machine learning. These models integrate techniques from computer vision and natural language processing (NLP) to transform textual inputs into corresponding digital images. Despite these advancements, users have increasingly noticed the limitations of these models in practical applications. Nonetheless, we are still far from being able to say that the limited capacity of text-to-image and image-to-text models is fully realized.

There are various successful models for text-to-image (e.g., Stable Diffusion [1], DALL-E [2]) and image-to-text (e.g., BLIP [3]) tasks. However, validating the semantics of the generated images remains challenging without human intervention. Therefore, it is essential to ensure that the input prompt provided to a text-to-image model and the output text generated by an image-to-text model are consistent in order to evaluate the performance of the generated image. This manuscript focuses on evaluating the consistency between the validated input prompt and the output text.

This study aims to bridge the gap between human language and machine-interpreted prompts by exploring the interaction between generative AI models and both textual and visual modalities. We propose a self-referential multimodal model in which an image is generated from an initial natural language prompt, followed by an automatic textual description of the generated image. This new description is then reused as a prompt for subsequent image generation. The iterative feedback contributes to understanding how well semantic consistency

*Corresponding author

Onur DOĞAN; Aydın Adnan Menderes University, Faculty of Engineering, Department of Computer Engineering, Aydın, Türkiye, email: 191805029@stu.adu.edu.tr; 0009-0001-5083-0163
Almila ALTINTAŞ; Aydın Adnan Menderes University, Faculty of Engineering, Department of Computer Engineering, Aydın, Türkiye, email: 191805009@stu.adu.edu.tr; 0009-0008-7955-3789
Buse YÜCETÜRK; Aydın Adnan Menderes University, Faculty of Engineering, Department of Computer Engineering, Aydın, Türkiye, email: 191805044@stu.adu.edu.tr; 0009-0003-3078-4352
Doğa AYDIN; Aydın Adnan Menderes University, Faculty of Engineering, Department of Computer Engineering, Aydın, Türkiye, email: 191805026@stu.adu.edu.tr; 0009-0000-7782-0830
Fatih SOYGAZİ*; Aydın Adnan Menderes University, Faculty of Engineering, Department of Computer Engineering, Aydın, Türkiye, email: fatih.soygazi@adu.edu.tr; 0000-0001-8426-2283
Yılmaz KILIÇASLAN; Mudanya University, Faculty of Engineering, Architecture and Design, Dpt. of Comp. Engineering, Mudanya, Bursa, Turkey, email: yilmaz.kilicaslan@mudanya.edu.tr; 0000-0002-5020-6547

can be preserved across modalities by measuring the similarity between human-generated descriptions and machine prompts to enhance prompt engineering. This will assist users to improve interactions with image generation systems.

2. Related Work

The rapid advancements in both NLP and image generation have fueled a surge of research exploring the connection between visual and textual content. Studies integrating these fields have become increasingly prevalent. The relationship between visual and textual information has been investigated in previous studies using various methodologies. For instance, Reed et al. [4] successfully generated meaningful and visually coherent images from textual descriptions using deep learning models. This research is considered a significant milestone in understanding the connection between textual and visual content. This study focused on text-based image generation using Generative Adversarial Networks (GANs) and is regarded as an important resource for investigating the relationship between text and images.

Kiros et al. [5] developed a model that combines visual and linguistic data using multimodal deep learning techniques to learn the connection between visual and textual data. Their research aimed to create a more comprehensive and holistic representation by integrating visual and textual elements.

Lin et al. [6] focused on providing word-level explanations to analyze biases in text-to-image models. The authors developed methods to elucidate how specific words in the text input contribute to particular visual features in the model's output. This approach enables the identification of words that may lead the model to make biased decisions. The primary aim of this study is to develop visual-semantic embedding methods by combining images and texts in a common representation space, allowing for the creation of more robust and meaningful models that utilize both visual and textual information. Bias in text-to-image models can stem from imbalances in the training datasets and may manifest in the models' outputs as gender, race, or other social biases. This study is important for detecting and correcting such biases in text-to-image models.

3. Methodology

Our study examines the connection between visual and textual information using a novel methodology. Firstly, it utilizes the API of Stable Diffusion [1], which generates images from textual expressions. We then retrieve textual descriptions of these images using the Salesforce BLIP API [3], which is an AI model that generates text from photographs. Finally, we determine the image with the highest similarity ratio by comparing the sentences generated for visual material created from the user's initial input. This process allows us to measure the similarity ratio between human language and machine language. This method provides a deeper insight into the interaction between textual and visual content by combining the strengths of several AI models to offer a more comprehensive solution. By leveraging both text-to-image and image-to-text models, our approach offers a unique perspective on the bidirectional relationship between linguistic and visual representations in AI systems.

3.1. Problem definition

One notable issue is that models generating text from images, such as BLIP [3], sometimes fail to capture spatial details or semantic coherence, as shown in Figure 1. On the other hand, text-to-image models like Stable Diffusion [1] and Dall-E [2] can struggle with accurately understanding and interpreting text, as illustrated in Figure 2. The effectiveness of these models heavily depends on the prompts used—text-based instructions provided by users to guide the model's output. Basic or vague prompts, such as "dog" or "park," may result in undesired images, whereas detailed prompts like "a scene of a dog playing ball in a park" lead to more accurate and intended results.

In the context of Generative AI, a significant challenge is understanding and improving the interaction between visual and textual information. Users often find that existing models, while capable of generating images from text or creating textual descriptions from images, do not always produce results that accurately reflect the intended meaning. For instance, the images generated from textual prompts may not fully capture the user's intent, and the text descriptions derived from images may lack detail or coherence. A clear example of this limitation is observed when a prompt such as "a man brought his money to his friend on the bank and gave him the money" is provided to an image generation model (Figure 3). Due to the polysemy of the word "bank"—which can refer to a financial institution or a physical bench near a river or park—the model produced two different visual interpretations: one depicting a man handing over money inside a formal banking environment, and another showing a casual exchange of money on an outdoor bench. This demonstrates how language ambiguity, especially in homonyms, can lead to divergent and unintended visual outputs. Such cases highlight the importance of semantic grounding

and contextual disambiguation in multimodal AI systems. This issue highlights the need for a method that effectively bridges the gap between human language and machine-generated content, ensuring that AI systems can produce outputs that align more closely with user expectations.

Figure 1. *Sample Text-to-Image Model Output using Stable Diffusion*

Figure 2. *Sample Text-to-Image Model Output Using Dall-E*

Figure 3. *Sample Images Generated from a Vague Prompt*

3.2. Proposed approach

To address this problem, our study proposes a novel methodology that combines the strengths of multiple AI models to enhance the alignment between textual and visual information. The solution involves three key stages:

Text-to-image stage: We use the Stable Diffusion [1] model to generate images based on user-provided textual descriptions. This stage is crucial for converting textual input into visual content, as shown in Figure 4.

Image-to-text stage: The generated images are then processed using the Salesforce BLIP API [3], an AI model that produces textual descriptions from pictures. This step converts the visual content back into text, aiming to capture the image's details and context, as depicted in Figure 5.

Comparison stage: Finally, we use a comparison module to evaluate the similarity between the text generated by the image-to-text model and the original user prompt. This comparison assesses how well the image and its description match the initial text, providing insights into the effectiveness of the prompt and the generated content.

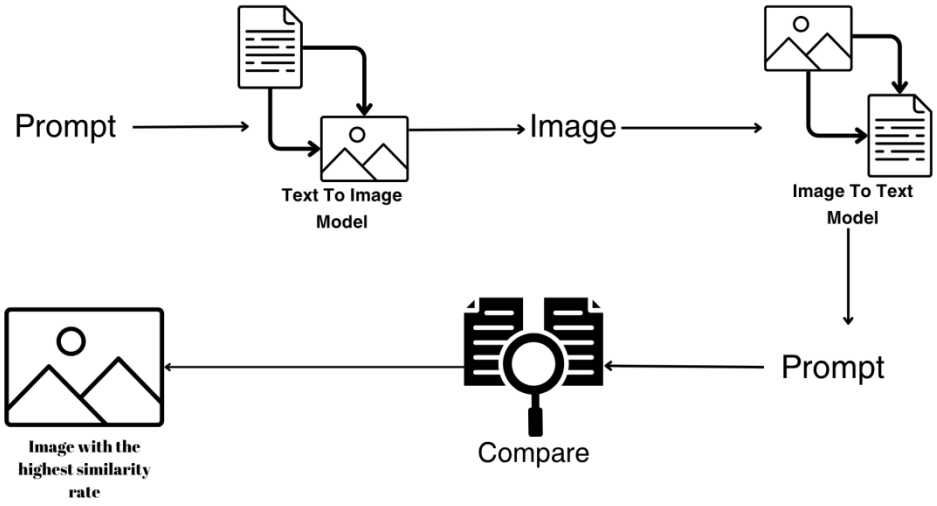

Figure 4. *Proposed Approach*

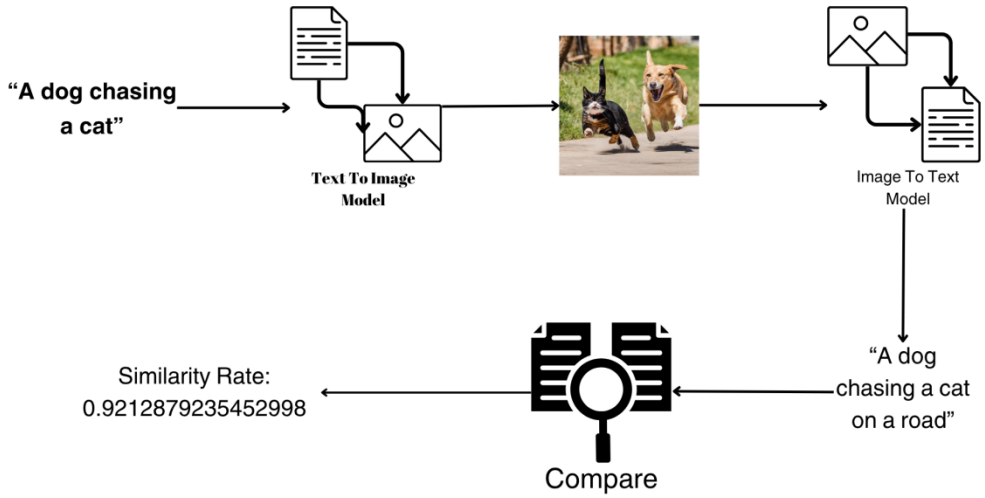

Figure 5. *Applied Workflow of the Proposed Approach*

Figure 4 and Figure 5 illustrate the workflow from text input to image generation and back to text output, highlighting how each stage contributes to improving the interaction between textual and visual representations.

3.2.1. Text-to-image

The "Text-to-Image" screen enables users to generate images from text descriptions using the Stable Diffusion [1] model, selected for its ability to produce high-quality, realistic images. This model operates through a noise reduction diffusion process involving forward and reverse propagation steps. During forward propagation, Gaussian noise is progressively added to the original image, effectively corrupting it. In reverse propagation, the model learns to remove this noise, reconstructing the image from the corrupted version [7]. This process helps the model map points in the latent space to realistic images, allowing for accurate image generation from textual prompts. Figure 6 and Figure 7 illustrate this diffusion process. Figure 6 shows the noise addition during forward propagation, while Figure 7 depicts the image reconstruction during reverse propagation, providing a visual representation of how the model transforms noisy images into clear, high-quality outputs.

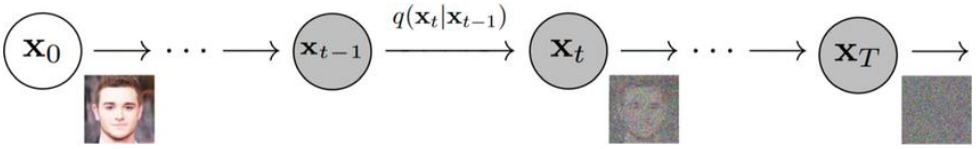

Figure 6. *Forward propagation process of diffusion models [8]*

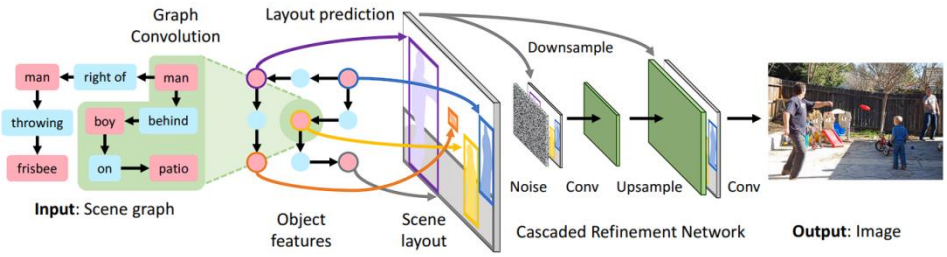

Figure 7. *Image reconstruction process [9]*

3.2.2. Image-to-text

HuggingFace platform offers several open-source methodologies under the name "Image-to-Text," which use image analysis to generate relevant textual descriptions. In this study, Salesforce's BLIP [3] model was used. This model can recognize objects, actions, and scenes in photographs and convert this information into text using NLP and image processing algorithms. At the core of the BLIP model, encoder-decoder based models are used [3]. The working principle of BLIP model is shown in Figure 8.

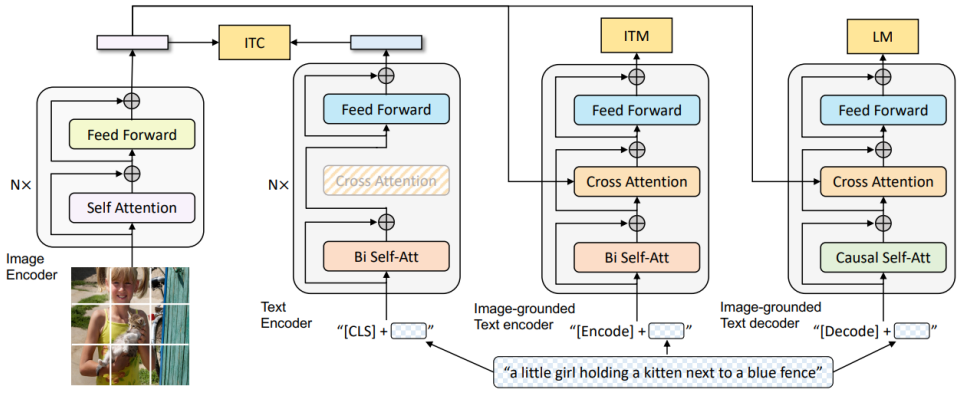

Figure 8. *BLIP Working Principle [3]*

In our study, the "Image-to-Text" module provides an interface for converting images into text. First, for example, the sentence "A dog is chasing a cat" is entered. In the second stage, on top of each image generated in the first stage, the user will see the prompt generated by the machine based on that image. Although prompts are generated by machine learning algorithms over images, the initial prompt is provided by humans. This comparison provides an opportunity to evaluate the differences between prompts given by machines and humans. Thus, a way has been created to see similarities and differences between prompts generated by AI and humans. Hence, the user can understand under what conditions different results are produced based on inconsistencies or differences between human comments and texts generated by AI.

3.2.3. Text similarity

Initially, we utilized the SpaCy[1] library to rigorously evaluate the similarity between the natural language prompts provided by humans and the prompts generated by AI models. SpaCy is one of the most widely used NLP libraries in open source and commercial applications. It is particularly known for being fast and efficient and is used in many NLP tasks such as text mining, information extraction, language model building and many more using deep learning methods. Similarity calculation based on word embeddings is used by SpaCy.

[1] https://spacy.io/

Similarity is a concept that expresses the relationship between vector dimensions representing the properties of two objects. Basically, the concept of similarity is used to measure how similar or different data objects are. In this paper, the similarity score calculated is based on the cosine similarity between the vector representations of the two texts. Cosine similarity measures the cosine of the angle between two vectors and provides a value between -1 and 1. A cosine similarity score of 1 indicates that the vectors are identical, while a score of 0 indicates that they are orthogonal (no similarity), and a score of -1 indicates that they are diametrically opposed.

In the similarity calculation stage, we calculate the similarity between two texts by transforming them into numerical representations (vectors) that capture their semantic meanings and contexts within a large corpus of text data. These vectors are derived from word embeddings, where words with similar meanings have vectors that are closer together in a multi-dimensional space.

4. Evaluation

In this study, we initially generated images from text (text-to-image) using model Stable Diffusion, and then retrieved textual descriptions of these images with model Salesforce's BLIP (image-to-text). Then we compared these AI-generated descriptions with the original human-provided prompts to assess how closely the AI's understanding aligns with human desire. The phases during evaluation are:

a) *Vectorization:* Each text (human prompt and AI-generated prompt) is tokenized and converted into a vector representation using SpaCy's pre-trained word embeddings.

b) *Cosine similarity calculation:* The cosine similarity between the vector representations of the human prompt *H* and the AI-generated prompt *A* is computed to obtain the similarity score *sim(H,A)*, which quantifies the degree of semantic similarity between the two prompts.

c) *Evaluation using similarity comparison:* Higher similarity scores indicate that the AI-generated prompt closely matches the human-provided prompt in terms of semantic content and context. Lower scores suggest discrepancies or differences in interpretation between the AI model and human users (Figure 9). Figure 9 represents that text-to-image model creates various number images and the image-to-text model translates each image to a machine generated text. Hence, the similarity score calculation assists us to find out the best fitted image of the prompt given by the human to text-to-image model.

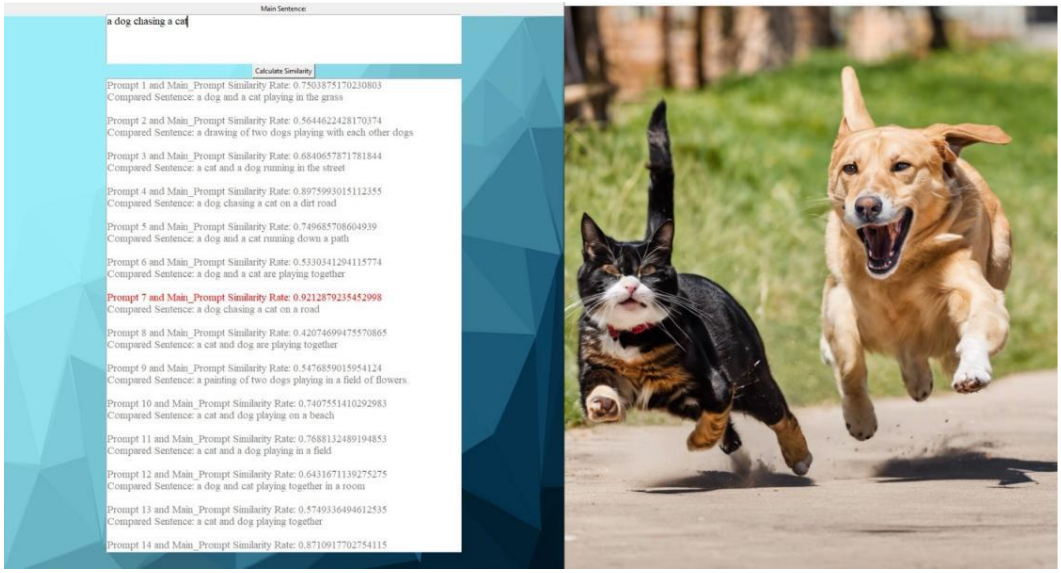

Figure 9. *User Interface of the Proposed Approach for Capturing the Best-Fitting Image based on a Human-Generated Prompt*

Figure 9 demonstrates that the prompt “a dog chasing a cat” generates numerous images. The prompt “a dog chasing a cat on a road” when evaluated using the text-to-image model, produced the image on the right, which is deemed the most appropriate. The comparison of the similarity scores provides valuable insights into the effectiveness of AI models in interpreting and generating content based on textual prompts. It helps in evaluating how well AI understands and replicates human requests from textual prompts to visual outputs. This evaluation procedure is crucial for improving AI model performance and enhancing user interaction with AI-driven systems.

Our study aims to elucidate the impact of prompt design and specificity on the accuracy and relevance of AI-generated outputs by checking their semantic consistency. Understanding the nuances lead to better guidelines for users to craft effective prompts and for developers to enhance AI models' capabilities in generating outputs that align closely with user expectations. This detailed comparison methodology not only contributes to advancing the field of generative AI but also underscores the importance of human-AI collaboration in refining and optimizing AI systems for practical applications.

While the similarity score provides an indication of how well the AI-generated content aligns with the user' prompt, it also offers insight into how deep neural networks capture semantic and contextual information to enable accurate generation. Therefore, in Section 5, we applied Explainable Artificial Intelligence (XAI) techniques to illustrate the models' behavior for the given prompts, aiming to provide a detailed technical understanding.

5. Transformer Explainer-based Prompt Interpretation

Transformer Explainer [10] is an interactive visualization tool designed to help non-experts understand Transformers [11]. Transformer Explainer aids users in comprehending complex Transformer concepts by providing an integrated model overview and facilitating seamless transitions across different levels of abstraction. It assists users to experiment with their own input and observe how the internal components and parameters of the Transformer collaborate to predict subsequent tokens.

Transformer models use token embedding and positional encoding to transform text into a numerical representation. In this stage, each word (token) is first converted into an embedding vector. This vector represents how the word is interpreted by the model. Positional encoding is then added to determine the order of each word in the sentence. This process allows the model to take into account not only the meaning of the words, but also their place in the sentence [11].

As shown in Figure 10, the model processes each word in the sentence "A dog chasing a cat on a road" by sorting them by position. The token IDs shown next to each word represent the numerical equivalent of that word in the model's language dictionary. For example, the word "A" is assigned an ID of 32, while the word "dog" is assigned an ID of 3290. The positional encoding indicates the order of the word in the sentence and allows the model to better decode the language meaning. This process allows the Stable Diffusion model to produce more meaningful and consistent results when generating images from text.

Figure 10. *Token Embeddings and Positional Encodings of Stable Diffusion model using Transformer Explainer*

In the Transformer model, Query, Key and Value (QKV) vectors are computed to make sense of each word's relationship with other words [11]. As shown in Figure 11, the model first transforms the text into numeric vectors (embedding) and then works with these vectors to perform the QKV calculations.

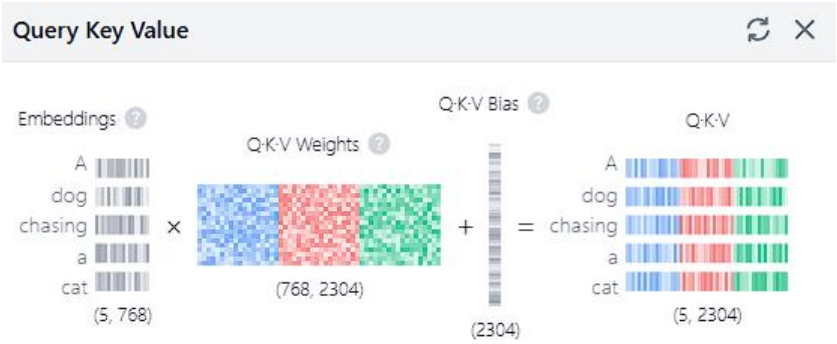

Figure 11. Demonstration of QKV Calculation for the Given Prompt

The embedding vector of each word is multiplied by the QKV weights and then a bias is added. As a result of these operations, the model learns how each word is related to other words. For example, in the sentence "A dog chasing a cat on a road", the relationship between the words "dog" and "cat" is strengthened by these QKV calculations. QKV computations are the most fundamental component that drives the attention mechanism of language models in particular. This process enables the model to produce more accurate results by determining the relationship of each word with other words. Especially in image-to-text models such as BLIP, these calculations play an important role in text generation from images.

The self-attention mechanism, one of the most important components of the Transformer model, defines the relationship of each word to other words. As shown in Figure 12, the model determines how much attention each word in the sentence "A dog chasing a cat on a road" pays to other words. This process helps the model to understand which words are focused on and which words have a stronger link between them.

Figure 12. Demonstration of Self-Attention Mechanism for the Given Prompt

The attention mechanism calculates the importance of the relationships between words, allowing the model to focus on the most critical words. For example, the words "dog" and "cat" have a stronger association, while a conjunction like "on" has a weaker association. In this context, the model is enabled to make sense of the text and identify which words are more connected to the image. This process helps the model to produce more accurate and meaningful results both when generating images from text and when extracting text from images.

During the self-attention mechanism, the attention level of each word to other words is calculated by various mathematical operations. As shown in Figure 13, the model first determines the correlations between words by performing a dot product operation. This measures how much attention each word pays to other words.

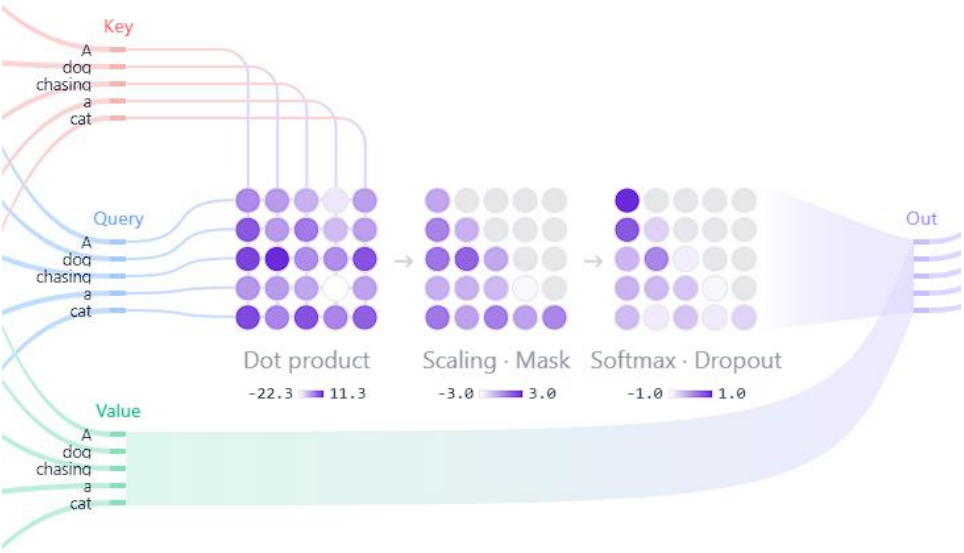

Figure 13. Internal Structure of the Self-Attention Mechanism

The values obtained after dot product operation are then processed by scaling and masking. Scaling stabilizes the attention values, while masking prevents the model from paying attention to future words. In the final stage, the softmax operation is used to distribute attention. This allows the model to evaluate the relationships between words in a more meaningful way. For example, in the sentence “A dog chasing a cat on a road”, we can see that there is a stronger attentional connection between the words “dog” and “chasing”.

When following the model’s internal structure with the given prompt, Transformer Explainer enhances semantic and contextual understanding. Additionally, the obtained similarity scores can be validated by assessing the similarities or proximity of the words within each prompt.

6. Discussion

This study highlights the critical role of prompt detail in the performance of text-to-image and image-to-text models. Specifically, we examine how models compensate for incomplete prompts by generating statistically probable completions and how these additions affect semantic similarity scores. For instance, when a user provides a simple prompt like "A dog chasing a cat" the text-to-image model may generate an image depicting "A dog chasing a cat on a road", a scenario derived from its training data. When this image is then fed into an image-to-text model and converted back to text, the model might return the prompt not only as "A dog chasing a cat" but possibly as "A dog chasing a cat on a road" based on its analysis of the image. This leads to decreases in the similarity score when the two sentences are compared using NLP tools such as Spacy, as the phrase "on a road" is not present in the first prompt.

Our findings suggest that detailed and precise prompts significantly enhance both the quality of generated images and the consistency of similarity scores. Users should aim for specificity in their inputs to minimize unintended model completions and improve output alignment. This finding highlights the importance of detailed prompts when using text-to-image and image-to-text models and suggests that future studies should investigate the impact of prompt design and elaboration on model performance in more depth.

While SpaCy offers a fast and efficient method for computing semantic similarity using word embeddings, its approach has notable constraints. Specifically, it struggles to capture nuanced contextual meanings, particularly in longer or more complex sentences. Unlike transformer-based models such as Sentence-BERT (SBERT) [12] or BERTScore [13], which leverage deep bidirectional attention to analyze entire sentence contexts, SpaCy relies on static word embeddings and averaging techniques. Consequently, it may miss subtle semantic distinctions, disregard word order, or inadequately represent syntactic structure. Future work could address these limitations by incorporating contextualized embedding models like SBERT, which may improve the accuracy of similarity assessments, especially when evaluating semantic alignment in generative AI outputs.

In addition to the similarity score-based comparison of the text-to-image and image-to text models, Transformer Explainer aids in understanding how the relatedness of words is captured in a deep learning model. It also provides a visual interpretation of the representation of these similarity scores.

7. Conclusion

This study presents a novel method for analyzing semantic consistency between textual and visual modalities by leveraging text-to-image (Stable Diffusion) and image-to-text (BLIP) models. Meaningful visual information is obtained from various words using the Stable Diffusion model to create visual content. Similarly, using BLIP model that generate text from images, meaningful language can be produced for visual information. Our findings demonstrate that prompt specificity significantly impacts output quality and similarity scores, as validated through cosine similarity metrics and Transformer Explainer-based interpretations. When similarity analysis is conducted using NLP model, it has been found that sentences generated for visual content exhibit a high similarity rate with the prompt provided by the user. This demonstrates the potential for improving the understanding and interpretation of textual and visual content through generative AI models. This potential improvement of the understandability of the models have been enhanced by interpreting the textual contents in each phase of the Transformer models using Transformer Explainer.

The findings of this paper can advance research in this field and contribute to the development of new techniques for enhancing the understanding of interactions between textual and visual content in newly developed models. To further strengthen the evaluation of our approach, future work will include benchmarking against standardized datasets (e.g., COCO [14] or Conceptual Captions [15]) to quantitatively compare performance with existing methods. Such experiments will provide deeper insights into the robustness and generalizability of our framework. These advancements will contribute to refining prompt engineering guidelines and enhancing the interpretability of generative AI systems.

Declaration of Interest

As authors, we declare that we have no conflict of interest with anyone related to our work.

References

[1] 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 (CVPR), pp. 10684-10695, 2022.

[2] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li ... and A. Ramesh, “Improving image generation with better captions”, Computer Science, 2(3), 8, https://cdn. openai. com/papers/dall-e-3.pdf, 2023.

[3] J. Li, D. Li, C. Xiong and S. Hoi, “Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation”, In International conference on machine learning, PMLR, pp. 12888-12900, 2022.

[4] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele and H. Lee, “Generative adversarial text to image synthesis”, In Proceedings of the 33rd International Conference on Machine Learning (ICML), PMLR, pp. 1060-1069, 2016.

[5] R. Kiros, R. Salakhutdinov and R. S. Zemel, “Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models”, arXiv preprint arXiv:1411.2539, 2014.

[6] A. Lin, L. Monteiro Paes, S. H. Tanneru, S. Srinivas and H. Lakkaraju, “Word-Level Explanations for Analyzing Bias in Text-to-Image Models”, arXiv preprint arXiv:2302.06578, 2023.

[7] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, and M. H. Yang, “Diffusion models: A comprehensive survey of methods and applications”. ACM Computing Surveys, 56(4), 1-39, 2023.

[8] J. Ho, A. Jain and P. Abbeel, “Denoising diffusion probabilistic models”, Advances in neural information processing systems, 33, 6840-6851, 2020.

[9] J. Johnson, A. Gupta and L. Fei-Fei, “Image generation from scene graphs.”, In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp. 1219-1228, 2018.

[10] A. Cho, G. C. Kim, A. Karpekov, A. Helbling, Z. J. Wang, S. Lee, ... and D. H. P. Chau, “Transformer Explainer: Interactive Learning of Text-Generative Models”, In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, no. 28, pp. 29625-29627, 2025.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez ... and I. Polosukhin, “Attention is all you need”. Advances in neural information processing systems, 30, 2017.

[12] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks”. arXiv preprint arXiv:1908.10084, 2019.

[13] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “Bertscore: Evaluating text generation with bert”. arXiv preprint arXiv:1904.09675, 2019.

[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, “Microsoft coco: Common objects in context”, In European Conference on Computer Vision (ECCV), pp. 740–755, Springer, 2014.

[15] 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 the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2556-2565, 2018.