ABSTRACT:

This work explores the integration of BLIP (Bootstrapped Language-Image Pre-training) and CLIP (Contrastive Language–Image Pretraining) to advance image-to-text generation and prompt optimization tasks. The framework is designed to align image features with meaningful textual prompts, leveraging the robust multimodal capabilities of both models. Using CLIP embeddings, the system refines prompts generated by BLIP by optimizing their semantic relevance while removing extraneous or less informative elements. This is particularly useful in applications such as prompt engineering for Stable Diffusion 2.0 and other text-to-image generative models.The proposed method involves a novel adaptation of CLIP Interrogator, which enhances the alignment between image features and textual descriptions, resulting in highly context-aware prompts. By refining the descriptive precision of generated prompts, the approach contributes to improving reverse engineering tasks, where the goal is to infer original prompts based on generated images. The integration demonstrates competitive performance on leaderboards, achieving a score of 0.45836, and highlights its potential for use in other vision language tasks.

The study underscores the importance of combining pre-trained vision-language models to solve complex multimodal challenges. The use of complementary architectures like BLIP for generating prompts and CLIP for refining semantic consistency ensures robust and efficient processing. This framework can be extended to other applications requiring bidirectional image

text understanding, such as automated captioning, content generation, and visual search.

Overall, the approach marks a significant step forward in leveraging state-of-the-art pre-trained models for prompt inversion and multimodal alignment. By combining the strengths of pre trained models like BLIP for generating prompts and CLIP for ensuring semantic coherence, the framework bridges a critical gap in multimodal AI. The proposed system not only improves the efficiency of generating high-quality prompts but also sets a foundation for further advancements in reverse-engineering generated media. With broad applicability across vision language domains, this work marks a significant milestone in leveraging state-of-the-art technologies for image-text alignment and retrieval tasks.

PROPOSED WORK:

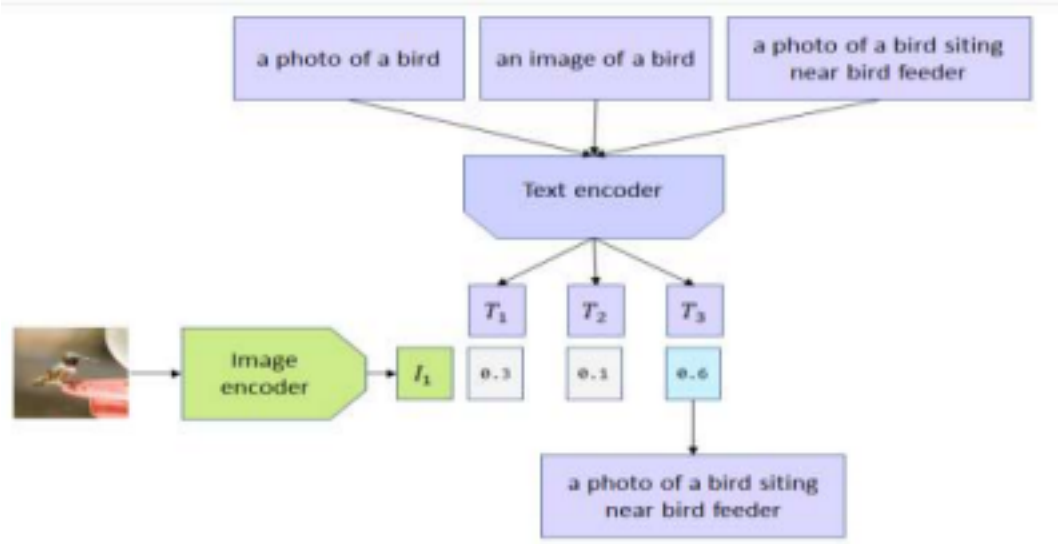

FIGURE 1- Architecture of CLIP model

This study proposes a novel framework to enhance image-to-prompt generation by integrating two state-of-the-art multimodal models, BLIP and CLIP. The proposed system aims to address challenges in reverse-engineering image-generation processes by optimizing semantic alignment between images and textual prompts.The methodology involves utilizing BLIP for generating initial descriptive prompts from visual inputs, capturing key contextual and semantic features. Subsequently, CLIP embeddings refine these prompts by enhancing alignment, eliminating extraneous elements, and ensuring semantic coherence with the corresponding images. A modified CLIP Interrogator mechanism is introduced to improve prompt generation quality further, facilitating robust image-prompt mapping.

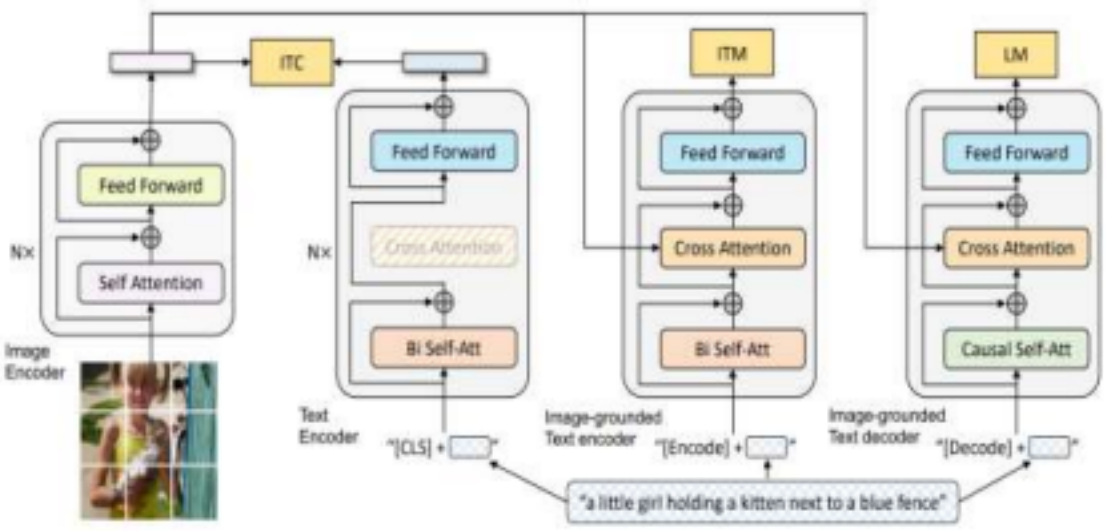

FIGURE 2- Architecture of BLIP model

The framework will be evaluated on publicly available datasets and benchmarked using standard metrics such as BLEU, METEOR, and CIDEr, ensuring the system's effectiveness in generating concise and meaningful prompts. The proposed work aims to achieve state-of-the-art performance in tasks like reverse-prompt generation, image captioning, and visual search.This system's innovative integration of BLIP and CLIP highlights the potential of combining complementary pre-trained vision-language models for robust and efficient multimodal applications. The study further explores potential extensions in tasks such as automated captioning and text-to-image synthesis, ensuring broader applicability in real-world scenarios.