© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716248

Postforge: A Unified, Locally Deployable Multimodal AI Platform for Social Media Content Generation

HARSHIT RASAM[1], AYUSH MAURYA[2], SHUBHAM SURYAWANSHI[3], SANGITA NIKUMBH[4]
[1,2,3]*Dept. of Artificial Intelligence and Data Science New Horizon Institute of Technology and Management Thane, India*

Abstract- ***The exponential growth of social media platforms has created a demand for high-quality, optimized content that exceeds the capacity of manual creation methods. While Arti¬ficial Intelligence (AI) offers a solution, current tools are often fragmented, requiring users to navigate multiple applications for image generation, captioning, and hashtag optimization. Furthermore, reliance on third-party APIs raises concerns regarding cost, latency, and data privacy. This paper presents PostForge, a unified web application designed to streamline social media content creation by integrating specialized AI models. Unlike monolithic multimodal models, PostForge em-ploys a modular architecture leveraging distinct state-of-the-art models for specific tasks: Latent Diffusion Models for text-to-image generation, BLIP for image captioning, and transformer-based models for hashtag generation. The system is designed for local deployment, ensuring data sovereignty and reducing operational costs. Experimental evaluation demonstrates the system's efficacy, achieving a ROUGE-L score of 0.457 and BLEU score of 0.041, validating the feasibility of a unified, privacy-centric approach to content automation.***

Index Terms- ***Multimodal AI, Content Generation, Natu¬ral Language Processing, Stable Diffusion, Image Captioning, social media.***

I. INTRODUCTION

In the contemporary digital landscape, social media has evolved from a communication utility into a critical driver of global commerce and personal branding. This shift has placed immense pressure on content creators and marketers to produce visually appealing and contextually relevant material at high velocity. While Artificial Intelli¬gence (AI) has advanced significantly in domains such as Computer Vision (CV) and Natural Language Processing (NLP), the workflow for content creation remains ineffi¬cient.

Currently, creators rely on a fragmented ecosystem of tools. A typical workflow might involve using one appli¬cation for image generation, a separate service for writing captions, and yet another for trend analysis. This lack of integration results in disjointed outputs, increased latency, and higher operational costs due to subscription fees for multiple APIs. Furthermore, the reliance on cloud-based third-party APIs introduces significant privacy risks, as proprietary images and draft content must be uploaded to external servers.

To address these challenges, this paper proposes post-forge, a unified web application that consolidates the content creation pipeline. PostForge distinguishes itself through a" best-of-breed" architectural approach. Rather than utilizing a single, generalized monolithic model that may lack depth in specific tasks, PostForge integrates specialized models optimized for individual components of the workflow.

The primary contributions of this work are:

- The design and implementation of a unified multi¬modal pipeline capable of Text-to-Image, Image-to-Caption, and Text-to-Hashtag generation.
- A locally deployable architecture that eliminates de¬pendency on external APIs, thereby ensuring data privacy and reducing costs.
- An evaluation of the system's performance using standard NLP metrics, demonstrating the viability of the proposed modular integration.

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716248

II. REVIEW OF LITERATURE

The development of PostForge draws upon recent ad¬vancements in three key areas of deep learning: image captioning, text-to-image synthesis, and hashtag genera¬tion.

A. Image Captioning

Image captioning requires bridging the semantic gap between visual features and natural language. Zhang et al. [1] proposed the Adaptive Semantic-Enhanced Trans¬former, which integrates visual and semantic features via a weakly supervised module. This approach yields captions that are contextually rich, a feature essential for social media engagement. Similarly, Zeng et al. [2] introduced the S2 Transformer, emphasizing spatial- and scale-awareness to preserve spatial information in generated text. These works inform the image-to-caption module of PostForge, ensuring that generated descriptions are both semantically accurate and spatially grounded.

B. Text-to-Image Synthesis

Generative Adversarial Networks (GANs) and Diffusion Models have revolutionized image synthesis. Zhang et al. [3] presented XMC-GAN, utilizing cross-modal contrastive learning to enhance semantic alignment between text prompts and generated images. While effective, the field has increasingly shifted towards Latent Diffusion Models (LDMs) for their superior fidelity and diversity. PostForge leverages these advancements by implementing diffusion-based architectures for its text-to-image module, ensuring high-quality visual outputs from textual prompts.

C. Hashtag Generation

Hashtags are critical for content discoverability. Yu et al. [4] approached this as a guided generation task, aug¬menting input features with trending signals to improve relevance. This contrasts with static keyword extraction methods. PostForge adapts this methodology by integrat¬ing NLP models that analyze caption context to generate hashtags that are not only relevant but also trend-aware.

III. SYSTEM ARCHITECTURE

PostForge follows a modular client-server architecture designed for extensibility and local deployment. The sys¬tem comprises three distinct layers: the User Interface, the Application Server, and the Model Layer.

A. Model Layer

The core intelligence of PostForge resides in the Model Layer, which hosts the pre-trained deep learning models.

- Text-to-Image Module: This module utilizes Latent Diffusion Models (specifically Stable Diffusion 1.5). The process begins with a text encoder (CLIP ViT-L/14) transforming the user's prompt into an em¬bedding. A U-Net then iteratively denoises a latent representation, which is subsequently decoded into a high-resolution pixel image.
- Image-to-Caption Module: We employ the BLIP (Bootstrapping Language-Image Pre-training) model. It uses a Vision Transformer (ViT) to extract visual features from an input image. These features are passed to a text decoder that generates a descriptive caption autoregressively.
- Hashtag Generation Module: This module utilizes transformer-based models (T5/BART) to refine cap¬tions and extract key terms. Term Frequency-Inverse Document Frequency (TF-IDF) is used alongside se¬mantic embeddings to map keywords to high-impact hashtags.
- Pipeline Automation and Workflow: Beyond individ¬ual model performance, the utility of PostForge lies in its ability to chain these models into coherent work¬flows. The system implements a pipeline architecture where the output of one module serves as the input for the next. For instance, a user can initiate a 'Text-to-Post' workflow where a prompt first generates an image via the Latent Diffusion module; this generated image is automatically passed to the BLIP captioning module, and the resulting text is subsequently ana¬lyzed for hashtag relevance. To mitigate the latency inherent in sequential inference, the pipeline

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716248

employs asynchronous task queuing, ensuring that the user interface remains responsive during the generation process.

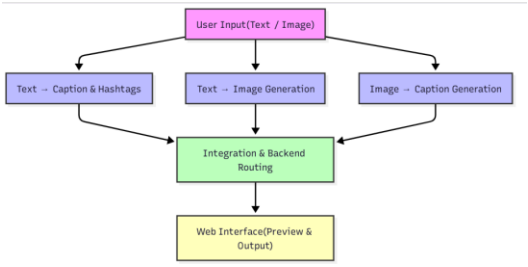

Fig. 1. Block Diagram

B. Application Layer

The backend is implemented using FastAPI, chosen for its asynchronous capabilities and high performance. It acts as an orchestrator, routing HTTP requests to the appro¬priate model endpoints. It manages the pipeline logic; for example, in the automated workflow, the output of the image generator is automatically passed to the caption generator, creating a seamless end-to-end experience.

C. Presentation Layer

The frontend is built using Gradio, providing an intu¬itive web interface. It handles multi-modal inputs (text and image uploads) and displays the generated outputs in a unified dashboard, abstracting the complexity of the underlying AI models from the user.

IV. EXPERIMENTAL SETUP AND RESULTS

A. Implementation Details

The system was implemented on a standard workstation to validate the feasibility of local deployment. The hard¬ware configuration included an Intel CPU, 16GB RAM, and an NVIDIA GPU with 8GB VRAM. The software stack utilized Python 3.10, PyTorch, and the Hugging Face Transformers and Diffusers libraries. Datasets from MS COCO and InstaCaptions were used for evaluation benchmarks.

B. Data Preprocessing and Tokenization

To ensure optimal performance across the diverse input modalities, specific preprocessing steps were employed. Input images for the BLIP and Diffusion models were resized to a standard resolution of 512x512 pixels and normalized using mean and standard deviation values corresponding to the ImageNet dataset. For textual in-puts, the tokenization process utilized the SentencePiece tokenizer associated with the T5 architecture, truncating sequences to a maximum length of 512 tokens to manage memory constraints on the local hardware. This standard¬ization allows for seamless interoperability between the text and image modules without requiring extensive model retraining.

C. Evaluation Metrics

To quantitatively assess the performance of the Image-to-Caption module, we employed standard Natural Lan¬guage Generation (NLG) metrics: BLEU (Bilingual Eval¬uation Understudy) and ROUGE (Recall-Oriented Under¬study for Gisting Evaluation).

- BLEU: Measures the precision of n-grams between the generated caption and reference captions.
- ROUGE: Measures recall, focusing on the longest common subsequence (ROUGE-L), which is crucial for capturing the structure of the sentence.

D. Results

The evaluation was conducted on a test subset of the dataset. The results for the caption generation module are summarized in Table I.

TABLE I
Performance Evaluation of Image-to-Caption Module

Metric	Score
ROUGE-1	0.580
ROUGE-L	0.458
BLEU	0.041

The ROUGE scores indicate a strong recall capability, meaning the generated captions successfully capture the key objects and actions present in the images. The BLEU score, while lower, reflects the strict n-gram matching criteria which can often penalize valid but structurally different phrasing. Qualitatively, the system demonstrated the ability to generate contextually relevant hashtags and high-fidelity images from textual prompts within an

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716248

average latency of 5 seconds per image on the specified hardware.

E. Discussion and Limitations

While the quantitative results indicate satisfactory per¬formance, the system's modularity presents a trade-off between flexibility and inference speed. Running distinct models sequentially—specifically the diffusion-based im¬age generator followed by the captioning module—in-troduces latency that may be prohibitive for real-time applications. Furthermore, the reliance on local hardware limits the batch processing capability; scaling this solution for enterprise-level throughput would require optimization techniques such as model quantization or the integration of TensorRT accelerators to reduce memory footprint and improve frame rates.

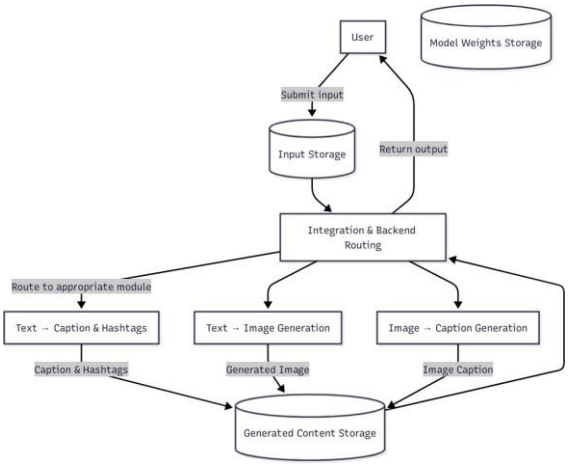

Fig. 2. Data Flow.

V. CONCLUSION

PostForge presents a viable solution to the fragmenta¬tion and privacy issues plaguing current AI-driven content creation tools. By integrating specialized models for image synthesis, captioning, and hashtag generation into a locally deployable web application, the system offers a unified, efficient, and private alternative to existing fragmented workflows. The experimental results confirm that the mod¬ular approach achieves satisfactory performance metrics while maintaining the flexibility to upgrade individual components as technology evolves. Future work will focus on optimizing inference speeds for lower-end hardware and expanding the system's capabilities to include video content generation.

VI. ACKNOWLEDGMENT

The authors would like to thank our guide, Mrs. Sangita Nikumbh for providing support and guidance throughout our development of the project, as well as the Department of Artificial Intelligence and Data Science for creating a welcoming and educational space that made this work possible.

REFERENCES

[1] J. Zhang, Z. Fang, H. Sun, and Z. Wang, "Adaptive Semantic-Enhanced Transformer for Image Captioning," IEEE Transac-tions on Neural Networks and Learning Systems, vol. 35, no. 2, 2024.

[2] P. Zeng, H. Zhang, J. Song, and L. Gao, "S2 Transformer for Image Captioning," in Proc. 31st Int. Joint Conf. Artificial Intelligence (IJCAI), 2022, pp. 1608–1614.

[3] H. Zhang, J. Y. Koh, J. Baldridge, H. Lee, and Y. Yang, "Cross-Modal Contrastive Learning for Text-to-Image Generation," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recog¬nition (CVPR), 2021, pp. 833–842.

[4] T. Yu et al., "Generating Hashtags for Short-form Videos with Guided Signals," in Proc. 61st Annual Meeting Assoc. Computational Linguistics (ACL), 2023, pp. 9482–9495.

[5] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," in Proc. Int. Conf. Machine Learning (ICML), 2021.

[6] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Un¬derstanding and Synthesis," in Proc. 39th Int. Conf. Machine Learning (ICML), 2022, pp. 12888–12900.

[7] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Mod-els," in Proc. IEEE/CVF Conf. Computer

Vision and Pattern Recognition (CVPR), 2022, pp. 10684–10695.

[8] C. Raffel et al., “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer,” J. Machine Learning Research, vol. 21, no. 140, pp. 1–67, 2020.

[9] A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” in Proc. 9th Int. Conf. Learning Representations (ICLR), 2021.

[10] A. Radford et al., “Learning Transferable Visual Models From Natural Language Supervision,” in Proc. 38th Int. Conf. Ma¬chine Learning (ICML), 2021, pp. 8748–8763.