ABSTRACT
MULTIMODAL CONTENT GENERATION

Bashaarat Nawaz Mohammad, Dilip Kumar Kasina, Soumiya Rao Thada, Vaishnavi Samboji, Venkata Kiran Reddy Kotha

This project targets a combined need in practice: fast creation of high-quality social media content and intelligent support for learning from multimodal scientific material. Social platforms demand visually appealing posts with matching background music, while students and instructors increasingly work with science questions that mix diagrams and text plus long, dense research papers. Our objective is to build a single, realistic system that can generate images, music, multimodal science answers, and concise research summaries, and to demonstrate measurable improvements over off-the-shelf models on each of these tasks.

The work decomposes the problem into four modelling tasks—text-to-image, text-to-music, multimodal ScienceQA, and scientific summarization—and assigns one model to each: Stable Diffusion, CLAP-guided MusicGen-small, Qwen2-VL fine-tuned on ScienceQA, and T5-small fine-tuned on an arXiv summarization corpus. We implement data pipelines for ScienceQA, arXiv articles/abstracts, and GTZAN-style music clips; design training scripts using PEFT/LoRA for parameter-efficient Qwen2-VL adaptation and standard seq2seq fine-tuning for T5-small; and build evaluation routines for accuracy, ROUGE/BLEU-style metrics, CLAP/FAD_CLAP, and qualitative inspection. These model services are then integrated into a modular backend and a simple web UI that exposes four user workflows: social image generation, social music generation, multimodal science QA, and scientific summarization.

The resulting prototype is a working multimodal content studio deployed as a single application. Fine-tuned T5-small achieves substantial gains over the base model across ROUGE, BLEU, METEOR, BERTScore, FactCC and perplexity; Qwen2-VL gains 2–3 percentage points in ScienceQA accuracy with stronger visual grounding; and CLAP-guided MusicGen-small attains lower FAD_CLAP and higher CLAP alignment than both base MusicGen-small and a MusicGen-large baseline, while Stable Diffusion consistently produces prompt-relevant images.

The system can support content creators and small teams in producing coherent bundles of thumbnails, short background tracks, and descriptions for reels, shorts, and educational posts with minimal manual editing. In education, it can help students and instructors answer diagram-based science questions and obtain quick but informative summaries of long research articles, lecture notes, or project documents. Because each capability is encapsulated as a model service, the architecture can be extended to other domains (for example, marketing, tutoring, or domain-specific document assistants) by swapping datasets and fine-tuned checkpoints, making this project a reusable template for building future multimodal AI assistants.