Multimodal Generative AI: The Future of Human-AI Creativity

Abstract

Multimodal Generative AI is an emerging field of artificial intelligence that enables machines to understand and generate content across various data types such as text, images, sketches, audio, and video. This unified approach supports cross-modal creativity and fosters enhanced human-AI collaboration. Tools like OpenAI's GPT-4o, Google's Gemini, and Sora demonstrate the powerful capabilities of these systems in storytelling, education, entertainment, and design.

This seminar explores how such models function using modality-specific encoders (e.g., transformer for text, CNN for images, spectrograms for audio), fusion via attention mechanisms, and decoding through diffusion or transformer-based decoders. Technologies like Stable Diffusion, ControlNet, CLIP, and LLaVA facilitate high-quality, prompt-based generation, while LoRA and LCMs ensure fast and efficient performance. Applications include sketch-to-image design assistants, visual content generation, and multimodal educational platforms.

Looking ahead, the field is poised to advance with explainable multimodal AI (XMAI), real-time AI agents on edge devices, and integration with AR/VR interfaces. As the gap between imagination and output narrows, Multimodal Generative AI is set to redefine how we create, communicate, and innovate with intelligent systems.

Keywords: Multimodal AI, Generative Models, Diffusion, LLM, Cross-modal Generation, GPT-4o, Sketch-to-Image

Reference

X. Wang et al., "Multi-modal Generative AI: Multi-modal LLMs, Diffusions and the Unification," *IEEE TPAMI*, 2025.

Name: Abhinand S
Reg No: MES22AD002
Guide: Mr. Sherikh KK

Roll No: 02[17548]
Branch: ADS, MES College of Engineering
Date: 07-08-2025