Next-Gen AIGC: A Review of Multimodal Foundation Models for Text-to-Media Innovations

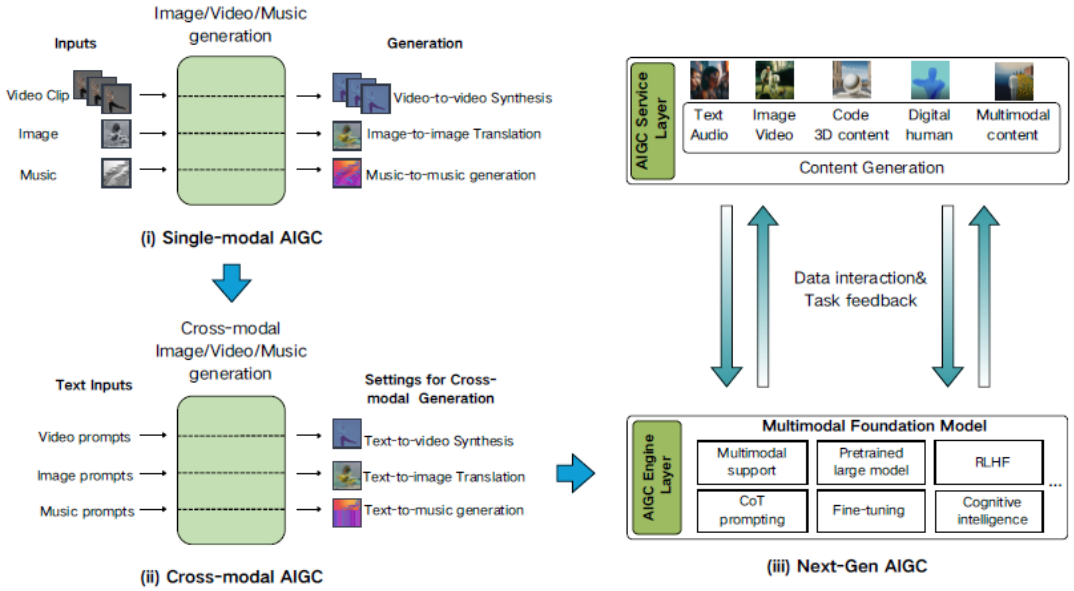
Cong JIN, Jingru FAN, Jinfa HUANG, Jinyuan FU, Tao MEI, Li YUAN, Jiebo LUO

Frontiers of Computer Science, DOI: 10.1007/s11704-025-51171-9

Problems & Ideas

Problems-Multimodal Foundation Models (MFMs), which encompass diffusion models and multimodal large language models, have attracted significant attention due to their scalable capabilities in visual and vision-language understanding and generation tasks. Although research on the technological evolution of such models continues to expand, a systematic review of their applications in text-to-multimodal content generation remains notably lacking.

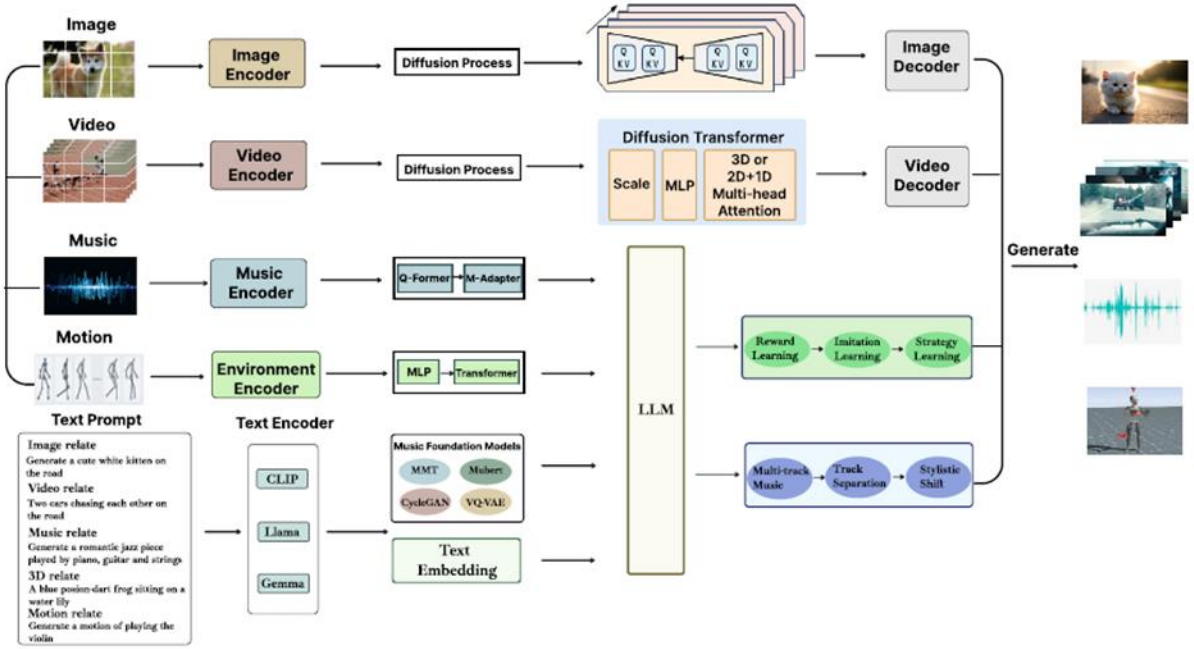
Ideas-The study focuses on four major types of artificial intelligence-generated content (AIGC) tasks based on textual prompts: text-to-image, text-to-video, text-to-music, and text-to-motion generation.



The Transition Roadmap from traditional AIGC to Next-Gen AIGC

Main Contributions

- Contributions:
  - Systematic Integration of Multimodal Foundation Models (MFMs) for Text-to-Media (text-to-image, video, music, and motion) Generation;
  - Evolutionary Perspective from Traditional AIGC to Next-Gen AIGC; This includes an emphasis on emerging capabilities such as cross-modal reasoning, zero-shot generation, and human feedback integration (e.g., RLHF and Chain-of-Thought).
  - Identification of Challenges and Future Pathways for Scalable and Controllable AIGC



General Framework of Text-to-Media Multimodal Foundation Models