# Cross-Modal Extensions of LayerComposer: Integrating Audio-Visual Inputs for Enhanced Text-to-Image Generation

Research Team

October 24, 2025

**Abstract**

This paper explores the extension of the LayerComposer framework to support cross-modal generation tasks, focusing on integrating audio and visual inputs with text-to-image (T2I) generation. By leveraging the strengths of audio cues and visual context, the proposed framework aims to enhance the fidelity and coherence of generated images while maintaining high personalization capabilities. We introduce novel techniques for audio-visual synchronization and multimodal conditioning, and demonstrate their effectiveness through comprehensive experiments.

## 1 Introduction

The LayerComposer framework has demonstrated significant capabilities in personalized text-to-image (T2I) generation by allowing users to control the spatial composition of multiple subjects within a scene. However, the integration of additional modalities such as audio can provide richer generative experiences. This paper presents extensions to LayerComposer that incorporate audio inputs, enabling synchronized audiovisual content generation. Our contributions include novel techniques for multimodal conditioning and synchronization of audio-visual cues with text prompts.

## 2 Related Work

Cross-modal generative models have gained attention in recent years, with notable advancements in integrating audio and visual data. Previous works have explored audio-driven image generation and visual storytelling, but challenges remain in achieving seamless integration. LayerComposer's advancements in spatially-aware layered canvases provide a foundation for addressing these challenges. We review related techniques in audio-visual synchronization and multimodal fusion, highlighting their relevance to our approach.

## 3 Cross-Modal LayerComposer Framework

The extended LayerComposer framework introduces an audio processing pipeline alongside the existing visual and textual inputs. We extract audio features using a pre-trained audio encoder and align them temporally with the visual content. The framework employs a multimodal transformer to fuse these features, ensuring coherent generation.

## 4 Multimodal Conditioning Techniques

To effectively combine audio, visual, and textual inputs, we employ attention mechanisms within a transformer-based architecture. The audio features are encoded into a latent space and combined with visual features from the layered canvas. We introduce a multimodal attention layer that dynamically weights the contribution of each modality based on context, ensuring that the generated content remains coherent and contextually relevant.

# 5 Experiments and Results

We conducted experiments using a dataset comprising synchronized audio and visual content. The evaluation metrics included image quality, audio-visual coherence, and user satisfaction. Our framework outperformed existing cross-modal models, demonstrating superior fidelity and synchronization.

# 6 Applications and Use Cases

The cross-modal LayerComposer framework has potential applications in interactive storytelling, multimedia content creation, and virtual reality environments. By integrating audio inputs, creators can produce more immersive experiences, enhancing engagement and interactivity. Use cases include virtual tours, educational content, and entertainment applications in AR/VR settings.

# 7 Conclusion and Future Work

We have presented a cross-modal extension to the LayerComposer framework, capable of integrating audio-visual inputs for enhanced T2I generation. Our approach improves the coherence and fidelity of generated content, offering new opportunities for multimedia applications. Future work will explore further enhancements in real-time processing and scalability.

# References

[1] G. G. Qian et al., "LayerComposer: Interactive Personalized T2I via Spatially-Aware Layered Canvas," *arXiv preprint arXiv:2510.20820v1*, 2025. `http://arxiv.org/pdf/2510.20820v1`