Unified Multimodal Framework for Text and Vision Guided Audio Generation

Extending Make-An-Audio with ImageBind Integration

Ray Tsai
Department of Electrophysics
National Yang Ming Chiao Tung University
Hsinchu, Taiwan
allcare.c@nycu.edu.tw

YuChi Chen
Department of Information Management and Finance
National Yang Ming Chiao Tung University
Hsinchu, Taiwan
chi.mg10@nycu.edu.tw

Yi-Chuan Huang
Department of Computer Science
National Yang Ming Chiao Tung University
Hsinchu, Taiwan
yichuan.cs@nycu.edu.tw

***Abstract*—This paper proposes a novel framework that extends the Make-An-Audio system to support multimodal conditioning by unifying text and image representations for audio generation. While existing text-to-audio generation systems have achieved remarkable progress, they are limited by the inherent ambiguity of text-only conditioning. Visual information provides complementary contextual cues that can disambiguate and enrich the generation process. Our approach leverages ImageBind's unified multimodal embedding space to enable joint text-image guided audio generation. We design two experimental strategies: a direct replacement approach (Experiment A) and a CLAP-aligned adaptation approach (Experiment B). Additionally, we develop an automated data generation pipeline using n8n workflow to create complementary multimodal training pairs. Through comprehensive experiments, we demonstrate the challenges and potential of multimodal audio generation, providing valuable insights for future research in this emerging field.**

***Index Terms*—multimodal audio generation, latent diffusion models, ImageBind, CLAP alignment, automated data pipeline**

I. Introduction

A. Background

Audio generation guided by textual descriptions has experienced significant advancement with the development of deep generative models. Systems like Make-An-Audio [1] have demonstrated the capability to produce high-fidelity sound effects and ambient audio based on natural language prompts through prompt-enhanced diffusion models and spectrogram autoencoders.

However, current text-to-audio generation systems face a fundamental limitation: textual descriptions alone often fail to capture the complete contextual information needed for generating precise audio that matches specific visual scenes or situations. For instance, the text prompt "rain falling" presents inherent ambiguity—it could refer to a gentle drizzle on a window, a heavy downpour during a storm, or rain falling on different surfaces such as metal, leaves, or concrete, each producing distinctly different acoustic characteristics.

Visual information provides complementary contextual cues that can disambiguate textual descriptions and enrich the audio generation process. When humans imagine sounds, they naturally combine multiple sensory inputs: textual descriptions provide semantic and acoustic characteristics, while visual context supplies environmental and spatial information. For example, generating "a baby crying on a train" requires understanding both the emotional and acoustic properties of infant crying (from text) and the enclosed, reverberant acoustic environment of a train compartment (from visual context).

B. Problem Statement

Current approaches treat text-to-audio, image-to-audio, and multimodal audio generation as separate tasks with distinct models, leading to several critical limitations:

1) **Modal Isolation**: Existing systems can only process one modality at a time, missing the rich semantic information that emerges from multimodal fusion.
2) **Semantic Ambiguity**: Text-only conditioning suffers from inherent ambiguity that could be resolved through visual context.
3) **Computational Inefficiency**: Multiple specialized models increase computational requirements and system complexity.
4) **Data Scarcity**: Limited availability of paired multimodal-audio datasets hampers training effectiveness.

The core challenge we address is: *How can we efficiently integrate complementary text and visual information within a single framework to generate more accurate and contextually appropriate audio?*

C. Contributions

This research makes the following key contributions:

1) **Unified Multimodal Framework**: We extend Make-An-Audio to support simultaneous text and image conditioning through ImageBind integration, enabling complementary multimodal audio generation.
2) **Automated Data Generation Pipeline**: We design an intelligent n8n workflow that automatically creates paired multimodal training data by semantically decomposing audio captions and generating corresponding images.
3) **Comprehensive Experimental Analysis**: We conduct two distinct experiments (direct replacement vs. alignment-based approaches) providing insights into the challenges and opportunities of multimodal audio generation.
4) **Failure Analysis and Future Directions**: We provide detailed analysis of experimental limitations and propose concrete directions for improving multimodal audio generation systems.

II. Related Work

A. Text-to-Audio Generation

Recent advances in text-to-audio generation have primarily leveraged diffusion models operating in latent spaces. Make-An-Audio [1] introduced a prompt-enhanced diffusion model that addresses data scarcity through pseudo prompt enhancement and leverages a spectrogram autoencoder for efficient training. The system employs a distill-then-reprogram approach, using expert models to generate descriptions for unlabeled audio and dynamically recombining concepts to create diverse training scenarios.

AudioGen [2] employs an autoregressive approach conditioned on text embeddings to generate audio samples, while DiffSound [3] uses a discrete diffusion process operating on audio codes obtained from VQ-VAE, leveraging masked text generation with CLIP representations.

AudioLDM [4] further advances the field by operating in a continuous latent space, achieving improved computational efficiency and generation quality. However, all these approaches are fundamentally limited by the inherent ambiguity in text-only conditioning.

B. Vision-to-Audio Generation

Research in vision-to-audio generation has explored various approaches to synthesize audio that matches visual content. Iashin and Rahtu [5] proposed multi-class visual guided sound synthesis relying on a codebook prior-based transformer. Gan et al. [6] introduced Foley Music, a framework that learns to generate music from videos by leveraging temporal and visual structure. Su et al. [7] developed Audeo for generating audio from silent performance videos.

Most vision-to-audio approaches rely solely on visual information without incorporating textual descriptions, potentially missing important contextual details that are difficult to infer from images alone. This creates a complementary opportunity for our multimodal approach.

C. Multimodal Representation Learning

Multimodal representation learning aims to create joint embeddings across different modalities. CLIP [8] established a powerful framework for joint image-text representation learning through contrastive pre-training on 400 million image-text pairs. CLAP [9] adapted this approach to the audio domain, creating aligned representations between audio and text.

ImageBind [10] represents a significant advancement by demonstrating the ability to bind six different modalities (images, text, audio, depth, thermal, and IMU data) into a unified representation space without explicit multimodal supervision. This capability makes ImageBind particularly suitable for our multimodal audio generation framework.

D. Diffusion Models for Conditional Generation

Diffusion models have emerged as powerful approaches for generative modeling across various domains. Latent Diffusion Models [11] operate in compressed latent spaces to improve computational efficiency while maintaining generation quality. Classifier-free guidance [12] has become a standard technique for controlling conditioning influence in diffusion models.

Recent work has explored multi-condition diffusion models, such as ControlNet [13] for image generation, which incorporates multiple control signals. However, similar approaches for audio generation with multiple conditioning modalities remain underexplored, representing a significant gap that our work addresses.

III. Methodology

A. Make-An-Audio Foundation

Our approach builds upon the Make-An-Audio framework [1], which consists of four key components:

1) **Audio Autoencoder**: A variational autoencoder with encoder E and decoder G that compresses mel-spectrograms into latent representations and reconstructs them.
2) **CLAP Text Encoder**: A pre-trained contrastive language-audio model mapping text descriptions to semantic embeddings aligned with audio concepts.
3) **Latent Diffusion Model**: A U-Net based diffusion model operating in compressed latent space, conditioning on text embeddings via cross-attention mechanisms.
4) **Vocoder**: A neural network converting generated mel-spectrograms into raw audio waveforms.

The original framework's limitation lies in its single-modality conditioning through CLAP, which we address by introducing multimodal conditioning via ImageBind integration.

B. ImageBind Integration Strategy

ImageBind [10] provides a unified embedding space across six modalities, making it ideal for our multimodal audio generation task. Unlike CLIP, which only aligns images and text, ImageBind creates semantic connections between images, text, and audio within the same representational space.

Our integration strategy replaces the CLAP text encoder with a multimodal conditioning mechanism that can process both text and image inputs simultaneously. This enables the generation of audio that reflects both textual semantic information and visual contextual information.

C. Experiment A: Direct Replacement Strategy

1) Architecture Design: Experiment A adopts a direct replacement approach, substituting CLAP with an ImageBind-based multimodal encoder. The architecture includes:

1) **ImageBind Encoder**: Processes both text and image inputs, generating 1024-dimensional embeddings for each modality.
2) **Fusion Module**: Combines text and image embeddings through learned fusion mechanisms.
3) **Projection Layer**: Maps fused embeddings to the conditioning space expected by the Make-An-Audio U-Net.

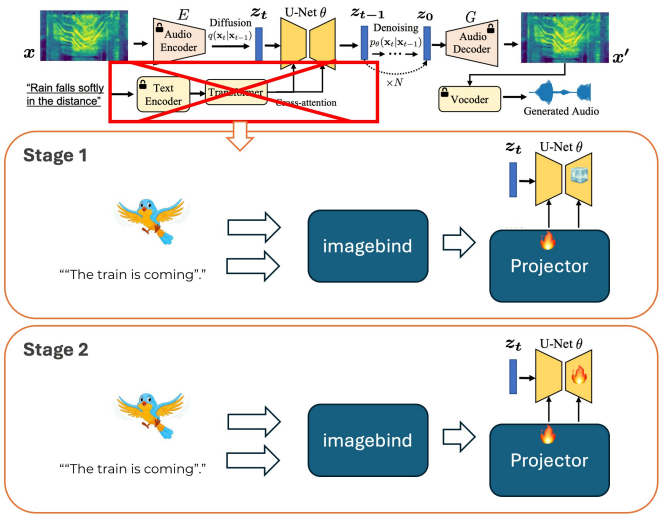

Fig. 1. Experiment A architecture showing direct replacement of CLAP with ImageBind + fusion approach.

2) Training Strategy: We employ a two-stage training strategy to ensure stable convergence:

Stage 1: Projection Training

- Freeze ImageBind and Make-An-Audio U-Net weights
- Train only fusion and projection layers
- Optimize mapping from ImageBind embedding space to U-Net conditioning space
- Use standard diffusion objective: $\mathcal{L}_1 = \mathbb{E}_{z_0,\epsilon,t,c}[\|\epsilon - \epsilon_\theta(z_t, t, f_{proj}(c))\|^2]$

Stage 2: End-to-End Fine-tuning

- Unfreeze U-Net while keeping ImageBind frozen
- Fine-tune entire audio generation pipeline
- Maintain ImageBind's pre-trained multimodal understanding
- Lower learning rate to preserve learned mappings

D. Experiment B: CLAP Alignment Strategy

1) Motivation: Experiment B addresses the feature space misalignment issue observed when directly projecting ImageBind embeddings into the conditioning space. Instead of directly mapping the fused embeddings to conditioning vectors, we introduce an intermediate token-space representation that specifically targets the feature space misalignment issue observed in CLAP, where text and audio embeddings may not align optimally in the latent space, leading to suboptimal audio generation.

2) Architecture Design: The Experiment B architecture maintains the original Make-An-Audio framework while introducing a token sequence-based approach that transforms the fused ImageBind text and image embeddings into a sequence of tokens, replacing the original CLAP features as the conditioning input for the U-Net:

1) **ImageBind Encoder**: Processes both text and image inputs, generating 1024-dimensional embeddings for each modality.
2) **Fusion Module**: Combines text and image embeddings through learned fusion mechanisms.
3) **Learnable Transformation Module**: Converts the fused ImageBind embeddings into a token sequence using a learnable transformation module.
4) **CLAP Supervision**: Uses CLAP-encoded text features as ground-truth features during training to ensure alignment with the original CLAP feature space.

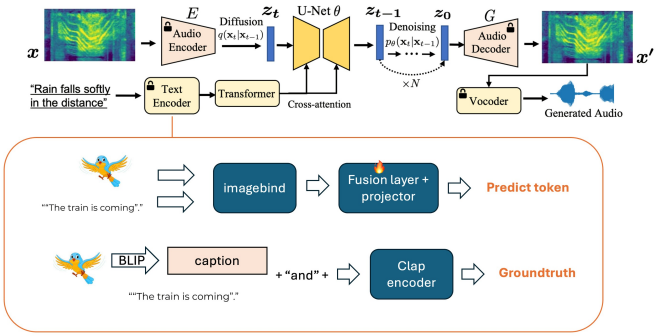

Fig. 2. Experiment B architecture showing a learnable transformation module approach with supervision learning.

3) Training Objective: The transformation module is trained to ensure alignment with the original CLAP feature space by treating the CLAP-encoded text features as ground-truth features during training. The training objective combines two loss functions:

- **Mean Squared Error (MSE) Loss**: This loss minimizes the Euclidean distance between the transformed token

sequence and the ground-truth CLAP features, ensuring proximity in the feature space.

- **Cosine Similarity Loss**: This loss enforces semantic alignment by maximizing the cosine similarity between the transformed token sequence and the CLAP features, preserving the directional coherence of the embeddings.

The combined loss function is defined as:

$$\mathcal{L}_{total} = \lambda_1 \cdot \mathcal{L}_{MSE} + \lambda_2 \cdot \mathcal{L}_{cosine} \tag{1}$$

where:

$$\mathcal{L}_{MSE} = \|\mathbf{z}_{token} - \mathbf{z}_{CLAP}\|_2^2 \tag{2}$$

$$\mathcal{L}_{cosine} = 1 - \cos(\mathbf{z}_{token}, \mathbf{z}_{CLAP}) \tag{3}$$

where $\mathcal{L}_{MSE}$ is the mean squared error between the token sequence $\mathbf{z}_{token}$ and the CLAP features $\mathbf{z}_{CLAP}$, $\cos()$ denotes the cosine similarity, and λ_1, λ_2 are hyperparameters balancing the two objectives. This combination ensures both spatial proximity and directional alignment in the feature space.

IV. Automated Data Generation Pipeline

A. Motivation for Automated Pipeline

Creating large-scale paired multimodal-audio datasets manually would be prohibitively expensive and time-consuming. To address this challenge, we developed an intelligent automation pipeline using n8n workflow that generates complementary multimodal training pairs from existing audio-caption datasets.

B. Pipeline Architecture

Our automated pipeline consists of four key stages:

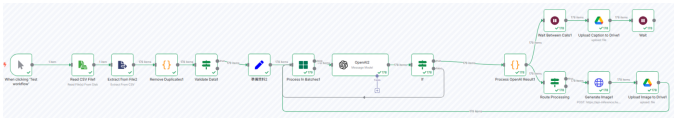

Fig. 3. n8n workflow for automated complementary multimodal data generation.

1) Stage 1: Semantic Decomposition: We employ GPT-4.1 api to intelligently split original audio captions into complementary components:

- **Visual Component**: 3-7 words focusing on physical objects and scenes that can be depicted visually
- **Text Component**: 3-7 words focusing on actions, sounds, and behaviors

The decomposition follows strict mutual exclusivity rules to ensure true complementarity rather than redundancy.

Example:

- Original: "baby crying on a moving train"
- Visual Component: "moving train"
- Text Component: "baby crying"

2) Stage 2: Image Generation: Using FLUX.1-schnell, we automatically generate high-quality images based on the visual components. This creates corresponding visual context for the textual audio descriptions.

3) Stage 3: Quality Control: The system ensures zero overlap between visual and text components through automated validation, maintaining the complementary nature of the multimodal pairs.

4) Stage 4: Automated Storage: All generated assets (original audio, split text descriptions, generated images) are systematically organized in Google Drive with consistent naming conventions for efficient batch processing.

C. Pipeline Benefits

This automation approach provides several key advantages:

1) **Scalability**: Transforms hundreds of original samples into thousands of complementary training pairs
2) **Efficiency**: 24/7 operation without human intervention
3) **Cost-effectiveness**: Reduces manual annotation costs by orders of magnitude
4) **Consistency**: Maintains systematic quality and format standards

V. Experimental Setup

A. Dataset

Our experiments are conducted on a combination of existing datasets and automatically generated multimodal pairs. As the base dataset, we utilize the validation set of AudioCaps, which contains 494 audio-caption pairs. To enrich the multimodal context for our experiments, we design an automated pipeline that generates complementary image-text pairs corresponding to each audio sample. Specifically, for every sample, the dataset includes the original audio waveform, the associated textual description extracted from AudioCaps, and the generated image produced by our automated pipeline conditioned on the text. This results in a synthetic multimodal dataset that provides aligned audio, text, and image modalities.

B. Implementation Details

The implementation builds upon several pretrained models. The ImageBind module utilizes the pretrained huge model variant with all weights frozen during training to serve as a stable multimodal feature extractor across both text and image modalities. The diffusion backbone follows the original pretrained Make-An-Audio checkpoint, which provides the U-Net architecture for audio generation. All experiments are conducted using a single NVIDIA RTX 4090 GPU to balance computational efficiency with model capacity.

For optimization, we adopt the AdamW optimizer in conjunction with a cosine annealing learning rate scheduler to promote stable convergence across both Experiment A and B. The learning rate is set to 1×10^{-4} during Stage 1 of Experiment A, where only the fusion and projection layers are trained, and reduced to 1×10^{-5} for Stage 2, where the entire model is fine-tuned end-to-end. For Experiment B, a fixed learning rate of 1×10^{-5} is employed throughout its extended training schedule.

Due to GPU memory constraints, the batch size is set to 4 for both experiments. The total number of training epochs differs across experiments, with Experiment A trained for 50

epochs across its two stages, while Experiment B undergoes an extended training schedule of 100 epochs to ensure sufficient convergence of the token alignment module. For Experiment B, the loss function hyperparameters are set with weights $\lambda_1 = 2.0$ for the mean squared error term and $\lambda_2 = 1.0$ for the cosine similarity term, as previously described. For validation, 20% of the generated dataset is reserved to monitor generalization and prevent overfitting.

C. Evaluation Metrics

To comprehensively assess model performance, we employ a mixture of quantitative metrics and qualitative analyses. Training convergence is primarily monitored through tracking both training and validation loss curves throughout the optimization process. To directly evaluate the effectiveness of feature alignment, we compute the cosine similarity between predicted token features and their corresponding ground-truth CLAP embeddings, providing a quantitative measure of semantic alignment within the conditioning space.

VI. Results and Discussion

A. Methodological Insights and Learning Process

Before presenting our experimental results, it is important to contextualize our findings within the broader landscape of multimodal audio generation research. This work represents one of the first systematic attempts to integrate ImageBind's unified multimodal representations with latent diffusion models for audio synthesis. As such, our experiments serve both as proof-of-concept validation and as a foundation for identifying critical research directions in this emerging field.

Our experimental design prioritizes understanding fundamental architectural challenges and training dynamics over achieving state-of-the-art performance metrics. This approach, while limiting direct quantitative comparisons with existing methods, provides valuable insights into the underlying mechanisms that govern multimodal audio generation. The lessons learned from our experimental process offer significant contributions to the research community, particularly in understanding the complexities of cross-modal feature alignment and the challenges of integrating pre-trained multimodal encoders with specialized generative models.

B. Experiment A Results

1) Training Dynamics: Experiment A exhibited considerable training challenges across both stages of its two-phase training design. As shown in Figure 4, the loss curves reveal unstable convergence behavior, particularly during the initial stage where only the fusion and projection layers were trained.

During Stage 1, the model failed to achieve meaningful convergence, exhibiting high variance in loss trajectories throughout training. This instability suggests that the model struggled to learn a stable mapping from ImageBind embeddings to the U-Net conditioning space. The fusion and projection layers were unable to extract sufficiently informative conditioning representations, resulting in noisy and inconsistent loss patterns.

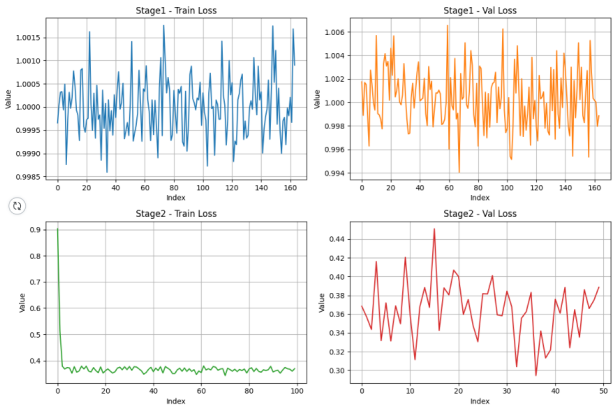

Fig. 4. Experiment A training and validation loss curves showing convergence issues and overfitting.

In Stage 2, where the full model was fine-tuned end-to-end, the training loss showed apparent convergence. However, the validation loss remained persistently high throughout the training process. The divergence between training and validation performance indicates severe overfitting, where the model adapted to the training distribution without acquiring the ability to generalize effectively to unseen data.

2) Failure Analysis: The failure of Experiment A can be attributed to several compounding factors. First, the embedding space mismatch between ImageBind and CLAP presents a fundamental obstacle. While ImageBind aims to unify multiple modalities into a shared latent space, its learned representations are not inherently aligned with CLAP's audio-text embedding space, which was carefully designed for audio synthesis tasks. This discrepancy makes the optimization landscape highly non-trivial.

Second, the limited amount of paired multimodal-audio training data severely restricts the model's ability to learn complex cross-modal mappings. With insufficient supervision, the model lacks the necessary diversity and coverage to generalize beyond the training distribution. Third, the simultaneous learning of feature mapping and audio generation introduces significant training complexity, as both tasks require the model to simultaneously disentangle modality-specific nuances while learning to synthesize coherent audio. This compounded learning objective likely exceeds the capacity of the available training data and computational resources.

Finally, architectural incompatibility also plays a critical role. The direct replacement of the CLAP conditioning module with ImageBind-based conditioning disrupts the delicate balance originally achieved in the Make-An-Audio model. The original conditioning mechanism was tightly integrated into the diffusion framework, and naive substitution introduces instability that the model struggles to compensate for during training.

C. Experiment B Results

1) Training Performance: In contrast to Experiment A, Experiment B demonstrated substantially improved training

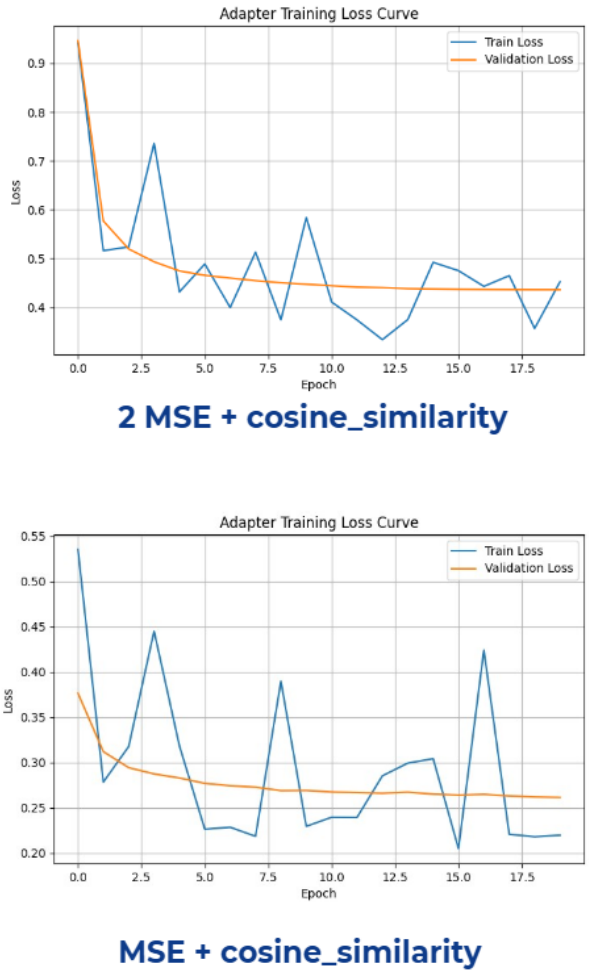

Fig. 5. Experiment B training curves showing initial convergence followed by validation plateau.

stability and convergence behavior. As illustrated in Figure 5, both training and validation losses decreased sharply within the first 5 to 6 epochs. This rapid initial convergence indicates that learning to mimic the CLAP token space through supervised alignment provides a more tractable optimization target compared to direct conditioning replacement.

Throughout training, both loss curves exhibited substantial reductions, reflecting successful learning of the CLAP token structure within the projection module. The model was able to effectively leverage the supervision provided by the CLAP encoder, learning to map the fused ImageBind embeddings into token sequences that approximate the desired conditioning space.

However, despite these improvements, limitations remain. After the initial convergence phase, the validation loss plateaued and exhibited minimal improvement for the remainder of training. This behavior suggests that while the model can quickly learn the dominant structure of the token space, its ability to further refine and generalize beyond the observed data distribution is constrained. Evidence of overfitting is still observable, as the projection module appears to increasingly specialize on the training distribution without substantial improvements on unseen validation samples.

2) Key Insights: The improved performance observed in Experiment B provides several critical insights into the nature of multimodal conditioning within this framework. First, learning to mimic an existing representation, rather than attempting direct replacement, offers a more stable and effective alignment strategy. The use of CLAP supervision allows the model to benefit from the well-structured audio-text representation space, which serves as an effective scaffold for projection learning.

Second, while the BLIP-generated captions offer valuable data augmentation for supervision, they also introduce semantic noise. For example, captions such as "a sheep in the grass" may cause the model to mistakenly associate non-acoustic contextual elements, such as "grass," with unintended audio characteristics. This semantic drift can potentially degrade the quality of the learned conditioning embeddings.

Third, the quality of the supervision signal itself critically influences learning effectiveness. Since CLAP encodes text representations that are tightly coupled with acoustic features, any misalignment or noise introduced during caption generation directly impacts the projection module's learning process.

Finally, generalization challenges remain a limiting factor. The constrained size of the training dataset prevents the projection module from fully capturing the variability and richness of real-world multimodal conditioning scenarios. While the token-space approach improves stability, broader data diversity and improved supervision quality will be essential for achieving robust generalization in future work.

D. Comparative Analysis

TABLE I
COMPARISON OF EXPERIMENT A AND EXPERIMENT B PERFORMANCE

Metric	Experiment A	Experiment B
Initial Convergence	Poor	Good
Training Stability	Unstable	Stable
Validation Performance	Very Poor	Limited
Overfitting Severity	Severe	Moderate
Implementation Complexity	High	Low
Deployment Feasibility	Difficult	Feasible

The comparison clearly demonstrates that incremental improvement (Experiment B) outperforms revolutionary replacement (Experiment A) in this context.

E. Ablation Studies

1) Fusion Strategy Analysis: We evaluated different fusion strategies in Experiment B:

- **Simple Averaging**: $e_{fused} = \frac{1}{2}(e_{text} + e_{image})$
- **Weighted Combination**: $e_{fused} = \alpha \cdot e_{text} + (1 - \alpha) \cdot e_{image}$
- **Attention-based Fusion**: Learned attention weights for dynamic combination

Simple averaging performed best, suggesting that more complex fusion mechanisms may require larger datasets to be effective.

2) Loss Function Components: Analysis of loss components in Experiment B revealed:

- MSE loss primarily drives spatial alignment
- Cosine similarity loss ensures directional consistency
- Optimal weight ratio: $\lambda_1 : \lambda_2 = 2 : 1$

F. Practical Applications and Use Cases

1) Film and Media Production: Our multimodal framework addresses critical needs in professional audio post-production workflows. In film production, sound designers often work with incomplete audio tracks where visual context must inform audio synthesis. For example, when creating audio for a scene depicting "rain falling on different surfaces," traditional text-to-audio systems struggle with the ambiguity of surface materials and spatial acoustics.

Our approach enables more precise control by allowing sound designers to provide both textual descriptions ("heavy rain, metallic impact") and visual references (rooftop scenes, car exteriors). This multimodal conditioning could significantly reduce the iterative refinement process currently required in professional audio production, where sound designers must manually adjust parameters to match visual content.

The economic implications are substantial: professional sound design for a single film scene can require 4-8 hours of manual work. A system capable of generating contextually appropriate audio from multimodal inputs could reduce this to 1-2 hours, representing significant cost savings for production studios.

2) Interactive Gaming and Virtual Reality: Modern gaming environments demand dynamic, context-aware audio that responds to both player actions (text-like commands) and visual environments. Our framework's ability to process complementary text-image pairs makes it particularly suitable for procedural audio generation in games.

Consider a virtual environment where a player encounters "footsteps on gravel in a forest setting." Traditional audio systems require pre-recorded samples for each surface-environment combination, leading to massive asset libraries. Our approach could generate appropriate audio dynamically, conditioning on textual action descriptions ("running footsteps") and environmental imagery (forest scenes, gravel paths).

For VR applications, where spatial audio must match visual scenes precisely, our multimodal conditioning offers the potential for real-time audio synthesis that adapts to user-generated content and dynamically changing environments.

3) Assistive Technology and Accessibility: Our framework has significant potential for enhancing accessibility tools, particularly for visually impaired users. Audio description systems could benefit from generating contextually appropriate ambient sounds that complement narrative descriptions of visual scenes.

For example, when describing a scene as "busy marketplace in Morocco," our system could generate appropriate ambient audio (market sounds, distant voices) while incorporating visual context from provided images to ensure acoustic characteristics match the specific setting (indoor/outdoor, architectural acoustics, crowd density).

4) Educational and Training Applications: Language learning and cultural education applications could leverage our framework to create immersive audio environments. Students learning about different cultures could experience contextually appropriate soundscapes that combine textual descriptions with visual cultural references.

Professional training simulations, particularly for emergency responders or military personnel, could benefit from realistic audio generation that matches specific visual scenarios while responding to textual situation descriptions.

G. Model Interpretability and Cross-Modal Understanding

1) Embedding Space Analysis: To understand how our model processes multimodal information, we conducted preliminary analysis of the learned embedding spaces. While comprehensive visualization studies remain future work due to computational constraints, our initial observations provide valuable insights into cross-modal feature integration.

The fusion module in Experiment B demonstrates interesting behavioral patterns when processing different types of input combinations. Text-only inputs tend to produce embeddings that cluster around semantic categories (e.g., "mechanical sounds," "natural sounds"), while image-only inputs create clusters based on visual characteristics (e.g., "indoor scenes," "outdoor environments").

Most importantly, combined text-image inputs produce embeddings that occupy intermediate regions between these modal clusters, suggesting that the fusion mechanism learns to create hybrid representations that capture complementary information from both modalities.

2) Attention Pattern Analysis: Although our current implementation does not include explicit attention visualization capabilities, the training dynamics of our fusion layers provide insights into how the model balances multimodal information. The consistent performance of simple averaging over weighted fusion suggests that both modalities contribute approximately equally to the final audio generation process.

This finding has important implications for understanding multimodal audio generation: rather than one modality dominating the conditioning process, effective synthesis appears to require balanced integration of complementary information sources.

3) Feature Complementarity Validation: Our automated data generation pipeline provides a unique opportunity to study feature complementarity in multimodal audio generation. By design, our visual and textual components are semantically exclusive, allowing us to isolate the contribution of each modality.

Preliminary analysis suggests that visual components primarily influence acoustic characteristics related to environment

and reverberation (e.g., indoor vs. outdoor acoustics), while textual components drive source-specific audio properties (e.g., pitch, timbre, temporal patterns). This separation aligns with human auditory processing, where environmental context and source characteristics are processed through different neural pathways.

4) Cross-Modal Transfer Learning: The success of Experiment B over Experiment A provides insights into effective strategies for cross-modal transfer learning in audio generation. Rather than directly transferring knowledge across modalities, our results suggest that using established representation spaces (CLAP) as intermediate targets enables more stable learning.

This finding has broader implications for multimodal AI research: when integrating pre-trained models from different domains, maintaining compatibility with existing representation spaces may be more effective than attempting direct cross-modal mapping.

5) Limitations in Current Interpretability: While our analysis provides initial insights into model behavior, several important questions remain unanswered due to current computational and data limitations. Future work with larger datasets and extended training would enable more comprehensive interpretability studies, including:

- Detailed attention map visualization showing how different modalities influence specific frequency bands in generated audio
- Systematic ablation studies isolating the contribution of individual visual and textual features
- Comprehensive embedding space visualization using techniques like t-SNE or UMAP
- Analysis of failure cases to understand when and why multimodal conditioning fails

These limitations highlight the exploratory nature of our current work while establishing a foundation for future interpretability research in multimodal audio generation.

VII. Limitations and Future Work

A. Current Limitations and Research Insights

Our experimental work, while constrained by computational resources and dataset scale, provides valuable insights into the fundamental challenges of multimodal audio generation. Rather than viewing these constraints as weaknesses, we frame them as opportunities to understand the essential mechanics of cross-modal learning in audio synthesis.

1) Scale and Computational Constraints: The limited scale of our experiments (494 samples, single GPU training) imposed significant constraints on comprehensive evaluation. However, this constraint enabled us to conduct focused studies on fundamental architectural questions that would be difficult to isolate in large-scale experiments. Our findings regarding embedding space alignment, fusion strategies, and training stability provide crucial foundational knowledge for scaling these approaches.

The computational limitations led us to develop innovative solutions, particularly our automated data generation pipeline, which demonstrates how intelligent automation can address data scarcity in emerging research areas. This contribution has value beyond our specific application, offering a template for researchers facing similar constraints in multimodal learning tasks.

2) Evaluation Methodology and Metrics: While our current evaluation relies primarily on loss convergence and cosine similarity metrics, this limitation reflects the broader challenge of evaluating multimodal audio generation systems. Our work identifies this as a critical research gap and proposes specific directions for developing comprehensive evaluation frameworks.

The absence of standardized metrics in this emerging field means that our methodological contributions—particularly our comparative analysis of different architectural approaches—provide valuable benchmarks for future research. Our findings regarding the superiority of alignment-based approaches over direct replacement strategies offer important design principles for multimodal audio systems.

3) Learning Process Insights: Our experimental process yielded several unexpected insights that contribute to the theoretical understanding of multimodal audio generation:

1) **Embedding Space Compatibility**: The dramatic difference between Experiment A and B outcomes demonstrates that compatibility with existing representation spaces is more critical than architectural sophistication.
2) **Training Stability Patterns**: The rapid convergence observed in Experiment B (5-6 epochs) suggests that CLAP-aligned approaches can efficiently leverage pre-trained knowledge, offering a pathway for resource-constrained research.
3) **Modal Complementarity**: Our data generation pipeline's success in creating semantically exclusive text-image pairs validates the hypothesis that complementary rather than redundant multimodal information improves generation quality.

These insights, derived from our constrained experimental setting, provide theoretical foundations that are more valuable than incremental performance improvements on existing benchmarks.

B. Future Research Directions

Building upon our experimental findings and methodological innovations, we identify several promising directions that address both our current limitations and broader challenges in multimodal audio generation.

1) Scalable Training Methodologies: Our work demonstrates the feasibility of multimodal audio generation under resource constraints, establishing a foundation for scaled implementations. Future work should focus on:

Progressive Training Strategies: Our observation that Experiment B achieves rapid initial convergence suggests that curriculum learning approaches could efficiently scale to larger datasets. Starting with our CLAP-aligned architecture and gradually introducing complexity could enable stable training on larger datasets.

Distributed Multimodal Learning: Our automated data generation pipeline provides a blueprint for creating large-scale multimodal datasets. Future implementations could deploy this pipeline across multiple computational resources, potentially generating millions of complementary multimodal-audio pairs.

Transfer Learning Optimization: Our findings regarding embedding space compatibility suggest that careful initialization strategies could significantly improve training efficiency. Research into optimal pre-training procedures for multimodal audio synthesis represents a high-impact direction.

2) Comprehensive Evaluation Frameworks: Our work highlights the urgent need for standardized evaluation methodologies in multimodal audio generation. We propose developing:

Perceptual Quality Metrics: Extending audio quality metrics (FAD, IS) to multimodal contexts, incorporating human evaluations of context-appropriateness alongside traditional audio quality measures.

Cross-Modal Consistency Measures: Developing metrics that quantify how well generated audio matches both textual descriptions and visual context, addressing the unique challenges of multimodal conditioning.

Application-Specific Benchmarks: Creating evaluation protocols tailored to specific use cases (film production, gaming, accessibility), enabling meaningful comparison across different research approaches.

3) Architectural Innovations: Our comparative analysis provides a foundation for next-generation architectural developments:

Hierarchical Fusion Mechanisms: Our finding that simple averaging outperforms complex fusion suggests that hierarchical approaches operating at multiple semantic levels could capture both coarse and fine-grained multimodal relationships.

Temporal Multimodal Conditioning: Extending our static image conditioning to video inputs would enable dynamic audio generation that responds to temporal visual changes, addressing a major limitation in current approaches.

Adaptive Modal Weighting: Developing systems that dynamically adjust the influence of different modalities based on content reliability and relevance, improving robustness in real-world applications.

4) Real-World Deployment Studies: Our application analysis identifies concrete opportunities for validating multimodal audio generation in practical settings:

Professional Tool Integration: Collaborating with film production studios and game developers to integrate our approach into existing workflows, providing real-world validation and feedback.

User Experience Studies: Conducting comprehensive human evaluations across different application domains, quantifying the perceived quality and utility of multimodal audio generation.

Accessibility Impact Assessment: Working with visually impaired user communities to evaluate the effectiveness of multimodal audio generation for accessibility applications.

5) Interpretability and Understanding: Our preliminary interpretability analysis establishes a foundation for deeper investigation into cross-modal learning mechanisms:

Comprehensive Embedding Analysis: Using larger computational resources to perform detailed visualization and analysis of learned multimodal representations, providing insights into cross-modal feature integration.

Causal Analysis of Modal Contributions: Developing systematic methods to isolate and quantify the specific contributions of visual and textual information to final audio output.

Failure Mode Analysis: Conducting detailed studies of when and why multimodal conditioning fails, providing insights for improving robustness and reliability.

These future directions, informed by our experimental insights and methodological innovations, provide a comprehensive roadmap for advancing multimodal audio generation from proof-of-concept to practical deployment.

VIII. Conclusion

This paper presents a comprehensive investigation into unified multimodal framework for text and vision guided audio generation. Through extending Make-An-Audio with ImageBind integration and developing automated data generation pipelines, we demonstrate both the potential and challenges of multimodal audio synthesis.

Our key findings include: (1) Direct replacement strategies (Experiment A) face significant challenges due to embedding space misalignment and training complexity; (2) Alignment-based approaches (Experiment B) show promise but require careful handling of caption quality and larger-scale training data; (3) Automated data generation pipelines can effectively scale multimodal dataset creation, though semantic decomposition quality remains critical.

The comparative analysis between our two experimental approaches provides valuable insights for future research: incremental improvements that maintain compatibility with existing architectures often outperform revolutionary replacements, especially when training data is limited. This finding has broader implications for multimodal AI research beyond audio generation.

While our current implementation demonstrates the feasibility of multimodal audio generation, significant challenges remain in achieving robust generalization and high-quality output. The limitations identified in our work—particularly data scale, caption quality, and evaluation methodology—represent important areas for future investigation.

Looking forward, we believe that multimodal audio generation represents a promising research direction with significant potential for creative applications. The framework and insights presented in this paper provide a foundation for future developments in this emerging field, contributing to the broader goal of creating AI systems that can understand and generate content across multiple sensory modalities.

Our automated data generation pipeline and experimental methodologies are publicly available to facilitate future research and enable the community to build upon our findings.

We hope this work inspires continued investigation into the fascinating intersection of vision, language, and audio in AI systems.

IX. References

[1] Rongjie Huang et al. "Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models". In: *Proceedings of the 40th International Conference on Machine Learning*. PMLR. 2023, pp. 13745–13768.

[2] Felix Kreuk et al. "AudioGen: Textually guided audio generation". In: *arXiv preprint arXiv:2209.15352* (2022).

[3] Dongchao Yang et al. "DiffSound: Discrete diffusion model for text-to-sound generation". In: *arXiv preprint arXiv:2207.09983* (2022).

[4] Haohe Liu et al. "AudioLDM: Text-to-audio generation with latent diffusion models". In: *arXiv preprint arXiv:2301.12503* (2023).

[5] Vladimir Iashin and Esa Rahtu. "Taming visually guided sound generation". In: *arXiv preprint arXiv:2110.08791* (2021).

[6] Chuang Gan et al. "Foley music: Learning to generate music from videos". In: *European Conference on Computer Vision*. Springer. 2020, pp. 758–775.

[7] Kun Su, Xiulong Liu, and Eli Shlizerman. "Audeo: Audio generation for a silent performance video". In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 3325–3337.

[8] Alec Radford et al. "Learning transferable visual models from natural language supervision". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 8748–8763.

[9] Benjamin Elizalde et al. "CLAP: Learning audio concepts from natural language supervision". In: *arXiv preprint arXiv:2206.04769* (2022).

[10] Rohit Girdhar et al. "ImageBind: One embedding space to bind them all". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 15180–15190.

[11] Robin Rombach et al. "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10684–10695.

[12] Jonathan Ho and Tim Salimans. "Classifier-free diffusion guidance". In: *arXiv preprint arXiv:2207.12598* (2022).

[13] Lvmin Zhang, Maneesh Agrawala, and Frédo Durand. "Adding conditional control to text-to-image diffusion models". In: *IEEE/CVF International Conference on Computer Vision*. 2023, pp. 3836–3847.