

# Title

Cross-Modal Audio Grounding with Visual Hallucination (CAVH): Enhancing LLMs' Audio Understanding through Visual Imagination

## Problem Statement

Large Language Models (LLMs) often struggle to ground abstract audio concepts in concrete visual representations, limiting their ability to reason about audio in relation to the physical world. This challenge hinders LLMs' performance in tasks requiring a deep understanding of the physical implications of audio content.

## Motivation

Existing approaches typically rely on pre-trained visual-audio models or simple cross-modal embeddings, which don't fully leverage the LLM's generative capabilities for visual imagination. By enabling LLMs to generate and reason with visual hallucinations based on audio input, we could dramatically improve their understanding of the physical world implications of audio content. This approach could bridge the gap between abstract audio concepts and concrete visual representations, potentially leading to more robust and versatile audio-visual reasoning capabilities in LLMs.

## Proposed Method

We propose Cross-Modal Audio Grounding with Visual Hallucination (CAVH), a novel framework that allows LLMs to generate and reason with visual representations of audio content. CAVH consists of four main components: (1) An audio encoder that extracts rich acoustic features, (2) A visual hallucination generator that produces abstract visual representations based on audio features, (3) A cross-modal attention mechanism that allows the LLM to attend to both audio features and generated visual representations, and (4) A visual-audio consistency module that ensures coherence between the audio input and visual hallucinations. During training, we use a multi-task objective that includes audio-to-image generation, image-to-audio retrieval, and audio-visual question answering. We also introduce a novel 'imagination consistency' loss that encourages the model to generate similar visual hallucinations for semantically similar audio inputs. At inference, the model can generate and manipulate visual hallucinations to support reasoning about abstract audio concepts.

## Step-by-Step Experiment Plan

### Step 1: Data Preparation

Create a new Audio-Visual Reasoning benchmark dataset. This dataset should include: (a) Audio clips with corresponding text descriptions, (b) Audio-visual question-answer pairs, (c) Audio-to-image generation pairs. Use existing datasets like AudioSet, VGGSound, and Flickr-SoundNet as starting points. Ensure a diverse range of sounds, from simple object sounds to complex soundscapes.

### Step 2: Model Architecture

Implement the CAVH architecture: (a) Audio encoder: Use a pre-trained audio model like VGGish or PANNs, (b) Visual hallucination generator: Implement a transformer-based generator that takes audio features as input and outputs visual features, (c) Cross-modal attention mechanism: Implement a multi-head attention layer that allows the LLM to attend to both audio and visual features, (d) Visual-audio consistency module: Implement a contrastive loss between the generated visual features and the original audio features.

## Step 3: Training

Train the model using the multi-task objective: (a) Audio-to-image generation: Use a GAN-based loss, (b) Image-to-audio retrieval: Use a contrastive loss, (c) Audio-visual question answering: Use a cross-entropy loss, (d) Imagination consistency: Implement a new loss that encourages similar visual hallucinations for semantically similar audio inputs. Use curriculum learning, starting with simple sound-object associations and gradually increasing complexity.

## Step 4: Evaluation

Evaluate CAVH on the Audio-Visual Reasoning benchmark: (a) Accuracy on audio-visual reasoning tasks, (b) Quality of generated visual hallucinations (use FID score), (c) Retrieval performance for image-to-audio and audio-to-image tasks, (d) Human evaluation of the coherence and relevance of generated visual hallucinations.

## Step 5: Baselines

Implement and evaluate baseline models: (a) LLM with simple audio feature concatenation, (b) LLM with pre-trained audio-visual embeddings, (c) LLM with external visual grounding (e.g., CLIP).

## Step 6: Ablation Studies

Conduct ablation studies to understand the contribution of each component: (a) Remove visual hallucination generator, (b) Remove cross-modal attention, (c) Remove visual-audio consistency module, (d) Remove imagination consistency loss.

## Step 7: Analysis

Perform qualitative analysis: (a) Visualize generated hallucinations for different audio inputs, (b) Analyze attention patterns in the cross-modal attention mechanism, (c) Investigate cases where CAVH outperforms baselines and vice versa.

# Test Case Examples

## Baseline Prompt Input

Q: You hear a loud crashing sound followed by the tinkling of glass. What might have happened?

## Baseline Prompt Expected Output

Based on the audio description, it sounds like something might have fallen and broken. The loud crashing sound suggests a significant impact, while the tinkling of glass indicates that a glass object was likely involved. Some possible scenarios could be: 5. A glass shower door or glass partition fell and broke. Without more context, it's difficult to determine the exact cause, but these are plausible explanations for the described sounds.

- A glass vase or decorative item fell from a shelf or table and shattered on the floor.
- A window was broken, perhaps by a thrown object or due to strong winds.
- A glass-topped table or furniture piece collapsed or was knocked over.
- Someone dropped a tray carrying glass items, such as glasses or dishes.
- A glass shower door or glass partition fell and broke.

## Proposed Prompt Input

Q: You hear a loud crashing sound followed by the tinkling of glass. What might have happened? Generate a visual hallucination based on this audio and use it to reason about the situation.

## Proposed Prompt Expected Output

Visual Hallucination: I'm imagining a scene in a living room. There's a tall bookshelf against a wall, and on one of the upper shelves, there's a large, ornate glass vase. The vase is teetering on the edge of the shelf, and then it falls. As it falls, I see it tumbling through the air, catching the light as it rotates. It then strikes the hardwood floor with a loud crash, shattering into countless pieces that scatter across the floor, some sliding under nearby furniture. Reasoning based on the visual hallucination: Based on this visual imagination triggered by the audio description, here's what might have happened: 1. Object Involved: The sound suggests a large, fragile object made primarily of glass. In this case, it appears to be a decorative glass vase. 2. Initial Loud Crash: This would correspond to the moment the vase hits the hard surface of the floor. The loudness indicates it fell from a significant height, which aligns with it being on an upper shelf of a tall bookcase. 3. Tinkling of Glass: After the initial impact, the vase would shatter into many pieces. These pieces would continue to move and settle, creating the tinkling sound as they slide across the floor and potentially hit other objects. 4. Possible Causes: The vase might have been precariously placed on the shelf, or perhaps it was knocked by vibrations (like a door slamming or someone walking heavily nearby). It's also possible that a draft from an open window or air conditioning could have disturbed its balance. 5. Aftermath: There would likely be a wide spread of glass shards on the floor, possibly extending under nearby furniture, making cleanup challenging and potentially dangerous. This visual hallucination helps to create a concrete scenario that fits the audio description, allowing for more detailed and grounded reasoning about what might have occurred.

- Object Involved: The sound suggests a large, fragile object made primarily of glass. In this case, it appears to be a decorative glass vase.
- Initial Loud Crash: This would correspond to the moment the vase hits the hard surface of the floor. The loudness indicates it fell from a significant height, which aligns with it being on an upper shelf of a tall bookcase.
- Tinkling of Glass: After the initial impact, the vase would shatter into many pieces. These pieces would continue to move and settle, creating the tinkling sound as they slide across the floor and potentially hit other objects.
- Possible Causes: The vase might have been precariously placed on the shelf, or perhaps it was knocked by vibrations (like a door slamming or someone walking heavily nearby). It's also possible that a draft from an open window or air conditioning could have disturbed its balance.
- Aftermath: There would likely be a wide spread of glass shards on the floor, possibly extending under nearby furniture, making cleanup challenging and potentially dangerous.

## Explanation

The CAVH method generates a rich visual hallucination based on the audio input, which allows for more detailed and grounded reasoning. The baseline method provides a general interpretation without specific visual details, while the CAVH method creates a vivid scene that explains the sounds more comprehensively and provides additional context for reasoning about the situation.

## Fallback Plan

If the proposed CAVH method doesn't significantly outperform baselines, we can pivot the project in several ways: 1) Conduct an in-depth analysis of where and why the visual hallucinations fail to improve reasoning. This could involve categorizing different types of audio inputs and analyzing performance across these categories. 2) Investigate whether the generated visual hallucinations, even if not improving task performance, provide interpretable insights into the model's understanding of audio concepts. This could turn the project into an analysis of LLM audio comprehension. 3) Explore whether the visual hallucinations can be used to augment training data for audio-visual tasks,

potentially improving performance indirectly. 4) Investigate whether the approach works better for specific subsets of audio-visual reasoning tasks, which could lead to insights about which types of audio concepts benefit most from visual grounding.

Ranking Score: 5