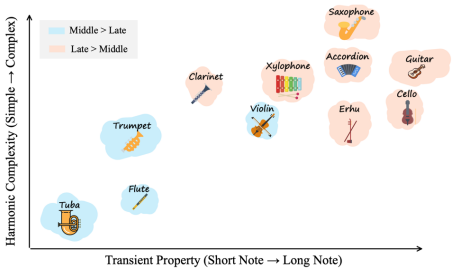
Audio-Visual Separation with Hierarchical Fusion and Representation Alignment

Han Hu*, Dongheng Lin*, Qiming Huang, Yuqi Hou, Hyung Jin Chang, Jianbo Jiao
The Mix Group, School of Computer Science, University of Birmingham, UK

BMVC

1. Motivation

Problem:
Self-supervised audio-visual source separation leverages natural correlations between audio and vision modalities to separate mixed audio signals.

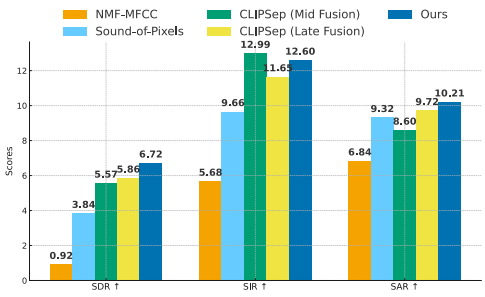


Key Observations:
- **Middle fusion** is better suited for handling short, transient sounds
- **Late fusion** is more effective for capturing sustained and harmonically rich sounds

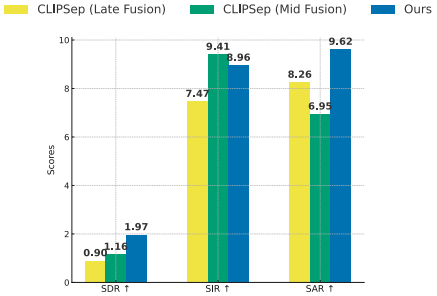
? Key Question:
Can we design a fusion strategy that combines the strengths of both approaches to improve separation across a wide range of sound characteristics?

3. Quantitative Results

Audio Separation on MUSIC Dataset

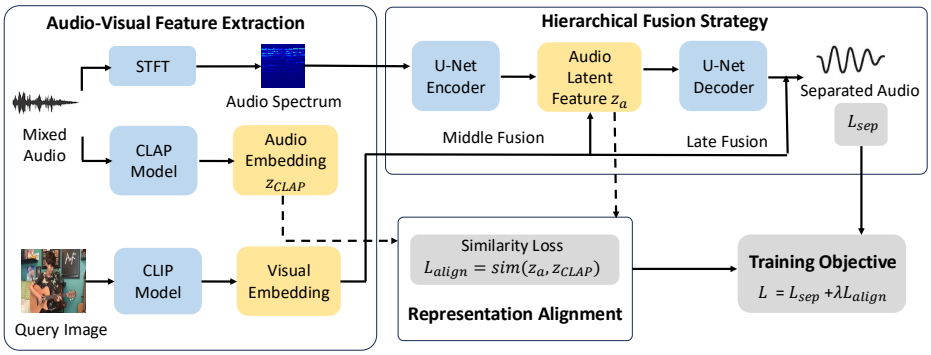


Audio Separation on VGGSound Dataset



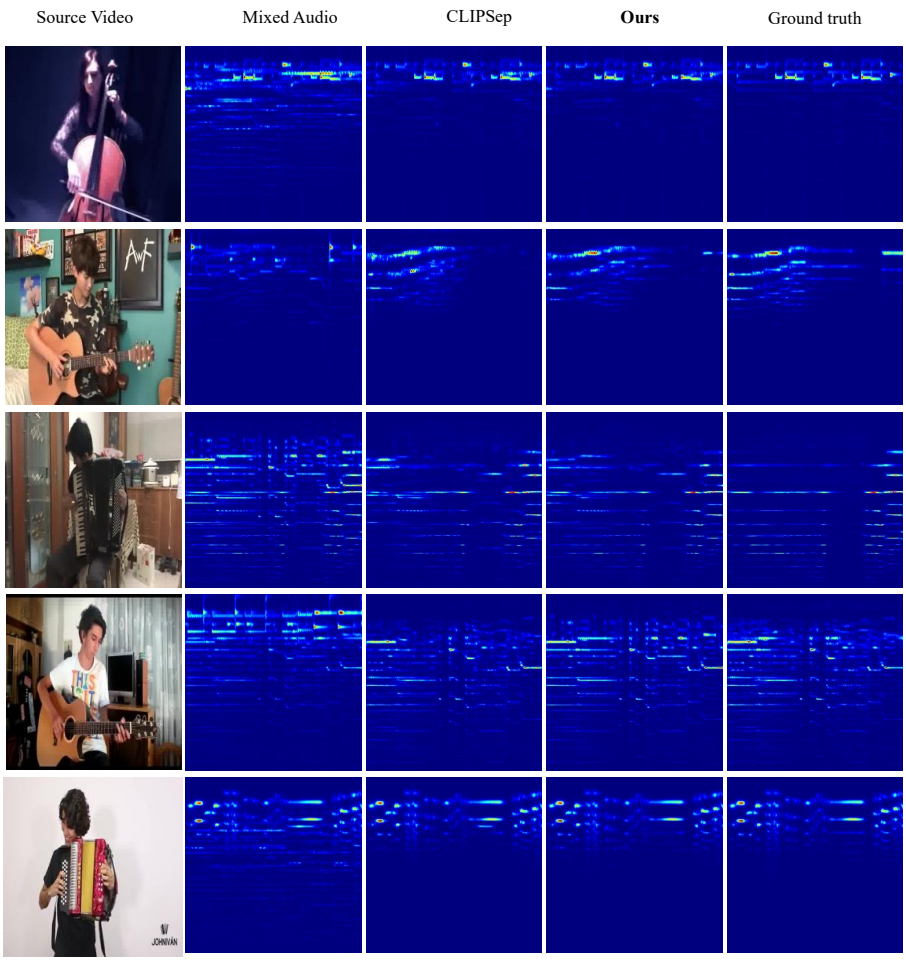
Takeaway: Our method got SOTA performance on mixed audio separation task for both datasets.

2. Method



Key Design: Align latent space with pre-trained embeddings
Combines middle and late fusion to leverage complementary strengths

4. Quanlitative Results



Takeaway:
- We reveal the relationship between acoustic properties and fusion strategies, proposing hierarchical fusion that adapts to both short transient and sustained harmonic sounds.
- Audio representation alignment reduces modality gap and enhances semantic richness, achieving state-of-the-art performance

Project Page: https://happy-new-bears.github.io/AudioSep-HFRA