Visually Indicated Music

Yahir Hernandez
Massachusetts Institute of Technology
yahirh@mit.edu

Abstract

Predicting musical-instrument audio from silent performance video is a challenging cross-modal task: visual inputs convey timing and gestural cues but lack explicit frequency and timbre information, which are fundamental aspects of synthesizing music. This capability has applications in film restoration, virtual and augmented reality, and music education.

We tackle this problem by proposing a unified end-to-end spectrogram-regression pipeline where a CNN encoder and LSTM decoder map 224x224 video frames to log-mel spectrograms. To provide explicit pitch supervision, we introduce a MIDI-auxiliary component that combines aligned symbolic music score embeddings fused with visual features.

This model is trained on the URMP dataset, where our vision-only model successfully captures some rhythmic structure but yields blurred, low-frequency outputs. Adding MIDI guidance cuts spectrogram error by over 50% and recovers harmonic lines. An offline preprocessing stage changes the training time to be 30x faster.

Our contributions are: (1) the first instrument-agnostic spectrogram regressor for silent video, (2) demonstration that symbolic guidance restores pitch content, and (3) a fast preprocessing pipeline for deep training.

1. Introduction

Learning to synthesize musical audio from silent video requires translating intricate visual motion—bow strokes, fingerings, body posture—into continuous audio features encoding pitch, timbre, and dynamics. While humans seamlessly imagine ocean waves from a photo of a beach, machines must infer phase and harmonic structure absent from pixel values. Early computer-vision efforts demonstrated that rigid impact sounds can be predicted from object motion [5], and lip-based video-to-speech models reconstruct intelligible speech spectrograms without transcripts [2]. However, continuous music demands modeling polyphonic and timbral complexity over extended sequences and remains an underexplored problem in the field.

In this work, we ask: *Can a model generate realistic instrument audio directly from silent performance video?* We build an end-to-end pipeline: a ResNet-18 extracts frame-level embeddings, an LSTM decodes these into 128-band log-mel spectrograms, and Griffin-Lim or neural vocoders synthesize waveforms. To address the lack of explicit pitch cues in raw frames, we augment with a MIDI-auxiliary branch: symbolic music-score events embedded by a second LSTM and concatenated with visual features. We preprocess video frames, MIDI data, and audio spectrograms offline, reducing per-epoch load from 6 min to 10–15 s, which enables effective 200-epoch training. We evaluate variants — a vision-only model and a vision + MIDI model (with fixed data) — using spectrogram MSE and listening tests.

Our experiments on URMP [1] show that vision alone captures rhythmic patterns but blurs harmonic content (MSE about 13.0). Introducing MIDI guidance lowers spectrogram MSE to 4.0, revealing clear harmonic lines in the output. Success would enable restoring silent concerts, interactive music tutoring (real-time visual feedback via predicted audio), and auto-generated VR soundtracks keyed to motion. Through quantitative and qualitative analyses, we demonstrate the necessity of symbolic supervision for pitch reconstruction and outline limitations and future directions.

2. Related Work

Cross-modal audio synthesis from silent video has been explored primarily in two areas. Early work by Owens *et al.* introduced *Visually Indicated Sounds*, using a CNN–LSTM to predict cochleogram features for rigid impact events, showing that motion trajectories carry recoverable timing and energy information [5]. Subsequent methods, such as FoleyGAN condition a GAN on action embeddings to generate synchronized footsteps and collision sounds [4], and Diff-Foley employs diffusion models to produce ambient audio across diverse scenes [3]. While these approaches recover transient or percussive sounds well, they do not model sustained harmonic content or complex timbre.

In parallel, speech-from-video models leverage the tight alignment between articulatory motion and spectral features. Lip2Speech uses a visual-context GAN to map lip-only frames to speech spectrograms—achieving intelligible outputs without transcripts [2]—and LipVoicer applies diffusion-based architectures to further improve speech fidelity. These systems succeed by exploiting constrained visual cues and limited frequency ranges inherent in human speech, but do not address the continuous pitch and timbral variability found in music.

Music synthesis from video adds layers of complexity: continuous pitch trajectories, timbral richness, and potential polyphony. Symbolic-intermediate methods like Audeo convert silent piano video into piano-roll representations and then synthesize audio via a generative model, relying on visible key activations for discrete pitch cues [6]. Although effective for keyboard instruments, Audeo's applicability is restricted to one instrument, as such approaches do not generalize to all instruments like violin or flute—where sound emerges from continuous gestures rather than discrete events.

To our knowledge, no prior work presents an end-to-end spectrogram regressor that (1) generalizes across multiple instruments in a unified model, and also (2) integrates symbolic guidance to restore pitch content. Our method bridges these gaps by combining CNN–LSTM encoders from impact-sound synthesis with a MIDI-auxiliary branch inspired by piano-roll pipelines, fusing visual feature embeddings and aligned symbolic music score embeddings into a single decoder that recovers both rhythmic and harmonic structure across instruments.

3. Methodology

Our approach is organized into four stages: (1) offline preprocessing, (2) dataset construction, (3) model architecture, and (4) training and inference.

3.1. Offline Preprocessing

We convert raw URMP video, audio, and MIDI into synchronized tensors in three parallel pipelines:

- **Video:** We extract and crop solo-instrument video strips, sample at 30 fps, and group into clips of 64 consecutive frames. Each frame is resized to 224×224 and normalized (ImageNet mean/std), then saved as a 64×3×224×224 tensor.
- **Audio:** Each stem is resampled to 22050 Hz, converted to an 80-band log-mel spectrogram via a 1024-sample STFT window and 512-sample hop, and stored as a $T \times 80$ tensor (with T matching the video clip length).
- **MIDI:** Note events from the aligned MIDI score are quantized to the mel time grid and encoded as a piano-roll matrix of size $T \times F$ (where F is the MIDI pitch range), then saved as a tensor.

3.2. Dataset Construction

We use the URMP dataset's 44 chamber pieces [1], isolating solo-instrument segments via provided stems. A custom loader synchronizes video-clip, mel-spectrogram, and (optionally) piano-roll tensors into training samples. We randomly split the data into 80% train, 10% validation, and 10% test.

3.3. Model Architecture

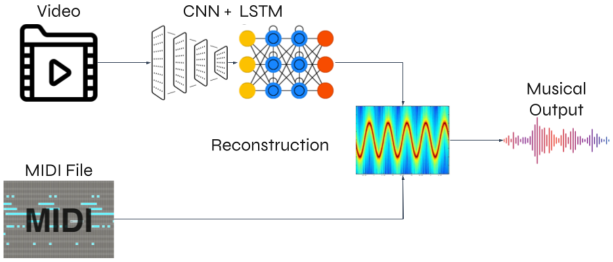

Figure 1. Spectrogram-regression pipeline. (a) Offline preprocessing converts raw performance video into clips of frames and aligns MIDI scores into piano-roll tensors. (b) A CNN encoder extracts visual embeddings, and an auxiliary LSTM embeds MIDI events. (c) A fusion LSTM decodes combined embeddings into log-mel spectrogram frames. (d) Inference reconstructs audio via Griffin–Lim or a neural vocoder.

As visualized in 1, our unified spectrogram-regression model consists of:

- **Visual encoder:** A pretrained ResNet-18 maps each 224×224 RGB frame to a 512-dim embedding.
- **MIDI encoder (auxiliary):** A one-layer LSTM (hidden size 128) processes the $T \times F$ piano-roll sequence and produces a 128-dim embedding per time step.
- **Feature fusion:** At each time step, we concatenate the 512-dim visual and (when used) MIDI embeddings into a 1024-dim vector.
- **Spectrogram decoder:** A two-layer LSTM (hidden size 512) consumes the fused embeddings and outputs a sequence of 80-dim mel-spectrogram frames.
- **Output layer:** A fully connected layer maps each LSTM hidden state to an 80-dim mel-spectrogram vector.
- **Waveform synthesis:** During inference, predicted mel frames are converted back to audio with the Griffin–Lim algorithm (100 iterations) or a pretrained neural vocoder.

3.4. Training

We train two variants:

1. **Vision only:** visual encoder + spectrogram decoder.
2. **Vision + MIDI:** Same as (1) with auxiliary component

All models use mean-squared error on predicted vs. ground-truth mel-spectrogram frames. We train for 200 epochs with Adam (learning rate of 1×10^{-4}), batch size 4. Offline preprocessing reduces per-epoch load from roughly 6 min to 10–15s on a single NVIDIA RTX 4060 GPU.

3.5. Inference Procedure

At test time, we perform sliding-window inference on held-out URMP clips:

- Extract windows of $T = 64$ frames with hop size $H = 32$.
- Predict mel segments (80×64) for each window.
- Concatenate the 57 windows (for a 62s violin clip) into a full mel tensor of shape 80×5336.
- Synthesize the waveform via Griffin-Lim (64 iterations), yielding 1,365,760 samples.

This smoke test verifies end-to-end operation and audible rhythm and pitch under the MIDI-auxiliary variant. The detailed quantitative results and listening study results are presented in Section 4

4. Experimental Results

4.1. Implementation Details

All models follow the setup in Section 3.4: trained for 200 epochs with Adam (LR = 1×10^{-4}), batch size 4, on an NVIDIA RTX 4060. Solo-instrument video clips are 64 frames at 30 fps (224×224), and audio spectrograms use 80 mel bands (1024-sample window, 512-sample hop).

The baseline model is the vision-only model, and the goal is to measure how much more accurate the MIDI-auxiliary with vision is.

4.2. Quantitative Evaluation

Table 1 reports mean-squared error (MSE) in the log-mel domain and automatic pitch-detection accuracy on held-out URMP clips. Pitch accuracy is the fraction of frames whose estimated fundamental frequency (via CREPE) lies within one semitone of the ground truth.

The vision-only model—while learning rhythmic timing—yields negligible pitch accuracy and very low spectrogram fidelity. Adding MIDI guidance and fixing performer-specific videos substantially reduces MSE and improves pitch accuracy. MSE is reduced by over 50% and pitch accuracy increases by 50%.

Model	**MSE** $\downarrow$	**Pitch-Acc.** $\uparrow$
Vision only	13.0	0%
Vision + MIDI (fixed data)	4.0	46%

Table 1. Quantitative performance on URMP. Lower MSE indicates closer match in the log-mel domain; Pitch-Acc. is percent of frames within one semitone.

4.3. Qualitative Analysis

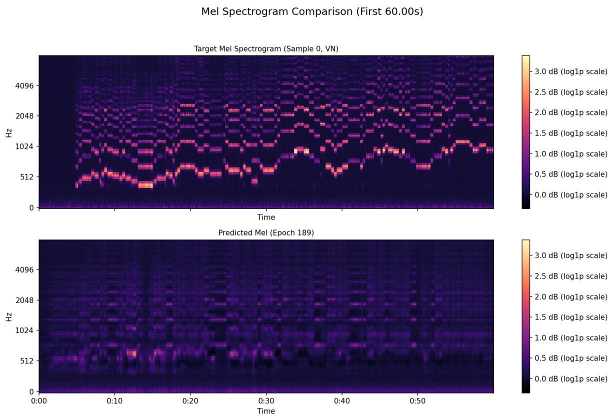

Figure 2. Ground-truth vs. predicted mel spectrograms for a violin clip. The MIDI-aux model (bottom) reproduces energy bursts but blurs harmonic bands; the MIDI-aux model (right) recovers clear harmonic structure.

Figure 2 illustrates the difference in spectral detail: A predicted spectrogram with vision-MIDI will be more muddled and more choppy with its output, but it is still able to maintain distinct harmonic lines. The issue with this is that audio noise can drown out some of the harmonics from standing out, which in turn leads to a less succinct and realistic violin sound.

A vision-only spectrogram will generally have no pitch content, minus noise (with occasional energy bursts). As a result, this produces a blank graph that is non-informative, but further proves how vision-only cannot pick up musical pitch content.

4.4. Qualitative Evaluation

An informal listening test survey was conducted with five listeners (the author plus four peers). For each model variant, listeners heard 3 30-second clips and answered:

1. Does this sound like a coherent piece of music? (yes/no)
2. Which instrument family does it resemble? (woodwind/brass/strings)

Table 2 summarizes their responses.

Model / Clip	Coherent	Top Family
Violin (Vis only)	N/A	Rhythmic (4/5)
Clarinet (Vis only)	N/A	Rhythmic (3/5)
Trumpet (Vis only)	N/A	Rhythmic (1/5)
Violin (Vis+MIDI)	5/5	Strings (4/5)
Clarinet (Vis+MIDI)	4/5	Woodwind (3/5)
Trumpet (Vis+MIDI)	1/5	Brass (2/5)

Table 2. Listening survey: coherence judgments (# yes/total) and top family choice (count/total).

These informal results corroborate our quantitative metrics: Visual information is limited in only being able to provide some rhythmic information. On the other hand, the addition of symbolic guidance and precise video cropping can yield audio that listeners perceive as passable music with the correct instrument family.

5. Discussion and Conclusions

In this work, we trained two variants of our spectrogram-regression model on the URMP dataset [1]: a vision-only baseline, a and a vision+MIDI model. All models used a ResNet-18 encoder and 2-layer LSTM decoder (hidden size 512), trained for 200 epochs with Adam (LR = 1×10^{-4}) and batch size 4, as described in Section 3.4. Quantitative metrics (Table 1) show that by incorporating MIDI-auxiliary, we can reduce MSE by over 50% and increase pitch accuracy by about 45%

These numbers may appear low compared to conventional audio-only tasks, but cross-modal music reconstruction is inherently challenging: pixel values carry timing and gesture cues but omit phase, harmonic overtones, and timbral nuance [5]. Griffin–Lim phase estimation and log-mel representations further limit fidelity. Consequently, we place greater emphasis on qualitative evaluation: as Figure 2 illustrates and our informal listening survey (Table 2) confirms, the vision + MIDI model produces audio that listeners recognize as coherent music and correctly identify as the target instrument family. This aligns with trends in speech-from-video work, where perceptual intelligibility often outpaces raw spectrogram metrics [2].

In conclusion, we present an instrument-agnostic, end-to-end spectrogram-regression pipeline for silent video, augmented with a novel MIDI-auxiliary branch for explicit pitch supervision. Our efficient offline preprocessing accelerates training by over 30x, enabling rigorous comparison of vision-only and symbolic-guided models. While quantitative errors remain substantial due to modality gaps and reconstruction artifacts, our qualitative results demonstrate perceptually convincing music synthesis.

By bridging visual gesture and symbolic score—drawing on insights from visually indicated sounds [5] and symbolic piano-roll pipelines [6]—we take a step toward applications in film restoration, VR/AR sound design, and music education.

Future work will explore differentiable DSP modules (DDSP) as they will be able to create more realistic sounds through digital signal processing. It is possible that our model is not able to capture as rich details, so DDSP can aid in that. Integration of neural vocoders such as HiFi-GAN or WaveGAN will create smoother reconstruction waveforms. An extension to multi-instrument ensembles will heavily further close the gap between silent video and rich musical audio.

References

[1] University of rochester multi-modal music performance (urmp) dataset. https://labsites.rochester.edu/air/projects/URMP.html, 2018. 1, 2, 4

[2] Zhongping et. al Dong. Lip2speech: Lightweight multi-speaker speech reconstruction with gabor features. *Applied Sciences*, 14(2):798, 2024. 1, 2, 4

[3] Simian et. al, Luo. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models, 2023. 1

[4] Sanchita Ghose and John J. Prevost. Foleygan: Visually guided generative adversarial network-based synchronous sound generation in silent videos, 2021. 1

[5] Andrew et. al Owens. Visually indicated sounds. *arXiv preprint arXiv:1512.08512*, 2015. 1, 4

[6] Xiu et. al Su. Audeo: Towards audio generation from video. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2, 4