A. Architecture Details

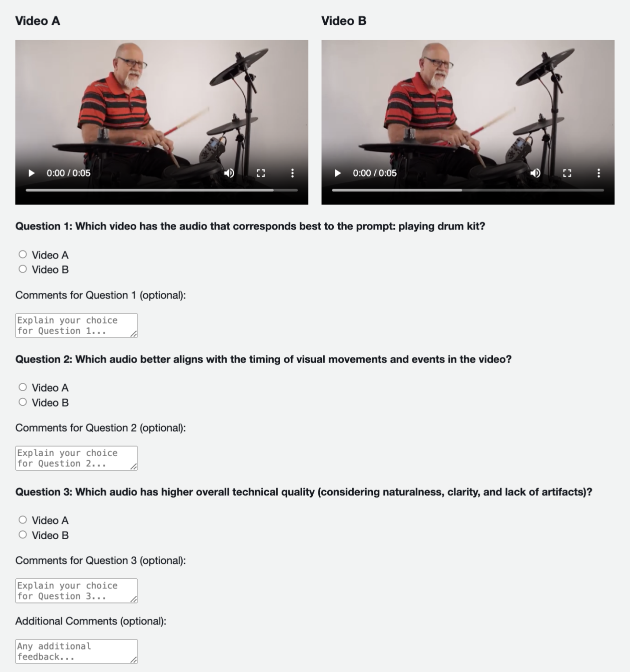
This section provides additional details about parts of the architecture not covered in the main paper.

AVCLIP preprocessing To align the resulting representation with the noised latent space z_t, we pre-processed the output of the AVCLIP encoder using two sequential transformation blocks. Each block consists of a fully connected (fc) layer, a ReLU activation, layer normalization, and dropout with $d = 0.1$. In the first block, the fully connected layer maintains the feature dimension at 768, while in the second block, it expands the feature dimension from 768 to 1024.

CLIP preprocessing Given a video V of 10 seconds at 25 FPS, we uniformly sampled 5 frames per second and computed their CLIP representations, resulting in feature vectors of shape $(1, 768)$ for each frame. We then applied linear interpolation, producing a signal of shape $(200, 768)$. To ensure alignment with the latent space, we symmetrically padded the signal on both sides so that the 10-second duration corresponds to 216 samples. Next, we processed the signal through a preprocessing block consisting of a fully connected (FC) layer, an ReLU activation, layer normalization, and dropout with $d = 0.1$. In this block, the FC layer preserves the feature dimension of the CLIP encoder output at 768. We then summed the result with the output of the first AVCLIP processing block and passed it through an additional preprocessing block, where the FC layer expands the feature dimension from 768 to 1024, ensuring that the final representation is aligned with the dimensions of z_t.

B. User Study Form

Figure 4 shows the user study interface where participants compared audio outputs from different models, evaluating quality, temporal alignment with video, and adherence to textual prompts.



Video A **Video B**

Question 1: Which video has the audio that corresponds best to the prompt: playing drum kit?

○ Video A
○ Video B

Comments for Question 1 (optional):

Explain your choice for Question 1...

Question 2: Which audio better aligns with the timing of visual movements and events in the video?

○ Video A
○ Video B

Comments for Question 2 (optional):

Explain your choice for Question 2...

Question 3: Which audio has higher overall technical quality (considering naturalness, clarity, and lack of artifacts)?

○ Video A
○ Video B

Comments for Question 3 (optional):

Explain your choice for Question 3...

Additional Comments (optional):

Any additional feedback...

Figure 4. User study form

C. Training Details

Model	Steps/Epochs	Hardware	Batch Size	GPU Hours
CAFA	81k steps	A100 (40GB)	16	∼48
CAFA-base	48k steps	A100 (40GB)	16	∼24
MMAudio	300k steps	H100	-	304
FoleyCrafter	164+80 epochs	-	128	-
Frieren	2.4M steps	2 × RTX 4090	-	-
VATT	“3 days” of training	A100 (80GB)	-	∼72
MultiFoley	650k steps	-	128	-

Table 6. Comparison of training costs across different models. CAFA employs a two-stage training approach (VGGSound followed by VisualSound fine-tuning), while CAFA-base uses single-stage training. FoleyCrafter uses separate semantic (164 epochs) and temporal (80 epochs) training stages. Frieren employs a three-stage approach, and MultiFoley uses a two-stage training method.

D. Asymmetric CFG (A-CFG)

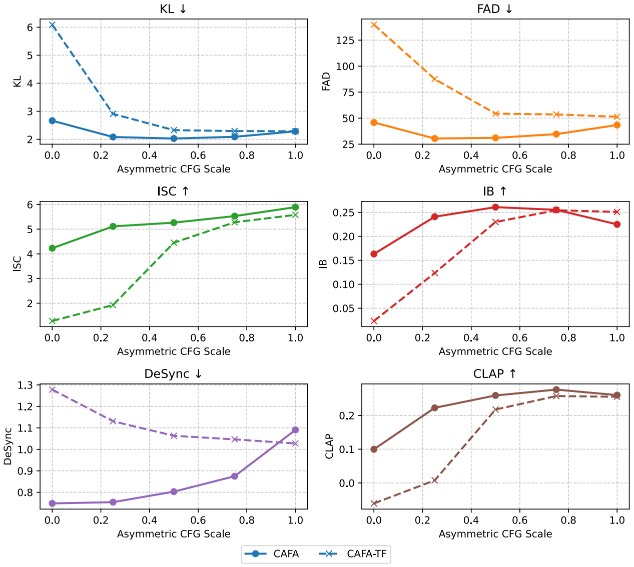


Figure 5. **Comparison of Asymmetric CFG Scaling Values.** CAFA denotes the default implementation using StableAudio-Open [9], while CAFA-TF refers to our adapter applied to TangoFlux [16]. For CAFA, applying CFG naively (e.g., $\gamma = 7$, $\alpha = 1$) yield high-quality audio with strong text alignment but reduce temporal consistency (DeSync). Lowering α to approximately 0.5 improves temporal alignment with minor effect on text fidelity or audio quality. While the effect is less pronounced for CAFA-TF, A-CFG remains a principled approach for tuning guidance.

E. Additional Figures

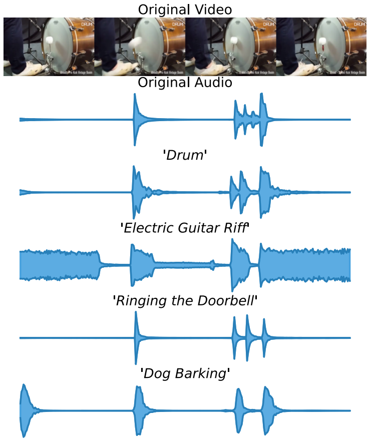


Figure 6. **CAFA creative control.** We demonstrate our method's ability to generate diverse, high-quality Foley sounds for videos through text prompts, ensuring temporal synchronization between audio and visual elements.

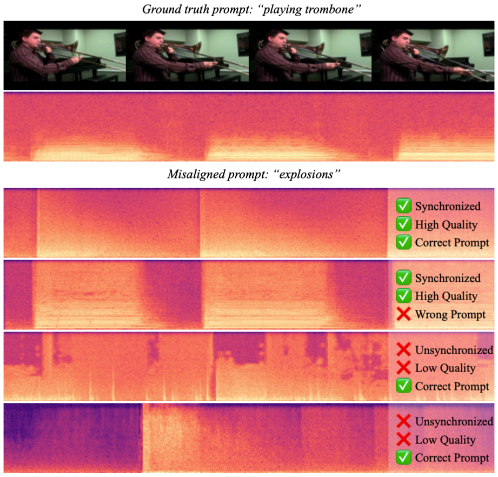


Figure 7. **Qualitative Comparison of Text-Video Disentanglement.** A video originally captioned as "playing trombone" (top) is paired with a mialigned prompt, "explosions", to generate audio. We analyze the results from four TV2A models, CAFA (ours), MMAudio, ReWaS, and FoleyCrafter, conditioned on the video and the new text prompt. The Evaluation focuses on synchronization, audio quality, and text-audio alignment. CAFA consistently demonstrates strong performance across all criteria, producing high-quality, well-synchronized audio accurately adhere to the requested target caption.