Conditional Generation of Audio from Video via Foley Analogies

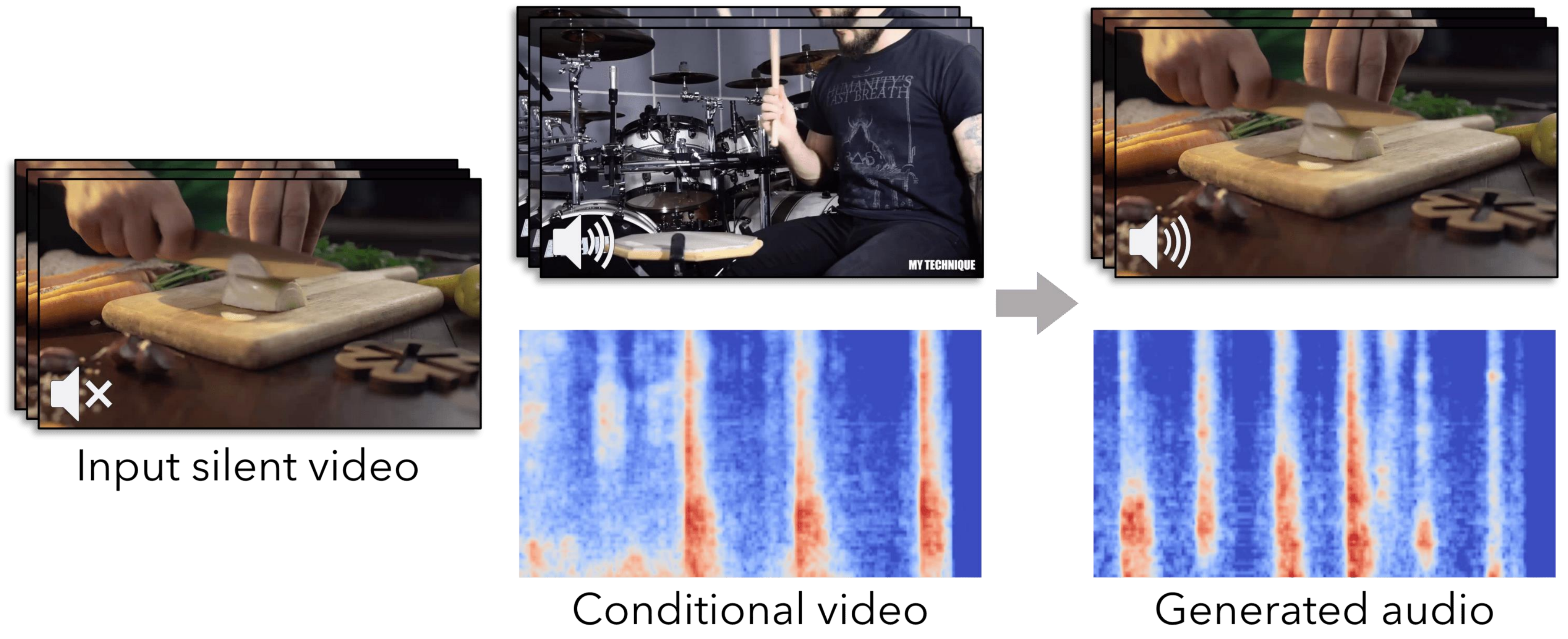
Yuexi Du[1,2] Ziyang Chen[1] Justin Salamon[3] Bryan Russell[3] Andrew Owens[1]
[1]University of Michigan [2]Yale University [3]Adobe Research

Webpage

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

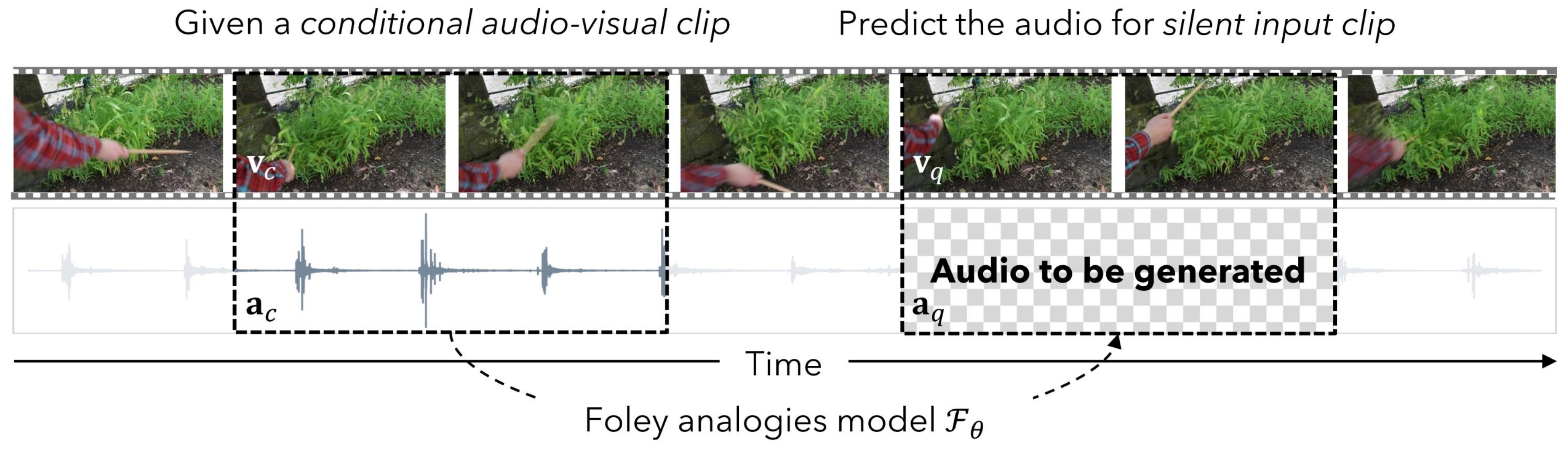
Introduction & Goal

Make silent input video sounds like the conditional video



Pretext task: Foley analogies

Idea: Natural videos tend to contain **repeated events** that produce closely related sounds.


Given a *conditional audio-visual clip*
Predict the audio for *silent input clip*
v_c
v_q
a_c
a_q
Audio to be generated
Time
Foley analogies model $\mathcal{F}_\theta$

Silent input video: $\mathbf{v}_q$

Conditional audiovisual clip: $(\mathbf{a}_c, \mathbf{v}_c)$

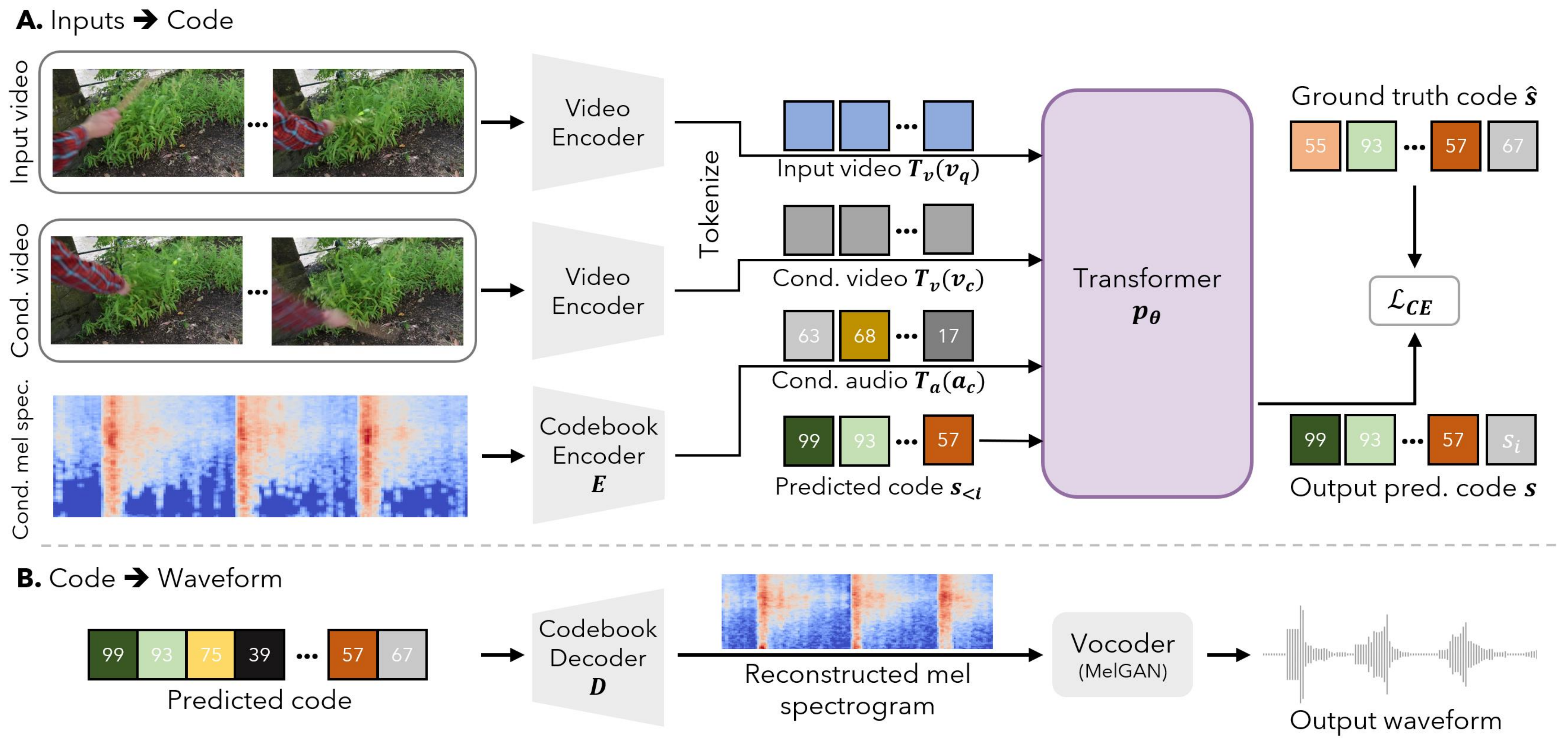
$\xrightarrow{\mathcal{F}_\theta}$ Output audio: $\mathcal{F}_\theta(\mathbf{v}_q, \mathbf{a}_c, \mathbf{v}_c)$

Learning Foley analogies:

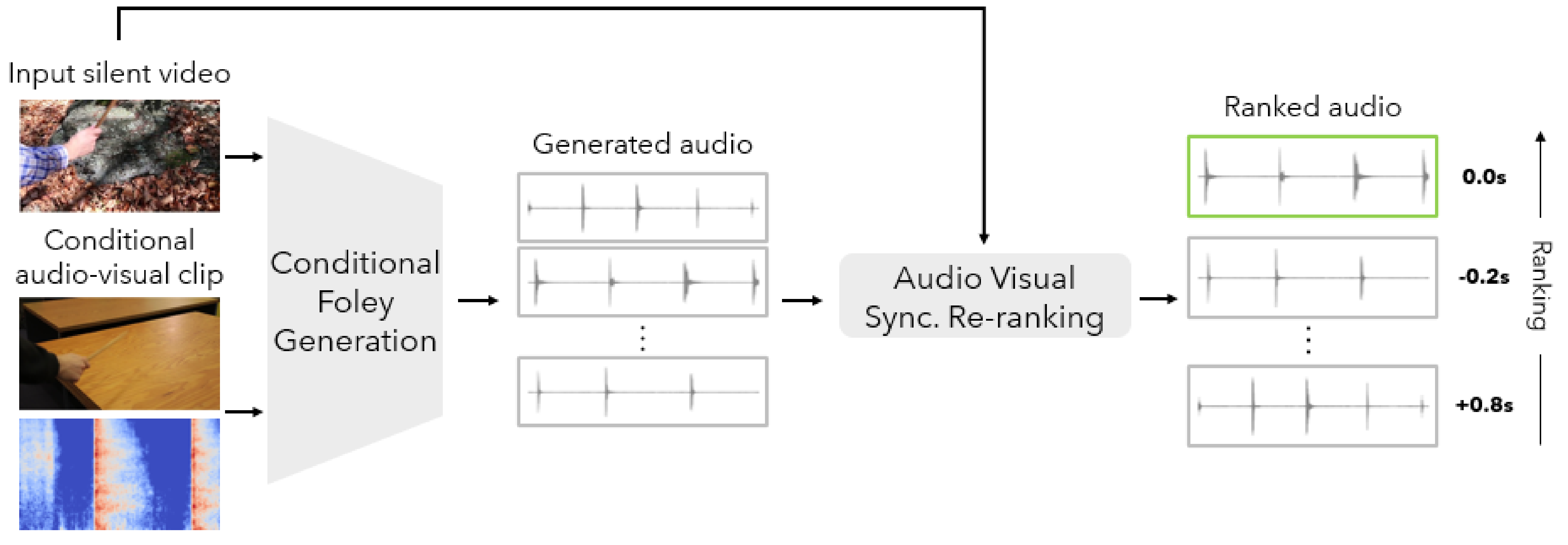
$$\arg\min_\theta \mathcal{L}\left(\mathbf{a}_q, \mathcal{F}_\theta(\mathbf{v}_q, \mathbf{a}_c, \mathbf{v}_c)\right)$$

Conditional audio prediction model

Idea: Predicted audio tokens auto-regressively with a GPT-2 transformer inspired by Spec-VQGAN (Iashin *et.al*, 2022) with silent input video and conditional audio-visual clip.

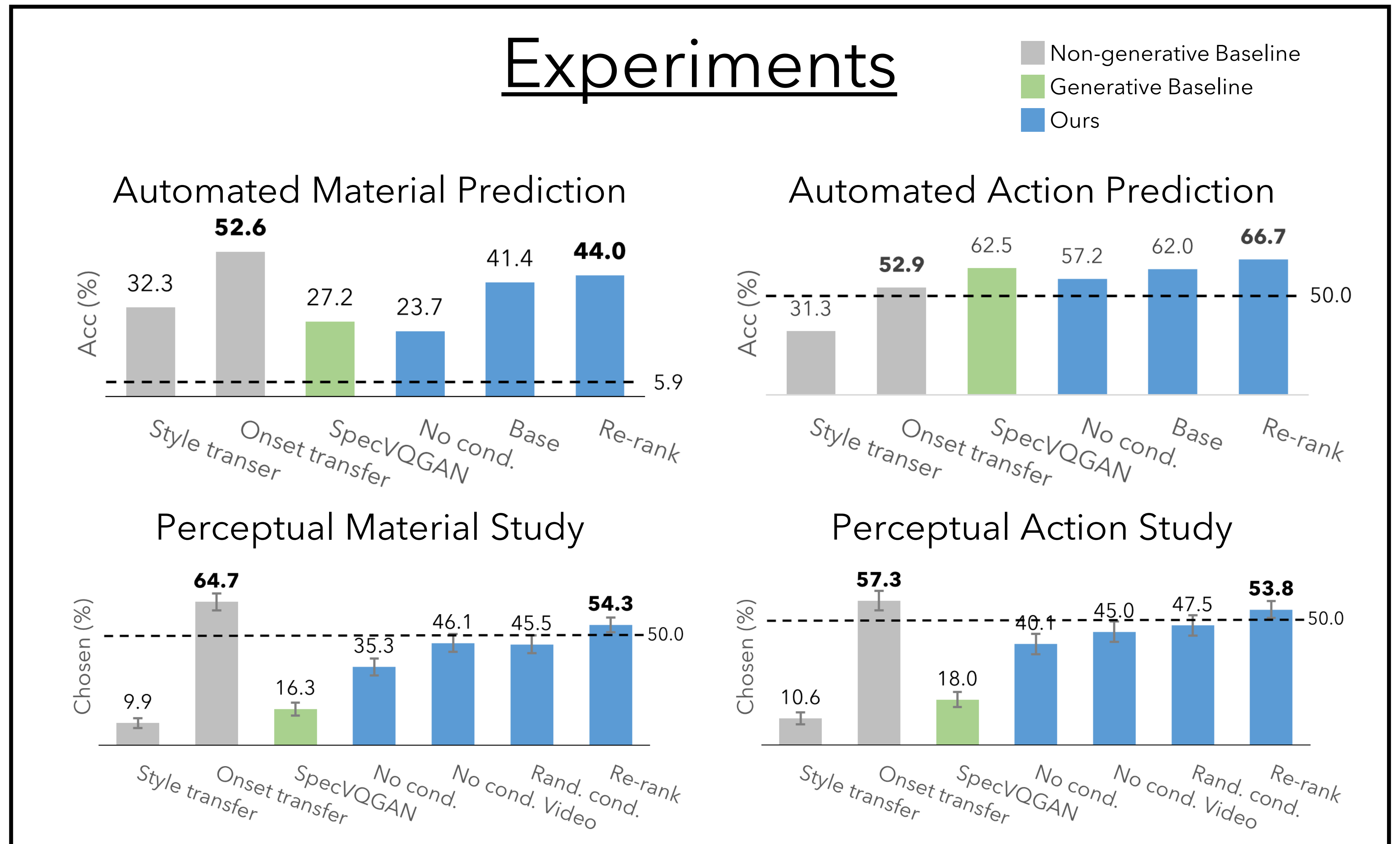


Inference-time audio re-ranking


Input silent video
Conditional audio-visual clip
Conditional Foley Generation
Generated audio
Audio Visual Sync. Re-ranking
Ranked audio
0.0s
-0.2s
+0.8s
Ranking

We re-rank multiple generated audio according to their **temporal alignment** with the input video predicted by off-the-shelf audio-visual synchronization model.

Experiments



Qualitative results (*Please check our website for video results*)