Looking Similar, Sounding Different:

Leveraging Counterfactual Cross-Modal Pairs for Audiovisual Representation Learning

Nikhil Singh
nsingh1@mit.edu

Chih-Wei Wu
chihweiw@netflix.com

Iroro Orife
iorife@netflix.com

Mahdi Kalayeh
mkalayeh@netflix.com

CVPR
JUNE 17-21, 2024
SEATTLE, WA

Summary

Dubbed movie scenes have varied speech while looking the same, mimicking real-world variation (e.g. different conversations in the same restaurant). We leverage this variation in speech to learn more robust and language-invariant features, showing this improves performance on various audio-visual tasks.

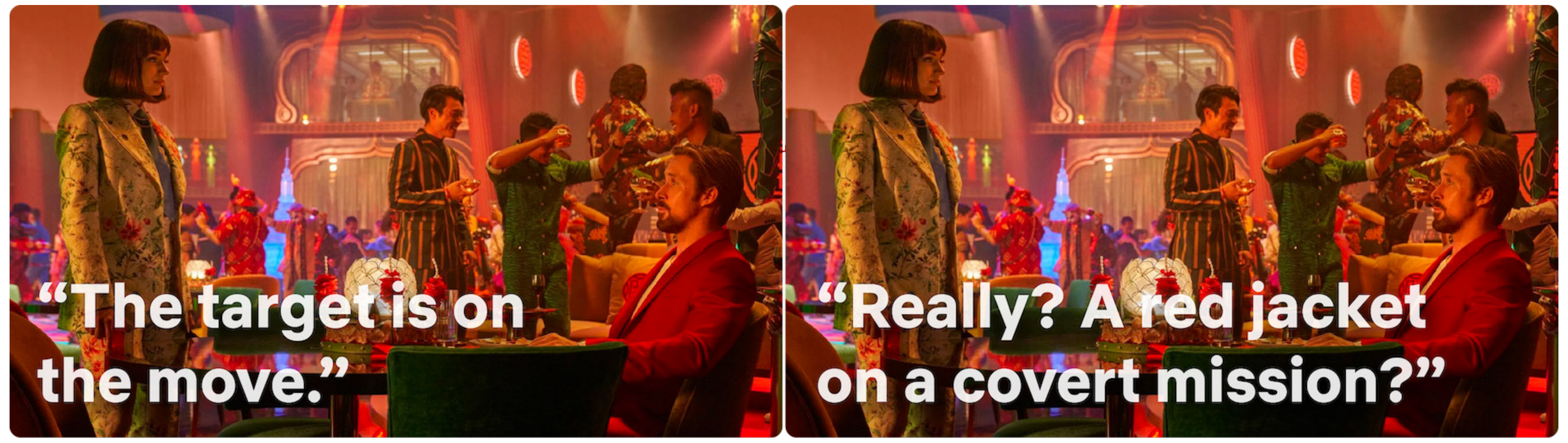
"The target is on the move."
"Really? A red jacket on a covert mission?"

Figure 1 (Counterfactual Pairs): Consider the pictured scene. Which of these dialog examples is more likely? Both are plausible within the scene, yet their phonetic-acoustic characteristics would create differences in the soundtrack. However, we do not have data that looks like this at scale!

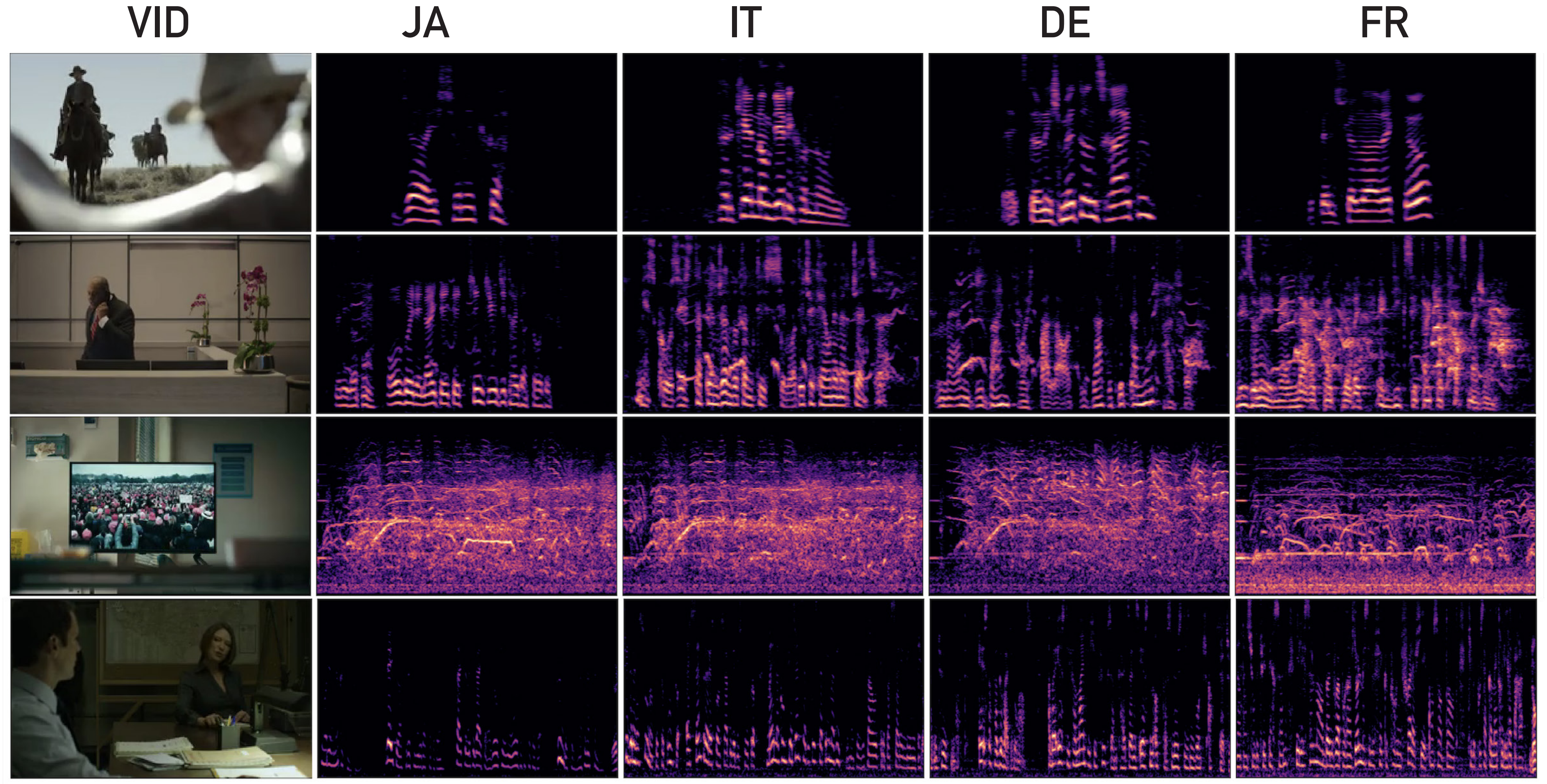

Figure 2 (Our Approach, Dubbed Audio): Movie dubs contain diverse audiovisual scenes, with varied speech content while preserving scene semantics. We leverage these to learn robust audiovisual representations.

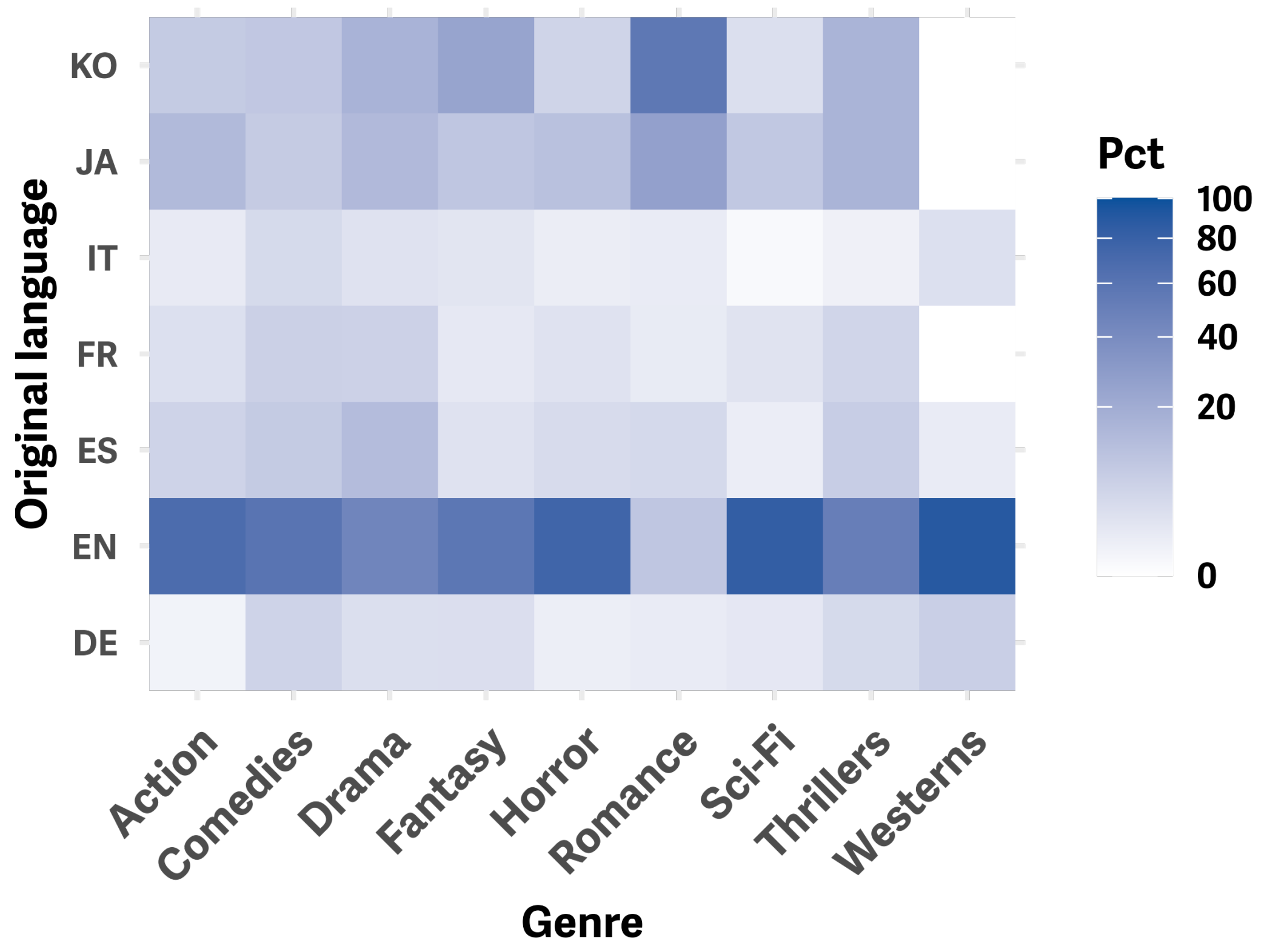

Figure 3 (Data Distribution): Movies & TV episodes in our pretraining dataset are chosen from diverse languages and genres. We aim to minimize potential content/story biases that could impact our self-supervised models. Beyond curating the dataset, we do not use this metadata for representation learning.

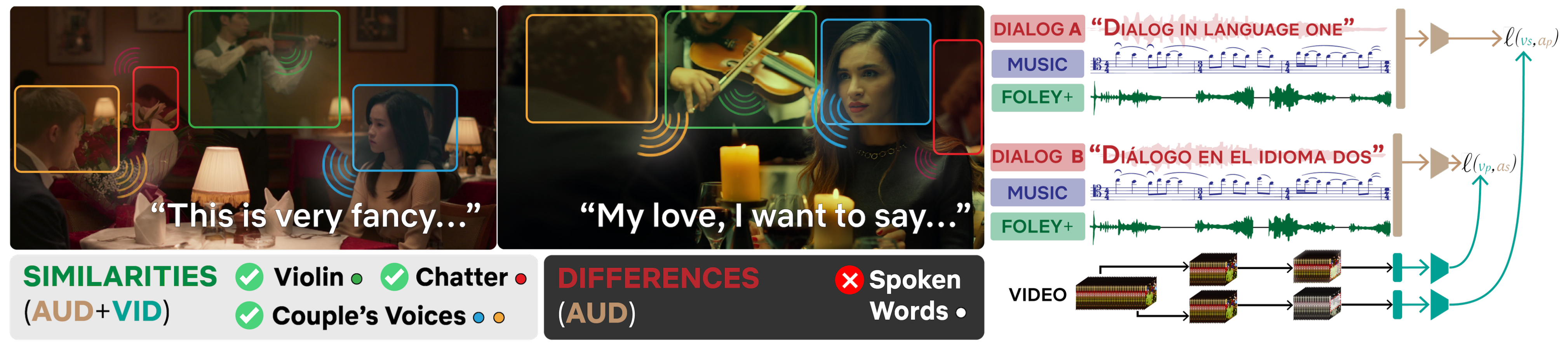

Figure 4 (Contrastive Training): We train Multiscale Vision Transformers (MViTs) for video and audio, using augmented and temporally jittered samples. We rely mainly on cross-modal training, with little to no within-modal contrast (weights for within-modal term are 0–0.2 depending on the model variant). Overall, we produce 11 model variants to investigate effects of data scale, data diversity, and model architecture depth.

Our training objective for minibatch of video v and audio a is given below in Eq. (1) & (2) and is a cross-modal variant of NT-Xent. Eq. (3) shows within-modal terms between augmented pairs. p = primary, s = secondary, z = embedding.

$$\ell_i(v,a) = -\log\left(\frac{e^{((z_v^i)^\top(z_a^i))/\tau}}{e^{((z_v^i)^\top(z_a^i))/\tau} + \sum_{(z_v',z_a')\in\mathcal{N}_i} e^{((z_v')^\top(z_a'))/\tau}}\right) \qquad (1)$$

$$\mathcal{L} = \frac{1}{2}\sum_{i=1}^{B}\Big(\ell_i(v_p,a_s) + \ell_i(v_s,a_p)\Big) \qquad (2)$$

$$\mathcal{L}_v = \sum_{i=1}^{B}\ell_i(v_p,v_s), \quad \mathcal{L}_a = \sum_{i=1}^{B}\ell_i(a_p,a_s) \qquad (3)$$

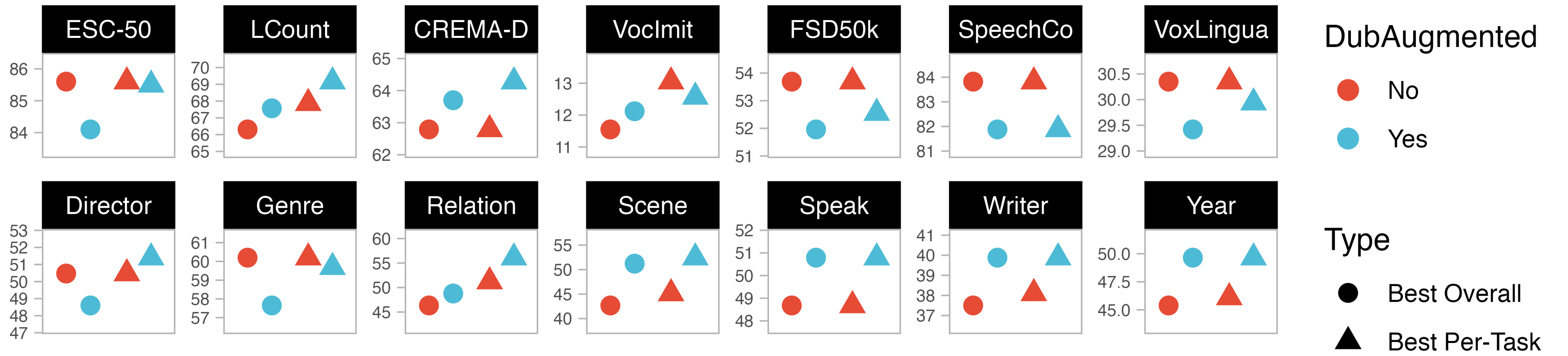

Figure 7 (Ablation Study): Average score without and with dubs in training, showing both the best overall model and the best per-task among our trained variants. Includes tasks for which we hypothesize improvements, and for which we hypothesize potential trade-offs (e.g. semantic speech tasks like Speech Commands and VoxLingua for which language-invariant training might pose challenges). Across the video tasks (bottom), the dubs lend an improvement on average. On audio tasks (top), improvements are also available on some tasks on average. All tasks are top-1 accuracy except mAP for FSD50k and VocImit.

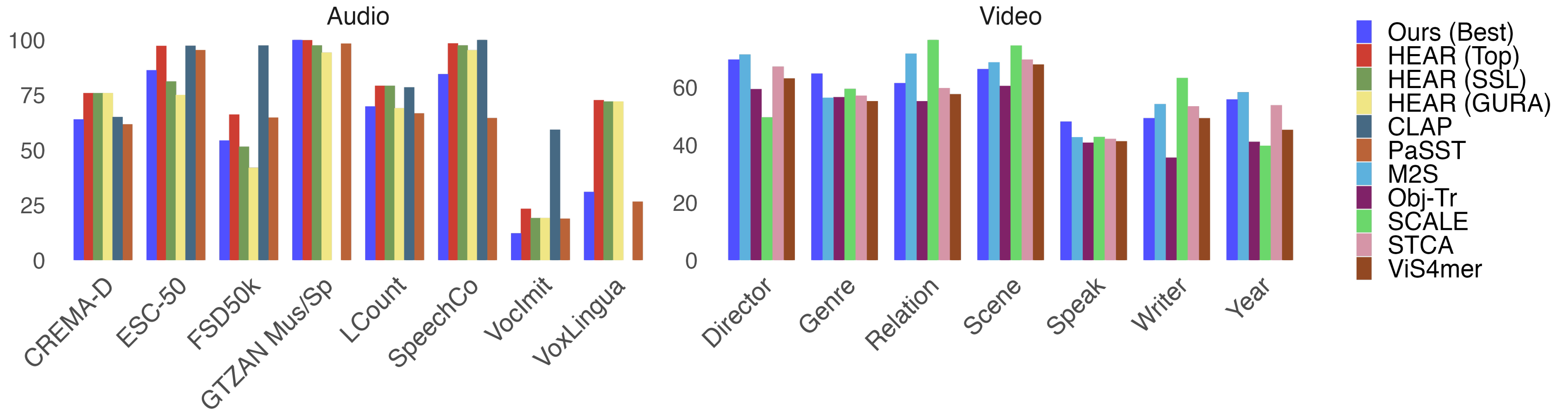

Figure 8 (SOTA Comparison): Models trained with our approach compare favorably to state-of-the-art results on a variety of audio tasks, and LVU video tasks. Ours (Best) is per-task best score.

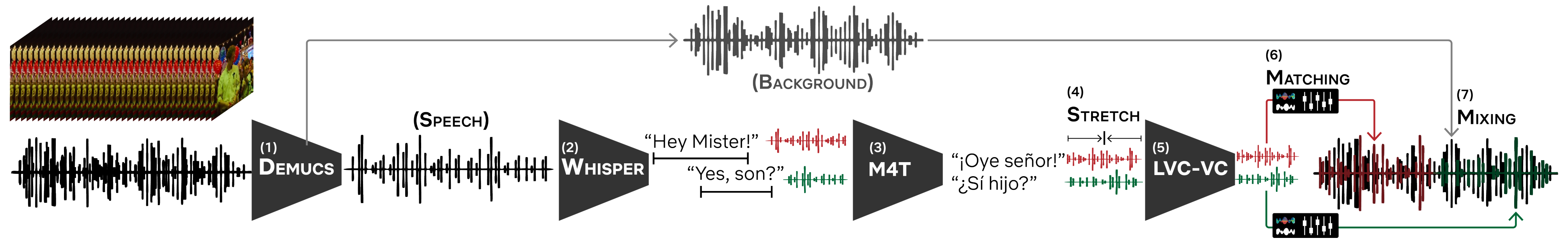

Figure 9 (Synthetic Counterfactual Pair Pipeline): We also propose a pipeline to produce synthetic counterfactual pairs from input video content (e.g. videos from LVU). This combines speech recognition, translation, alignment, voice conversion, and audio matching and mixing together, allowing experimenting with synthetic counterfactual pairs openly and at scale.