Supplementary Material
Robust Audio-Visual Instance Discrimination

1. Parametric studies

We provide a parametric study of key Robust-xID hyper-parameters.

Weight function shape parameter δ One critical parameter of Weighted-xID is the shape parameter δ, which specifies the mid-point location of the weight function. For example, when $\delta = -2$, the midpoint is located at $\mu - 2\sigma$ where μ and σ are the sample mean and standard deviation of the scores $\bar{\mathbf{v}}_i^T \bar{\mathbf{a}}_i$. This means that for $\delta = -2$, the majority of samples will have a weight of 1, and only a small fraction will have a weight close to $w_{\min}$. As δ increases, the proportion of samples that are down-weighted also increases. To study the impact of δ, we trained several models using Weighted-xID with different values of δ and for different amounts of injected faulty positives n_0. Other hyper-parameters were kept at their default values $w_{\min} = 0.25$ and $\kappa = 0.5$. The transfer performance is shown in Figure 1. As can be seen, the proposed robust xID procedure is not very sensitive to this hyper-parameter. This suggests that Robust-xID can help representation learning as long as clear faulty positives are suppressed.

Soft-xID: Mixing coefficient The mixing coefficient λ specifies the degree to which the one-hot targets of instance discrimination are softened in Soft-xID. The one-hot instance discrimination targets are used when $\lambda = 0$. As λ increases, the softening scores $S(j|i)$ are increasingly used to adjust the one-hot targets. To study the impact of the mixing coefficient λ, we trained several models using Soft-xID with various values of λ. Cycle consistent targets were used as the softening strategy. Figure 2 shows the transfer performance of the learned models on UCF and HMDB under the fine-tuning and retrieval protocols. The trend is consistent across the two datasets and two evaluation protocols. Softening the instance discrimination targets enhances representation learning, with the optimal performance achieved with a mixing coefficient between 0.25 and 0.5. However, as the mixing coefficient increases substantially $\lambda > 0.65$, the targets are derived from the model prediction alone and disregard instance labels. In this case of large λ, the pre-training fails completely,

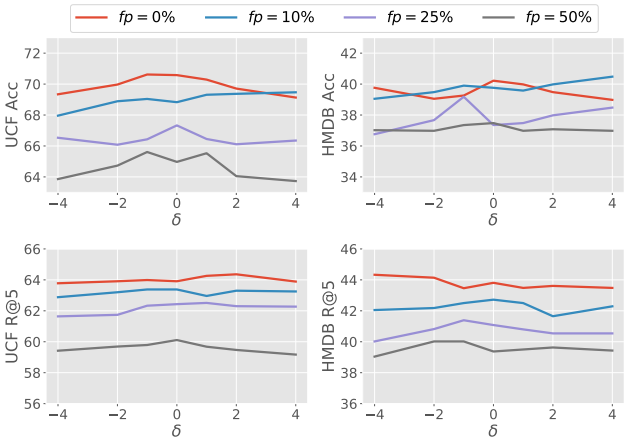

Figure 1: Effect of shape parameter δ in Weighted-xID. Transfer learning performance is evaluated on two datasets (UCF and HMDB) under two protocols (full finetuning and retrieval). For the fine-tuning protocol, we report final accuracy of video level predictions. For the retrieval protocol, we report R@5.

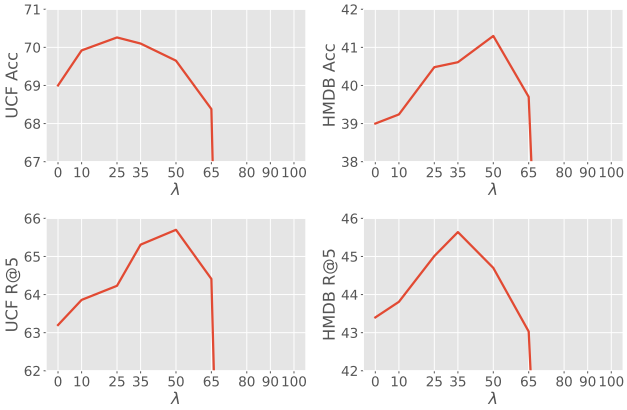

Figure 2: Effect of mixing coefficient λ in Soft-xID. Transfer learning performance is evaluated on two datasets (UCF and HMDB) under two protocols (full finetuning and retrieval). For the fine-tuning protocol, we report final accuracy of video level predictions. For the retrieval protocol, we report R@5.

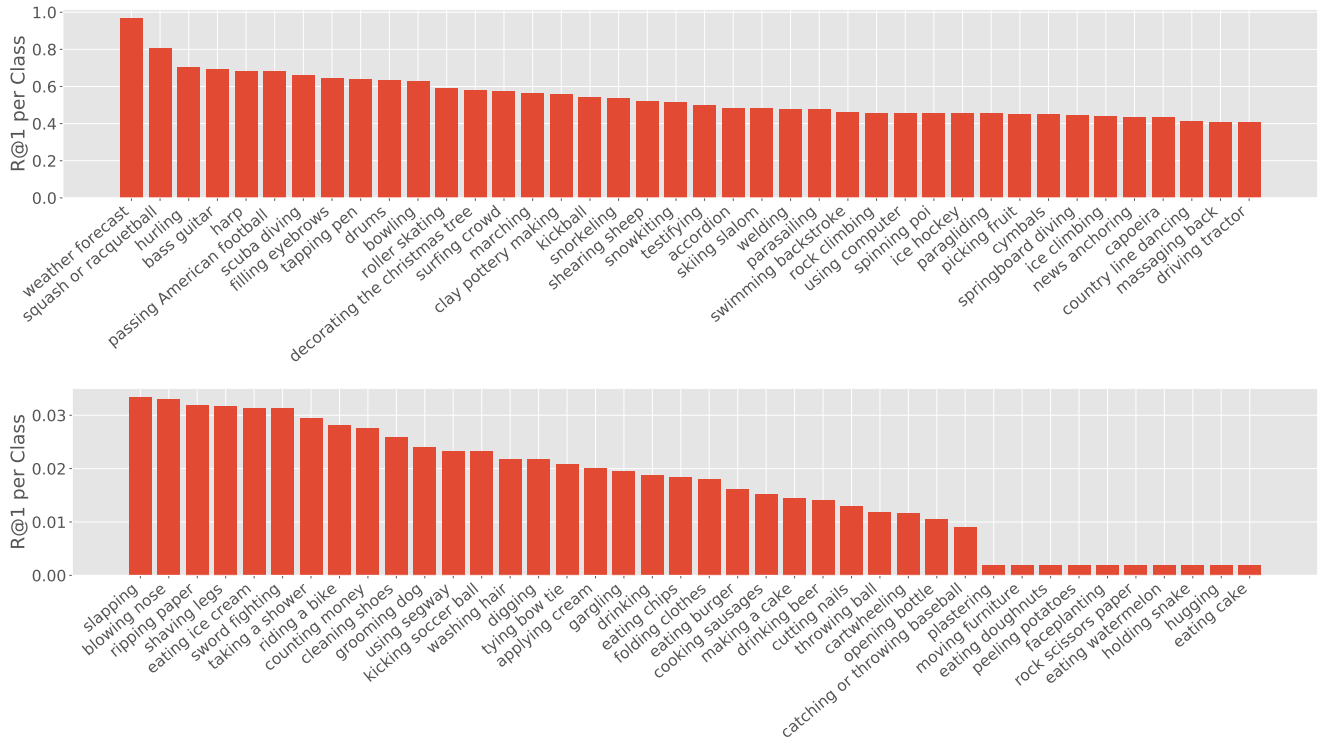

Figure 3: Best and worse Kinetics classes. For each class, we depict the top-1 retrieval performance (R@1) averaged across all images of each class. The plot above shows the top 40 classes and the plot below the bottom 40 classes.

i.e., the learned representations have very low transfer performance.

2. Additional analysis

The proposed approach learns high-quality feature representations that can be used to discriminate several action classes. This was shown in the main paper by reporting transfer learning results. We now provide additional qualitative evidence and analysis.

Retrieval For each video, we extracted $4 \times 4 \times 512$ feature maps from the video encoder learned using Robust-xID on the full Kinetics dataset. Figure 4 depicts the top 4 closest videos for several query samples. As can be seen, Robust-xID produces highly semantic features, enabling correct retrievals for a large number of videos spanning a large number of classes. Furthermore, even when a video of a different class is retrieved, the errors are intuitive (for example, the confusion between 'American football' and 'Hurling' in the third row). Failure cases also seem to be correlated with classes that are hard to distinguish from the audio alone (eg, different types of kicking sports or swimming strokes).

Class-based analysis To better understand which classes are better modeled by the Robust-xID framework, we measured the top-1 retrieval performance (R@1) averaged across all images of each class. Similar to the analysis above, each video is represented by a $4 \times 4 \times 512$ feature map extracted from a video encoder learned using Robust-xID on the full Kinetics dataset. Figure 3 depicts a list of Kinetics classes sorted by their average R@1 score. As can be seen, action classes which are often accompanied by long and distinctive sounds (*e.g.*, squash, harp, drums, accordion, or scuba diving) tend to be more easily distinguished from others. In contrast, classes with less distinctive audio (*e.g.*, making a cake, eating cake, or hugging) or classes where distinctive sounds are short-lived (*e.g.*, blowing nose, gargling or kicking ball) are harder to model using a cross-modal audio-visual framework. As a result, the features learned for such classes are less discriminative.

Faulty positive detection performance To obtain a rough estimate of performance of the faulty positive detection procedure, we randomly sampled 100 videos from the 10000 most likely faulty positives, as identified by Robust-xID trained on the full Kinetics

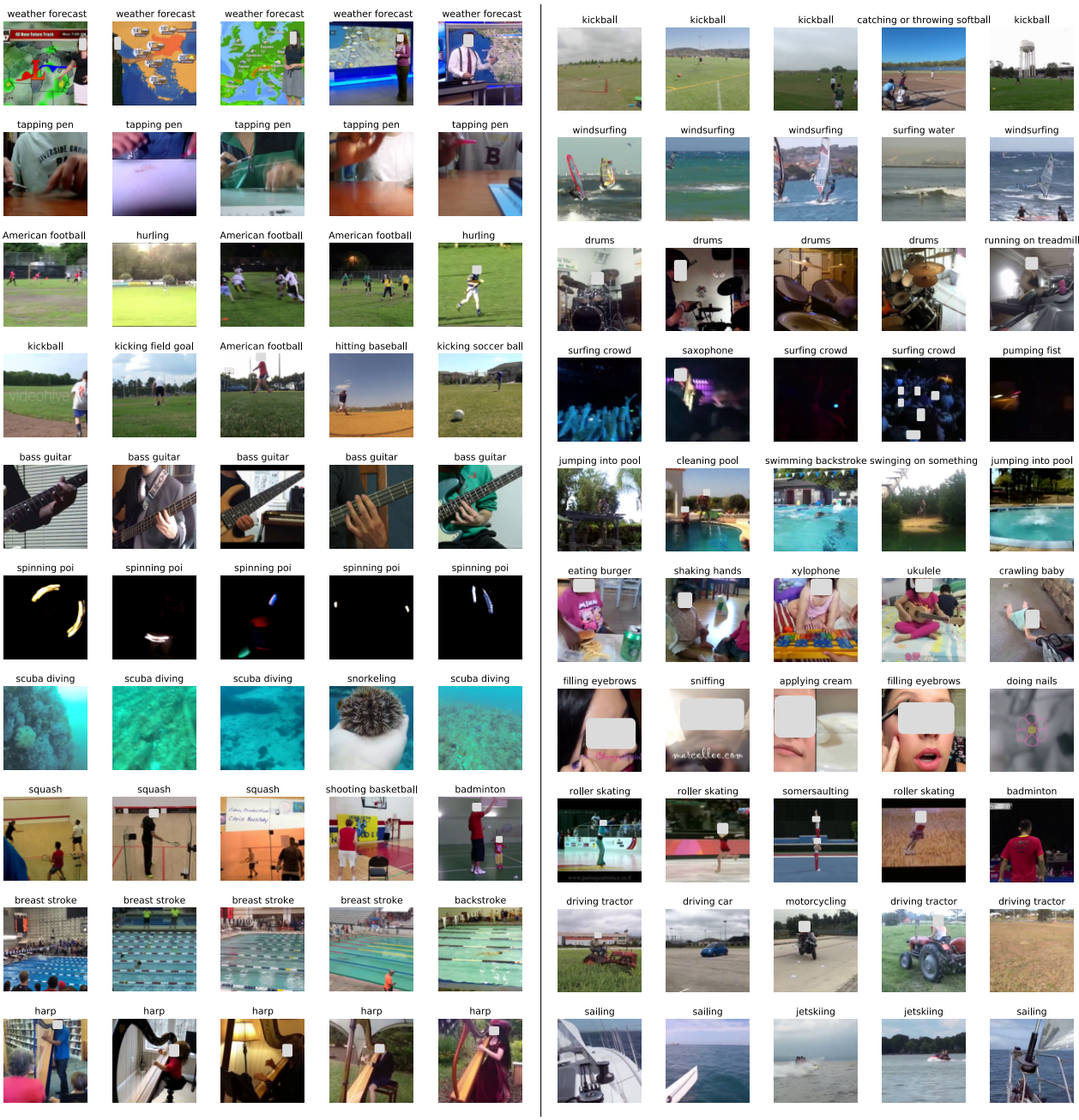

Figure 4: Retrievals. In each row, the first image depicts the query video, and the following four images depict the top 4 retrievals. The corresponding Kinetics class description is provided above each frame. Each video is represented by a $4 \times 4 \times 512$ feature map produced by the video encoder learned using Robust-xID on the full Kinetics dataset. Euclidean distance is used to determine video similarity.

dataset. We then manually labeled them according to how related their audio and visual signals are. From those, 67 were clear faulty pairs; 24 contained narrative voice-overs (*i.e.*, required natural language understanding to link the two modalities); and 9 samples were clearly misidentified.