# Supplementary Document
# See the Sound, Hear the Pixels

## S.1 Accuracies of individual event categories

The performance of each individual event category was observed for three variants of our proposed model (supervised learning) that uses: only audio, only visual and audio-assisted visual with audio features respectively. The comparison of accuracies of the three variants of our model is given in table S.1.

Our proposed model that uses audio-assisted visual with audio features, outperforms other model that uses only audio or only visual features for 22 out of 28 event categories. Out of the remaining 6 event categories, our model that uses only audio features gives superior results for 4 event categories (man, ukulele, guitar, mandolin) while that which uses only visual features dominates for 2 event categories (truck, horse).

However, even for those 6 event categories, the accuracies obtained by our model that uses audio-assisted visual with audio features, are close to that of the best performing models (that use only audio or only visual features) in most cases. On the whole, our proposed model that uses audio-assisted visual with audio features ensures best overall results.

## S.2 Qualitative results

Figures S.1 to S.5 show the attention maps obtained using our proposed method that uses Audio Visual Triplet Gram Matrix Loss (AVTGML) function. This function facilitates learning the attention in an unsupervised way. The attention maps are shown for videos taken across different event categories.

Figure S.5 shows an example where the attention maps generated by our unsupervised algorithm aren't precise. The frames in the figure are obtained from a video which belongs to the 'horse' event category. The audio heard is that of the horse softly neighing and its hooves clicking against the ground. The attention maps are inaccurate because of the sound being intermittent and also of low quality.

But in most of the cases, the attention maps are quite precise even without the use of event labels, which is evident from the impressive visual results (figures S.1-S.4). Also, the attention maps are consistent in accurately pointing to the object producing sound in the scene, across all segments containing the audio-visual event (that is both audible and visible).

| Model | bell | man speaking | dog | plane | racing /car | woman speaking | helicopter |
|---|---|---|---|---|---|---|---|
| Only audio | 98.6 | 80.2 | 52.8 | 58.1 | 46.9 | 75.6 | 48.9 |
| Only visual | 97.1 | 16.7 | 21.3 | 72.5 | 60.9 | 50.4 | 37.6 |
| Aud + Aud-ass. visual | 98.7 | 77.8 | 66.3 | 86.9 | 85.1 | 78.0 | 64.7 |

| Model | violin | flute | ukulele | frying | truck | shofar | motorcycle |
|---|---|---|---|---|---|---|---|
| Only audio | 70.9 | 90.7 | 75.6 | 77.0 | 31.4 | 54.5 | 61.3 |
| Only visual | 53.0 | 41.0 | 25.0 | 88.5 | 93.5 | 42.0 | 66.7 |
| Aud + Aud-ass. visual | 86.1 | 95.0 | 72.6 | 92.2 | 76.7 | 59.1 | 92.0 |

| Model | guitar | train | chainsaw | banjo | goat | bus | baby crying |
|---|---|---|---|---|---|---|---|
| Only audio | 69.7 | 72.1 | 80.1 | 72.0 | 46.1 | 8.3 | 52.4 |
| Only visual | 64.6 | 81.0 | 84.5 | 57.1 | 51.7 | 63.9 | 39.7 |
| Aud + Aud-ass. visual | 67.4 | 93.7 | 96.2 | 74.9 | 51.7 | 77.9 | 63.5 |

| Model | clock | cat | horse | toilet | rodent | accordian | mandolin |
|---|---|---|---|---|---|---|---|
| Only audio | 76.2 | 9.1 | 11.6 | 61.5 | 41.2 | 72.0 | 73.5 |
| Only visual | 85.6 | 2.0 | 65.1 | 64.2 | 47.5 | 81.3 | 17.0 |
| Aud + Aud-ass. visual | 90.6 | 33.3 | 39.5 | 87.1 | 63.8 | 95.3 | 64.7 |

Table S.1: **Performance (in %) comparison of individual event categories** on three variants of our proposed model (that uses supervised learning for event localization) that uses: only audio, only visual, audio-assisted visual plus audio features respectively. In most of the cases, our model that uses audio-assisted visual with audio features outperforms other models that use only audio or only visual features.

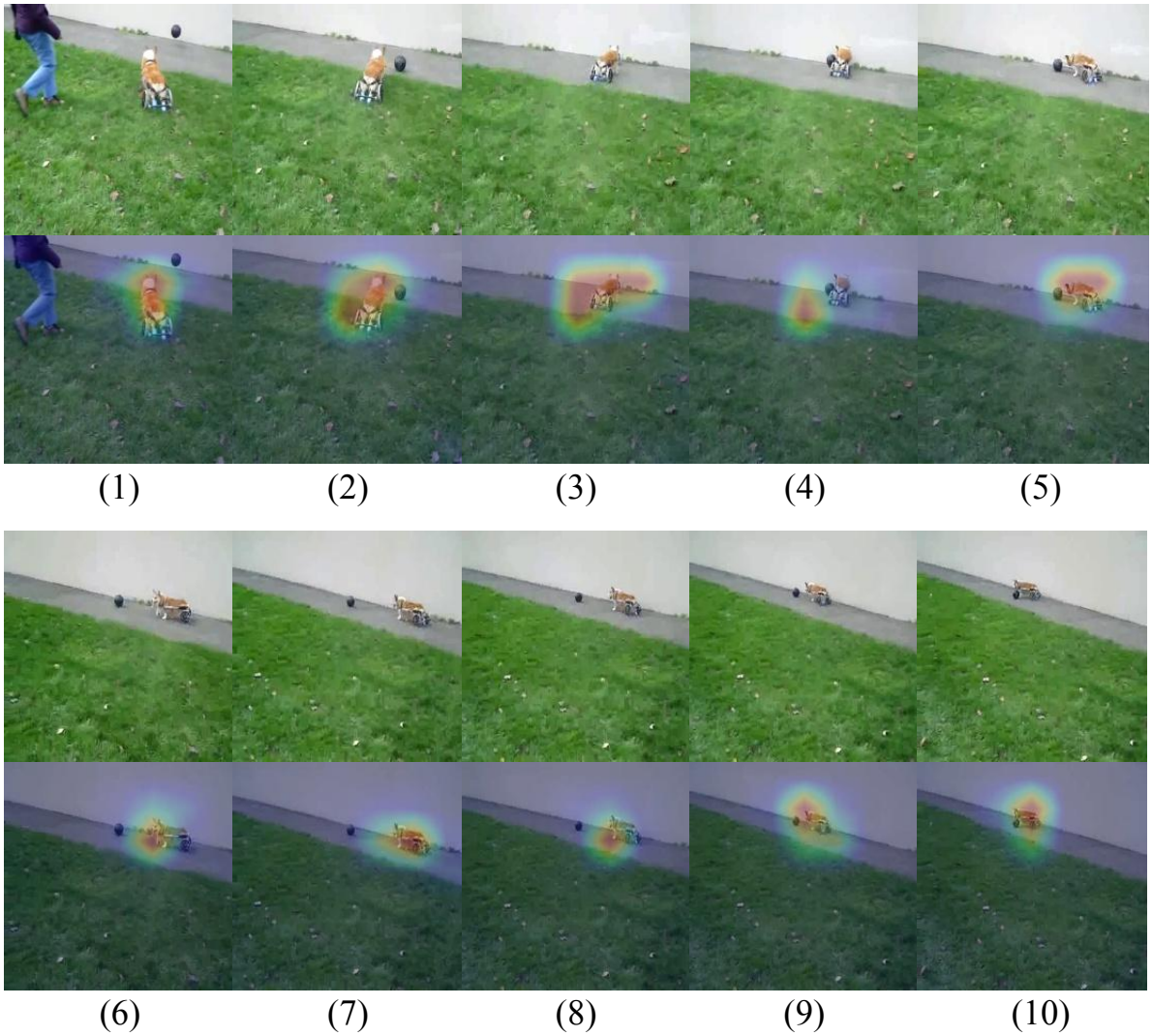


Figure S.1: Input frames along with their attention maps are shown for each of the 10 segments of a video belonging to the **event category "dog"**. The attention maps are obtained from the unsupervised sound source localization task that uses our proposed Audio Visual Triplet Gram Matrix Loss (AVTGML) function.

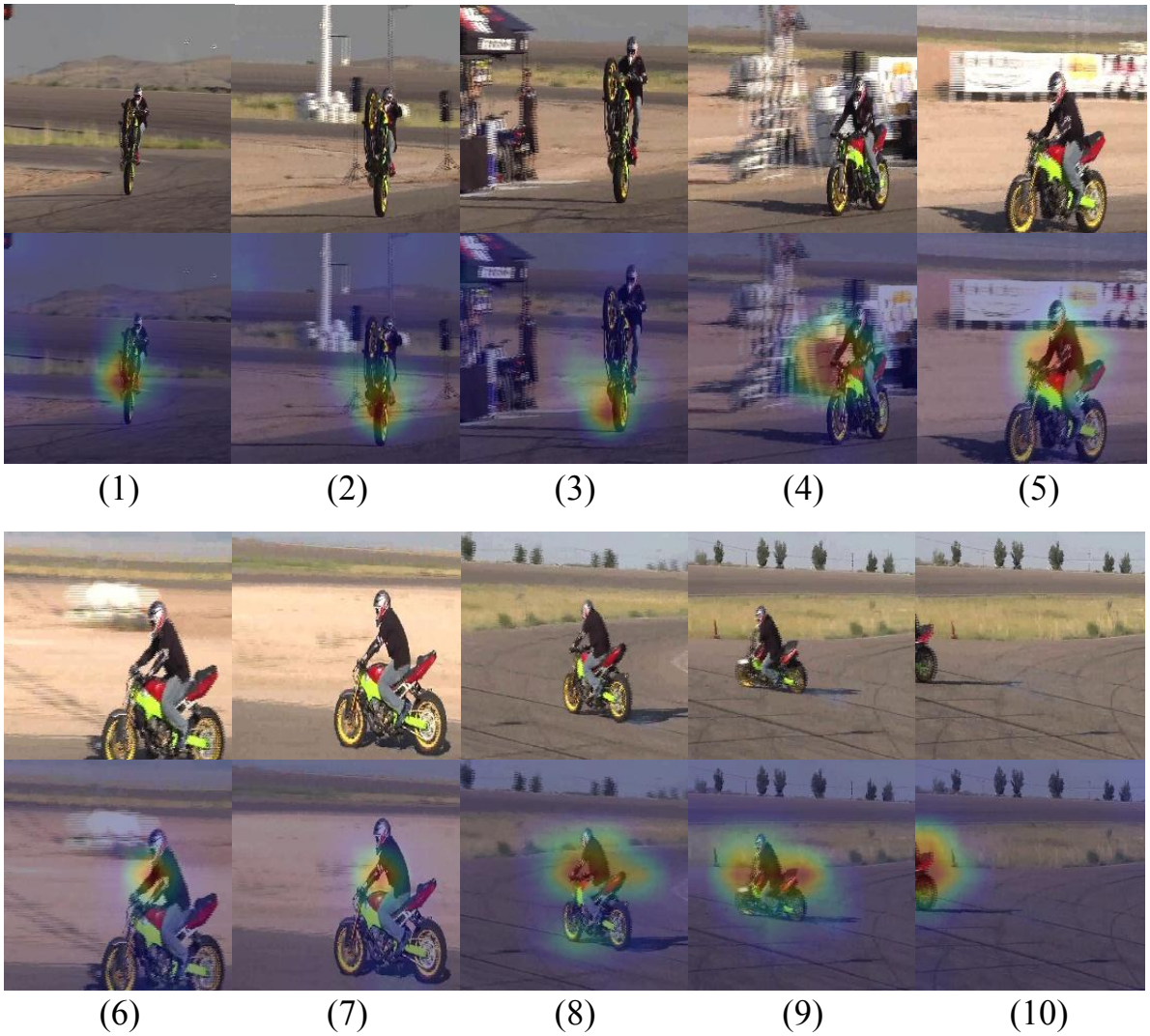


Figure S.2: Input frames along with their attention maps are shown for each of the 10 segments of a video belonging to the **event category "racing"**. The attention maps are obtained from the unsupervised sound source localization task that uses our proposed Audio Visual Triplet Gram Matrix Loss (AVTGML) function. (See equations 13-16 in the main document.)

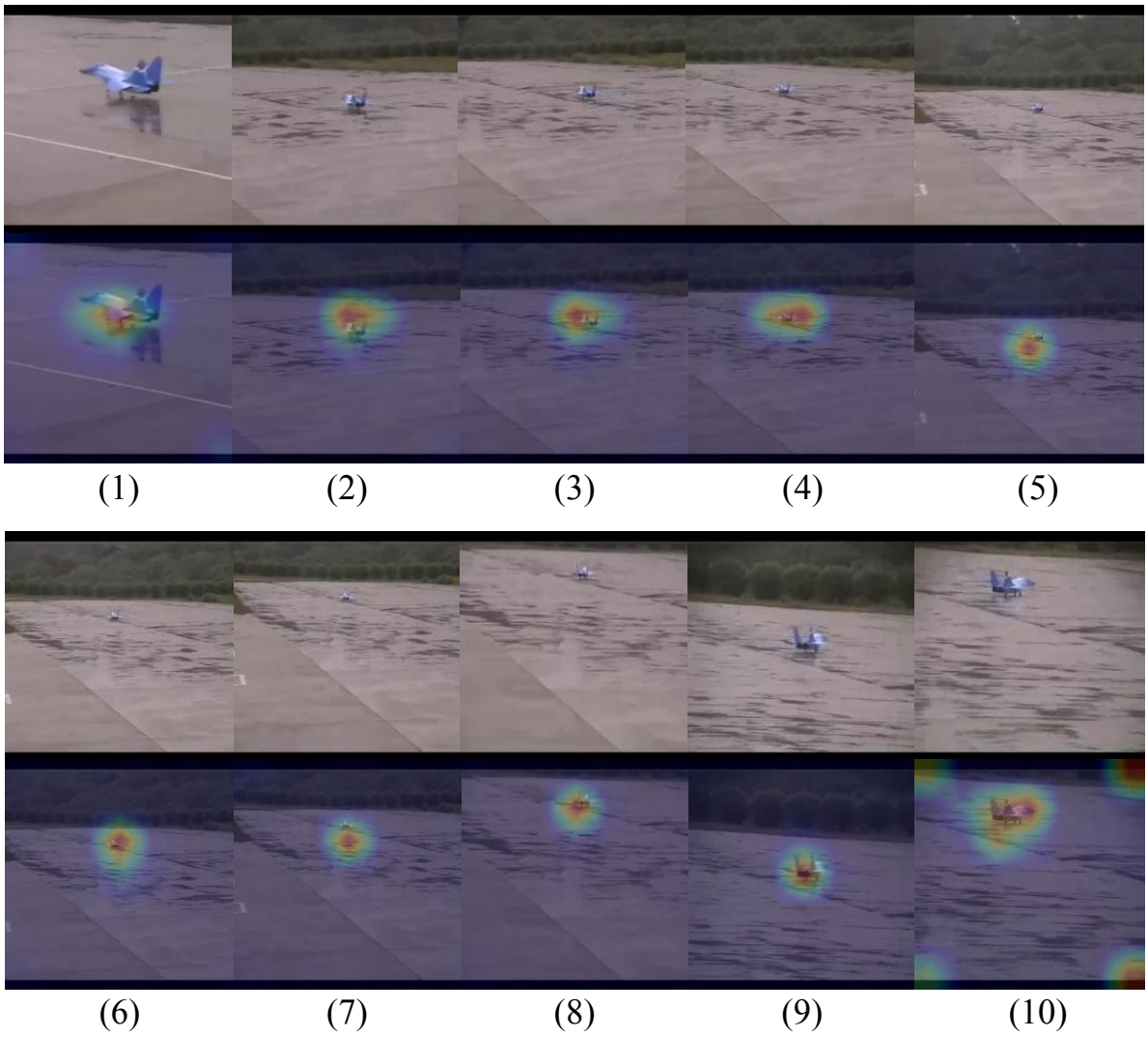


Figure S.3: Input frames along with their attention maps are shown for each of the 10 segments of a video belonging to the **event category "plane"**. The attention maps are obtained from the unsupervised sound source localization task that uses our proposed Audio Visual Triplet Gram Matrix Loss (AVTGML) function. (See equations 13-16 in the main document.)

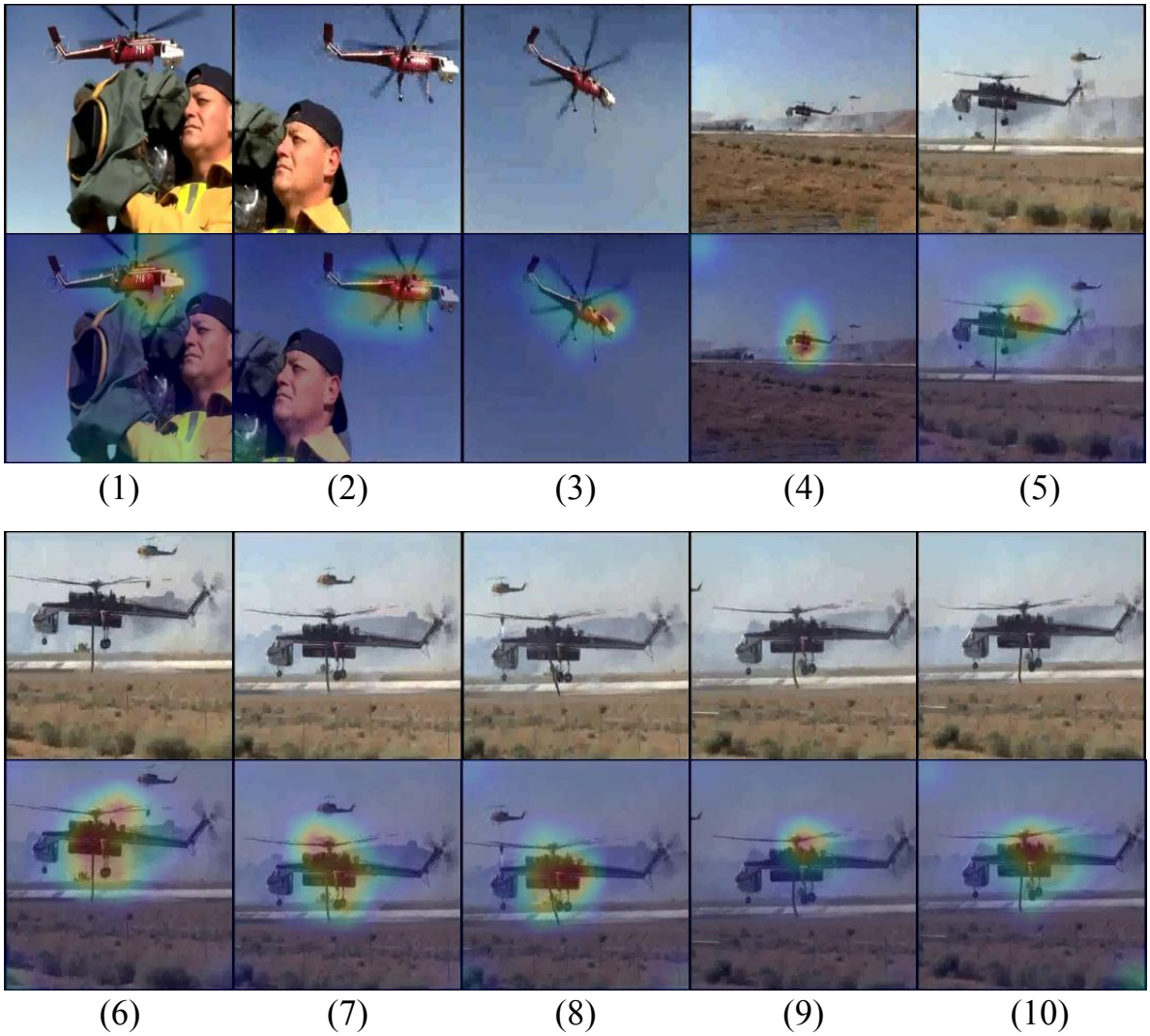


Figure S.4: Input frames along with their attention maps are shown for each of the 10 segments of a video belonging to the **event category "helicopter"**. The attention maps are obtained from the unsupervised sound source localization task that uses our proposed Audio Visual Triplet Gram Matrix Loss (AVTGML) function. The attention maps also show that the model has learnt to localize sound source by using both audio and visual content and not based on the salient objects present in the scene.

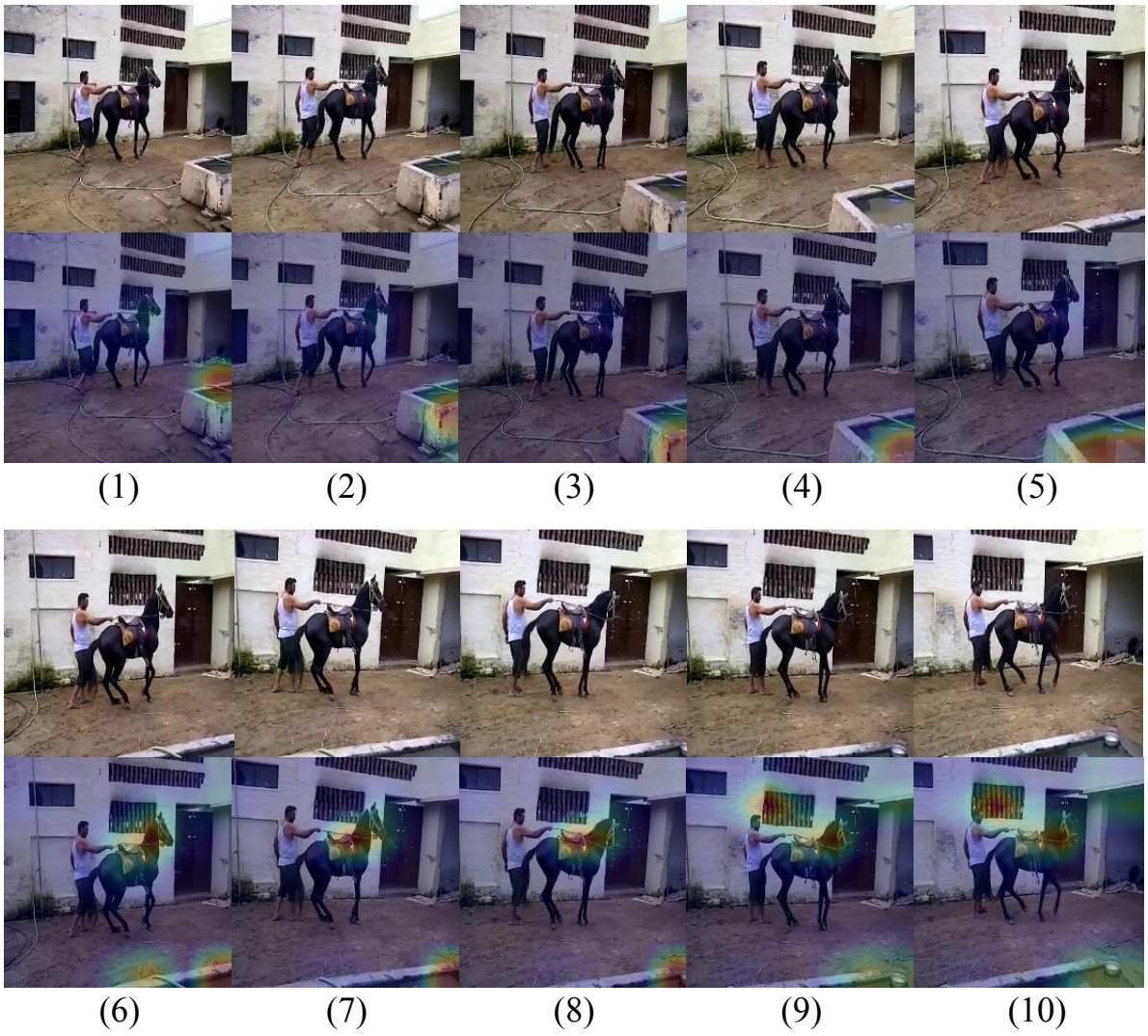


Figure S.5: Input frames along with their attention maps are shown for each of the 10 segments of a video belonging to the **event category "horse"**. The attention maps are obtained from the unsupervised sound source localization task that uses our proposed Audio Visual Triplet Gram Matrix Loss (AVTGML) function. The inaccurate attention maps are due to the sound being intermittent and of low quality.