Supplementary Material: MovieCuts

Alejandro Pardo[1], Fabian Caba Heilbron[2], Juan León Alcázar[1], Ali Thabet[1,3], and Bernard Ghanem[1]

[1] King Abdullah University of Science and Technology, KAUST
{alejandro.pardo,juancarlo.alcazar,ali.thabet,bernard.ghanem}@kaust.edu.sa
[2] Adobe Research {caba}@adobe.com
[3] Facebook Reality Labs {thabetak}@fb.com

1 MovieCuts Attributes Details

We leverage CLIP [3] to extract visual attributes from all instances in MovieCuts. We follow the zero-shot setup described in [3]. To do so, we create language queries using relevant classes, for each attribute type, as a set of candidate text-visual pairs and use CLIP's dual encoder to predict the most probable pair (the most probable tag). Thus, we compute an image embedding for the visual frames, and a text embedding for all candidate text queries (attribute tags) to then compute the cosine similarity between the L2-normalized embedding pairs. Instead of simply passing the tags to the language encoder, we augment the text queries using the following template: "a photo of a *subject attribute*", and "an *location attribute* photo" for the subject and location attributes, respectively. We retrieve tags for each of its shots by sampling a random frame before and after the shot transition from each cut.

Actions that trigger Cuts. Our goal is to find correlations between action tags and cut types. To do so, we first build a zero-shot action classifier based on CLIP [3]. Since the zero-shot action classifier did not offer us a high accuracy, we limit our analysis with the most confident tags only. Such tags allow us to find the most common co-occurrences between actions and cut-types. Figure 1 showcases three common action/cut pairs. These patterns are common across different movie scenes and editors' styles. These empirical findings reaffirm the theory of the film grammar [1,4], which suggests that video editing follows a set of rules more often than not.

2 Additional Results

Distribution-Balanced Loss Experiments. We experiment with the Distributed-Balanced Loss (DB Loss) [6] introduced by Wu *et al.* . The DB Loss was designed to tackle datasets with multiple labels per sample that follow a long-tail distribution. It proposes a modification to the standard binary cross-entropy loss by adding two terms to handle multiple labels and long-tail distributions. For

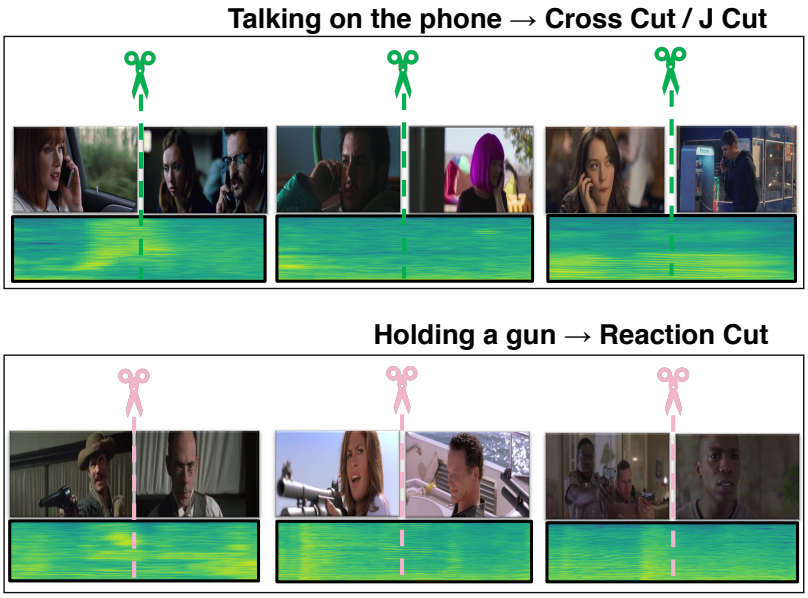


Fig. 1: **Actions that trigger cuts.** Some actions predominantly co-occur with a particular cut type. For instance, a Talking on the phone action is often edited via the Cross Cut and J Cuts. Another common pattern emerges when someone is Holding a gun. The predominant edit is the Reaction Cut, which first shows an actor holding the gun, and the next shot highlights a face reaction of another subject.

Model	Sampling	**mAP**	**CA**	**CW**	**CC**	**EC**	**MC**	**SC**	**RC**	**LC**	**JC**	**SC**
AV+Scaled GB	Uniform	47.2	64.3	62.7	32.3	31.2	2.6	23.5	83.0	44.5	51.6	76.3
AV+Scaled GB	Gaussian	47.4	64.0	62.7	34.4	**32.0**	2.1	23.8	82.7	43.7	**51.9**	77.0
AV+Scaled GB	Fixed	**47.9**	**65.6**	**63.0**	**34.9**	31.8	**2.3**	**24.4**	**83.3**	**45.0**	51.6	**77.1**

Table 1: **Window Sampling Results**. Different window sampling strategies, using the Audio-Visual + Scaled Gradient Blending Model [5]. All the reported numbers are % AP. Showing classes: Cutting on Action (**CA**), Cut Away (**CW**), Cross Cut (**CC**), Emphasis Cut (**EC**), Match Cut (**MC**), Smash Cut (**SC**), Reaction Cut (**RC**), L Cut (**LC**), J Cut (**JC**), Speaker Chance (**SC**).

further details, including the loss formulation, please refer to the original publication [53]. For the experiments shown in Table 2 we upgrade our base BCE loss by the DB Loss. For a fair comparison, we scale the original naive combination by 3.0 (to put the losses' magnitudes around the same scale). We observe that the DB Loss helps the base model and improves the mAP on most of the classes.

Model	**mAP**	**CA**	**CW**	**CC**	**EC**	**MC**	**SC**	**RC**	**LC**	**JC**	**SC**
AV Scaled	47.4	65.1	62.7	33.1	31.4	1.8	23.0	83.0	45.4	50.7	77.6
AV Scaled+DB Loss	47.8	65.5	63.0	35.0	31.7	1.9	23.4	83.1	45.7	50.8	77.7
AV+Scaled GB+DB Loss	**47.9**	**65.7**	63.7	**34.8**	**31.5**	1.9	**24.0**	**83.2**	**45.0**	**51.3**	**77.4**
AV+Scaled GB	**47.9**	**65.6**	**63.0**	**34.9**	**31.8**	**2.3**	**24.4**	**83.3**	**45.0**	**51.6**	**77.1**

Table 2: **DB Loss Results**. We show the performance of different experiments using DB Loss on the validation set with Fixed sampling. We use Audio-Visual Model combined with Gradient Blending [5] and DB Loss [6]. The reported number is % mAP. Showing classes: Cutting on Action (**CA**), Cut Away (**CW**), Cross Cut (**CC**), Emphasis Cut (**EC**), Match Cut (**MC**), Smash Cut (**SC**), Reaction Cut (**RC**), L Cut (**LC**), J Cut (**JC**), Speaker Chance (**SC**).

However, when combined with the Scaled GB weights, there is no significance difference between using the standard BCE Loss and the DB Loss.

Window Sampling. We show the results on each one of the classes for the Window Sampling study in Table 1.

Test Set Results. Finally, Table 3 presents the results of the best performing model (AV + Scaled GB) on the test set.

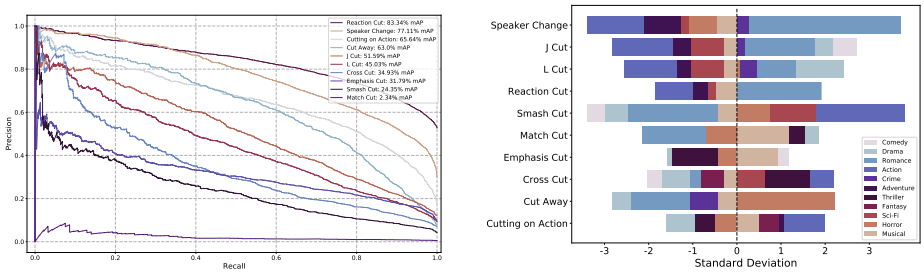
Model	Sampling	**mAP**	**CA**	**CW**	**CC**	**EC**	**MC**	**SC**	**RC**	**LC**	**JC**	**SC**
AV + Scaled GB	Gaussian	47.7	66.0	63.0	32.2	32.6	2.8	26.5	82.8	43.5	50.7	76.7

Table 3: **Test Set Results**. We show the performance of the fine-tuned models from table 1 of the main paper, evaluated on the test set. The reported number is % mAP.

Precision-Recall Curves. Besides, in Figure 2a we showcase the Precision-Recall (PR) curves for our best model on the validation set as an additional metrics to the ones shown in the main manuscript. We observe the Precision and Recall values for different confident thresholds for each one of the classes in MovieCuts.

3 Additional Statistics

Figure 2b summarizes the difference in labels' distribution across genres. To do this visualization we first calculate the average numbers of cuts for each of the classes, then, we plot the standard deviation from the classes' mean for each of the genres. Thus, we visualize how frequent or infrequent is each of the classes depending on the movie genre. For instance, we observe that for genres like Romance, and Drama the classes Speaker Change, J-cuts and L-cuts are more frequent as compared to Action, and Adventure movies. However, for Action, and Adventure, Cross-cuts and Cuts on Action are more frequent.



(a) **Precision vs Recall.** AV + Scaled GB on validation set.

(b) **Cut-types' distribution per genre.**

Fig. 2: **Figure 2a** shows the Precision vs Recall Curves for every class in the dataset. **Figure 2b** shows the summary of the cut-tupe's distribution per genre.

Additional to Figure 2b, in Figures 4 and 5 we show the distribution of classes for the most represented genres for the different splits, train 4a and 5a, validation 4b and 5b, and testing 4c and 5c. We see that the distributions across splits are independent and identically distributed (iid).

3.1 Qualitative Results

We showcase representative qualitative results for the Cutting on Action class in Figure 3. We observe that the first two cuts are correctly classified as cutting on action, since the cut happens right after the action is performed (gunshot and boxing punch). The third example is a false positive. The model wrongly predicts it as a Reaction Cut. The model fails gracefully though; the shot focuses on the face of the actor right before the action, which is similar to what happens in a Reaction Cut. At the end, the actor is not reacting but is performing an action across the cut.

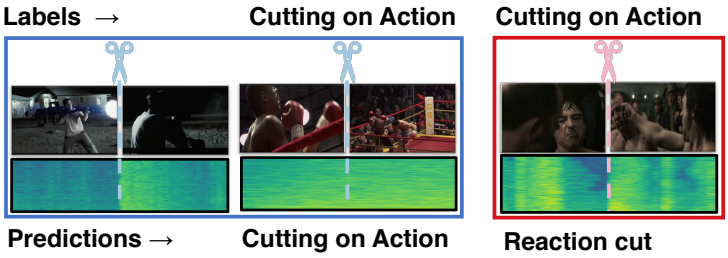


Fig. 3: **Qualitative results.** We showcase three examples of Cutting on Action. The blue box indicates True Positives, while the red box indicates a False Positive.

4 Machine-assisted Video Editing with MovieCuts

Section 4.5 presents the results of our model acting as an editor. We clarify the implementation details of such experiments and the user study.

Audio-Visual Model as a video editor. The task we are trying to solve is the creation of an edited sequence **V** based on two raw sequences **A** and **B**. The professional editor creates **V** by alternating between sequences **A** and **B** in the right places to cut (as explained in Section 1). However, professional editors use different types of heuristics and rules to define these cut places. We argue that the Audio-Visual model trained on MovieCuts has some knowledge of these cut triggers and can perform cuts between the two sequences. Thus, we collected 11 edited sequences from EditStock.com with their corresponding unedited shots. We choose sequences that alternate between two shots. Thus, the task is to find the best places to transition between **A** and **B**. After aligning the raw sequences, we create all possible cuts (transitions) from one shot to the other. Then, we score these possible cuts with our best model. We assign the maximum class score to each cut. Using these scores, we use only the top-k as good places to cut. Finally, we use these scores to perform cuts alternating between **A** and **B**.

We ask 63 AMT turkers to pick among the edits done by professionals vs the automatic methods. The results are shown on table 4b. The AV Model trained on MovieCuts was the one picked the most over the professionals, showing that the edits made by it are preferred by users over the other methods. Moreover, we use the edited sequences to create the ground-truth cut places and evaluate how close were the different automatic edits compared to these professionally edited sequences. In table 4a, we measure Purity, Coverage, and F1, implemented by [2]. Yet again, our method outperforms all the other automatic methods when comparing them with the professionally edited sequences. Please, be reminded that this editing process was done without training for it, but just using the model trained on MovieCuts as a scoring function for cut places. We argue that improvements in Cut-type recognition tasks can translate into advances in tasks related to machine-assisted video editing.

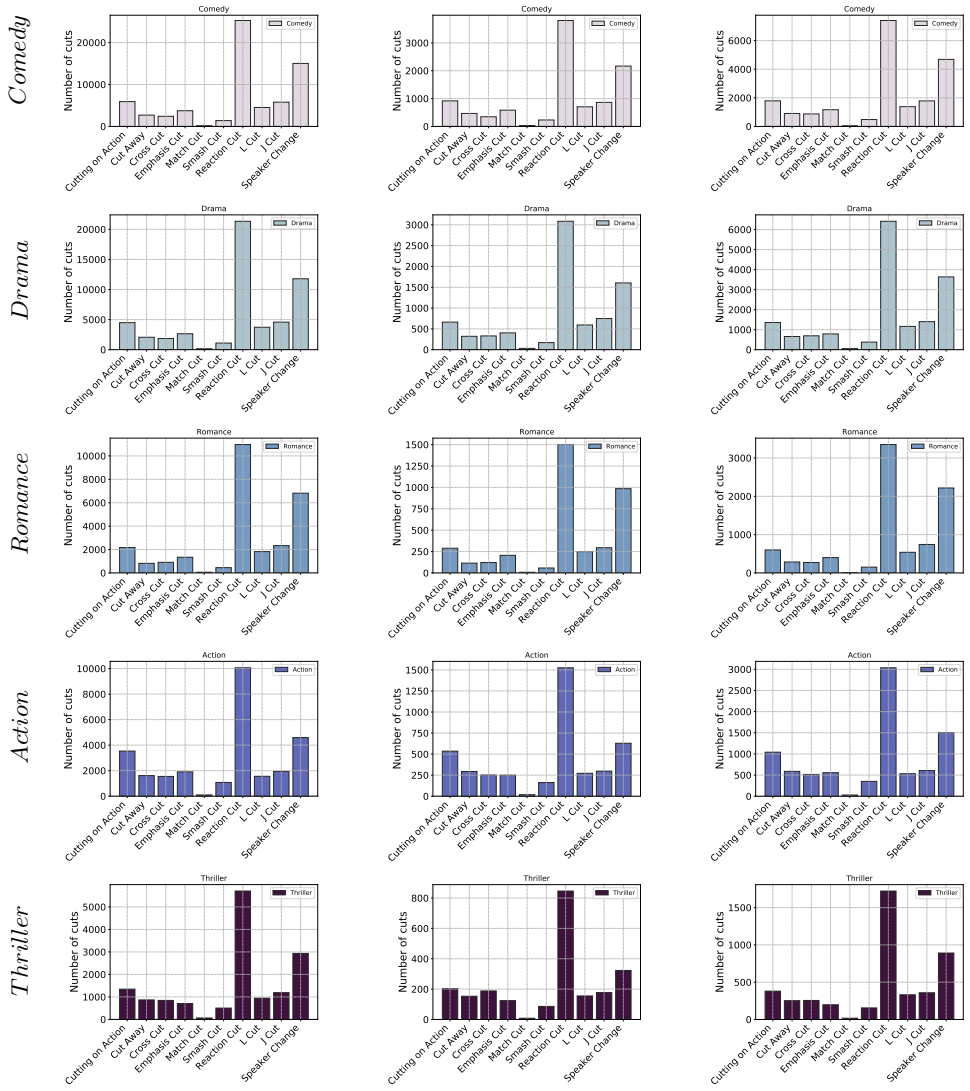
Method	Purity	Coverage	F1
Random Frame	99	10	18
Random Snippet	87	52	63
Biased Random	74	82	77
MovieCuts AV	80	82	**81**

(a) **Quantitative Results.** Results are reported in %. We use the human editor as ground truth and evaluate Purity, Coverage and F1 implemented in [2].

Method	vs Human Editor
Random Frame	1.8
Random Snippet	15.7
Biased Random	34.5
MovieCuts AV	**38.1**

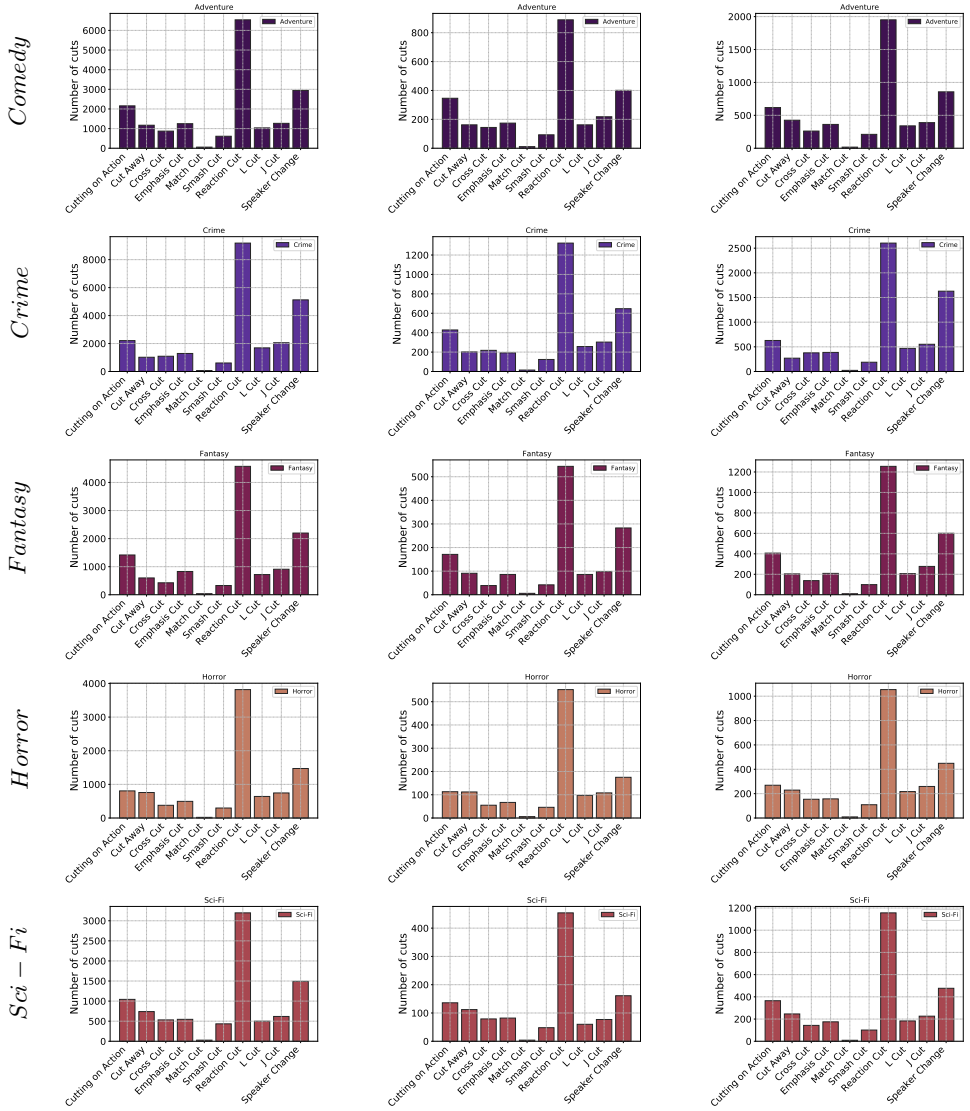
(b) **Qualitative Results.** Results are report in % of times that humans pick such method over the professional editing.

Table 4: **Video Editing Results**. Results of MovieCuts' automated video editing. Our method performs better than a set of baselines.



(a) Train (b) Validation (c) Test

Fig. 4: **Cut-type distribution across movie genres and MovieCuts splits.** Comedy, Drama, Romance, Action, Thriller.



(a) Train (b) Validation (c) Test

Fig. 5: **Cut-type distribution across movie genres and MovieCuts splits.** Aventure, Crime, Fantasy, Horror, Sci-Fi.

References

1. Arijon, D.: Grammar of the film language. Focal Press London (1976)
2. Bredin, H.: pyannote.metrics: a toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems. In: Interspeech 2017, 18th Annual Conference of the International Speech Communication Association. Stockholm, Sweden (August 2017), http://pyannote.github.io/pyannote-metrics
3. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021)
4. Smith, T.J.: The attentional theory of cinematic continuity. Projections **6**(1), 1–27 (2012)
5. Wang, W., Tran, D., Feiszli, M.: What makes training multi-modal classification networks hard? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12695–12705 (2020)
6. Wu, T., Huang, Q., Liu, Z., Wang, Y., Lin, D.: Distribution-balanced loss for multi-label classification in long-tailed datasets. In: European Conference on Computer Vision. pp. 162–178. Springer (2020)