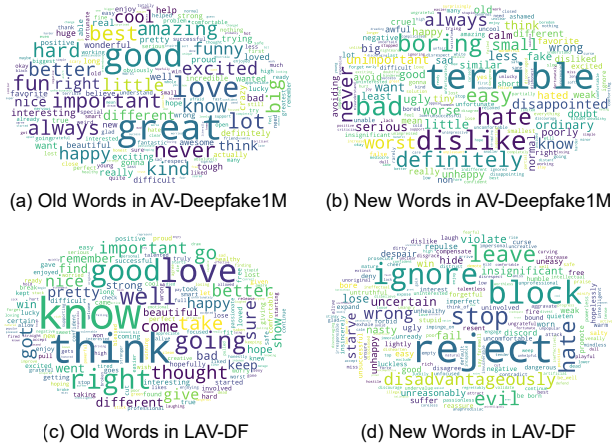


## AV-Deepfake1M: A Large-Scale LLM-Driven Audio-Visual Deepfake Dataset

### Supplementary Material



**Figure 6: Qualitative comparison of transcript modifications in AV-Deepfake1M and LAV-DF.** (a) The old words before the manipulations in AV-Deepfake1M. (b) The new words after the LLM-driven manipulations in AV-Deepfake1M. (c) The old words before manipulations in LAV-DF. (d) The new words after the rule-based manipulations in LAV-DF.

## A TRANSCRIPT MANIPULATION

In addition to the quantitative comparison of transcript modifications in AV-Deepfake1M and LAV-DF [6] (see Section 3.1.1), here we also present a qualitative one. Figure 6 illustrates word clouds for `old_word(s)` and `new_word(s)` for both datasets. A comparison between the new words generated by the rule-based strategy utilized in LAV-DF and our LLM-driven generation further demonstrates that the latter results in more natural and diverse transcript manipulations.

## B HUMAN QUALITY ASSESSMENT

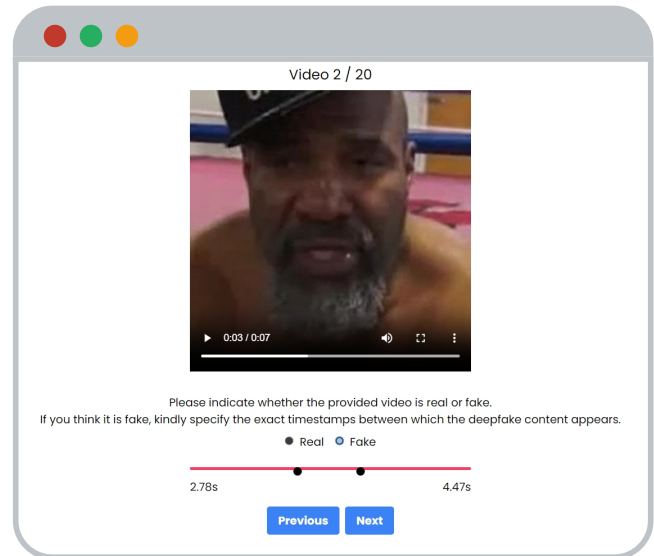
Here we provide further details on the user study (see Section 3.3) that aims to evaluate humans’ performance in detecting the highly realistic deepfake content in AV-Deepfake1M.

### B.1 Data

The data used in the user study are 200 videos randomly sampled from the *test* set of AV-Deepfake1M and LAV-DF [6], with the aim to maximize the number of unique identities. Please note that the user study setup ensures each participant cannot see a duplicated identity. The videos include 50 real videos from AV-Deepfake1M, 50 fake videos from AV-Deepfake1M, 50 real videos from LAV-DF, and 50 fake videos from LAV-DF. For fair comparison with LAV-DF, the fake videos contain only one audio-visual *replacement* (see Section 3).

## B.2 Participants

We randomly group the participants into 10 groups where each group evaluates 10% of the videos (i.e., 20 videos including 5 real videos



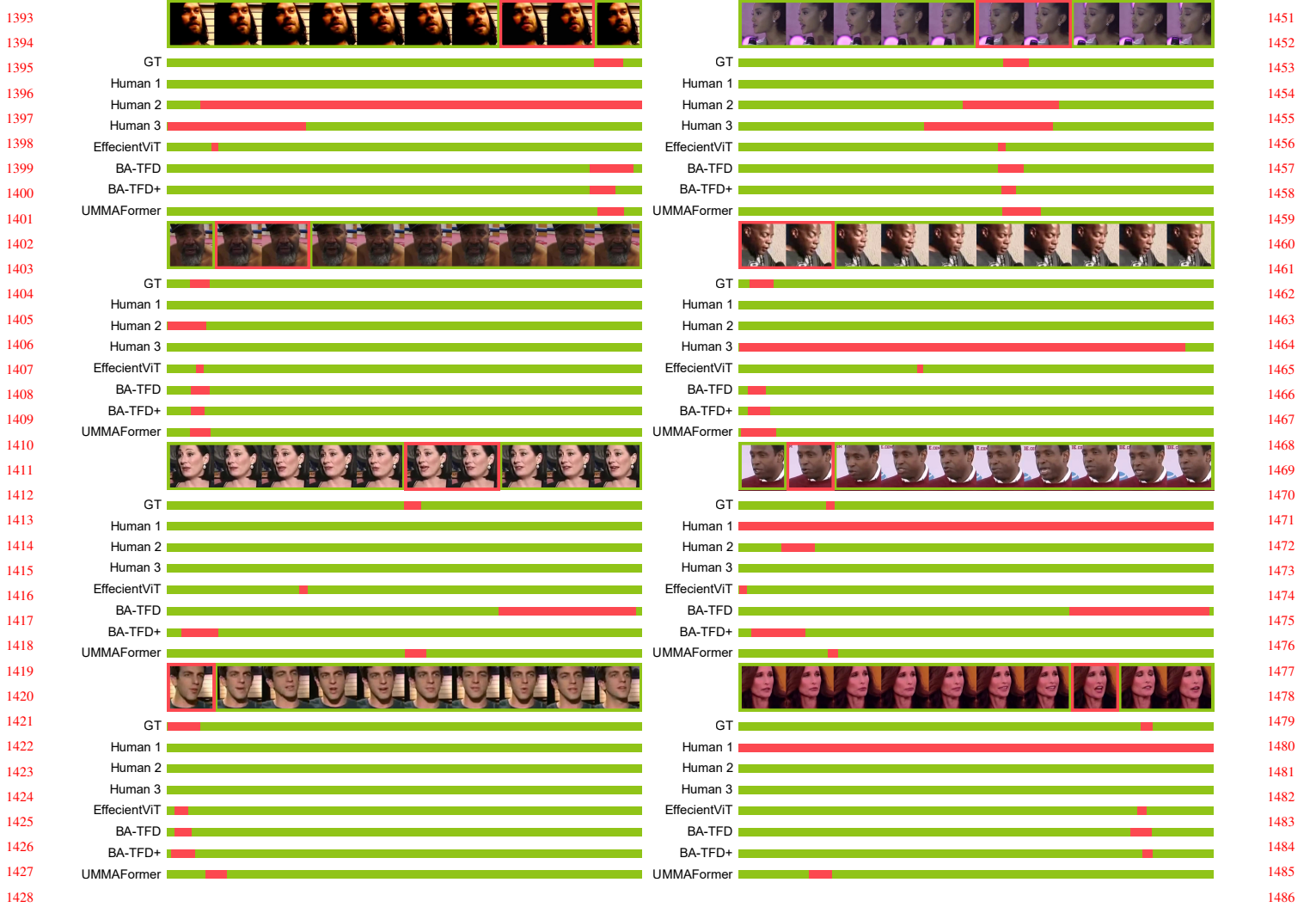
**Figure 7: Screenshot of the user study interface.** On the top is the video with audio, the middle is the textual description of the task, and the bottom is the participant’s controls to 1) Select whether the video is *real* or *fake* and 2) If the participant selects *fake*, use a slider to specify the begin and end of the fake segment.

from AV-Deepfake1M, 5 fake videos from AV-Deepfake1M, 5 real videos from LAV-DF, and 5 fake videos from LAV-DF). We utilize a random non-overlapping selection of videos for each participant, meaning that each participant evaluates videos for 20 out of the 200 videos. After watching each video, the participants first answer whether the video is *real* or *fake*, and if they think the video is *fake*, the participants can choose the start and end timestamps for the fake segment. A screenshot of the developed user study interface based on the React<sup>2</sup> framework is shown in Figure 7.

### B.3 Evaluation and Analysis

Among the 25 participants that took part in the user study, the binary deepfake detection/classification accuracy is 64.84% for AV-Deepfake1M. This low performance indicates that the deepfake content in AV-Deepfake1M is very challenging for humans to detect. A similar pattern is observed for the temporal localization of fake segments. Similarly to Table 5, here we report and compare average precision (AP) and average recall (AR) scores in Table 12 and extend that comparison with the state-of-the-art methods using the same subset of videos. The AP score for 0.5 IoU is 01.92. Thus, we reduced the AP threshold to 0.1 IoU, improving the AP score to 15.32. Figure 8 illustrates a similar qualitative comparison. The low human performance in each aspect indicates that to detect highly realistic deepfake content, we need more sophisticated detection and localization methods.

<sup>2</sup><https://react.dev/>



**Figure 8: Examples of user study results and comparison with the state-in-the-art in temporal deepfake localization.** Green color represents *real* segments and red color represents *fake* segments. GT: Ground truth.

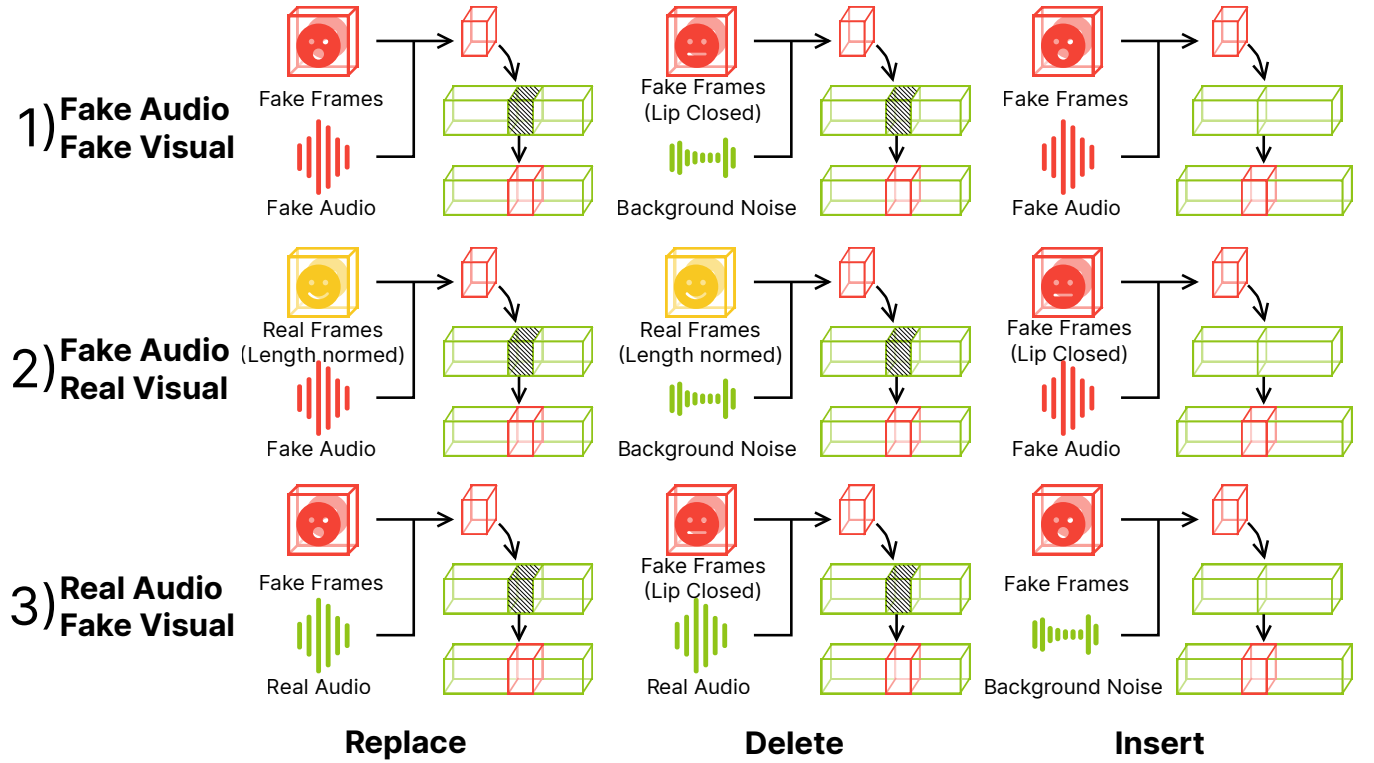
**Table 12: User study results compared with the state-in-the-art in temporal deepfake localization.**

Dataset Method	LAV-DF [6]				AV-Deepfake1M			
	Acc.	AP@0.1	AP@0.5	AR@1	Acc.	AP@0.1	AP@0.5	AR@1
Xception [12]	96.00	69.33	41.75	30.40	77.00	58.78	24.26	12.20
BA-TFD [6]	-	95.37	80.33	66.44	-	59.69	44.87	21.27
BA-TFD+ [4]	-	98.00	98.00	87.60	-	65.44	51.41	23.26
UMMAFormer [65]	-	98.00	98.00	97.80	-	69.77	53.72	38.39
Human	84.03	36.80	14.17	10.04	68.64	15.32	01.92	02.54

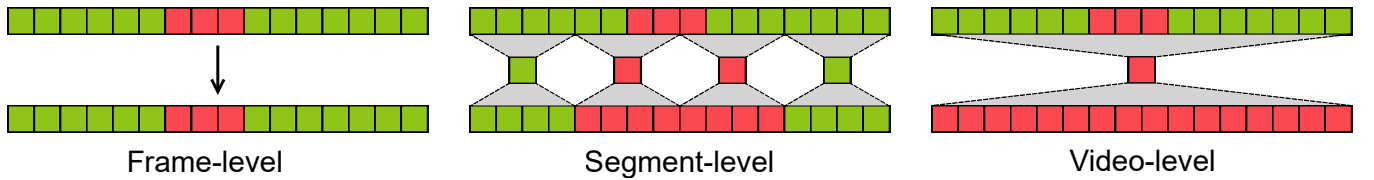
Considering LAV-DF [6], we observed similar patterns - human performance is lower than the state-of-the-art detection and localization methods. Comparing the human performance between AV-Deepfake1M (Acc. 68.64, AP@0.1 15.32) and LAV-DF (Acc. 84.03, AP@0.1 36.80), we find that AV-Deepfake1M is more challenging than LAV-DF for humans.

## C AUDIO AND VIDEO GENERATION

Here we provide complete details on the manipulations in AV-Deepfake1M (see Section 3). Figure 9 provides visualizations corresponding to each of the three modifications and the resulting deepfake content. Please note that for example for **Fake Audio** and **Real Visual** in the cases of *deletion* and *insertion*, there are slight modifications in the visual signal as well. The reason we regard the visual



**Figure 9: Details of the audio-visual content generation.** Here, we show the audio-visual content manipulation strategy in three setups i.e. fake audio fake video, fake audio real video and real audio fake video. We believe that these three variations of fake content generation add more challenge in the temporal localization task.



**Figure 10: Complete details on the label access for training.** Green color represents the *real* and red color represents *fake* content. The top row represents the original *frame-level* labels in a video. The middle row represents the segment- and video-level labels based on whether the segment/video contains any *fake* frames. For fair comparison across different methods, the bottom row represents the mapped segment- and video-level labels to frame-level labels.

signal as *real* is the fact that words were not *inserted* or *deleted* in that modality. Similarly for **Real Audio** and **Fake Visual**.

## D LABEL ACCESS FOR TRAINING

Figure 10 provides complete details on the label access during training (see Section 5.2).

- In the *frame-level* configuration, the models are trained using the ground truth labels for each frame in the video.
- In the *segment-level* configuration, if the segment contains any *fake* frames, it is labelled as *fake* otherwise it is labelled as *real*. For the segment-based methods MARLIN [5] and MDS [13], we used the *segment-level* labels during training. For a fair comparison when training the frame-based methods

Meso4 [1] and MesoInception4 [1] we mapped the *segment-level* labels to *frame-level*.

- In the *video-level* configuration, if the video contains any *fake* frames, it is labelled as *fake* otherwise it is labelled as *real*. Similarly to the *segment-level* configuration, for a fair comparison when training the frame-based methods Meso4 [1] and MesoInception4 [1] we mapped the *video-level* labels to *frame-level*.