ExDDV: A New Dataset for Explainable Deepfake Detection in Video (Supplementary)

Vlad Hondru[1], Eduard Hogea[2], Darian Onchiş[2], Radu Tudor Ionescu[1,*]
[1]University of Bucharest, Romania, [2]West University of Timişoara, Romania

1. Annotation Process and Video Resolutions

The GUI (shown in Fig. 1) contains two video players, side by side. The deepfake video is played on the left side of the window, while the corresponding real video is played on the right side. By default, the real video is not played. If the annotator needs to play the real video along with the deepfake one, they could simply press the button below the player screen. When the fake video is playing, the user can click anywhere on the frame to indicate the location of artifacts. Behind the scene, the application records the relative pixel location and the timestamp (*i.e*. frame index) of the click. Under the fake video player, there is a text box in which the details describing the visual issues can be written by the user. In the bottom right, there are three radio buttons, which allow the user to indicate the difficulty level of identifying deepfake evidence. We ask users to label deepfake videos as *hard*, when they need to play the deepfake video at least two times, or when they need to activate the real video to observe artifacts. In a similar manner, we instruct them to label videos as *easy*, if they are able to identify more than one artifact with a single play of the deepfake video. For real videos, we do not collect clicks.

In Figure S1, we plot a bar chart showing the various resolutions that comprise ExDDV. The bar chart clearly shows that the first three resolutions are significantly more frequent than the others.

In Figure S2, we present more annotated samples from ExDDV, having different levels of difficulty, click locations, and explanatory text lengths.

2. Additional Results

Figure S3 contains a comprehensive diagram with qualitative samples for all possible training scenarios applied on LLaVA [3]. The examples include both relevant explanations as well as wrong explanations, *e.g*. identifying artifacts on real videos.

In Figure S4, we showcase some examples of how the ViT-based click predictor compares with the ground-truth click locations. We observe that the predictor is able to precisely locate visual artifacts.

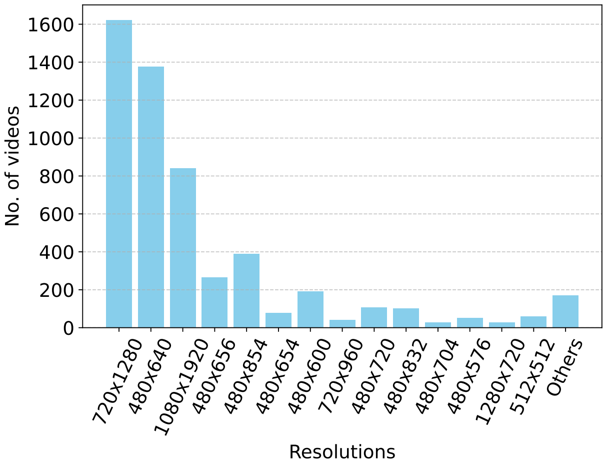


Figure S1. The number of videos for each frame resolution in ExDDV.

3. Training Environments

We have worked on multiple environments for our experiments. For the in-context learning experiments, we used a Tesla V100-SXM2 GPU with 32GB VRAM. Phi-3-Vision [1] and LLaVA [3] were fine-tuned on a single H100 GPU with 80GB VRAM. BLIP-2 [2] was fine-tuned using an RTX 4090 GPU with 24GB VRAM. The same training environment as for BLIP-2 was used for the click predictors.

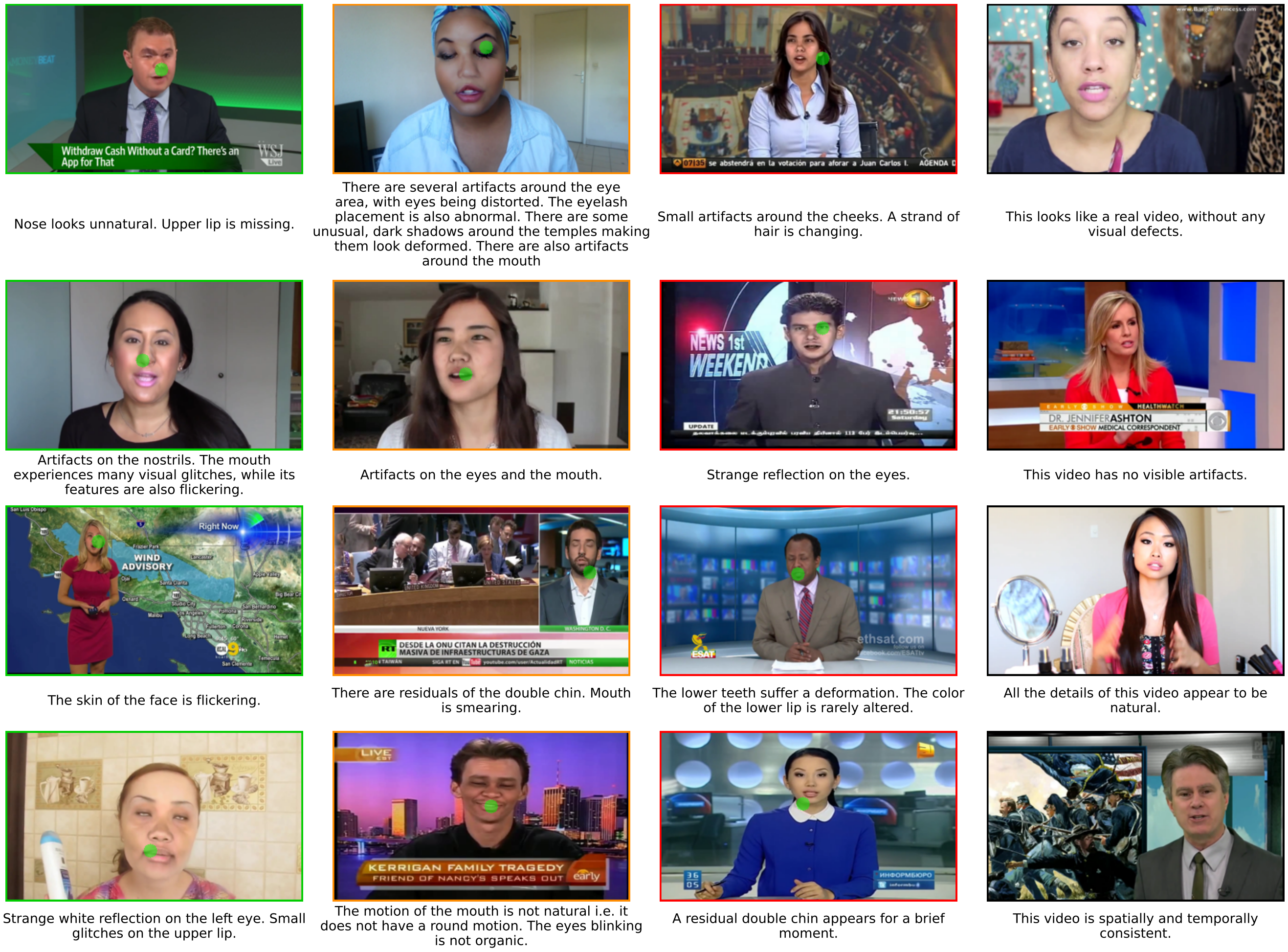
References

[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 1

[2] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proceedings of ICML*, pages 19730–19742, 2023. 1

[3] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *Proceedings of NeurIPS*, pages 34892–34916, 2023. 1

[*] Corresponding author: raducu.ionescu@gmail.com.


Withdraw Cash Without a Card? There's an App for That
WSJ Live
07|35 se abstendrá en la votación para aforar a Juan Carlos I.
AGENDA
NEWS 1st WEEKEND
UPDATE
21:50:57 Saturday
HEALTHWATCH
DR. JENNIFER ASHTON
EARLY SHOW MEDICAL CORRESPONDENT
Right Now
WIND ADVISORY
NUEVA YORK
WASHINGTON D.C
DESDE LA ONU CITAN LA DESTRUCCIÓN MASIVA DE INFRAESTRUCTURAS DE GAZA
ESAT
ethsat.com
LIVE
KERRIGAN FAMILY TRAGEDY
FRIEND OF NANCY'S SPEAKS OUT
early
ИНФОРМБЮРО

Nose looks unnatural. Upper lip is missing.

There are several artifacts around the eye area, with eyes being distorted. The eyelash placement is also abnormal. There are some unusual, dark shadows around the temples making them look deformed. There are also artifacts around the mouth

Small artifacts around the cheeks. A strand of hair is changing.

This looks like a real video, without any visual defects.

Artifacts on the nostrils. The mouth experiences many visual glitches, while its features are also flickering.

Artifacts on the eyes and the mouth.

Strange reflection on the eyes.

This video has no visible artifacts.

The skin of the face is flickering.

There are residuals of the double chin. Mouth is smearing.

The lower teeth suffer a deformation. The color of the lower lip is rarely altered.

All the details of this video appear to be natural.

Strange white reflection on the left eye. Small glitches on the upper lip.

The motion of the mouth is not natural i.e. it does not have a round motion. The eyes blinking is not organic.

A residual double chin appears for a brief moment.

This video is spatially and temporally consistent.

Figure S2. More examples of video frames from ExDDV with text and click annotations. Clicks are represented as large green dots. Real videos are not annotated with clicks or difficulty levels. The border color indicates the difficulty level: green=easy, orange=medium, red=hard, black=real. Best viewed in color.

Ground-truth	There are some distortions around the face. The beard on the chin is faded. The nose skin is not uniform.	The mouth is heavily distorted and some artifacts are present.	This looks like a real video, without any visual artifacts.	The video does not appear to have any appearance or motion defects.	The right eyebrow is thicker than the left one. The interior of the mouth is smearing.
In-context learning	The image is blurry and has artifacts, making it difficult to discern the details of the person's face.	The image has significant distortion and artifacts, making it difficult to discern the features of the person's face.	The image has a blurry face with a strange light around the eyes.	The face in the image appears unnatural and has an artifact on the mouth with mouth movement that looks strange.	The image has visual artifacts, causing distortion around the eyes and eyebrows.
In-context learning Hard mask	The face in the image has several deepfake artifacts, including distortion of facial features, flickering hair, and inconsistent skin blending. The mouth does not open when talking, and the face appears swollen.	The face in the image is affected by deepfake artifacts, which include unnatural looking mouth with appearing and disappearing teeth, irregular skin color difference in the left upper part of the face, artifacts on the eyes, an additional pair of eyebrows fluctuating above the original ones, unnatural gaze, and eye twitching.	The image shows a woman with a blurry face, distorted eyebrows, and inconsistent color on her face. There are artifacts around the edges of her face, and her mouth experiences multiple visual glitches. The skin corresponding to the mandible is bleached, and the irises of her eyes are darting or changing. The eyebrows are darting up and down, and a residual set of eyebrows is visible.	The face in the image has several inconsistencies and artifacts, including glitches on the mouth, distortion near the right ear, and different sizes for the eyes. The mouth does not open when talking, and the face is blurry. These artifacts suggest that the image may have been manipulated or altered.	The face in the image has several deepfake artifacts, including distorted eyebrows, unnatural-looking eyes, and a flickering chin. The eyebrows are wavering above the eyes, and there are residual artifacts around the mouth causing distortion. Additionally, the eyes are being closed when in a neutral position, and there are artifacts on the left and right sides of the face around the ear.
In-context learning Soft mask	The image shows a man's face with a blurry background, and the face appears to be a deepfake. The inconsistencies in the image include the flickering hair on the left temple, the small white noise artifacts on the nostrils, and the blurred mouth.	The face in the image has significant distortion, with artifacts all over it and the teeth appearing too big for the mouth. The irises are changing colors, and there are distortions of the mouth.	The image shows a woman with a bleached jawline, which is an inconsistency in the face's appearance. The eyes have small deformations, and there are artifacts on the nostrils, which are signs of deepfake manipulation. The outer line of the face is also blurred.	The face in the image appears blurry and distorted, with color discrepancy between the face and the forehead. There are artifacts on the forehead, and the light is flickering on the face, which suggests that the image might be a deepfake or manipulated in some way.	The face in the image appears to be a deepfake, as it has been manipulated to include distorted eyebrows and a wobbling mouth. The eyebrows are not in their natural position and are wavering above the face, while the mouth is also distorted, causing a discrepancy in the overall appearance.
Fine-tuned Baseline	The face is distorted, especially the nose and cheeks being deformed.	The person's appearance and movements seem to be adequate.	There are small deformations of the face, with the outer line being blurred. There are artifacts on the nostrils.	Artifacts on the mouth	This looks like a real video, without any visual artifacts.
Fine-tuned Hard mask	There are some unusual shadows on the right temple. The nose also gets deformed at the end and the tip shifts towards the right. The mouth also glitches and the teeth are not well defined.	The mouth is heavily distorted and some artifacts are present.	The person in this video appears to have natural features and movements.	The person's appearance and movements seem to be adequate.	The eyebrows are being altered.
Fine-tuned Soft mask	Artifacts on the nose. Inconsistent color on the face.	The lips experience many visual glitches.	The face is not well rendered and does not look natural, with the features being blurry and distorted.	There are no appearance and motion defects in this video.	The eyebrows are distorted and also wobble slightly. There are artifacts around the mouth causing distortion.

Figure S3. Qualitative examples for LLaVA in all training scenarios: pre-trained, in-context learning with and without masking, and fine-tuned with and without masking.

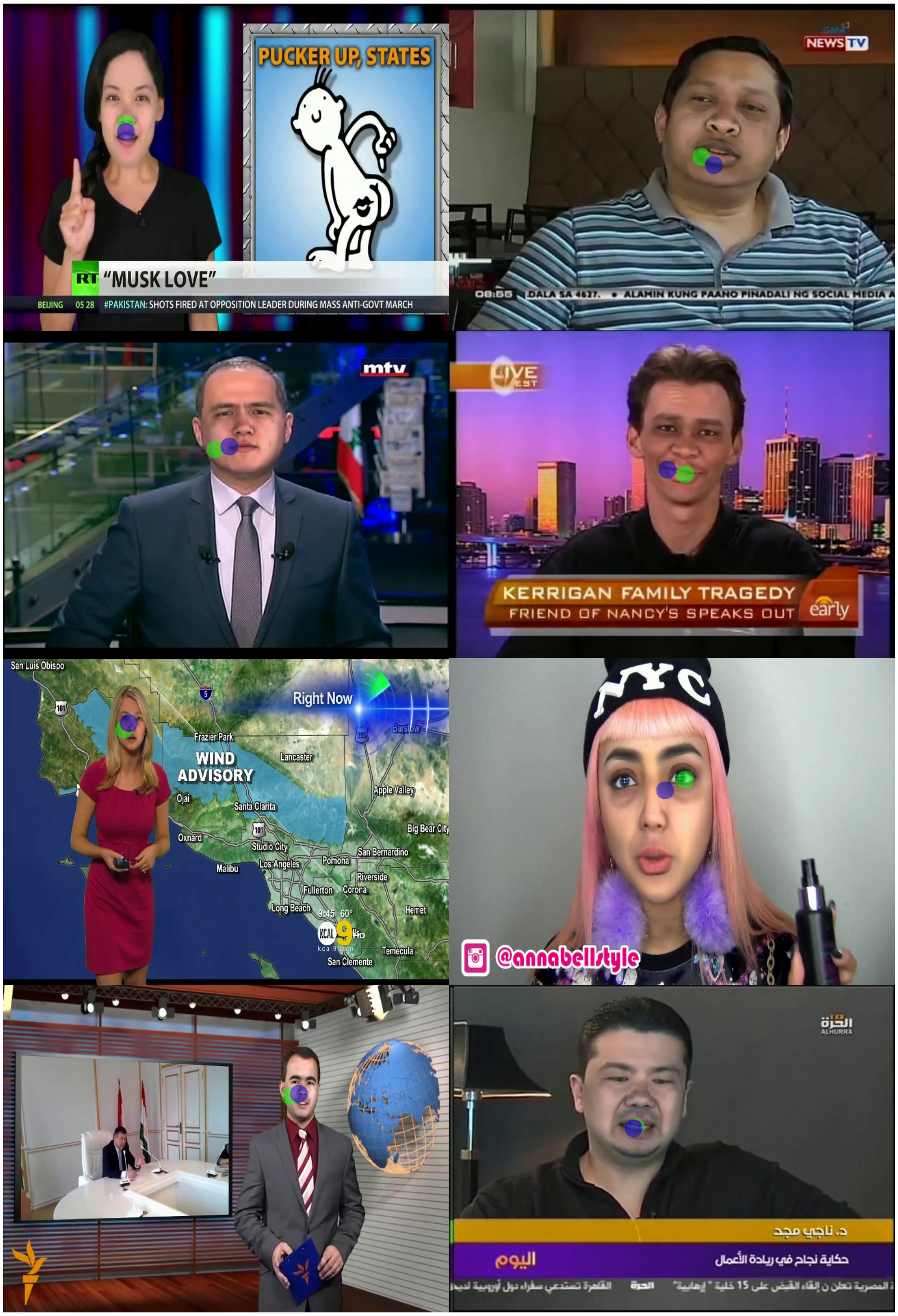


Figure S4. Qualitative examples of our click predictor. The green dots represent the ground-truth locations and the blue dots represent the corresponding predictions. Best viewed in color.