

ExDDV: A New Dataset for Explainable Deepfake Detection in Video

Vlad Hondru¹, Eduard Hoge², Darian Onchiș², Radu Tudor Ionescu^{1,*}

¹University of Bucharest, Romania, ²West University of Timișoara, Romania

Abstract

The ever growing realism and quality of generated videos makes it increasingly harder for humans to spot deepfake content, who need to rely more and more on automatic deepfake detectors. However, deepfake detectors are also prone to errors, and their decisions are not explainable, leaving humans vulnerable to deepfake-based fraud and misinformation. To this end, we introduce ExDDV, the first dataset and benchmark for Explainable Deepfake Detection in Video. ExDDV comprises around 5.4K real and deepfake videos that are manually annotated with text descriptions (to explain the artifacts) and clicks (to point out the artifacts). We evaluate a number of vision-language models on ExDDV, performing experiments with various fine-tuning and in-context learning strategies. Our results show that text and click supervision are both required to develop robust explainable models for deepfake videos, which are able to localize and describe the observed artifacts. Our novel dataset and code to reproduce the results are available at <https://github.com/vladhondru25/ExDDV>.

1. Introduction

Online fraud and misinformation based on deepfake videos reached unprecedented expansion rates in recent years. A recent forensics report suggests that identity fraud rates nearly doubled, showing a significant rise in the prevalence of deepfake videos between 2022 and 2024, from 29% to 49%¹. This rise of deepfake content is primarily caused by recent advances in generative AI, especially with the emergence of highly capable diffusion models [15, 47, 71, 72, 75, 77, 78, 82]. The high quality and realism of generated videos put online users in difficulty of telling the difference between real and fake content. In this context, humans can turn to state-of-the-art automatic deepfake detectors for help [10, 26, 30, 33, 45, 57, 76, 79]. To come in handy, such models need to be robust and trustworthy, while also providing explainable decisions that would enable users to gain important insights regarding the kinds of artifacts they should look for in a video. Yet, Croitoru

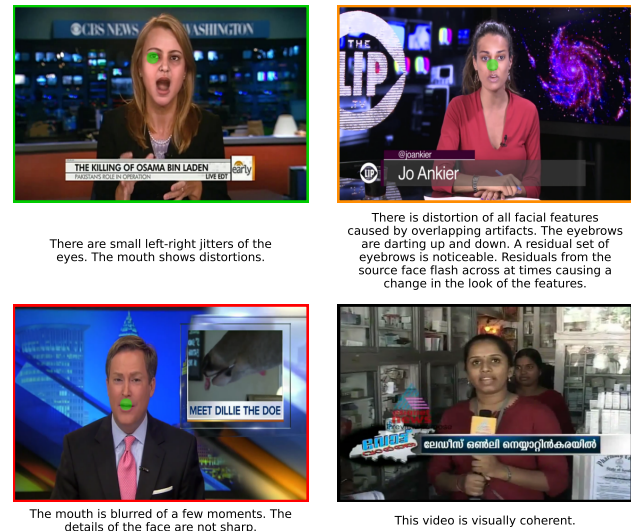


Figure 1. Examples of video frames from ExDDV with text and click annotations. Clicks are represented as large green dots. Real videos are not annotated with clicks or difficulty levels. The border color indicates the difficulty level: green=easy, orange=medium, red=hard, black=real. Best viewed in color.

et al. [19] showed that state-of-the-art deepfake detectors fail to generalize to content generated by new generative models (not seen during training). Moreover, to the best of our knowledge, the task of generating textual descriptions to explain the artifacts observed in deepfake videos has not been explored so far.

To this end, we introduce the first dataset and benchmark for Explainable Deepfake Detection in Video, called ExDDV. Our new dataset comprises approximately 5.4K real and deepfake videos that are manually annotated with text descriptions, clicks and difficulty levels. The text descriptions explain the artifacts observed by human annotators, while the clicks provide precise localizations of the described artifacts, as shown in Fig. 1. The annotated videos are gathered from a broad set of existing datasets for video deepfake detection, including DeeperForensics [37], FaceForensics++ [68], DeepFake Detection Challenge [22] and BioDeepAV [19], to enhance the diversity of our collection. The inter-annotator agreement (0.6238 cosine similarity in Sentence-BERT space) confirms that the collected annota-

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

¹<https://regulaforensics.com/news/deepfake-fraud-doubles-down/>

tions are consistent and of high quality. ExDDV comes with an official split into training, validation and test, which facilitates reproducibility of results and future comparisons.

We further evaluate a number of vision-language models (VLMs) on ExDDV, comparing various architectures, training procedures and supervision signals. In terms of architectures, we experiment with BLIP-2 [41], Phi-3-Vision [1] and LLaVA-1.5 [46]. In terms of training strategies, we consider pre-trained versions, as well as versions based on in-context learning and fine-tuning. Regarding the supervision signals, we consider text descriptions alone or in combination with clicks. For click supervision, we study two alternative approaches, namely soft and hard input masking. Our empirical results show that fine-tuning provides the most accurate explanations for all VLMs, confirming the utility of ExDDV in developing robust explainable models for deepfake videos. Moreover, we find that both text and click supervision signals are required to jointly localize and describe the observed artifacts, as well as to generate top-scoring explanations.

In summary, our contribution is threefold:

- We introduce the first dataset for explainable deepfake detection in video, comprising 5.4K videos that are manually labeled with descriptions, clicks and difficulty levels.
- We study various VLM architectures and training strategies for explainable deepfake detection, all leading to a comprehensive benchmark.
- We publicly release our dataset and code to reproduce the results and foster future research.

2. Related Work

Deepfake detection in video. Nowadays, deepfakes have started to pose a real threat, since generative methods have significantly evolved and their number has increased. Nevertheless, substantial effort has been made to develop detection methods [19] and counter the misuse of generative AI technology. Early methods for deepfake detection in video were based on convolutional networks [2, 3, 18, 25, 26, 28, 33, 45, 54, 57, 69, 83]. To handle both spatial and temporal dimensions, two different strategies are commonly adopted. The first is to apply 2D convolutions on individual frames and subsequently combine the resulting latent representations either by using basic operations (such as pooling or concatenation) [2, 25] or by employing recurrent neural networks [3, 28, 33, 45, 54, 57, 69]. The second strategy involves extending the convolutions to 3D to capture spatio-temporal features [18, 26, 83].

To learn more robust feature representations or enhance the detection of frame-level inconsistencies and motion artifacts, researchers have incorporated attention mechanisms into deepfake detectors [10, 76, 79]. Some recent works [17, 27, 84] employed the transformer architecture [23, 74]. Such methods have a superior ability to capture long-range

dependencies and are effectively applied to detect temporal inconsistencies.

To the best of our knowledge, current video deepfake detectors do not have intrinsic capabilities to explain their decisions. This is primarily caused by the lack of deepfake datasets providing explanatory annotations for the video content.

Deepfake video datasets. The task of deepfake detection has been extensively studied, and thus, there are many datasets that are now publicly available. Among these, the most notable are LAV-DF [11], GenVideo [13], DeepFake Detection Challenge [22], DeeperForensics [37], FakeAVCeleb [38], Celeb-DF [42], FaceForensics++ [68], WildDeepfake [86] and MAVOS-DD [20]. Although such datasets contain the binary label (real or fake) associated with each video, they do not provide other kinds of annotations about the video content. Unlike existing datasets, we provide a dataset for deepfake detection with textual explanations for the artifacts observed by human annotators, along with clicks (points) that indicate artifact locations. To the best of our knowledge, our novel dataset is the first to provide explanatory annotations for deepfake video content.

Explainable deepfake detection. The research community has extensively explored explainable AI (XAI), proposing various approaches [35, 39, 48, 49, 56, 67, 70, 85]. While there are many established methods for XAI, such as Gradient-weighted Class Activation Mapping (Grad-CAM) [70], SHapley Additive exPlanations (SHAP) [49] or LIME [67], explainable AI has barely been applied to deepfake detection. Ishrak et al. [34] implemented a binary classification model to detect whether video frames are artificially generated or not. Then, they employed Grad-CAM [70] to estimate the salient regions that could explain the prediction, subsequently verifying if these regions overlap with the area of the face. Grad-CAM is a generic framework that back-propagates the gradients at various layers and computes a global average to obtain saliency maps.

Other methods opted for detecting deepfakes focusing only on a specific factor. Using only convolutional-based architectures, Haliassos et al. [29, 30] designed a detection method that concentrates on mouth movements, while Demir and Çiftçi [21] proposed an approach that analyzes face motions. Due to the focus on specific factors, these methods can inherently provide some limited level of explainability.

To the best of our knowledge, none of the existing explainable AI methods are specifically designed to explain deepfake videos. However, we acknowledge that researchers have studied the explainability of generated images. For example, the WHOOPS dataset [9] provides explanations for why synthetic images defy common sense. The authors note that such images can be easily identified by humans, raising the question if machines can do

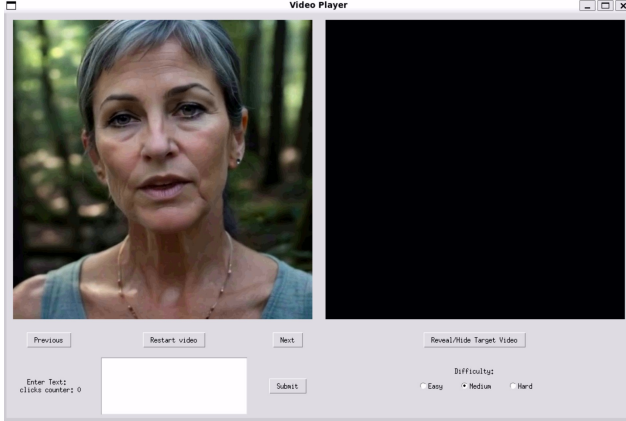


Figure 2. A screenshot of the application used to annotate ExDDV.

the same. In contrast, we focus on deepfake video content instead of generated images. We emphasize that deepfake content is potentially harmful to humans. Since deepfake videos are typically generated with a harmful intent in mind, they are not aimed to defy common sense (on the contrary). We thus consider the study of Bitton-Guetta et al. [9] as a complementary work to our own.

3. Proposed Benchmark

The main contribution of our work is to introduce ExDDV, a benchmark specifically designed to facilitate human-interpretable explanations of deepfake videos. The novelty of ExDDV stems not only from being the first of its kind, but also from its comprehensive exploration of the task.

Video collection. The fake videos are collected from four different sources: DeeperForensics [37], FaceForensics++ [68], DeepFake Detection Challenge (DFDC) [22] and BioDeepAV [19]. This ensures that the dataset covers a wide range of generative methods and includes various video resolutions and durations. As real video samples, we use the original movies from DeeperForensics and FaceForensics++. From FaceForensics++, we include all videos generated by the Face2Face, FaceSwap, and FaceShifter methods. From the other two datasets (DFDC and BioDeepAV), we randomly select a number of videos. In Tab. 1, we present the breakdown of source datasets from which we gather videos for ExDDV. We provide an official split of the video collection into 4,380 training videos, 482 validation videos and 485 test videos.

Annotation. To annotate the dataset, we developed a simple application from scratch, with a custom graphical user interface (GUI), which is shown in Fig. 2. The annotation efforts are carried out by two paid human annotators, who are experts in the field. Each dataset is equally split in two subsets, such that each method and each possible kind of visual issue is described from two distinct points of view. More information about the annotation process can be found in the supplementary. A couple of annotated ex-

Source Dataset	Method	#Samples
DeeperForensics	Real	1,000
FaceForensics++	DF-VAE	1,000
FaceForensics++	Face2Face	1,000
	FaceSwap	1,000
	FaceShifter	1,000
DeepFake Detection Challenge	Multiple	269
BioDeepAV	Talking-face	100
Total		5,369

Table 1. Source datasets and generative methods for the deepfake videos included in ExDDV. Our dataset contains 1,000 real videos and 4,369 fake videos generated by various methods.

Measure	Average Score
Sentence-BERT	0.6238
BERTScore	0.3857
BLEU	0.0808
METEOR	0.2349
ROUGE-1	0.3023
ROUGE-2	0.1037
ROUGE-L	0.2434

Table 2. Inter-annotator agreement among textual descriptions. Sentence-BERT and BERTScore indicate semantic alignment, while BLEU, METEOR, and ROUGE capture n-gram overlaps. Higher scores indicate a better alignment.

amples are shown in Fig. 1 and Fig. S2.

Post-processing. We post-process the text annotations to correct spelling errors. We employ an automatic spell checker [6] to find the misspelled words, and then manually correct them. We also change the forms of some words in order to consistently use American English instead of British English (*e.g.* changing “colour” to “color”).

Inter-annotator agreement. To estimate the inter-annotator agreement and guarantee the consistency of the annotations, we provide 100 deepfake videos to both annotators. All these videos have explanations and clicks from both humans.

We first evaluate the similarity between the text descriptions provided by the two annotators in terms of two semantic similarity metrics based on Sentence-BERT [65] and BERTScore [81], respectively. We hereby emphasize that conventional scores, such as BLEU [61], are not able to capture if two sentences have the same meaning when they contain different words [12], *e.g.* annotators use synonyms to refer to the same concept. We therefore consider the similarities based on Sentence-BERT [65] and BERTScore [81] as more reliable. The results shown in Tab. 2 demonstrate a close semantic alignment between textual descriptions, with an average cosine similarity in the Sentence-BERT embedding space of 0.6238. The discrepancy be-

Temporal Window	Spatial Window	Agreement (%)		
		Temporal	Spatial	Joint
30	50×50	54.70	79.60	35.10
60	75×75	66.14	89.69	53.82
120	100×100	84.54	93.94	75.87

Table 3. Spatial and temporal inter-annotator agreement for the provided clicks. We report the percentage of clicks that match in specific temporal, spatial and joint (spatial-temporal) windows. Higher percentage points indicate a better agreement.

Statistic	Min	Max	Average
FPS	8	60	27.66
Frames	50	1814	477.5
Length (seconds)	2.01	72.56	17.54
Clicks	1	35	4.9

Table 4. Summary of statistics for ExDDV.

tween the higher Sentence-BERT and BERTScore values and the lower BLEU, METEOR and ROUGE scores indicates that the annotators rather used different words to explain the same artifacts in many cases.

We also assess the temporal and spatial agreement for the clicks provided by the two annotators. We measure the temporal alignment in terms of the percentage of click pairs that match inside a temporal window of at most 30, 60 and 120 frames, respectively. Since the average FPS is 30, the window lengths correspond to 1, 2 and 4 seconds, respectively. Similarly, we measure the spatial alignment in terms of the percentage of click pairs that match inside a spatial window of 50, 75, and 100 pixels. Finally, we compute the joint spatio-temporal agreement as the percentage of click pairs situated inside the same temporal and spatial windows. As shown in Tab. 3, more than 53% of the clicks made by the annotators are paired within a reasonable spatio-temporal window of 60 frames (about 2 seconds) and 75×75 pixels. This high percentage of matching clicks indicates a significant agreement in terms of the artifact locations pointed out by the two annotators through clicks.

Overall, the various inter-annotator agreement measures indicate that the annotations are consistent. In addition, we also visually inspected the collected annotations and confirmed that they are of high quality.

Statistics. The dataset consists of a total of 5,369 videos and 2,553,148 frames, with an average number of frames per video of approximately 477. We report several statistics about ExDDV in Tab. 4. The frames per second (FPS) rate varies from 8 to 60, 90% of all videos having either 25 or 30 FPS. The total number of clicks is 21,282. On average, each movie is annotated with around 4.9 clicks, the number of clicks per video ranging from 1 to 35. ExDDV comprises a wide range of video resolutions, from 480×272 to 1080×1920 pixels, with most videos having 720×1280 , 480×640 or 1080×1920 pixels (see Fig. S1 from the sup-

plementary for the number of videos per resolution). Each video is annotated with a single text description.

4. Explainable Methods

We consider various vision-language models as candidates for deepfake video XAI methods. Although there have been many efforts on visual question answering (VQA) [4, 8, 24, 51–53, 62, 66, 80], it was only with the significant growth of large language models (LLMs) [36, 44, 55, 58, 59, 63, 73] that capable vision-language models (VLMs) have recently emerged. For our task, we consider three different model families: Bootstrapping Language-Image Pre-training (BLIP) [40, 41], Large Language and Vision Assistant (LLaVA) [46] and Phi-3 [1].

4.1. Baseline architectures

All chosen VLMs involve three components: an image encoder, a text encoder, and a decoder. First, the input image and the query are encoded with their corresponding modules. The second step is to project the visual tokens into the same vector space as the text tokens. Finally, all tokens are combined and decoded to generate an answer.

BLIP-2. The original BLIP architecture is designed to train a unified VLM end-to-end. It bootstraps from noisy image-text pairs by generating synthetic captions and filtering out noise. The training regime combines contrastive learning, image-text matching and autoregressive language modeling. BLIP-2 avoids end-to-end learning by taking a different route. More specifically, it introduces a lightweight Querying Transformer (Q-Former) that serves as a bridge between the image encoder and the LLM. The Q-Former employs a fixed set of learnable query tokens to extract the most relevant features from the image encoder, resulting in an efficient design with minimal computational costs. We opt for the BLIP-2 model based on an Open Pretrained Transformer (OPT) backbone with 2.7B parameters. For efficiency, we use the version optimized through `int8` quantization.

Phi-3-Vision. The Phi-3 [1] suite of models ranges in functionality from solving pure language tasks to multimodal (vision and language) tasks. For our task, we employ the Phi-3-Vision model, which adopts a data-centric approach rather than simply increasing model size or applying distinct training regimes. The model is pre-trained using a curated pipeline that combines heavily filtered public web data with synthetic data generated by LLMs. This pre-training strategy first develops a broad language understanding, then enhances the model’s reasoning and logical inference capabilities. We select the architecture based on 128K tokens, which uses flash attention.

LLaVA-1.5. The LLaVA [46] family seamlessly integrates visual inputs with an LLM to generate natural language descriptions for visual tasks. When an image is provided as input, a pre-trained vision encoder extracts the relevant fea-

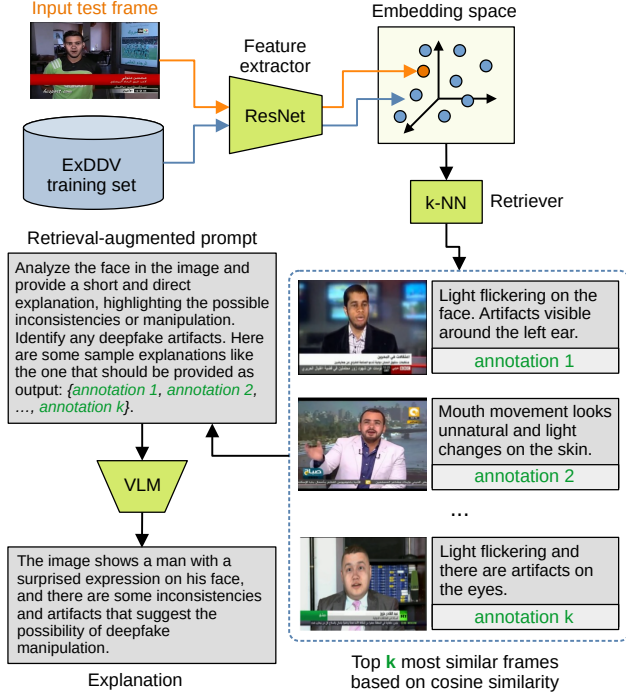


Figure 3. Overview of the in-context learning pipeline, which retrieves deepfake annotations from visually similar training frames using a k-NN based on a ResNet backbone. Best viewed in color.

tures. These are further projected into the LLM’s token space through a linear transformation that preserves both spatial and semantic information. The LLM processes this sequence auto-regressively, producing responses that are inherently grounded in visual input. In our application, this approach enables the model not only to accurately answer questions about deepfake detection, but also to provide detailed and interpretable explanations of its reasoning. We opt for the LLaVA-1.5 version, with 7B parameters.

Prompting. We prompt the pre-trained VLMs with the following query to obtain explanations about visual artifacts: “Analyze the face in the image. Identify any deepfake artifacts (if any), focusing specifically on the affected parts of the face. Provide a short and direct explanation highlighting the inconsistencies or manipulations”.

4.2. In-context learning

Pipeline. In Fig. 3, we illustrate our few-shot in-context learning pipeline. The training examples are chosen by a k-nearest neighbor (k-NN) model for each test video. The k-NN model extracts feature vectors with the CLIP [64] image encoder, which is based on the ResNet [31] architecture. Features are only extracted from the training frames that are annotated with clicks. The frames are further used independently by the k-NN. For efficient inference, we store the training feature embeddings along with the associated annotations. When a test image is provided, we obtain its fea-

tures via a similar process, extracting the ResNet features for the middle frame. Then, we search for the closest k training frames based on the cosine similarity between feature embeddings. The annotations of the nearest neighbors are used to enrich the custom prompt given to the VLM. The k-NN procedure is supposed to provide a set of k deepfake explanations that are potentially relevant for the test video. The custom prompt instructs the VLM to provide a similar explanation for the input test video.

Hyperparameters. We consider two alternative CLIP image encoders for the k-NN: ResNet-50 and ResNet-101. For the number of neighbors k , we test values in the set $\{1, 3, 5, 10\}$. The best choice is $k = 5$. To perform the k-NN search at various representation levels, we extract features at three different depth levels. We use low-level features right before the first residual block, mid-level features from an intermediate residual block (equally distanced from the input and output), and high-level features from the last layer. The features producing the best results are the high-level ones. We keep the default parameters of all VLMs during inference, except for the temperature, which is set to 0 to reduce the chance of generating hallucinations.

4.3. Fine-tuning

Pipeline. On the one hand, we follow the default fine-tuning methodology for the BLIP-2 model. On the other hand, given the increased memory requirements of Phi-3-Vision and LLaVA-1.5, we employ Low-Rank Adaptation (LoRA) [32] to fine-tune these two models.

To make a prediction for a whole video, the first step is to extract multiple frames from the video. Our strategy is to use equally-spaced frames between the first and last frames, while the number of frames depends on the model. Since BLIP-2 and Phi-3-Vision are lighter models, we provide 5 frames as input to these models. Phi-3-V supports multiple frames, while for BLIP-2, we process each frame with the visual encoder, then aggregate the resulting embeddings and provide the average embedding to the Q-former. Along with the input frames, we provide a unique query prompt: “What is wrong in the images? Explain why they look real or fake”. In our preliminary experiments, we tried to vary the prompt, considering even more complex prompts. However, we did not observe any significant differences, so we decided to stick with a single and relatively simple prompt for the presented experiments.

Hyperparameters. Phi-3-Vision and LLaVA are trained for 10 epochs with mini-batches of 16 and 32 samples, respectively. BLIP-2 is also fine-tuned for 10 epochs with a mini-batch size of 16, but with a gradient accumulation of 2. All models are optimized with AdamW, using a learning rate of $2 \cdot 10^{-4}$ and a cosine annealing scheduler to reduce the learning rate during training. For Phi-3-V, we use a LoRA rank of 64 and a dropout rate of 0.05. For LLaVA,

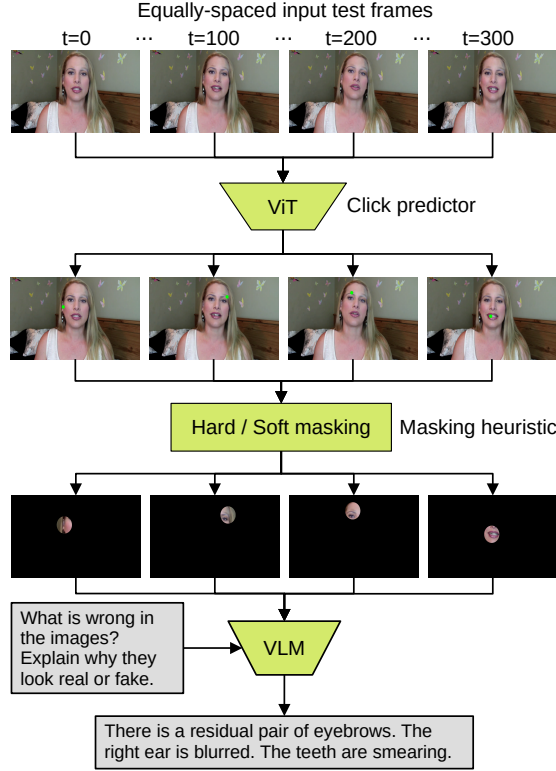


Figure 4. Our click supervision pipeline at inference time. A ViT-based click predictor estimates click coordinates for the input frames. A hard or soft masking is applied to mask the area outside the region of interest. The masked frames are given as input to a fine-tuned VLM. Best viewed in color.

we set the rank to 128 and do not employ dropout. The parameter α of LoRA is set to 256 for both models.

4.4. Click supervision

We harness the additional supervision signal in our dataset, namely the locations of the visual flaws represented by mouse clicks. While this idea was found to be useful in a number of vision tasks [7, 14, 16, 50, 60], to our knowledge, it has not been explored to train explainable video models. The rationale behind using click supervision is to guide the explainable model towards focusing on the region of interest (ROI) and consequently provide precise explanations.

In Fig. 4, we present our pipeline based on click-supervision, which comprises a click predictor based on ResNet, and a frame masking heuristic, which preserves the ROI, while masking the rest of the frame. To train the click predictor, we first extract all the training frames in which a visual glitch is annotated through a click, along with the coordinates of each click. The coordinates are further normalized between 0 and 1 to make the click predictor invariant to distinct frame resolutions. Next, we train a ViT-based regression model to predict click coordinates. The ViT [23] backbone is pre-trained on ImageNet. During inference, we

apply the regressor on each test frame to predict the clicks. Then, we mask the image area outside the ROI, which is defined as a round area of radius r around the click coordinates. We consider two alternative masking operations, soft and hard. Hard masking implies replacing the masked pixels with zero. Soft masking is based on a 2D Gaussian distribution centered in the predicted click location, where the $\sigma = r$. The masking operation is performed after the input images are normalized.

Hyperparameters. We consider two models pre-trained on ImageNet to predict clicks, a ViT-B and a ResNet-50. We replace their classification heads with a regression head to predict the two coordinates of a click. The click predictors are trained for 15 epochs with the AdamW optimizer and a learning rate of $9 \cdot 10^{-5}$. The learning rate is reduced on plateau by half, if the validation loss does not change for two consecutive epochs. An important hyperparameter for the masking heuristic is the mask radius r , which needs to be fixed to provide just enough context for the VLM. We tuned r between 50 and 150 pixels with a step of 5 pixels. The optimal radius is $r = 75$.

5. Experiments and Results

Research questions. Through our experiments, we aim to address the following research questions (RQs):

1. Are the collected annotations useful to train XAI models for deepfake videos?
2. What is the performance impact of click supervision?
3. How many data samples are required to train XAI models for deepfake videos?
4. Can the locations of artifacts be accurately predicted?

To address RQ1, we compare off-the-shelf (pre-trained) VLMs with VLMs based on two alternative training strategies, namely fine-tuning and in-context learning. To answer RQ2, in both types of learning frameworks, we harness the additional information provided through clicks and assess the performance gains brought by this supervision signal. To answer RQ3, we trained the top-scoring VLM with varying training set dimensions in the set $\{128, 256, 512, 1024, 2048, 4380\}$. To address RQ4, we report results with two click predictors that are both fine-tuned on ExDDV training data.

Evaluation measures. Although there are many measures to assess either the semantic similarity or the n-gram overlap of two text samples, such measures are not able to fully capture the similarity between texts. As a result, we evaluate the XAI models using a wide range of metrics, aiming to provide an extensive evaluation of the results. For semantic understanding, we employ Sentence-BERT [65] to embed both predicted and ground-truth descriptions, and compute the cosine similarity between the two. We also employ BERTScore [81] to compute the similarity at the token level. Every token in the ground-truth is greedily matched

Model	Masking	Sentence-BERT	BERTScore	BLEU	METEOR	ROUGE-1	ROUGE-2	ROUGE-L
BLIP-2 pre-trained		0.09	0.03	0.01	0.04	0.05	0.00	0.05
BLIP-2 in-context		0.44 $\pm$ 0.008	0.19 $\pm$ 0.004	0.03 $\pm$ 0.001	0.21 $\pm$ 0.001	0.14 $\pm$ 0.003	0.06 $\pm$ 0.001	0.14 $\pm$ 0.001
BLIP-2 in-context	hard	0.47 $\pm$ 0.002	0.20 $\pm$ 0.001	0.04 $\pm$ 0.001	0.21 $\pm$ 0.003	0.14 $\pm$ 0.005	0.06 $\pm$ 0.001	0.14 $\pm$ 0.002
BLIP-2 in-context	soft	0.46 $\pm$ 0.004	0.20 $\pm$ 0.003	0.04 $\pm$ 0.001	0.20 $\pm$ 0.001	0.14 $\pm$ 0.001	0.06 $\pm$ 0.001	0.14 $\pm$ 0.002
BLIP-2 fine-tuned		0.45 $\pm$ 0.009	0.29 $\pm$ 0.002	0.09 $\pm$ 0.010	0.14 $\pm$ 0.015	0.21 $\pm$ 0.010	0.09 $\pm$ 0.007	0.20 $\pm$ 0.009
BLIP-2 fine-tuned	hard	0.55 $\pm$ 0.005	0.36 $\pm$ 0.006	0.14 $\pm$ 0.005	0.22 $\pm$ 0.007	0.31 $\pm$ 0.004	0.38 $\pm$ 0.004	0.29 $\pm$ 0.004
BLIP-2 fine-tuned	soft	0.54 $\pm$ 0.006	0.36 $\pm$ 0.002	0.13 $\pm$ 0.002	0.21 $\pm$ 0.001	0.30 $\pm$ 0.004	0.14 $\pm$ 0.001	0.28 $\pm$ 0.003
Phi-3-V pre-trained		0.25	0.10	0.01	0.01	0.05	0.00	0.04
Phi-3-V in-context		0.30 $\pm$ 0.002	0.14 $\pm$ 0.003	0.03 $\pm$ 0.002	0.18 $\pm$ 0.005	0.15 $\pm$ 0.002	0.05 $\pm$ 0.002	0.13 $\pm$ 0.002
Phi-3-V in-context	hard	0.30 $\pm$ 0.002	0.14 $\pm$ 0.005	0.03 $\pm$ 0.004	0.18 $\pm$ 0.005	0.16 $\pm$ 0.003	0.05 $\pm$ 0.003	0.13 $\pm$ 0.005
Phi-3-V in-context	soft	0.30 $\pm$ 0.006	0.14 $\pm$ 0.003	0.03 $\pm$ 0.004	0.18 $\pm$ 0.004	0.15 $\pm$ 0.002	0.05 $\pm$ 0.003	0.13 $\pm$ 0.004
Phi-3-V fine-tuned		0.42 $\pm$ 0.005	0.30 $\pm$ 0.002	0.06 $\pm$ 0.001	0.20 $\pm$ 0.004	0.21 $\pm$ 0.005	0.07 $\pm$ 0.005	0.19 $\pm$ 0.006
Phi-3-V fine-tuned	hard	0.53 $\pm$ 0.004	0.38 $\pm$ 0.003	0.09 $\pm$ 0.004	0.27 $\pm$ 0.004	0.28 $\pm$ 0.002	0.11 $\pm$ 0.004	0.25 $\pm$ 0.003
Phi-3-V fine-tuned	soft	0.53 $\pm$ 0.003	0.38 $\pm$ 0.003	0.09 $\pm$ 0.008	0.26 $\pm$ 0.011	0.28 $\pm$ 0.009	0.11 $\pm$ 0.010	0.25 $\pm$ 0.009
LLaVA pre-trained		0.39	0.18	0.02	0.18	0.12	0.02	0.10
LLaVA in-context		0.48 $\pm$ 0.002	0.22 $\pm$ 0.008	0.03 $\pm$ 0.001	0.21 $\pm$ 0.001	0.12 $\pm$ 0.002	0.03 $\pm$ 0.002	0.10 $\pm$ 0.002
LLaVA in-context	hard	0.47 $\pm$ 0.002	0.29 $\pm$ 0.010	0.05 $\pm$ 0.001	0.22 $\pm$ 0.001	0.20 $\pm$ 0.002	0.04 $\pm$ 0.000	0.16 $\pm$ 0.002
LLaVA in-context	soft	0.50 $\pm$ 0.006	0.23 $\pm$ 0.004	0.04 $\pm$ 0.001	0.20 $\pm$ 0.002	0.12 $\pm$ 0.002	0.04 $\pm$ 0.001	0.11 $\pm$ 0.002
LLaVA fine-tuned		0.45 $\pm$ 0.003	0.33 $\pm$ 0.003	0.07 $\pm$ 0.002	0.23 $\pm$ 0.003	0.24 $\pm$ 0.001	0.08 $\pm$ 0.001	0.21 $\pm$ 0.001
LLaVA fine-tuned	hard	0.49 $\pm$ 0.011	0.35 $\pm$ 0.009	0.08 $\pm$ 0.005	0.24 $\pm$ 0.009	0.25 $\pm$ 0.011	0.09 $\pm$ 0.007	0.22 $\pm$ 0.008
LLaVA fine-tuned	soft	0.49 $\pm$ 0.006	0.35 $\pm$ 0.004	0.08 $\pm$ 0.002	0.25 $\pm$ 0.002	0.25 $\pm$ 0.003	0.09 $\pm$ 0.000	0.22 $\pm$ 0.003

Table 5. Results of various XAI models on ExDDV. Soft or hard input masking is based on click coordinates predicted by a ViT-based regressor. The top score for each model and evaluation metric is colored in blue.

with a token in the prediction to compute a recall. Similarly, every token in the prediction is matched with a token in the ground-truth to compute the precision. These are combined into an F1 score, called *BERTScore*. To assess n-gram overlaps, we adopt the most popular evaluation measures used in image captioning: BLEU [61], METEOR [5] and ROUGE [43]. For these metrics, we set the maximum n-gram length to $n = 2$. We evaluate click predictors in terms of the mean absolute error (MAE). We run each experiment three times and report the average scores and the standard deviations.

Quantitative results. The results of our experiments are shown in Tab. 5. We present the results for all models (BLIP-2, Phi-3-Vision and LLaVA-1.5) and learning scenarios (pre-training, in-context learning and fine-tuning). Click supervision is integrated via soft or hard masking, respectively. Consistent with the inter-annotator agreement scores, we observe that all models yield better scores in terms of semantic measures than n-gram overlap measures. The gap can be explained by the fact that models generate varied outputs, often using alternative phrases and words to express the same concept.

All fine-tuned models surpass the pre-trained versions by large margins. Although the metrics indicate that the pre-trained LLaVA is close to the fine-tuned models, a visual inspection of its generated answers indicates that they are very generic, in many cases just describing the videos and not providing any information about their authenticity. The

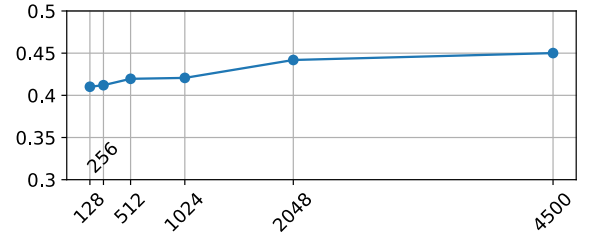


Figure 5. The performance of fine-tuned LLaVA (vertical axis) against the number of samples used for training (horizontal axis).

reported results offer strong evidence for RQ1, indicating that the collected annotations are useful to train XAI models, both via in-context learning and fine-tuning.

The reported scores also attest the advantages of integrating click predictions via soft and hard masking. This is observable for all three models, although the gains are somewhat lower for LLaVA. Both masking strategies appear to be equally effective. In response to RQ2, we find that using click supervision has a positive impact, boosting the performance of all tested VLMs.

In Fig. 5, we showcase the performance of the fine-tuned LLaVA model for different dimensions of the training set. The performance reaches a plateau after 2,000 training samples. This observation suggests that our dataset contains enough samples to train an explainable deepfake detection model, elucidating RQ3.

					
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.
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 6. Qualitative examples for LLaVA based on in-context learning and LLaVA fine-tuned with soft input masking. Best viewed in color.

Model	ViT	ResNet-50
MAE	0.0553	0.0595

Table 6. Results of click predictors.



Figure 7. 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.

In Tab. 6, we report the mean absolute errors for both ViT and ResNet-50 click predictors. The error of ViT is slightly lower, representing an average offset of only 12 pixels w.r.t. the ground-truth coordinates. These results confirm that the regression models can accurately localize visual artifacts, thus providing a positive answer to RQ4.

Qualitative results. Besides the quantitative measurements, we also present qualitative results. In Fig. 6, we illustrate some explanations provided by two variants of LLaVA. The examples include both relevant explanations as well as wrong explanations, *e.g.* identifying artifacts on real videos. We present examples for additional versions of LLaVA in Fig. S3 from the supplementary.

In Fig. 7, 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. We present additional click prediction examples in Fig. S4 from the supplementary.

6. Conclusion

In our work, we introduced a novel dataset for explainable deepfake detection in videos and made it publicly available. ExDDV consists of 5.4K manually annotated real and fake videos. In addition to the explanations for each video, the dataset also contains the locations of visual artifacts. We also explored different VLMs on the explainable deepfake detection task and evaluated their performance. The empirical results showed that the models are capable of learning to predict the source of visual errors in fake videos, while also detecting real videos. While the reported results are promising, we found that the tested VLMs are well below the inter-annotator agreement scores, suggesting that further exploration is required to build more capable models. In future work, we also aim to harness the difficulty labels, *e.g.* via curriculum learning, to boost performance.

With deepfake methods rapidly becoming more powerful and easier to access by the whole public, we consider our work as a stepping stone towards developing more robust and transparent detection models that will overcome the harms of deepfakes. We believe that our research will contribute to trustworthy AI systems that will only bring benefits to society, as well as reduce the skepticism around AI technology.

Acknowledgments

This work was supported by a grant of the Ministry of Research, Innovation and Digitization, CCCDI - UEFISCDI, project number PN-IV-P6-6.3-SOL-2024-2-0227, within PNCIDI IV. This research was also supported by the project “Romanian Hub for Artificial Intelligence - HRIA”, Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 351416.

References

- [1] Marah Abidin, Jyoti Aneja, Hany Awadallah, 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. 2, 4
- [2] Shruti Agarwal, Hany Farid, Tarek El-Gaaly, and Ser-Nam Lim. Detecting deep-fake videos from appearance and behavior. In *Proceedings of WIFS*, pages 1–6, 2020. 2
- [3] Irene Amerini and Roberto Caldelli. Exploiting prediction error inconsistencies through lstm-based classifiers to detect deepfake videos. In *Proceedings of IH&MMSec*, pages 97–102, 2020. 2
- [4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, Lawrence C. Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *Proceedings of ICCV*, pages 2425–2433, 2015. 4
- [5] Satantjeet Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of ACL*, pages 65–72, 2005. 7
- [6] Tyler Barrus. pypellchecker - Pure Python Spell Checking Library, 2024. Accessed: 2025-03-05. 3
- [7] Rodrigo Benenson, Stefan Popov, and Vittorio Ferrari. Large-scale interactive object segmentation with human annotators. In *Proceedings of CVPR*, pages 11700–11709, 2019. 6
- [8] Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samuel White, and Tom Yeh. VizWiz: nearly real-time answers to visual questions. In *Proceedings of UIST*, pages 333–342, 2010. 4
- [9] Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. Breaking common sense: WHOOPS! a vision-and-language benchmark of synthetic and compositional images. In *Proceedings of ICCV*, pages 2616–2627, 2023. 2, 3
- [10] Nicolo Bonettini, Edoardo Daniele Cannas, Sara Mandelli, Luca Bondi, Paolo Bestagini, and Stefano Tubaro. Video face manipulation detection through ensemble of CNNs. In *Proceedings of ICPR*, pages 5012–5019, 2021. 1, 2
- [11] Zhixi Cai, Kalin Stefanov, Abhinav Dhall, and Munawar Hayat. Do You Really Mean That? Content Driven Audio-Visual Deepfake Dataset and Multimodal Method for Temporal Forgery Localization. In *Proceedings of DICTA*, pages 1–10, 2022. 2
- [12] Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluating the Role of BLEU in Machine Translation Research. In *Proceedings of EACL*, pages 249–256, 2006. 3
- [13] Haoxing Chen, Yan Hong, Zizheng Huang, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Jun Lan, Huijia Zhu, Jianfu Zhang, Weiqiang Wang, and Huaxiong Li. DeMamba: AI-Generated Video Detection on Million-Scale GenVideo Benchmark. *arXiv preprint arXiv:2405.19707*, 2024. 2
- [14] Pengfei Chen, Xuehui Yu, Xumeng Han, Najmul Hassan, Kai Wang, Jiachen Li, Jian Zhao, Humphrey Shi, Zhenjun Han, and Qixiang Ye. Point-to-Box Network for Accurate Object Detection via Single Point Supervision. In *Proceedings of ECCV*, pages 51–67, 2022. 6
- [15] Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. EchoMimic: Lifelike Audio-Driven Portrait Animations through Editable Landmark Conditions. In *Proceedings of AAAI*, pages 2403–2410, 2025. 1
- [16] Bowen Cheng, Omkar Parkhi, and Alexander Kirillov. Pointly-supervised instance segmentation. In *Proceedings of CVPR*, pages 2617–2626, 2022. 6
- [17] Jongwook Choi, Taehoon Kim, Yonghyun Jeong, Seungryul Baek, and Jongwon Choi. Exploiting style latent flows for generalizing deepfake video detection. In *Proceedings of CVPR*, pages 1133–1143, 2024. 2
- [18] Davide Cozzolino, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. ID-Reveal: Identity-aware DeepFake Video Detection. In *Proceedings of ICCV*, pages 15108–15117, 2021. 2
- [19] Florinel-Alin Croitoru, Andrei-Iulian Hiji, Vlad Hondru, Nicolae Catalin Ristea, Paul Irofti, Marius Popescu, Cristian Rusu, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Deepfake Media Generation and Detection in the Generative AI Era: A Survey and Outlook. *arXiv preprint arXiv:2411.19537*, 2024. 1, 2, 3
- [20] Florinel-Alin Croitoru, Vlad Hondru, Marius Popescu, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. MAVOS-DD: Multilingual Audio-Video Open-Set Deepfake Detection Benchmark. *arXiv preprint arXiv:2505.11109*, 2025. 2
- [21] Ilke Demir and Umur Aybars Çiftçi. How Do Deepfakes Move? Motion Magnification for Deepfake Source Detection. In *Proceedings of WACV*, pages 4768–4778, 2024. 2
- [22] Brian Dolhansky, Joanna Bitton, Ben Pfau, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (DFDC) dataset. *arXiv preprint arXiv:2006.07397*, 2020. 1, 2, 3
- [23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of ICLR*, 2021. 2, 6
- [24] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question answering. In *Proceedings of NeurIPS*, pages 2296–2304, 2015. 4
- [25] Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, and Lizhuang Ma. Delving into the local: Dynamic inconsistency learning for deepfake video detection. In *Proceedings of AAAI*, pages 744–752, 2022. 2
- [26] Zhihao Gu, Taiping Yao, Yang Chen, Shouhong Ding, and Lizhuang Ma. Hierarchical contrastive inconsistency learning for deepfake video detection. In *Proceedings of ECCV*, pages 596–613, 2022. 1, 2
- [27] Jiazhi Guan, Hang Zhou, Zhibin Hong, Errui Ding, Jingdong Wang, Chengbin Quan, and Youjian Zhao. Delving into sequential patches for deepfake detection. In *Proceedings of NeurIPS*, pages 4517–4530, 2022. 2

- [28] David Güera and Edward J Delp. Deepfake video detection using recurrent neural networks. In *Proceedings of AVSS*, pages 1–6, 2018. 2
- [29] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don’t lie: A generalisable and robust approach to face forgery detection. In *Proceedings of CVPR*, pages 5037–5047, 2021. 2
- [30] Alexandros Haliassos, Rodrigo Mira, Stavros Petridis, and Maja Pantic. Leveraging real talking faces via self-supervision for robust forgery detection. In *Proceedings of CVPR*, pages 14930–14942, 2022. 1, 2
- [31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of CVPR*, pages 770–778, 2016. 5
- [32] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proceedings of ICLR*, 2022. 5
- [33] Juan Hu, Xin Liao, Jinwen Liang, Wenbo Zhou, and Zheng Qin. FInfer: Frame Inference-Based Deepfake Detection for High-Visual-Quality Videos. In *Proceedings of AAAI*, pages 951–959, 2022. 1, 2
- [34] Gazi Hasin Ishrak, Zalish Mahmud, MD Farabe, Tahera Khanom Tinni, Tanzim Reza, and Mohammad Zavid Parvez. Explainable Deepfake Video Detection using Convolutional Neural Network and CapsuleNet. *arXiv preprint arXiv:2404.12841*, 2024. 2
- [35] Sarthak Jain, Sarah Wiegrefe, Yuval Pinter, and Byron C. Wallace. Learning to faithfully rationalize by construction. In *Proceedings of ACL*, pages 4459–4473, 2020. 2
- [36] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023. 4
- [37] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deepforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of CVPR*, pages 2889–2898, 2020. 1, 2, 3
- [38] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S Woo. FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset. In *Proceedings of NeurIPS*, 2021. 2
- [39] Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. In *Proceedings of EMNLP*, pages 107–117, 2016. 2
- [40] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of ICML*, pages 12888–12900, 2022. 4
- [41] 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. 2, 4
- [42] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A large-scale challenging dataset for deepfake forensics. In *Proceedings of CVPR*, pages 3207–3216, 2020. 2
- [43] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of summaries. In *Proceedings of ACL*, page 10, 2004. 7
- [44] Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Daya Guo, et al. DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model. *arXiv preprint arXiv:2405.04434*, 2024. 4
- [45] Baoping Liu, Bo Liu, Ming Ding, Tianqing Zhu, and Xin Yu. TI2Net: Temporal Identity Inconsistency Network for Deepfake Detection. In *Proceedings of WACV*, pages 4691–4700, 2023. 1, 2
- [46] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *Proceedings of NeurIPS*, pages 34892–34916, 2023. 2, 4
- [47] Tao Liu, Feilong Chen, Shuai Fan, Chenpeng Du, Qi Chen, Xie Chen, and Kai Yu. AniTalker: Animate Vivid and Diverse Talking Faces through Identity-Decoupled Facial Motion Encoding. In *Proceedings of ACM MM*, pages 6696–6705, 2024. 1
- [48] Yibing Liu, Haoliang Li, Yangyang Guo, Chenqi Kong, Jing Li, and Shiqi Wang. Rethinking Attention-Model Explainability through Faithfulness Violation Test. In *Proceedings of ICML*, pages 13807–13824, 2022. 2
- [49] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of NeurIPS*, pages 4765–4774, 2017. 2
- [50] Junwei Luo, Xue Yang, Yi Yu, Qingyun Li, Junchi Yan, and Yansheng Li. PointOBB: Learning Oriented Object Detection via Single Point Supervision. In *Proceedings of CVPR*, pages 16730–16740, 2024. 6
- [51] Lin Ma, Zhengdong Lu, and Hang Li. Learning to answer questions from image using convolutional neural network. In *Proceedings of AAAI*, pages 3567–3573, 2016. 4
- [52] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Proceedings of NeurIPS*, pages 1682–1690, 2014.
- [53] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of ICCV*, pages 1–9, 2015. 4
- [54] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. Two-branch recurrent network for isolating deepfakes in videos. In *Proceedings of ECCV*, pages 667–684, 2020. 2
- [55] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large Language Models: A Survey. *arXiv preprint arXiv:2402.06196*, 2024. 4
- [56] Akash Kumar Mohankumar, Preksha Nema, Sharan Narasimhan, Mitesh M. Khapra, Balaji Vasan Srinivasan, and Balaraman Ravindran. Towards Transparent and Explainable Attention Models. In *Proceedings of ACL*, pages 4206–4216, 2020. 2
- [57] Daniel Mas Montserrat, Hanxiang Hao, Sri K. Yarlagadda, Sriram Baireddy, Ruiting Shao, János Horváth, Emily Bartu-

- siak, Justin Yang, David Guera, Fengqing Zhu, et al. Deepfakes detection with automatic face weighting. In *Proceedings of CVPR*, pages 668–669, 2020. 1, 2
- [58] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A Comprehensive Overview of Large Language Models. *arXiv preprint arXiv:2307.06435*, 2023. 4
- [59] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Proceedings of NeurIPS*, pages 27730–27744, 2022. 4
- [60] Dim P. Papadopoulos, Jasper R.R. Uijlings, Frank Keller, and Vittorio Ferrari. Training object class detectors with click supervision. In *Proceedings of CVPR*, pages 6374–6383, 2017. 6
- [61] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL*, pages 311–318, 2002. 3, 7
- [62] A.J. Piergiovanni, Wei Li, Weicheng Kuo, Mohammad Safar, Fred Bertsch, and Anelia Angelova. Answer-Me: Multi-Task Open-Vocabulary Visual Question Answering. *arXiv preprint arXiv:2205.00949*, 2022. 4
- [63] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-Training. *OpenAI*, 2018. 4
- [64] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of ICML*, pages 8748–8763, 2021. 5
- [65] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of EMNLP*, pages 3982–3992, 2019. 3, 6
- [66] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. In *Proceedings of NeurIPS*, pages 2953–2961, 2015. 4
- [67] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “Why should I trust you?” Explaining the predictions of any classifier. In *Proceedings of SIGKDD*, pages 1135–1144, 2016. 2
- [68] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of ICCV*, pages 1–11, 2019. 1, 2, 3
- [69] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. *Proceedings of CVPR*, pages 80–87, 2019. 2
- [70] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128:336–359, 2020. 2
- [71] Michał Stypułkowski, Konstantinos Vougioukas, Sen He, Maciej Zięba, Stavros Petridis, and Maja Pantic. Diffused Heads: Diffusion Models Beat GANs on Talking-Face Generation. In *Proceedings of WACV*, pages 5091–5100, 2024. 1
- [72] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. EMO: Emote Portrait Alive – Generating Expressive Portrait Videos with Audio2Video Diffusion Model under Weak Conditions. In *Proceedings of ECCV*, pages 244–260, 2024. 1
- [73] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*, 2023. 4
- [74] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of NeurIPS*, pages 6000–6010, 2017. 2
- [75] Cong Wang, Kuan Tian, Jun Zhang, Yonghang Guan, Feng Luo, Fei Shen, Zhiwei Jiang, Qing Gu, Xiao Han, and Wei Yang. V-Express: Conditional Dropout for Progressive Training of Portrait Video Generation. *arXiv preprint arXiv:2406.02511*, 2024. 1
- [76] Tianyi Wang and Kam Pui Chow. Noise Based Deepfake Detection via Multi-Head Relative-Interaction. In *Proceedings of AAAI*, pages 14548–14556, 2023. 1, 2
- [77] Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Yao Yao, and Siyu Zhu. Hallo: Hierarchical Audio-Driven Visual Synthesis for Portrait Image Animation. *arXiv preprint arXiv:2406.08801*, 2024. 1
- [78] Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Baining Guo. VASA-1: Lifelike Audio-Driven Talking Faces Generated in Real Time. In *Proceedings of NeurIPS*, pages 660–684, 2025. 1
- [79] Zhiyuan Yan, Yandan Zhao, Shen Chen, Xinghe Fu, Taiping Yao, Shouhong Ding, and Li Yuan. Generalizing Deepfake Video Detection with Plug-and-Play: Video-Level Blending and Spatiotemporal Adapter Tuning. *arXiv preprint arXiv:2408.17065*, 2024. 1, 2
- [80] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of CVPR*, pages 21–29, 2016. 4
- [81] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of ICLR*, 2020. 3, 6
- [82] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimic-Motion: High-Quality Human Motion Video Generation with Confidence-aware Pose Guidance. *arXiv preprint arXiv:2406.19680*, 2024. 1
- [83] Yiru Zhao, Wanfeng Ge, Wenxin Li, Run Wang, Lei Zhao, and Jiang Ming. Capturing the persistence of facial expression features for deepfake video detection. In *Proceedings of ICICS*, pages 630–645, 2020. 2

- [84] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *Proceedings of ICCV*, pages 15044–15054, 2021. [2](#)
- [85] Ruiqi Zhong, Steven Shao, and Kathleen McKeown. Fine-grained sentiment analysis with faithful attention. *arXiv preprint arXiv:1908.06870*, 2019. [2](#)
- [86] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection. In *Proceedings of ACMMM*, pages 2382–2390, 2020. [2](#)