Why Fake ? Unveiling the Semantic Vocabulary of Deepfake Detectors

Vazgken Vanian* Alexandros Doumanoglou* Dimitris Zarpalas
vvanian@iti.gr aldoum@iti.gr zarpalas@iti.gr

Information Technologies Institute (ITI)
Centre For Research and Technology HELLAS (CERTH)

Abstract

Deepfake (DF) technology poses a significant threat to information integrity, driving the need for robust detection methods. Most DF detectors only consider predicting a binary label for whether the input is real or fake, lacking the justification required for real-world applications like legal proceedings. Explainable DF Detection has emerged to address this limitation, but existing techniques frequently fall short by either relying on human annotations for precise artifact localization or generating superficially plausible textual explanations without grounding. This work investigates the use of post-hoc explainable AI (XAI) to analyze the decision-making process of state-of-the-art black-box DF detectors. Specifically, we employ Encoding-Decoding Direction Pairs (EDDP), a technique suitable for uncovering the concept space of DF detectors (their semantic vocabulary) as well as the mechanism for writing and reading concept information to and from internal representations. Our analysis reveals previously hidden real and fake features learned implicitly during detector training, offering nuanced explanations unattainable through conventional methods. This enables global model understanding, spatially aware concept localization, and counterfactual what-if analysis, all contributing to a deeper comprehension of DF detection strategies.

1. Introduction

The proliferation of deepfake (DF) content presents an escalating challenge to information integrity across various domains, from journalism and politics to legal proceedings [7, 17]. As generative models continue to advance, realism has surpassed previously unimaginable thresholds, necessitating the development of robust detection methodologies. The majority of existing approaches focus on binary classification [1, 34]. While effective, this approach suffers from a critical limitation: a lack of transparency and justification. In contexts demanding accountability a simple *real* or *fake* label is insufficient. Instead, understanding *why* a piece of content is flagged as manipulated is paramount.

This need for explainability has spurred recent research into Explainable Deepfake Detection (XDFD). Existing approaches broadly fall into two categories: spatio-temporal localization [5, 15, 27] and textual explanation methods [14].

Spatio-temporal localization methods aim to identify manipulated regions by highlighting areas of the image or segments of the video that are presumed to contain forged content. The motivation is to move beyond binary prediction and provide visual evidence supporting the decision. However, in practice, these methods often produce coarse localization, frequently defaulting to highlighting the entire face rather than isolating specific manipulation artifacts. As a result, they provide limited insight into what precise cues led to the prediction.

Textual explanation methods instead generate natural language justifications describing why an image was classified as fake. While these explanations improve human interpretability, they often lack accurate spatial grounding and may not faithfully reflect the underlying evidence used by the detector. Consequently, despite their explanatory intent, existing XDFD methods primarily function as prediction tools with attached justification mechanisms, rather than systems that deliver precise and causally grounded explanations.

In this work, we explore XDFD from a different perspective by analyzing deepfake detectors through the lens of **post-hoc** Explainable Artificial Intelligence (XAI). Rather than modifying detectors to explicitly produce explanations, we aim to uncover and interpret the intrinsic features learned during standard training.

This perspective provides several advantages. First, explanations are derived directly from the representations that drive the detector's prediction, ensuring tight alignment between explanation and decision function and thereby improving faithfulness. Second, because our approach is purely post-hoc, it introduces no architectural modifications or auxiliary training objectives, avoiding potential trade-offs between interpretability and predictive performance. Third, the

*Equal contribution

framework can be applied to existing pretrained detectors without retraining or artifact-level supervision, making it deployable across models and datasets.

Our technical approach centers on Encoding-Decoding Direction Pairs (EDDP) [11], a recent technique that can unveil a) the concepts that a deep vision network uses to make predictions (its semantic vocabulary), and b) the network's mechanism for encoding and decoding these concepts in internal representations, under the linear representation hypothesis [2]. We argue that EDDP concepts are a natural fit for XDFD because access to the network's encoding-decoding mechanism for real and fake concepts enables: (a) global model understanding, by identifying which concepts drive predictions toward *real* or *fake*; (b) spatially aware, concept-based local explanations; and (c) counterfactual what-if analysis.

To the best of our knowledge, this is the first application of post-hoc, concept-based XAI within the deepfake detection domain, opening a new avenue for improving transparency and trust in automated content verification systems.

2. Related Work

2.1. DeepFake Generation

Deepfake synthesis methods can be broadly categorized into face-swapping, face-reenactment, and audio-driven lip-syncing.

Face-swapping replaces identity while preserving pose and expression. Early works relied on encoder–decoder frameworks and required person-specific training [21, 24]. Later many-to-many models such as SimSwap [6] enabled identity transfer across unseen subjects. Subsequent works improved temporal coherence and realism using transformers and motion–appearance decoupling strategies [9, 18, 25]. More recently, diffusion-based approaches have emerged [19, 38, 42, 45]. They leverage pretrained generative priors for high-fidelity synthesis.

Face-reenactment transfers motion and expression from a driving source to a target identity. Face2Face [37] pioneered 3D model-based reenactment, followed by learning-based approaches that emphasize generalization and one-shot transfer [3, 33]. Large-scale diffusion video models further improve realism and temporal stability [32].

Lip-syncing methods generate mouth motion from speech signals. Wav2Lip [30] introduced adversarial synchronization learning, while SadTalker [44] models 3D motion coefficients for natural head dynamics. Diffusion-based systems such as [12] extend this paradigm toward holistic audio-driven human video generation.

2.2. Deepfake Detection and Datasets

Deepfake detection has evolved alongside advances in generation techniques. Early methods relied on hand-crafted cues, such as abnormal eye blinking or head-pose inconsistencies [22, 41]. CNN-based approaches like MesoNet [1] and XceptionNet [34] captured subtle spatial artifacts, while frequency-aware methods such as F3-Net [31] exploited frequency-domain inconsistencies. Transformer-based architectures later integrated spatial and temporal frequency representations for improved robustness [20], and recent CLIP-based approaches leverage vision-language representations with parameter-efficient fine-tuning to enhance cross-dataset generalization [39, 43].

Datasets have grown in parallel to match these advances, from early benchmarks like UADFV [41] and Deepfake-TIMIT to FaceForensics++ (FF++) [34] and Celeb-DF [23], which improved realism and temporal consistency. Larger-scale collections such as DFDC [10], DeeperForensics-1.0 [16], ForgeryNet [13], and recent benchmarks DF40 [40] and DDL [28] provide greater diversity and coverage of modern generation techniques. These datasets form the foundation for evaluating robust and generalizable deepfake detectors.

2.3. Explainability in DeepFake Detection

Explainability for deepfake detection remains comparatively underexplored, with most existing works relying on post-hoc saliency maps like Grad-CAM [35] that provide limited insight into model reasoning. Recent research has introduced several distinct strategies to address this. ProtoExplorer [4] utilizes prototype-based reasoning within a human-in-the-loop framework, allowing predictions to be grounded in learned visual archetypes. In a different approach, the Locally-Explainable Self-Blended (LESB) detector [36] focuses on Local Feature Discovery to decompose global predictions into discrete, region-level contributions. This is achieved with part specific self-blending augmentations that focus on key areas like eyes,nose and mouth.

Multimodal transparency is explored in ExDDV [14], which pairs spatial annotations with textual justifications tailored for vision-language models. Furthermore, network dissection is utilized in [26] to quantify transparency by aligning individual internal CNN neurons with semantic facial concepts.

While these individual methods advance the interpretability of detection models, they often introduce significant operational overhead. For instance, the network dissection approach in [26] requires fine-tuning standard architectures, such as VGG-16, Inception V3, or ResNet-50, on forensic datasets to map internal activations to meaningful concepts. Similarly, methods like ProtoExplorer [4] impose specific architectural constraints through the requirement of specialized prototype layers. These dependencies suggest that high interpretability currently relies on additional supervision or task-specific retraining, leaving open the challenge of developing truly scalable, model-intrinsic explanation mechanisms that can be applied to diverse, unlabelled datasets

without architectural modification.

3. Unveiling DeepFake Concepts with Encoding-Decoding Direction Pairs (EDDP)

To understand how deepfake detectors represent and utilize information about real and fake content, we leverage a technique called **Encoding-Decoding Direction Pairs (EDDP)**. A core assumption underlying EDDP is the **linear representation hypothesis** [2]. This hypothesis posits that concepts, in this case, features indicative of real or fake images, are encoded within image representations as linear combinations of concept embeddings, i.e., vectors in the latent space, each associated with a concept.

3.1. Method

Given a feature representation $\boldsymbol{X} \in \mathbb{R}^{D\times H\times W}$ and its constituent patch embeddings $\boldsymbol{x_p} \in \mathbb{R}^D$, EDDP jointly learns the matrix of decoding directions $\boldsymbol{W} \in \mathbb{R}^{D\times I}$, the encoding directions $\boldsymbol{S} \in \mathbb{R}^{D\times I}$, and a vector of thresholds $\boldsymbol{b} \in \mathbb{R}^I$ for a predefined number of concepts $I \in \mathbb{N}^+$.

The learned decoding directions $\boldsymbol{W}$, coupled with thresholds $\boldsymbol{b}$, serve as deepfake related concept detectors. For any patch $\boldsymbol{x_p}$, the presence of concept i is determined by the condition:

$$\boldsymbol{w}_i^T \boldsymbol{x_p} - b_i > 0 \tag{1}$$

The encoding directions $\boldsymbol{S}$ represent the concept embeddings, which denote the specific directions in the latent space along which each concept is represented.

3.2. Concept Detection Maps and Concept Contribution Maps

We localize each concept within an image using concept detectors. To do this, we generate a **Concept Detection Map** or **Concepts Presence Map (CPMs)**: first, the detector is applied to each spatial element of the image representation $\boldsymbol{X}$, then a hard thresholding step creates a binary map, and finally, the map is up-sampled to match the original image resolution. While Concept Detection Maps identify the presence of a concept, they do not quantify that concept's actual influence on the model's final prediction.

To quantify this influence, we compute **Concept Contribution Maps (CCMs)**. CCMs are based on a sample's concept contributions and baseline concept contributions to the explanation logit, which is the difference between the network's class logit for an input image and a class logit corresponding to a baseline artificial representation from the uncertainty region, where concept detectors are at their decision threshold, as defined by [11]. CCMs provide a spatially aware breakdown and quantify the contributions of each concept across patches.

Visualization examples for CPMs and CCMs are shown in Figure 2.

4. Concept Analysis and Global Understanding

4.1. Experimental Setup

For our experimental framework, we employ an Xception [8] architecture pre-trained on the FaceForensics++ (FF++) [34] dataset. We apply the EDDP method to the activations of the model's 12th residual block, using the FF++ training set as our concept extraction source.

The selection of the 12th residual block is motivated by a trade-off between semantic depth and the dimensionality of the feature space. While deeper layers represent higher-level semantic information, they often undergo a "rank collapse" as the network converges toward a single classification decision. Our empirical analysis using Principal Component Analysis (PCA) indicates that the 12th block maintains a sufficiently high-rank feature space compared to subsequent layers. PCA reveals that this layer maintains a high-rank feature space (rank ≈ 25), capturing over 90% of the variance. We empirically set the number of concepts to 16 concepts after evaluating the trade-off between concept-transfer accuracy, misclassified samples correction success, and semantic simplicity across settings of 12 to 24 concepts.

4.2. Concept Identification

The objective of our concept identification process is to isolate human-interpretable concepts that govern the model's final decision-making process. To characterize the learned concepts, establish their semantic groundedness, and analyze their distribution across the dataset, we employ three validation procedures: Relative Concept Activation Vector (RCAV) sensitivity analysis, semantic mapping, and distribution analysis.

It is important to note that the concepts identified through this process are intrinsically coupled with both the architectural priors of the Xception backbone and the specific data distribution of FF++. Because EDDP decomposes the model's internal latent space, the resulting encoding and decoding directions represent the specific features and artifacts that the detector has learned to prioritize for this particular classification task. Consequently, these concepts reflect the intersection of the model's representational capacity and the visual cues present within the FF++ dataset

We evaluate the functional influence of each identified concept on the model's decision-making process using RCAV [29] sensitivity scores. As detailed in Table 1, these scores quantify the model's sensitivity to specific concepts c_i with respect to the "Real" and "Fake" class predictions. Our analysis reveals that a subset of concepts $(c_1, c_3, c_9, c_{10}, c_{12}, c_{14})$ consistently biases the model's output toward the "Fake" class, serving as primary indicators of manipulation. Conversely, concepts such as c_0, c_5, c_7, c_8, and c_{11} strongly correlate with the "Real" class prediction. Notably, certain concepts (c_2, c_4, c_6, c_{15}) exhibit negligible

Figure 1. Visualization of the discovered concept vocabulary. Each concept c_i is illustrated by representative face patches with high activation. The accompanying RCAV scores indicate the relative contribution of each concept toward a "real" or "fake" classification

impact on the final logit, suggesting they represent features that, while present, are not utilized as discriminative evidence by the detector.

Following this, we quantitatively map the concepts to specific facial regions using Interesection over Union (IoU) scores. These are calculated between concept presence maps and a ground-truth proxy derived from facial segmentation predictions. As detailed in Table 2, this mapping allows us to associate abstract concept directions with interpretable semantic classes (e.g. c_5 showing high IoU with the "nose" region).

Furthermore, we use the IoU metric, the RCAV sensitivity scores as well as qualitative inspection of the top-5 most strongly activated samples (as shown in Figure 1) and assign descriptive names to the identified concepts. A summary of this naming is provided in Table 3.

4.3. Concept Distribution and Manipulation Statistics

To understand how these concepts manifest across the dataset, we calculate the dataset-wide concept presence as shown in Table 4. This reveals the prevalence of specific features in real versus manipulated imagery. Further, we analyze the overlap between identified concepts and specific manipulation methods present in the FF++ dataset (FS,NT,DF,FF) in Table 5.

Notably, there are certain concepts that exhibit high specificity to certain deepfake types. For instance, c_3 and c_9 show significantly higher presence in DeepFakes (DF) and Face2Face (FF) compared to real images, suggesting these

concepts capture artifacts inherent to those specific generative processes. This distribution remains consistent across both training and testing splits, indicating the robustness of the identified concepts.

Table 1. RCAV sensitivity scores per concept on the test set. Higher positive scores indicate a stronger positive correlation with the respective class

	c0	**c1**	**c2**	**c3**	**c4**	**c5**	**c6**	**c7**
Real	0.73	-0.39	-0.32	-0.38	-0.43	0.94	-0.13	0.99
Fake	-0.29	0.88	0.36	0.97	0.19	-0.37	0.03	-0.43
	c8	**c9**	**c10**	**c11**	**c12**	**c13**	**c14**	**c15**
Real	0.98	-0.35	-0.32	0.84	-0.48	0.98	-0.20	0.09
Fake	-0.44	0.88	0.97	-0.35	1.00	0.79	0.96	-0.05

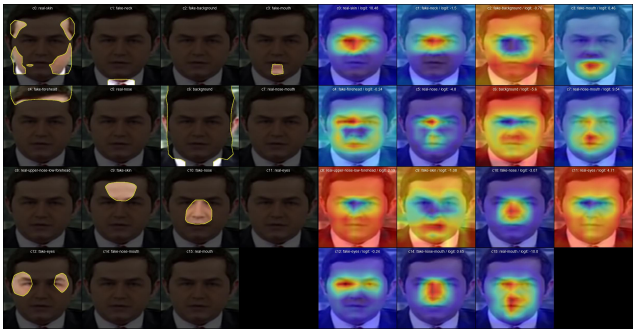

Figure 2. Visualization of Concept Presence Maps (left) and the corresponding Concept Contribution Maps (right). Presence maps indicate the spatial activation of each learned concept, while contribution maps highlight their influence on the model's prediction.

5. Explanation Faithfulness Assessment

To evaluate the faithfulness of the learned concept explanations, we introduce a concept cloning intervention that directly manipulates the classifier's internal representation and measures whether the resulting predictions change in a controlled and predictable manner. If the learned concepts are genuinely integrated into the model's decision process, then intervening on the concept coefficients should systematically affect the model output.

We begin by decomposing the intermediate representation at a selected layer. Specifically, we express the representation as a base component $\boldsymbol{X}_{bc}$, which lies in the uncertainty region of all concepts, and a coefficient vector $\boldsymbol{u}$ that quantifies the deviation from this concept-neutral point along each learned signal direction.

To assess faithfulness, we sample pairs of inputs $(\boldsymbol{X}_s, \boldsymbol{X}_t)$, referred to as source and target examples, and compute their respective decompositions:

$$\boldsymbol{X}_s = \boldsymbol{X}_{bc,s} + \boldsymbol{S}^\top \boldsymbol{u}_s, \tag{2}$$

$$\boldsymbol{X}_t = \boldsymbol{X}_{bc,t} + \boldsymbol{S}^\top \boldsymbol{u}_t. \tag{3}$$

We then construct a synthetic representation by combining the source base component with the target concept coefficients:

$$\boldsymbol{X}_{\text{syn}} = \boldsymbol{X}_{bc,s} + \boldsymbol{S}^\top \boldsymbol{u}_t. \tag{4}$$

This operation preserves the structural component of the source representation while transplanting the concept-specific deviations of the target. The synthetic representation is injected into the classifier, and faithfulness is quantified by measuring the agreement between the predictions of the synthetic example and those of the target.

In a complementary evaluation, we assess whether concept-level interventions can correct misclassified samples. For each misclassified instance, we apply targeted interventions to the concept coefficients by adding or removing specific concepts to support the flipping operation. The goal is to flip the prediction to the correct label and report the resulting correction accuracy.

The results of both evaluations are presented in Table 6. The concept-transfer experiment yielded 87.34% accuracy while the intervention on misclassified samples achieved a 99.8% success rate. These results suggest that the concept coefficients successfully capture the specific features the model uses for classification. Specifically, cloning these coefficients onto a different base representation reliably results in the target prediction. Furthermore, by performing targeted interventions on these coefficients, one can successfully steer the model's internal logic toward correcting misclassified samples. These two observations provide strong empirical evidence that the identified concepts are the primary drivers of the final output.

6. Local Concept-Based Explanations

To understand how individual concepts influence a specific classification, we decompose the model's logit with the help of CCMs. We focus on a test sample with a ground truth label of "Fake", which the model correctly classifies.

We employ CCMs to visualize the spatial influence of each concept on the final decision. In Figure 2 , we depict the concept presence maps (CPMs) (left) with their corresponding CCMs (right). We observe that the heatmaps are highly localized, for instance, the activations of c_{12} and c_3 are highly activated for regions around the eyes and mouth respectively. This confirms that the model is reacting to specific artifacts within these facial components rather than some global noise.

Furthermore in Figure 3 we provide the Grad-CAM [35] heatmap, the CPMs union, and the quantitative decomposition of the prediction where we decompose the total logit into individual concepts contributions and a residual term that is equivalent to the proportion of logits not explained by

Table 2. Semantic alignment of learned concepts via IoU. We report the IoU between concept activation maps and ground-truth facial semantic masks. The highest overlap for each concept is highlighted in blue, while non-zero overlaps are in red

	background	skin	nose	eye_g	l_eye	r_eye	l_brow	r_brow	l_ear	r_ear	mouth	u_lip	l_lip	hair	hat	ear_r	neck_l	neck	cloth
c0	0.03	**0.21**	0.01	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.01	0.01	0.05	0.00	0.01	0.00	0.04	0.03
c1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.22**	0.05
c2	**0.09**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.01	0.00	0.00	0.00	0.00
c3	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	**0.07**	0.00	0.00	0.00	0.00	0.01	0.00
c4	0.04	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.30**	0.01	0.00	0.00	0.00	0.00
c5	0.00	0.02	**0.19**	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
c6	**0.29**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.00	0.00	0.18	0.00	0.00	0.00	0.04	0.16
c7	0.00	0.03	**0.13**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.05	0.02	0.00	0.00	0.00	0.00	0.00	0.00
c8	0.00	**0.04**	0.01	0.00	0.00	0.00	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
c9	0.01	**0.04**	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
c10	0.00	0.03	**0.15**	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
c11	0.00	0.03	0.00	0.01	0.01	0.02	0.04	**0.07**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
c12	0.00	**0.03**	0.00	0.01	0.01	0.01	0.03	0.03	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
c13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
c14	0.03	0.02	**0.09**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.01	0.01	0.00	0.00	0.00	0.00	0.00
c15	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.05	**0.09**	0.00	0.00	0.00	0.00	0.00	0.00

Table 3. Concept ID to semantic label mapping. Labels are derived from a combination of the semantic IoU analysis 2 and qualitative inspection of high-activation patches

c0: real-skin	**c1**: fake-neck	**c2**: fake-background	**c3**: fake-mouth
c4: fake-forehead	**c5**: real-nose	**c6**: background	**c7**: real-nose-mouth
c8: real-upper-nose-low-forehead	**c9**: fake-skin	**c10**: fake-nose	**c11**: real-eyes
c12: fake-eyes	**c13**: N/A	**c14**: fake-nose-mouth	**c15**: real-mouth

Table 4. Global concept presence across the dataset. We report the percentage of images in which each concept c_i is active

	Train		Test	
	Real (%)	**Fake (%)**	**Real (%)**	**Fake (%)**
c0	21.8	78.2	21.6	78.4
c1	20.8	75.3	20.9	77.0
c2	10.7	38.3	11.2	38.6
c3	2.7	45.3	2.8	42.8
c4	21.7	78.1	21.6	78.4
c5	20.9	18.9	19.9	19.2
c6	21.8	78.2	21.6	78.4
c7	20.7	17.0	19.4	16.9
c8	19.5	30.1	18.2	31.5
c9	0.6	34.8	0.7	32.5
c10	0.1	35.0	0.2	33.3
c11	19.9	22.6	18.2	22.2
c12	1.7	43.3	1.9	40.4
c13	0.0	0.0	0.0	0.0
c14	4.1	34.5	5.1	35.2
c15	20.3	20.3	19.3	22.7

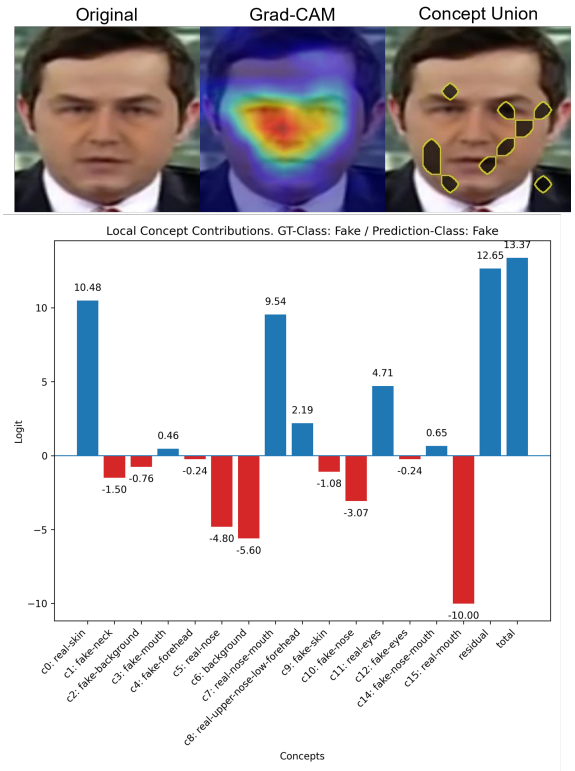

Figure 3. Visualization of the original image, it's Grad-CAM heatmap and the union of all Concept Presence Maps (top). Logit decomposition into individual concept contributions and the residual term (bottom)

any concept.

7. What-if-Analysis: Counter-Factual Explanations

To evaluate the causal influence of our identified concepts on the model's final decision, we perform a series of counterfactual interventions. By systematically adding or removing concepts, we can observe whether the model's prediction "flips".

7.1. Intervention methodology

We define a set of candidate concepts for intervention based on their RCAV sensitivity scores. Specifically, any concept

Table 5. Concept overlap by manipulation type and real samples. We break down concept activations across four specific forgery techniques: FaceSwap (FS), NeuralTextures (NT), DeepFake (DF), and Face2Face (FF).

	Train				
	Real (%)	**FS (%)**	**NT (%)**	**DF (%)**	**FF (%)**
c0	100.0	100.0	100.0	100.0	100.0
c1	95.7	95.6	97.3	96.0	96.4
c2	49.1	48.7	49.4	49.5	48.5
c3	12.4	30.7	17.4	90.3	79.3
c4	99.8	99.7	99.9	99.7	99.9
c5	96.2	10.0	91.5	3.8	2.1
c6	100.0	100.0	100.0	100.0	100.0
c7	95.3	9.4	84.6	1.4	1.6
c8	89.5	16.7	85.3	8.6	48.5
c9	2.7	4.3	3.9	74.5	79.0
c10	0.3	70.6	1.6	95.1	8.1
c11	91.6	6.8	88.3	3.5	24.5
c12	8.0	72.0	10.3	75.4	58.0
c13	0.0	0.0	0.0	0.0	0.0
c14	19.0	30.5	19.9	32.6	85.5
c15	93.4	28.3	71.0	4.1	10.0
	Test				
	Real (%)	**FS (%)**	**NT (%)**	**DF (%)**	**FF (%)**
c0	100.0	100.0	100.0	100.0	100.0
c1	96.8	97.6	98.8	98.9	97.9
c2	51.6	49.1	51.7	47.3	49.4
c3	13.0	31.5	14.9	81.7	78.7
c4	99.9	100.0	100.0	100.0	99.9
c5	91.8	14.3	83.5	5.4	3.9
c6	100.0	100.0	100.0	100.0	100.0
c7	89.5	13.2	76.4	2.8	2.8
c8	84.3	27.3	78.2	13.8	46.2
c9	3.4	5.9	4.0	66.4	75.6
c10	1.1	60.5	3.1	91.4	10.9
c11	84.3	12.5	80.2	6.5	21.0
c12	8.7	64.5	11.3	70.1	55.2
c13	0.0	0.0	0.0	0.0	0.0
c14	23.6	35.9	24.0	32.4	81.7
c15	89.4	33.1	71.6	6.7	13.1

Table 6. Concept Faithfulness Evaluation: Accuracy in Concept-Transfer and Correcting Miss-classified samples.

	Concept-Transfer	Correcting Misclassified
Accuracy:	87.34%	99.8%

c_i whose absolute sensitivity score satisfies Eq. 5 is considered a primary driver of the model's prediction and thus a candidate for manipulation. Formally,

$$|S(c_i)| > \tau, \tag{5}$$

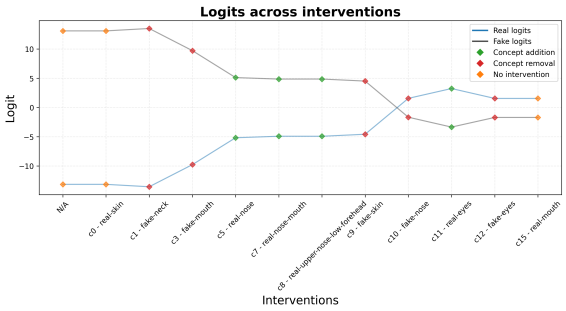

Figure 4. Logit shift under counterfactual concept intervention. We illustrate the causal impact of specific concepts on a single test sample. By starting from a baseline (N/A) and sequentially adding (green) or removing (red) semantic concepts c_i, we observe the dynamic shift in model logits

where $\tau = 0.9$ and $S(c_i) \in \mathbb{R}$ denotes the scalar RCAV sensitivity score associated with concept c_i.

Prior to intervention, we compute concept-specific activation statistics. For each concept, we estimate representative signal values using distributional statistics and compute the corresponding top and bottom quantiles, which are later used during intervention, similar to [11].

Interventions are performed by modifying the latent representation along the direction of the selected candidate concepts. We consider three types of interventions:

(i) **Concept addition**, where we randomly sample a mask from samples in which the concept is active and substitute the concept-related information using the top-quantile statistics to strengthen its presence; (ii) **Concept removal**, where we suppress a concept associated with the opposing class by substituting the representation with the bottom-quantile statistics; and (iii) **No intervention**, where the latent representation remains unchanged if a high-sensitivity concept is already present (or absent) in a manner consistent with the target class.

7.2. Single-Sample Intervention Trace

Figure 4 illustrates a successful counterfactual trajectory for a sample initially classified as "Fake". We apply a sequence of interventions, beginning with the removal of fake-associated concepts and followed by the addition of real-associated concepts.

As shown in the logit plot, the initial state exhibits a high "Fake" logit and a low "Real" logit. By removing concepts such as c_3 (fake-mouth) and c_9 (fake-skin), and adding c_{11} (real-eyes), each intervention progressively shifts the logits toward the target direction, ultimately flipping the prediction. This step-wise transition demonstrates that the model's decision is strongly tied to the identified concepts.

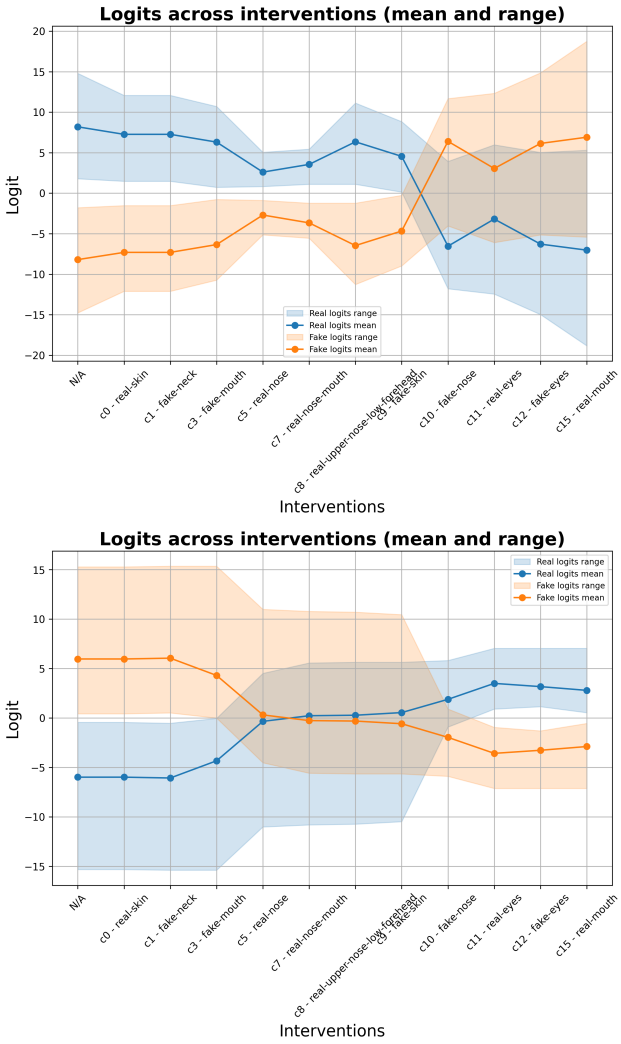

Figure 5. Statistical distribution of logits across sequential interventions. We report the mean (solid lines) and range (shaded areas) of logits for subsets of samples initially predicted as ”Real” (top) and ”Fake” (bottom)

7.3. Statistical Robustness of Interventions

To verify that the observed counterfactual behavior is not limited to a single example, we repeat the analysis on a balanced set of 20 samples: 5 real (ground-truth real, predicted real), 5 real (ground-truth real, predicted fake), 5 fake (ground-truth fake, predicted fake), and 5 fake (ground-truth fake, predicted real). Figure 5 reports the mean and range of logit changes across all samples, grouped by their initial prediction.

For samples initially classified as “Real”, interventions consistently decrease the mean “Real” logit while increasing the “Fake” logit. Conversely, samples initially classified as “Fake” exhibit the opposite trend, being steered toward “Real”. The consistency of these trends, reflected in the shaded ranges, suggests that the identified concepts are robust across samples and that the model systematically relies on them to distinguish between real and manipulated faces.

8. Discussion and Conclusion

In this study, we applied the EDDP method to provide post-hoc interpretability for deepfake detectors. We successfully identified a semantic vocabulary of internal concepts, effectively mapping abstract latent activations to interpretable human-understandable semantics.

Our faithfulness assessments confirmed that the learned concept directions unveil the internals of the model’s decision-making process. The ability to reliably reproduce target predictions by changing their coefficients suggests they capture the essential characteristics used for classification. This is further supported by the CCMs, where concepts and their logit contributions are grounded to facial areas. Finally, the counterfactual ”what-if” analysis provides causal evidence for the role of these concepts. By systematically manipulating concept presence and observing the resulting shifts in prediction logits, we demonstrated that the model’s internal logic can be effectively steered. This series of experiments add a level of transparency that is crucial for deepfake detectors.

Despite these strengths, several limitations define the current scope of this work. First, while EDDP does not require retraining the underlying deepfake detector, learning of the EDDP is still required on the representation space of the model to unveil the encoding and decoding directions of concepts. Second, the number of concepts remains a manually defined hyperparameter which may lead to redundant or insufficient signal directions if not selected correctly. Lastly, the identified concepts are inherently optimized for the specific model architecture and dataset used during training. These semantic directions are likely not directly transferable across different models or datasets, meaning a separate training process may be required for each model-dataset combination.

This work demonstrates that the ”black-box” nature of deepfake detection can be unraveled through concept-based interpretability. By grounding classification logic in identifiable facial regions and providing a mechanism to steer model predictions we enable more transparent and trustworthy detection.

9. Acknowledgments

This research has been supported by the European Commission funded program DETECTOR, under Horizon Europe Grant Agreement 101225942.

References

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*, pages 1–7. IEEE, 2018. 1, 2

[2] Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety–a review. *arXiv preprint arXiv:2404.14082*, 2024. 2, 3

[3] Stella Bounareli, Christos Tzelepis, Vasileios Argyriou, Ioannis Patras, and Georgios Tzimiropoulos. Hyperreenact: one-shot reenactment via jointly learning to refine and retarget faces. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7149–7159, 2023. 2

[4] Merel de Leeuw den Bouter, Javier Lloret Pardo, Zeno Geradts, and Marcel Worring. Protoexplorer: Interpretable forensic analysis of deepfake videos using prototype exploration and refinement. *Information Visualization*, 23(3):239–257, 2024. 2

[5] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18710–18719, 2022. 1

[6] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2003–2011, 2020. 2

[7] Bobby Chesney and Danielle Citron. Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.*, 107:1753, 2019. 1

[8] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 3

[9] Kaiwen Cui, Rongliang Wu, Fangneng Zhan, and Shijian Lu. Face transformer: Towards high fidelity and accurate face swapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 668–677, 2023. 2

[10] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020. 2

[11] Alexandros Doumanoglou, Kurt Driessens, and Dimitrios Zarpalas. Learning encoding-decoding direction pairs to unveil concepts of influence in deep vision networks. *arXiv preprint arXiv:2509.23926*, 2025. 2, 3, 7

[12] Jiazhi Guan, Kaisiyuan Wang, Zhiliang Xu, Quanwei Yang, Yasheng Sun, Shengyi He, Borong Liang, Yukang Cao, Yingying Li, Haocheng Feng, Errui Ding, Jingdong Wang, Youjian Zhao, Hang Zhou, and Ziwei Liu. Audcast: Audio-driven human video generation by cascaded diffusion transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[13] Yinan He, Bei Gan, Siyu Chen, Yichun Zhou, Guojun Yin, Luchuan Song, Lu Sheng, Jing Shao, and Ziwei Liu. Forgerynet: A versatile benchmark for comprehensive forgery analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4360–4369, 2021. 2

[14] Vlad Hondru, Eduard Hogea, Darian Onchis, and Radu Tudor Ionescu. Exddv: A new dataset for explainable deepfake detection in video. *arXiv preprint arXiv:2503.14421*, 2025. 1, 2

[15] Juan Hu, Xin Liao, Difei Gao, Satoshi Tsutsui, Qian Wang, Zheng Qin, and Mike Zheng Shou. Delocate: detection and localization for deepfake videos with randomly-located tampered traces. *arXiv preprint arXiv:2401.13516*, 2024. 1

[16] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2889–2898, 2020. 2

[17] Iqra Khan, Kashif Khan, and Arshad Ahmad. A comprehensive survey of deepfake generation and detection techniques in audio-visual media. *ICCK Journal of Image Analysis and Processing*, 1(2):73–95, 2025. 1

[18] Jiseob Kim, Jihoon Lee, and Byoung-Tak Zhang. Smooth-swap: A simple enhancement for face-swapping with smoothness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10779–10788, 2022. 2

[19] Kihong Kim, Yunho Kim, Seokju Cho, Junyoung Seo, Jisu Nam, Kychul Lee, Seungryong Kim, and KwangHee Lee. Diffface: Diffusion-based face swapping with facial guidance. *Pattern Recognition*, 163:111451, 2025. 2

[20] Taehoon Kim, Jongwook Choi, Yonghyun Jeong, Haeun Noh, Jaejun Yoo, Seungryul Baek, and Jongwon Choi. Beyond spatial frequency: Pixel-wise temporal frequency-based deepfake video detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11198–11207, 2025. 2

[21] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 3677–3685, 2017. 2

[22] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International workshop on information forensics and security (WIFS)*, pages 1–7. Ieee, 2018. 2

[23] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3207–3216, 2020. 2

[24] Kunlin Liu, Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Wenbo Zhou, and Weiming Zhang. Deepfacelab: Integrated, flexible and extensible face-swapping framework. *Pattern Recognition*, 141:109628, 2023. 2

[25] Xiangyang Luo, Ye Zhu, Yunfei Liu, Lijian Lin, Cong Wan, Zijian Cai, Yu Li, and Shao-Lun Huang. Canonswap: High-fidelity and consistent video face swapping via canonical space modulation. In *Proceedings of the IEEE/CVF Interna-*

tional Conference on Computer Vision, pages 10064–10074, 2025. 2

[26] Nazneen Mansoor and Alexander I Iliev. Explainable ai for deepfake detection. *Applied Sciences*, 15(2):725, 2025. 2

[27] Changtao Miao, Qi Chu, Zhentao Tan, Zhenchao Jin, Tao Gong, Wanyi Zhuang, Yue Wu, Bin Liu, Honggang Hu, and Nenghai Yu. Multi-spectral class center network for face manipulation detection and localization. *arXiv preprint arXiv:2305.10794*, 2023. 1

[28] Changtao Miao, Yi Zhang, Weize Gao, Man Luo, Weiwei Feng, Zhiya Tan, Jianshu Li, Ajian Liu, Yunfeng Diao, Qi Chu, et al. Ddl: A dataset for interpretable deepfake detection and localization in real-world scenarios. *arXiv preprint arXiv:2506.23292*, 2025. 2

[29] Jacob Pfau, Albert T Young, Jerome Wei, Maria L Wei, and Michael J Keiser. Robust semantic interpretability: Revisiting concept activation vectors. *arXiv preprint arXiv:2104.02768*, 2021. 3

[30] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 484–492, 2020. 2

[31] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, pages 86–103. Springer, 2020. 2

[32] Di Qiu, Zhengcong Fei, Rui Wang, Jialin Bai, Changqian Yu, Mingyuan Fan, Guibin Chen, and Xiang Wen. Skyreels-a1: Expressive portrait animation in video diffusion transformers. *arXiv preprint arXiv:2502.10841*, 2025. 2

[33] Andre Rochow, Max Schwarz, and Sven Behnke. Fsrt: Facial scene representation transformer for face reenactment from factorized appearance head-pose and facial expression features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7716–7726, 2024. 2

[34] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 3

[35] 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. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2, 5

[36] Elahe Soltandoost, Richard Plesh, Stephanie Schuckers, Peter Peer, and Vitomir Štruc. Extracting local information from global representations for interpretable deepfake detection. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 1629–1639, 2025. 2

[37] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016. 2

[38] Runqi Wang, Yang Chen, Sijie Xu, Tianyao He, Wei Zhu, Dejia Song, Nemo Chen, Xu Tang, and Yao Hu. Dynamicface: High-quality and consistent face swapping for image and video using composable 3d facial priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13438–13447, 2025. 2

[39] Zhiyuan Yan, Jiangming Wang, Zhendong Wang, Peng Jin, Ke-Yue Zhang, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Effort: Efficient orthogonal modeling for generalizable ai-generated image detection. *arXiv preprint arXiv:2411.15633*, 2, 2024. 2

[40] Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie Wang, Shouhong Ding, Yunsheng Wu, et al. Df40: Toward next-generation deepfake detection. *Advances in Neural Information Processing Systems*, 37:29387–29434, 2024. 2

[41] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 8261–8265. IEEE, 2019. 2

[42] Fulong Ye, Miao Hua, Pengze Zhang, Xinghui Li, Qichao Sun, Songtao Zhao, Qian He, and Xinglong Wu. Dreamid: High-fidelity and fast diffusion-based face swapping via triplet id group learning. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–10, 2025. 2

[43] Andrii Yermakov, Jan Cech, and Jiri Matas. Unlocking the hidden potential of clip in generalizable deepfake detection. *arXiv preprint arXiv:2503.19683*, 2025. 2

[44] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8652–8661, 2023. 2

[45] Wenliang Zhao, Yongming Rao, Weikang Shi, Zuyan Liu, Jie Zhou, and Jiwen Lu. Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8568–8577, 2023. 2