DF-CBM: Region-Aware Concept Bottleneck Models for Deepfake Detection

Georgios Tsoumplekas[1], Vazgken Vanian[2], Alexandros Doumanoglou[2], Panos K. Papadopoulos[2], Yannis Spyridis[3], Dimitrios Zarpalas[2], and Vasileios Argyriou[1]

[1] Department of Networks & Digital Media, Kingston University London, Kingston upon Thames, UK
{G.Tsoumplekas,Vasileios.Argyriou}@kingston.ac.uk
[2] Centre for Research and Technology Hellas, Thessaloniki, Greece
{vvanian,aldoum,ppapadop,zarpalas}@iti.gr
[3] Department of Computer Science, Kingston University London, Kingston upon Thames, UK
Y.Spyridis@kingston.ac.uk

Abstract. Deepfake detection methods have become increasingly effective yet most provide limited insight into the evidence behind their predictions. However, in forensic settings users also need to know which manipulation cues support the decision and where they appear. Existing explainability methods only partially address this need since localization-based approaches lack semantic descriptions while language-based explanation methods are only weakly grounded in visual evidence. In this work, we propose DF-CBM, a region-aware concept bottleneck model for explainable deepfake detection. DF-CBM builds a compact vocabulary of manipulation-related concepts from textual artifact annotations and links each concept to plausible facial and boundary regions. It then predicts these concepts from visual features using a concept-specific masked attention mechanism guided by parsed facial masks and the final real/fake decision is made from the predicted concept bottleneck. Our experiments show that DF-CBM outperforms concept-based baselines in concept prediction and deepfake classification while remaining competitive with state-of-the-art black-box detectors. Finally, qualitative results and intervention analyses demonstrate that DF-CBM provides spatially grounded concept evidence and enables counterfactual explanations of how individual manipulation concepts influence the final prediction.

Keywords: Deepfake Detection · Concept Bottleneck Model · Concept Extraction

1 Introduction

Recent advances in generative modeling [12, 22, 40, 51, 67] have enabled the synthesis and manipulation of highly realistic visual media commonly referred to as deepfakes [36]. While these technologies have beneficial applications in entertainment [6, 32], media production [70] and digital art [14], they also create

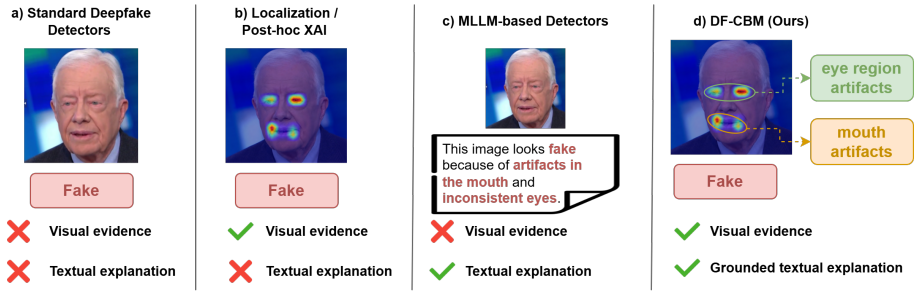


Fig. 1: Explanation capabilities of existing deepfake detection approaches and DF-CBM that provides both localized visual evidence and grounded manipulation concepts.

substantial opportunities for misuse including political disinformation, identity fraud, coordinated misinformation campaigns and non-consensual content generation [24, 53].

To address these risks, a wide range of deepfake detection methods have been developed to distinguish authentic from manipulated content [7, 29, 41, 44, 46]. Yet, despite substantial progress, many detectors remain limited by their black-box nature providing little insight into the visual evidence that drives their predictions. This lack of transparency is particularly problematic in forensic and high-stakes settings where reliable decision-making requires not only accurate predictions but also explanations that are transparent, verifiable and suitable for human oversight.

Existing explainable deepfake detection methods mainly follow two directions, as illustrated in Fig. 1. Localization-based approaches, such as saliency maps, attention visualizations and other post-hoc explanations, highlight image regions that influence a detector's prediction [4, 45, 54]. However, these methods typically indicate where the model focuses without providing a semantic description of the forensic evidence. In contrast, recent MLLM-based approaches generate textual explanations that describe possible manipulation artifacts in natural language [17, 48, 58, 69]. While more interpretable to humans, such explanations are often only weakly grounded in the image and may describe artifacts without establishing a precise correspondence to the supporting visual evidence. As a result, current methods either provide visual evidence without semantic concepts, or textual explanations without sufficiently reliable spatial grounding.

To address this gap, we introduce DF-CBM, a visually grounded concept bottleneck framework for explainable deepfake detection. DF-CBM first constructs a compact vocabulary of manipulation-related concepts and associates each concept with plausible facial or boundary regions. It then learns a region-aware concept bottleneck model that predicts these concepts from visual features while encouraging each concept to rely on the facial regions where that artifact is expected to appear. This way, each predicted manipulation concept is associated not only with a semantic label but also with localized visual evidence, improving the interpretability of the final deepfake prediction.

Notably, unlike conventional black-box detectors and post-hoc explanation methods, DF-CBM makes the intermediate forensic evidence explicit since the final prediction is obtained from predicted manipulation concepts rather than directly from unconstrained visual features. This enables counterfactual analyses and allows users to inspect which concepts contributed to the decision and where their supporting evidence appears in the image. We evaluate DF-CBM under intra-dataset and cross-dataset settings, comparing it against both state-of-the-art black-box detectors and concept-based baselines. Our results show that DF-CBM outperforms concept-based baselines in both concept prediction and real/fake classification while achieving competitive detection performance against state-of-the-art detectors. We further provide qualitative attention maps and contribution scores to visualize the spatial evidence and decision-relevant concepts and conduct a concept intervention analysis to examine how modifying individual concept activations affects the final prediction. Our contributions can be summarized as follows:

- We introduce DF-CBM, a region-aware concept bottleneck framework for explainable deepfake detection that predicts semantically meaningful manipulation concepts and grounds them in anatomically relevant facial regions.
- We develop an automated pipeline for constructing and grounding forensic concepts, combining text-based concept extraction with facial parsing masks and concept-specific attention to support each concept prediction with localized visual evidence.
- We demonstrate that DF-CBM outperforms concept-based baselines in both concept prediction and deepfake detection while achieving competitive detection performance against state-of-the-art black-box detectors and enhanced interpretability through attention maps, contribution scores and concept intervention analysis.

2 Related Work

2.1 Deepfake Detection Methods

Early deepfake detection methods focused on identifying visual inconsistencies introduced by face manipulation pipelines including physiological, geometric and boundary-related artifacts [29, 65]. With the release of large-scale benchmarks [10, 19, 31, 41, 71] CNN-based detectors became the dominant paradigm. Several influential methods observed that face manipulation often introduces blending artifacts near facial boundaries leading to approaches such as Face X-Ray [29] and Self-Blended Images [44] which improve generalization by learning manipulation-independent cues.

More recent work has shifted toward generalizable detection under unseen manipulation methods and datasets. UCF [62] disentangles common forgery cues from content- and method-specific artifacts, while LSDA [59] expands the latent forgery space to reduce overfitting to known manipulation patterns. Other approaches improve robustness through forgery augmentation [33], discrepancy

learning [66], frequency debiasing [23], frequency-guided adaptation [2] or adaptation of pretrained vision-language representations [7, 28, 46].

2.2 Explainability in Deepfake Detection

Explainability in deepfake detection has been explored through visual evidence, prototype-based reasoning and localized explanation methods. Dynamic Prototype Networks [52] use temporal prototypes to explain video-level deepfake dynamics, while ProtoExplorer [4] provides a visual analytics interface for prototype-based forensic models. Other works localize decision-relevant facial evidence using local explanations or post-hoc analysis [45, 54]. While such methods improve transparency over standard black-box detectors, they often explain predictions through saliency, prototypes or discovered concepts that are not explicitly constrained to correspond to predefined, human-interpretable manipulation concepts grounded in facial anatomy.

A complementary line of work uses language supervision or multimodal large models to generate richer explanations of manipulated content. Specifically, Visual-linguistic face forgery detection [48], DD-VQA [69] and ExDDV [15] introduce textual artifact descriptions, question-answering formulations and localized language annotations for explainable deepfake detection. Meanwhile, more recent methods and benchmarks [17, 20, 21, 49, 58] evaluate or exploit multimodal reasoning for detection, localization and explanation. However, free-form textual explanations may be weakly grounded in the image and can be difficult to verify spatially. In contrast, DF-CBM avoids unconstrained explanation generation by predicting a fixed set of manipulation concepts, each explicitly linked to anatomically meaningful facial regions.

2.3 Concept Bottleneck Models

Concept Bottleneck Models (CBMs) provide ante-hoc interpretability by constraining predictions to pass through an intermediate layer of human-understandable concepts [25, 26]. This enables concept-level explanations and test-time interventions but early CBMs required dense concept annotations, limiting their scalability. Subsequent work has relaxed this requirement through richer concept representations [11] and weaker supervision [56].

Recent CBM variants have further improved scalability and spatial grounding. Label-Free CBMs [38] and Post-hoc CBMs [68] leverage vision-language models and concept activation vectors to construct interpretable bottlenecks without dense manual annotations. Spatially-aware approaches, such as SALF-CBM [3] and locality-aware CBMs [18] extend CBMs from global concept activations to localized concept maps while CLIP-free extensions broaden their applicability beyond vision-language supervision [42]. Yet despite this progress, CBMs have not been systematically explored for explainable deepfake detection, where concepts must capture subtle manipulation artifacts and be grounded in anatomically meaningful facial regions.

3 Methodology

Given an input face image, DF-CBM predicts a set of semantically meaningful manipulation concepts and uses them as an intermediate representation for final real/fake classification. The framework follows a two-stage pipeline. In the first stage, we construct a compact vocabulary of manipulation-related concepts from text-enhanced deepfake datasets and associate each concept with plausible facial or boundary regions using large language models (LLMs). In the second stage, we train a region-aware concept bottleneck model to predict these concepts from frozen visual features while explicitly grounding each concept prediction in the corresponding facial regions.

3.1 Concept Vocabulary Extraction and Region Mapping

The first stage of DF-CBM aims to construct a compact set of manipulation-related concepts and associate each concept with the facial regions in which it is expected to appear. This produces a region-grounded concept vocabulary that is later used to guide the region-aware concept bottleneck.

We start from textual annotations in deepfake explanation datasets, including ExDDV [15], DD-VQA [69] and FakeClue [57]. These annotations describe visual evidence of manipulation types in natural language. Given each text description annotation, we prompt Qwen3.5-9B [64] to extract atomic manipulation concepts that are concise visual cues that correspond to individual artifacts rather than full sentence-level explanations. The goal is to decompose complex descriptions into simpler concept candidates, such as eye artifacts, mouth distortions, boundary inconsistencies and global facial abnormalities.

The initial concept set is then filtered and consolidated to remove noisy and redundant entries. We first deduplicate the extracted concepts and encode them using EmbeddingGemma-300M [55]. The resulting embeddings are clustered using HDBSCAN to group semantically equivalent concepts. For each cluster, we select the top-k entries closest to the cluster centroid and provide them to the same LLM which synthesizes a standardized concept label. The same top-k entries and concept label are then used by the LLM to assign each concept to a predefined set of facial and boundary regions, corresponding to the regions extracted by the facial parsing model in Sec. 3.2. This process yields a binary concept–region mapping $A \in \{0,1\}^{K \times R}$, where K is the number of concepts and R is the number of facial and boundary regions. Each entry $A_{k,r} = 1$ indicates that region r is anatomically relevant for concept k and may be used as visual evidence by the bottleneck model.

Finally, we remove rare concepts that occur in fewer than $N_{\min}$ training samples, with $N_{\min} = 750$ in our experiments. This filtering avoids training concept predictors for poorly supported concepts and improves the reliability of the final vocabulary. The vocabulary size is therefore governed by a trade-off between semantic granularity and supervisory density. Here, we favor a compact set of recurring manipulation cues that are sufficiently represented in the data,

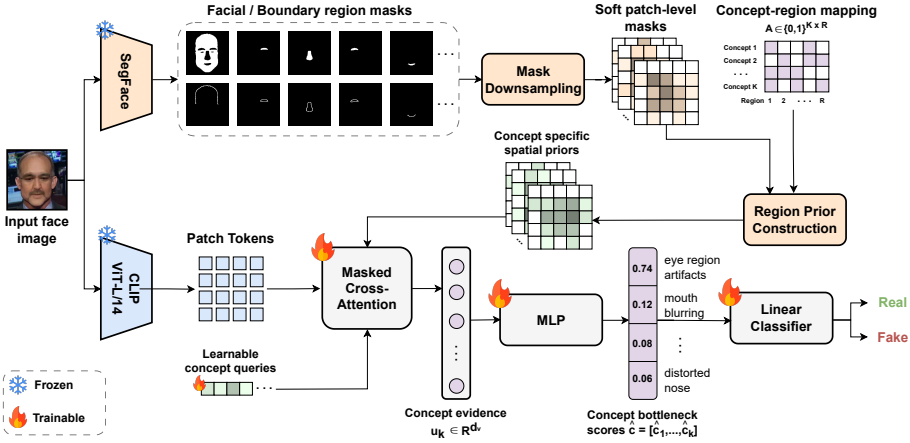


Fig. 2: Overview of the region-aware concept bottleneck model in DF-CBM.

improving concept prediction robustness while preserving interpretability. After applying this criterion, the resulting vocabulary contains $K = 6$ concepts and serves as the semantic basis for the second stage of DF-CBM. Importantly, the architecture does not impose a fixed vocabulary size, allowing larger concept sets when sufficiently dense and reliable supervision is available.

3.2 Region-Aware Concept Bottleneck Model

Fig. 2 illustrates the second stage of DF-CBM, which predicts manipulation-related concepts while grounding each prediction in anatomically relevant facial regions. Specifically, given a face image $x \in \mathbb{R}^{3\times H\times W}$ with height H and width W, the model outputs a concept bottleneck score $\hat{\mathbf{c}} = [\hat{c}_1, \ldots, \hat{c}_K]$, where each $\hat{c}_k \in [0, 1]$ denotes the predicted presence probability of concept k.

We first extract patch-level visual features using a frozen CLIP [39] image encoder. This produces a sequence of patch tokens $F = \{f_i\}_{i=1}^{N}$, with $f_i \in \mathbb{R}^D$, where $N = 256$ corresponds to the 16×16 patch grid and D denotes the CLIP feature dimensionality. In parallel, we use a frozen SegFace [37] facial parsing model to obtain semantic facial region masks at the same image resolution. For each facial region r, we denote the corresponding binary mask as $M_r^{\text{img}} \in \{0, 1\}^{H\times W}$.

In addition to the semantic facial regions, we derive boundary masks between predefined adjacent regions, such as skin–nose, skin–hair, mouth–upper lip and lower lip–upper lip. The goal of these masks is to capture transition areas between facial parts motivated by the fact that deepfake artifacts often appear near region transitions, blending seams or facial part boundaries [29]. Boundary masks are computed directly from the facial parsing output by selecting pixels lying on the boundary of two neighboring regions.

To align the facial masks with the CLIP patch tokens, each image-level mask M_r^{img} is downsampled to the 16×16 patch grid using average pooling with kernel size and stride equal to the CLIP patch size. This gives a soft patch-level mask $M_r = [m_{r,1}, \ldots, m_{r,N}]$, where $m_{r,i} \in [0, 1]$ denotes the fraction of patch i covered by region r. The use of soft masks avoids assigning each patch to a single region and preserves partial region coverage within each patch.

Next, for each concept k, we construct a concept-specific spatial prior $P_k = [p_{k,1}, \ldots, p_{k,N}]$ over the CLIP patch grid by combining the patch-level masks of the regions associated with that concept:

$$p_{k,i} = \max_{r:A_{k,r}=1} m_{r,i}. \tag{1}$$

Here, $A_{k,r} = 1$ is the concept-region mapping entry indicating that region r is relevant for concept k. We use the maximum so that a patch receives high prior weight if it overlaps with any region associated with the concept. This is done because many manipulation concepts may correspond to multiple plausible facial regions and averaging would unnecessarily weaken the prior when only one of these regions is present in a patch. Overall, the resulting concept-specific spatial prior specifies which patches are anatomically plausible evidence for predicting concept k.

However, the actual evidence for each concept must still be inferred from the visual features. We therefore use a concept-specific masked cross-attention mechanism to extract visual evidence from the CLIP patch tokens. Each concept k is represented by a learnable query vector $q_k \in \mathbb{R}^d$, initialized from the CLIP text embedding of the corresponding concept name. The query attends over the patch tokens while the spatial prior is added to the attention logits as a soft mask:

$$\alpha_{k,i} = \text{softmax}_i \left(\frac{q_k^\top W_K f_i}{\sqrt{d}} + \log(p_{k,i} + \epsilon) \right), \tag{2}$$

where $W_K \in \mathbb{R}^{d \times D}$ projects CLIP features into the query space, $p_{k,i} \in [0, 1]$ is the spatial prior value for concept k at patch i, d is the attention dimension and $\epsilon > 0$ is a small constant for numerical stability. The logarithmic prior acts as a soft masking mechanism that suppresses patches outside the relevant facial regions while still allowing soft attention within the valid support.

Using the resulting attention weights, we aggregate the value-projected patch features to obtain a concept-specific visual evidence vector:

$$u_k = \sum_{i=1}^{N} \alpha_{k,i} W_V f_i, \tag{3}$$

where $W_V \in \mathbb{R}^{d_v \times D}$ is a learned value projection and $u_k \in \mathbb{R}^{d_v}$. The vector u_k summarizes the visual evidence relevant to concept k by pooling information from the CLIP patches according to the concept-specific attention mechanism. Thus, different concepts can extract different evidence from the same image even though they share the same underlying patch tokens.

The evidence vector is then mapped to a shared concept feature space using:

$$h_k = \mathrm{ReLU}(W_s u_k + b_s), \tag{4}$$

where $W_s \in \mathbb{R}^{d_h \times d_v}$ and $b_s \in \mathbb{R}^{d_h}$ are shared across all concepts and $h_k \in \mathbb{R}^{d_h}$. A concept-specific prediction head then maps h_k to a scalar concept probability:

$$\hat{c}_k = \sigma(w_k^\top h_k + b_k), \tag{5}$$

where $w_k \in \mathbb{R}^{d_h}$ and $b_k \in \mathbb{R}$ are concept-specific parameters and $\hat{c}_k \in [0,1]$ denotes the predicted probability that concept k is present. Overall, this formulation also yields three interpretable outputs: the concept–region mapping A specifies where each concept can look, the attention weights $\alpha_{k,i}$ localize the supporting patches and the bottleneck scores $\hat{c}_k$ indicate which manipulation concepts are present.

Finally, the concept bottleneck scores are used for real/fake prediction through a linear classifier:

$$z = w_{\mathrm{cls}}^\top \hat{\mathbf{c}} + b_{\mathrm{cls}}, \tag{6}$$

where $w_{\mathrm{cls}} \in \mathbb{R}^K$ and $b_{\mathrm{cls}} \in \mathbb{R}$ are the classifier parameters and $z \in \mathbb{R}$ is the binary classification logit. Since the classifier operates directly on the concept bottleneck, each classifier weight links a concept activation to the final detection decision.

3.3 Model Training

Given an input image x, DF-CBM outputs a concept bottleneck score $\hat{\mathbf{c}}(x) = [\hat{c}_1(x), \ldots, \hat{c}_K(x)]$, where $\hat{c}_k(x) \in [0,1]$ denotes the predicted probability of concept k and classification logits $z(x)$ for real/fake prediction. During training, each image is associated with a binary label $y \in \{0,1\}$ where $y = 1$ denotes a fake image and a concept target encoded in a binary vector $\mathbf{c} \in \{0,1\}^K$. We train the concept bottleneck using a weighted binary cross-entropy loss:

$$\mathcal{L}_{\mathrm{con}}(x) = \frac{1}{K} \sum_{k=1}^{K} \omega_k(x)\, \mathrm{BCE}(\hat{c}_k(x), c_k), \tag{7}$$

where c_k is the target label for concept k and $w_k(x)$ is defined as:

$$\omega_k(x) = \begin{cases} \rho_k, & c_k = 1, \\ 1, & c_k = 0,\ y = 0, \\ \beta, & c_k = 0,\ y = 1. \end{cases} \tag{8}$$

Here, $\rho_k = N_k^- / N_k^+$ is a concept-specific positive weight computed from the training set, where N_k^+ and N_k^- denote the number of positive and negative labels

for concept k, respectively, used to account for the strong imbalance in concept annotations. The parameter $\beta \in [0, 1]$ down-weights negative concept labels on fake images to reflect asymmetric label reliability. Positive labels and negatives from real images are treated as reliable, whereas negative concept labels in fake images may correspond to unannotated artifacts. We therefore treat these entries as weak negatives, reducing their influence while preserving useful supervision.

The real/fake classifier is trained with a standard cross-entropy loss $\mathcal{L}_{\text{cls}}(x) = \text{CE}(z(x), y)$ over the classification logits $z(x)$ and the full training objective is:

$$\mathcal{L} = \mathbb{E}_{(x,y,\mathbf{c})} \left[\mathcal{L}_{\text{cls}}(x) + \lambda \mathcal{L}_{\text{con}}(x) \right], \tag{9}$$

where $\lambda > 0$ controls the strength of concept supervision. The region-aware concept bottleneck and the final classifier are optimized jointly under this objective.

4 Experimental Results

4.1 Experimental Setup

We train DF-CBM on the concept-annotated subset of FaceForensics++ [41] constructed using the concept extraction procedure in Section 3.1. Following DeepfakeBench [63], we use the `c23` version for all experiments. All images are cropped and resized to 224×224. DF-CBM uses CLIP ViT-L/14 [39] as the image encoder and SegFace [37] with a Swin-Base backbone for facial region parsing. The attention and value dimensions, as well as the shared concept-evidence projection, are all set to 768. The model is trained jointly for 15 epochs using AdamW [34] with a learning rate of 0.0002, batch size 128, concept loss weight $\lambda = 1.0$ and noise discount factor $\beta = 0.15$ on 2 RTX A6000 GPUs.

Following prior work [7, 46, 60], we report frame- and video-level AUC for real/fake detection. For concept prediction, we report balanced accuracy (B-Acc.), F_1, frame-level AUC (F-AUC) and video-level AUC (V-AUC). We evaluate DF-CBM in a cross-method intra-dataset setting, training on FaceForensics++ and testing on eight unseen manipulation methods from DF40 [61] and in a cross-dataset setting using CDF-v2 [31], DFD [8], DFDC [9], DFDCP [10] and UADFV [30]. We compare against the concept-based baselines Joint-CBM [26] and BotCL [56], trained on the same concept-annotated subset as DF-CBM, as well as state-of-the-art black-box detectors using the reported results from [7,46].

4.2 Concept Prediction Results

We first evaluate whether the learned bottleneck provides meaningful concept predictions. We compare DF-CBM with Joint-CBM [26] and BotCL [56], on the FaceForensics++ test set. Evaluation is performed on all real test samples and on fake samples annotated with at least one concept from our vocabulary resulting in approximately 4.5K real samples and 6.5K fake samples.

Table 1 reports the macro-averaged results across all concepts in terms of balanced accuracy, F_1, F-AUC and V-AUC. DF-CBM outperforms both Joint-CBM and BotCL on F_1, F-AUC and V-AUC, achieving the largest gain in macro-F_1, with a 25.3% improvement over Joint-CBM, while maintaining comparable balanced accuracy. Overall, these results show that grounding concept prediction in anatomically relevant facial and boundary regions leads to a more reliable and interpretable concept bottleneck.

Table 1: *Macro-averaged* concept prediction performance on the FF++ test set.

Model	B-Acc.	F_1	F-AUC	V-AUC
Joint-CBM [26]	**0.687**	0.310	0.739	0.753
BotCL [56]	0.651	0.300	0.689	0.693
DF-CBM	0.675	**0.563**	**0.743**	**0.772**

Table 2: Intra-dataset deepfake detection performance on FaceForensics++ using video-level AUC. DF-CBM is compared with black-box detectors (top-part) and concept-based baselines (bottom part) across eight manipulation methods. Best results among the concept-based methods are shown in **bold**.

Methods	UniFace	BlendFace	MobSwap	e4s	FaceDan	FSGAN	InSwap	SimSwap	Avg.
SBI [44]	0.724	0.891	0.952	0.750	0.594	0.803	0.712	0.701	0.766
UCF [62]	0.831	0.827	0.950	0.731	0.862	0.937	0.809	0.647	0.824
IID [16]	0.839	0.789	0.888	0.766	0.844	0.927	0.789	0.644	0.811
LSDA [59]	0.872	0.875	0.930	0.694	0.721	0.939	0.855	0.793	0.835
ProDet [5]	0.908	0.929	0.975	0.771	0.747	0.928	0.837	0.844	0.867
CDFA [33]	0.762	0.756	0.823	0.631	0.803	0.942	0.772	0.757	0.781
Effort [60]	0.962	0.873	0.953	0.983	0.926	0.957	0.936	0.926	0.940
DFD-HR [46]	0.988	0.956	0.986	0.993	0.970	0.984	0.981	0.970	0.978
Joint-CBM [26]	0.932	**0.897**	0.912	0.865	0.687	0.920	0.894	0.827	0.867
BotCL [56]	**0.954**	0.881	**0.942**	0.759	0.755	**0.936**	**0.937**	0.878	0.880
DF-CBM	0.912	0.878	0.919	**0.968**	**0.845**	0.928	0.835	**0.891**	**0.897**

4.3 Deepfake Detection Results

Intra-dataset evaluation. We first evaluate video-level deepfake detection performance across different manipulation methods based on images from FaceForensics++. Table 2 compares DF-CBM with both state-of-the-art black-box detectors and concept-based baselines. For the black-box detectors, we report the results from [46] under the same evaluation protocol. For the concept-based baselines, Joint-CBM [26] and BotCL [56] are trained on the same concept-annotated subset of FaceForensics++ used to train DF-CBM.

Among concept-based methods, DF-CBM achieves the best average video-level AUC, improving over both Joint-CBM and BotCL. Compared with black-box detectors, DF-CBM remains competitive, achieving the third-best average performance overall. Although the strongest fully black-box detectors still obtain higher AUC, this is expected because DF-CBM constrains the prediction to pass through a compact concept bottleneck. In return, DF-CBM provides explicit

Table 3: Cross-dataset deepfake detection performance using video-level AUC. Models are trained on FaceForensics++ and evaluated on unseen datasets. Best results among the concept-based methods (bottom part) are shown in **bold**.

Methods	CDF-v2	DFD	DFDC	DFDCP	UADFV	Avg.
SBI [44]	0.886	0.827	0.717	0.848	-	-
UCF [62]	0.837	0.867	0.742	0.770	-	-
IID [16]	0.838	0.939	0.700	0.689	-	-
LSDA [59]	0.875	0.881	0.701	0.812	-	-
ProDet [5]	0.926	0.901	0.707	0.828	-	-
CDFA [33]	0.938	0.954	0.830	0.881	-	-
Effort [60]	0.956	0.965	0.843	0.909	-	-
ForAda [7]	0.957	0.972	0.872	0.929	-	-
DFD-HR [46]	0.960	0.980	0.865	0.901	-	-
Joint-CBM [26]	0.646	0.674	0.625	0.633	0.974	0.710
BotCL [56]	0.746	0.789	0.699	0.676	0.953	0.773
DF-CBM	**0.891**	**0.927**	**0.758**	**0.816**	**0.995**	**0.877**

concept-level evidence and spatial grounding which are not available in standard black-box detectors.

Cross-dataset evaluation. Tables 3 and 4 report video-level and frame-level AUC, respectively, when testing the examined methods on unseen datasets. We report black-box detector results from prior work [7, 46]. Notably, DF-CBM substantially outperforms the concept-based baselines across all target datasets. This improved performance is particularly important in the cross-dataset setting, where models must generalize beyond the manipulation methods and data distribution seen during training indicating that region-aware concept prediction provides a more transferable bottleneck than standard concept-based alternatives. Compared with SOTA black-box detectors, DF-CBM remains competitive, although the strongest unconstrained detectors still achieve higher AUC on several datasets due to the inclusion of the interpretable concept bottleneck layer in DF-CBM.

4.4 Qualitative Analysis

To analyze the visual evidence produced by DF-CBM, we visualize concept attention maps. The attention weights $\alpha_{k,i}$ indicate where the model attends when predicting concept k for image x. Since attention alone does not quantify the influence of each concept on the final prediction, we also compute a concept contribution score using the linear classifier $s_k(x) = w_{\text{cls},k}\hat{c}_k(x)$ where $w_{\text{cls},k}$ is the classifier weight associated with concept k. This score measures how strongly each activated concept supports the fake prediction.

Fig. 3 shows two manipulated images containing two annotated manipulation concepts each. In both examples, DF-CBM correctly predicts the annotated concepts which also receive the highest contribution scores. This indicates that the final decision is driven by the relevant manipulation concepts rather than

Table 4: Cross-dataset deepfake detection performance using frame-level AUC. Models are trained on FaceForensics++ and evaluated on unseen datasets. Best results among the concept-based methods (bottom part) are shown in **bold**.

Methods	CDF-v2	DFD	DFDC	DFDCP	UADFV	Avg.
SBI [44]	0.813	0.774	-	0.799	-	-
UCF [62]	0.753	0.807	0.719	0.759	-	-
ED [1]	0.864	-	0.721	0.851	-	-
CFM [35]	0.828	0.915	-	0.758	-	-
FoCus [50]	0.720	-	0.669	0.778	-	-
LSDA [59]	0.830	0.880	0.736	0.815	-	-
DiffusionFake [47]	0.805	0.904	-	0.810	-	-
MoE-FFD [27]	0.867	0.904	-	-	-	-
Dual-Adapter [43]	0.717	-	0.727	-	-	-
UDD [13]	0.869	0.910	0.758	0.856	-	-
Effort [60]	0.901	0.923	0.798	-	-	-
FFTG [48]	0.832	0.948	-	-	-	-
ForAda [7]	0.900	0.933	0.843	0.890	-	-
DFD-HR [46]	0.910	0.953	0.843	-	-	-
Joint-CBM [26]	0.604	0.644	0.609	0.614	0.943	0.683
BotCL [56]	0.683	0.738	0.676	0.649	0.919	0.733
DF-CBM	**0.830**	**0.874**	**0.732**	**0.774**	**0.983**	**0.839**

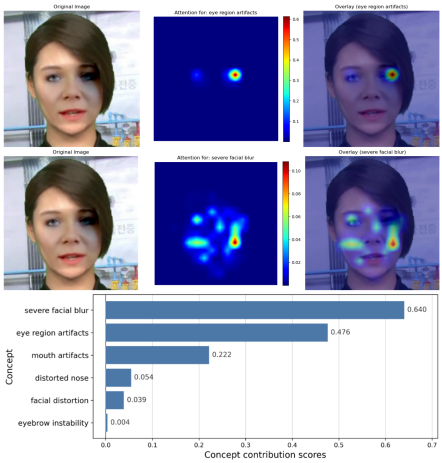
Table 5: Architectural component ablation on FaceForensics++ using video-level AUC.

Methods	UniFace	BlendFace	MobSwap	e4s	FaceDan	FSGAN	InSwap	SimSwap	Avg.
DF-CBM	**0.912**	**0.878**	0.919	**0.968**	0.845	**0.928**	0.835	**0.891**	**0.897**
w/o region prior	0.884	0.789	0.884	0.926	0.812	0.879	0.807	0.809	0.849
w/o concept-specific attention	0.830	0.810	0.866	0.936	0.811	0.896	0.768	0.843	0.845
w/o concept bottleneck	0.868	0.837	**0.927**	0.933	**0.877**	0.912	**0.844**	0.877	0.884

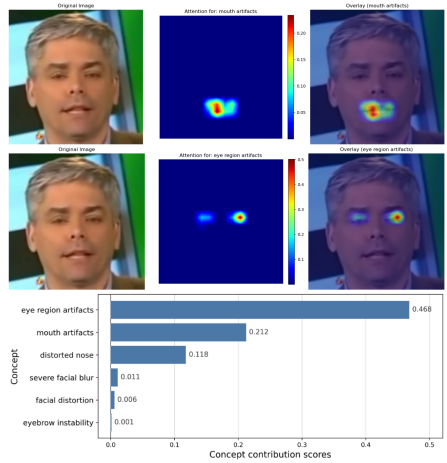
unrelated activations. In Fig. 3a, the model localizes the predicted concepts around the manipulated eye and cheek regions. Similarly, Fig. 3b shows fine-grained localization with higher attention on specific subregions such as the left side of the mouth and the right eye.

4.5 Ablation Study

Architectural component ablation. Table 5 ablates the main architectural components of DF-CBM on the intra-dataset evaluation set. Removing the region prior reduces the average AUC showing that the concept–region mapping provides useful spatial guidance for detecting manipulation concepts. Replacing concept-specific attention further lowers the average AUC to 0.845 indicating that allowing each concept to extract its own visual evidence is important for both concept prediction and final detection. Removing the concept bottleneck also degrades performance reducing the average AUC to 0.884 which suggests that explicit manipulation concepts provide useful semantic structure for detection.



(a) Eye artifacts and facial blur.



(b) Mouth and eye artifacts.

Fig. 3: Qualitative examples of DF-CBM explanations. For each manipulated image, we show concept-specific attention maps, their overlays on the input image and the corresponding concept contribution scores.

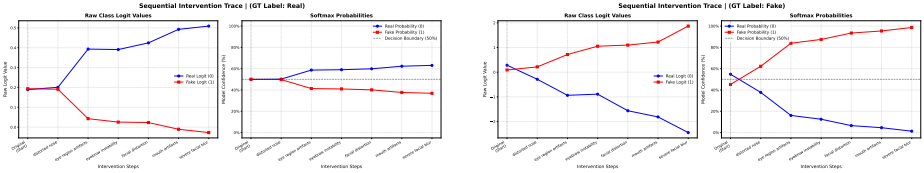
Table 6: Query design ablation for the masked cross-attention mechanism on FaceForensics++ using video-level AUC.

Methods	UniFace	BlendFace	MobSwap	e4s	FaceDan	FSGAN	InSwap	SimSwap	Avg.
DF-CBM	**0.912**	**0.878**	**0.919**	**0.968**	0.845	0.928	**0.835**	**0.891**	**0.897**
frozen text queries	0.911	0.867	**0.919**	0.967	**0.851**	**0.932**	0.829	0.887	0.895
random query initialization	0.827	0.811	0.875	0.955	0.805	0.907	0.784	0.825	0.849

Query design ablation. We further study the query design used in the masked cross-attention mechanism. Table 6 compares randomly initialized learnable queries, frozen queries initialized from CLIP text embeddings and the full DF-CBM with learnable CLIP-initialized queries. The largest gain comes from text-based initialization suggesting that concept semantics help the model associate manipulation concepts with relevant visual evidence while allowing the initialized queries to be further refined during training provides a small additional gain.

4.6 Interpretability via Intervention Analysis

Intervention mechanism. To evaluate the influence and interpretability of the concept bottleneck layer, we implement a sequential intervention mechanism. For each facial concept, we estimate the empirical distribution of its predicted logits over the training set. To simulate concept activation and suppression while remaining within the range observed during training, intervention values are anchored to extreme quantiles of the corresponding logit distribution. Specifically, concept absence and presence are represented by the 1^{st} and 99^{th} percentiles, respectively. We target misclassified samples and intervene on one concept at a



(a) Ground Truth: Real (Predicted: Fake) **(b)** Ground Truth: Fake (Predicted: Real)

Fig. 4: Sequential concept intervention traces for two misclassified samples: (a) ground-truth real, (b) ground-truth fake. At each step, a single concept is intervened on while retaining all previous interventions. The left panels show the evolution of the raw class logits, whereas the right panels show the corresponding softmax probabilities.

time, retaining each intervention before applying the next. For false real predictions concepts are sequentially set to their 99^{th} percentile intervention values, increasing evidence for deepfake artifacts and steering the prediction towards the *Fake* class. Conversely, for false fake predictions concepts are sequentially set to their 1^{st} percentile intervention values, suppressing artifact evidence and shifting the prediction towards the *Real* class.

Intervention traces. Fig. 4 illustrates the evolution of the raw class logits and corresponding softmax probabilities during sequential concept interventions for two randomly selected samples. As shown in Fig. 4a, a ground-truth real sample initially misclassified as fake exhibits little separation between the class logits after the first intervention. Subsequent concept interventions progressively decouple the logits, widening the decision margin and stabilizing the correct prediction. Conversely, Fig. 4b shows a ground-truth fake sample initially misclassified as real, where the first intervention immediately shifts the prediction across the decision boundary. Subsequent interventions further increase the target-class logit while suppressing the incorrect-class logit, resulting in steadily increasing prediction confidence.

5 Conclusion

In this work, we introduce DF-CBM, a region-aware concept bottleneck model for explainable deepfake detection. To address the gap between visually grounded evidence and language-based explanations, DF-CBM constructs a compact vocabulary of manipulation-related concepts, associates each concept with anatomically meaningful facial and boundary regions and learns a concept-specific masked attention mechanism that grounds concept predictions in localized visual evidence. Through intra-dataset and cross-dataset evaluations, we show that DF-CBM outperforms concept-based baselines in both concept prediction and deepfake classification while remaining competitive with state-of-the-art black-box detectors. Additionally, our qualitative and intervention analyses further provide interpretable evidence linking detected manipulation concepts to the final decision.

Acknowledgement

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

References

1. Ba, Z., Liu, Q., Liu, Z., Wu, S., Lin, F., Lu, L., Ren, K.: Exposing the deception: Uncovering more forgery clues for deepfake detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 719–728 (2024)
2. Badr, N.E.A., Liang, X., Nebel, J.C., Greenhil, D.: Frld-df: Frequency ring-guided lora adaptation of dinov2 vision transformer for generalizable deepfake detection. In: 2026 IEEE 20th International Conference on Automatic Face and Gesture Recognition (FG). pp. 1–11. IEEE (2026)
3. Benou, I., Raviv, T.R.: Show and tell: Visually explainable deep neural nets via spatially-aware concept bottleneck models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 30063–30072 (2025)
4. Bouter, M.d.L.d., Pardo, J.L., Geradts, Z., Worring, M.: Protoexplorer: Interpretable forensic analysis of deepfake videos using prototype exploration and refinement. Information Visualization **23**(3), 239–257 (2024)
5. Cheng, J., Yan, Z., Zhang, Y., Luo, Y., Wang, Z., Li, C.: Can we leave deepfake data behind in training deepfake detector? Advances in Neural Information Processing Systems **37**, 21979–21998 (2024)
6. Cui, J., Li, H., Zhan, Y., Shang, H., Cheng, K., Ma, Y., Mu, S., Zhou, H., Wang, J., Zhu, S.: Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21086–21095 (2025)
7. Cui, X., Li, Y., Luo, A., Zhou, J., Dong, J.: Forensics adapter: Adapting clip for generalizable face forgery detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19207–19217 (2025)
8. DFD: Contributing Data to Deepfake Detection Research. `https://ai.googleblog.com/2019/09/contributing-data-to-deepfakedetection.html` (2020), accessed: 2026-06-30
9. DFDC: Deepfake Detection Challenge. `https://www.kaggle.com/c/deepfake-detection-challenge` (2020), accessed: 2026-06-30
10. Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., Ferrer, C.C.: The deepfake detection challenge (dfdc) dataset. arXiv preprint arXiv:2006.07397 (2020)
11. Espinosa Zarlenga, M., Barbiero, P., Ciravegna, G., Marra, G., Giannini, F., Diligenti, M., Shams, Z., Precioso, F., Melacci, S., Weller, A., et al.: Concept embedding models: Beyond the accuracy-explainability trade-off. Advances in neural information processing systems **35**, 21400–21413 (2022)
12. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
13. Fu, X., Yan, Z., Yao, T., Chen, S., Li, X.: Exploring unbiased deepfake detection via token-level shuffling and mixing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 3040–3048 (2025)

14. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In: The Twelfth International Conference on Learning Representations (2024)
15. Hondru, V., Hogea, E., Onchis, D., Ionescu, R.T.: Exddv: A new dataset for explainable deepfake detection in video. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 4273–4284 (2026)
16. Huang, B., Wang, Z., Yang, J., Ai, J., Zou, Q., Wang, Q., Ye, D.: Implicit identity driven deepfake face swapping detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4490–4499 (2023)
17. Huang, Z., Hu, J., Li, X., He, Y., Zhao, X., Peng, B., Wu, B., Huang, X., Cheng, G.: Sida: Social media image deepfake detection, localization and explanation with large multimodal model. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28831–28841 (2025)
18. Jeon, S., Hwang, I., Lee, S., Zhang, B.T.: Locality-aware concept bottleneck model. In: UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models (2024)
19. Jiang, L., Li, R., Wu, W., Qian, C., Loy, C.C.: 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. pp. 2889–2898 (2020)
20. Jiang-Lin, J.Y., Huang, K.Y., Zou, L., Lo, L., Yang, S.P., Tseng, Y.W., Lin, K.H., Chen, C.L., Ta, Y.T., Wang, Y.T., et al.: Tridf: Evaluating perception, detection, and hallucination for interpretable deepfake detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17087–17098 (2026)
21. Jung, I., Choi, H., Le, B.M., Na, H., Woo, S.S.: A rich knowledge space for scalable deepfake detection. In: The Fourteenth International Conference on Learning Representations (2026)
22. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: International Conference on Learning Representations (2018)
23. Kashiani, H., Talemi, N.A., Afghah, F.: Freqdebias: Towards generalizable deepfake detection via consistency-driven frequency debiasing. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8775–8785 (2025)
24. Kirchengast, T.: Deepfakes and image manipulation: criminalisation and control. Information & Communications Technology Law **29**(3), 308–323 (2020)
25. Knab, P., Steinmann, D., Bartelt, C., Kersting, K., Schiele, B., Seidl, T., Schlegel, U., Stammer, W.: What's in the bottle? a survey and roadmap of concept bottleneck models. Transactions on Machine Learning Research (2026)
26. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: International conference on machine learning. pp. 5338–5348. PMLR (2020)
27. Kong, C., Luo, A., Bao, P., Yu, Y., Li, H., Zheng, Z., Wang, S., Kot, A.C.: Moe-ffd: Mixture of experts for generalized and parameter-efficient face forgery detection. IEEE Transactions on Dependable and Secure Computing (2025)
28. Koutlis, C., Papadopoulos, S.: Leveraging representations from intermediate encoder-blocks for synthetic image detection. In: European Conference on computer vision. pp. 394–411. Springer (2024)
29. Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., Guo, B.: Face x-ray for more general face forgery detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5001–5010 (2020)

30. Li, Y., Chang, M.C., Lyu, S.: In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In: 2018 IEEE International workshop on information forensics and security (WIFS). pp. 1–7. Ieee (2018)
31. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-df: A large-scale challenging dataset for deepfake forensics. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3207–3216 (2020)
32. Li, Z., Pun, C.M., Fang, C., Wang, J., Cun, X.: Personalive! expressive portrait image animation for live streaming. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18118–18128 (2026)
33. Lin, Y., Song, W., Li, B., Li, Y., Ni, J., Chen, H., Li, Q.: Fake it till you make it: Curricular dynamic forgery augmentations towards general deepfake detection. In: European conference on computer vision. pp. 104–122. Springer (2024)
34. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
35. Luo, A., Kong, C., Huang, J., Hu, Y., Kang, X., Kot, A.C.: Beyond the prior forgery knowledge: Mining critical clues for general face forgery detection. IEEE Transactions on Information Forensics and Security **19**, 1168–1182 (2023)
36. Mirsky, Y., Lee, W.: The creation and detection of deepfakes: A survey. ACM computing surveys (CSUR) **54**(1), 1–41 (2021)
37. Narayan, K., Vs, V., Patel, V.M.: Segface: Face segmentation of long-tail classes. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 6182–6190 (2025)
38. Oikarinen, T., Das, S., Nguyen, L.M., Weng, T.W.: Label-free concept bottleneck models. In: The Eleventh International Conference on Learning Representations (2023)
39. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
40. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
41. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1–11 (2019)
42. Sammani, F., Fischer, J., Deligiannis, N.: Clip-free, label free, unsupervised concept bottleneck models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3262–3272 (2026)
43. Shao, R., Wu, T., Nie, L., Liu, Z.: Deepfake-adapter: Dual-level adapter for deepfake detection. International Journal of Computer Vision **133**(6), 3613–3628 (2025)
44. Shiohara, K., Yamasaki, T.: Detecting deepfakes with self-blended images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18720–18729 (2022)
45. Soltandoost, E., Plesh, R., Schuckers, S., Peer, P., Štruc, V.: Extracting local information from global representations for interpretable deepfake detection. In: Proceedings of the Winter Conference on Applications of Computer Vision. pp. 1629–1639 (2025)
46. Sun, J., Yan, Z., Zhang, K.Y., Yao, T., Ding, S.: Dfd-hr: Generalizable deepfake detection via hierarchical routing learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13984–13995 (2026)

47. Sun, K., Chen, S., Yao, T., Liu, H., Sun, X., Ding, S., Ji, R.: Diffusionfake: Enhancing generalization in deepfake detection via guided stable diffusion. Advances in Neural Information Processing Systems **37**, 101474–101497 (2024)
48. Sun, K., Chen, S., Yao, T., Zhou, Z., Ji, J., Sun, X., Lin, C.W., Ji, R.: Towards general visual-linguistic face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19576–19586 (2025)
49. Tan, H., jun lan, Tan, Z., Shi, S., Liu, A., Song, C., Zhu, H., Wang, W., Wan, J., Lei, Z.: Veritas: Generalizable deepfake detection via pattern-aware reasoning. In: The Fourteenth International Conference on Learning Representations (2026)
50. Tian, J., Chen, P., Yu, C., Fu, X., Wang, X., Dai, J., Han, J.: Learning to discover forgery cues for face forgery detection. IEEE Transactions on Information Forensics and Security **19**, 3814–3828 (2024)
51. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. Advances in neural information processing systems **37**, 84839–84865 (2024)
52. Trinh, L., Tsang, M., Rambhatla, S., Liu, Y.: Interpretable and trustworthy deepfake detection via dynamic prototypes. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1973–1983 (2021)
53. Vaccari, C., Chadwick, A.: Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. Social media+ society **6**(1), 2056305120903408 (2020)
54. Vanian, V., Doumanoglou, A., Zarpalas, D.: Why fake? unveiling the semantic vocabulary of deepfake detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4101–4110 (2026)
55. Vera, H.S., Dua, S., Zhang, B., Salz, D., Mullins, R., Panyam, S.R., Smoot, S., Naim, I., Zou, J., Chen, F., et al.: Embeddinggemma: Powerful and lightweight text representations. arXiv preprint arXiv:2509.20354 (2025)
56. Wang, B., Li, L., Nakashima, Y., Nagahara, H.: Learning bottleneck concepts in image classification. In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition. pp. 10962–10971 (2023)
57. Wen, S., Feng, P., Kang, H., Wen, Z., Chen, Y., Wu, J., He, C., Li, W., et al.: Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. Advances in Neural Information Processing Systems **38**, 58972–59005 (2026)
58. Xu, Z., Zhang, X., Li, R., Tang, Z., Huang, Q., Zhang, J.: Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. In: International Conference on Learning Representations. vol. 2025, pp. 31186–31216 (2025)
59. Yan, Z., Luo, Y., Lyu, S., Liu, Q., Wu, B.: Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8984–8994 (2024)
60. Yan, Z., Wang, J., Jin, P., Zhang, K.Y., Liu, C., Chen, S., Yao, T., Ding, S., Wu, B., Yuan, L.: Orthogonal subspace decomposition for generalizable ai-generated image detection. In: International Conference on Machine Learning. pp. 70268–70288. PMLR (2025)
61. Yan, Z., Yao, T., Chen, S., Zhao, Y., Fu, X., Zhu, J., Luo, D., Wang, C., Ding, S., Wu, Y., et al.: Df40: Toward next-generation deepfake detection. Advances in Neural Information Processing Systems **37**, 29387–29434 (2024)

62. Yan, Z., Zhang, Y., Fan, Y., Wu, B.: Ucf: Uncovering common features for generalizable deepfake detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22412–22423 (2023)
63. Yan, Z., Zhang, Y., Yuan, X., Lyu, S., Wu, B.: Deepfakebench: A comprehensive benchmark of deepfake detection. In: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2023)
64. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
65. Yang, X., Li, Y., Lyu, S.: Exposing deep fakes using inconsistent head poses. In: ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 8261–8265. IEEE (2019)
66. Yang, Y., Qian, Z., Zhu, Y., Russakovsky, O., Wu, Y.: Dˆ 3: scaling up deepfake detection by learning from discrepancy. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23850–23859 (2025)
67. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. In: International Conference on Learning Representations. vol. 2025, pp. 83048–83077 (2025)
68. Yuksekgonul, M., Wang, M., Zou, J.: Post-hoc concept bottleneck models. In: The Eleventh International Conference on Learning Representations (2023)
69. Zhang, Y., Colman, B., Guo, X., Shahriyari, A., Bharaj, G.: Common sense reasoning for deepfake detection. In: European conference on computer vision. pp. 399–415. Springer (2024)
70. Zhou, M., Qin, S., Li, Y., Chen, Q., Jiang, P.: Autocut: End-to-end advertisement video editing based on multimodal discretization and controllable generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 37777–37787 (2026)
71. Zhou, T., Wang, W., Liang, Z., Shen, J.: Face forensics in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5778–5788 (2021)