

EXDF: EXPLAINABLE DEEFAKE DETECTION WITH VISION-LANGUAGE MODEL

Anonymous for submission

A. OVERVIEW OF THE EXDF DATASET

Figure 1 illustrates a diverse set of examples from the ExDF dataset, showcasing the variety and precision of the generated fake images. Each example in the figure features a specific type of facial attribute manipulation, accompanied by the corresponding ground truth (GT) mask and a detailed textual explanation of the modifications. The dataset includes a wide range of alterations, such as changes to facial features like the mouth, eyes, and hair, as well as the addition of accessories such as glasses and beards. These examples demonstrate both subtle and overt manipulations, reflecting the dataset’s intent to cover a broad spectrum of deepfake techniques. By encompassing such a variety of manipulations, the ExDF dataset ensures that it provides a rigorous testbed for evaluating the robustness and interpretability of modern deepfake detection models.

B. FACIAL ATTRIBUTE MODIFICATIONS

The facial attribute modifications are thoroughly detailed in Table 1. These modifications cover a broad spectrum of changes, including adjustments to the eyes, nose, mouth, eyebrows, skin texture, ears, and hair, as well as facial expressions and the addition of accessories such as glasses, beards, and earrings. The modifications also involve more subtle alterations, such as changing skin tone, applying makeup, reshaping facial structures, and even simulating the effects of aging. These comprehensive edits are designed to create either highly realistic or exaggerated facial features, offering a robust testbed for evaluating deepfake detection and model interpretability.

C. QUALITATIVE ANALYSIS

We provide additional qualitative analysis in Figure 3, comparing our results with GPT-4o [1] and InstructBLIP [2]. This analysis includes both facial attribute modifications and entirely AI-generated face images. The results clearly demonstrate that our model not only outperforms others in accurately detecting image manipulations but also excels in providing interpretable explanations for the altered regions. By focusing the model’s attention on specific manipulated features, we offer a clearer and more detailed understanding of how these changes were made. This enhanced explainability is critical for improving trust in deepfake detection systems and helping users better understand the underlying manipulations.

D. FAILED CASES IN TEXTUAL EXPLANATION

We present several cases of failure in the design of textual explanations, shown in Figure 2. Using only editing instructions often results in incomplete explanations, as the LLM may overlook other modified features. Providing just the fake image and instructions can lead to hallucinations and inaccuracies, with explanations not matching the actual manipulations. Even combining real and fake images with instructions can cause LLM hallucinations. We addressed these

limitations by integrating ground truth masks, textual attributes, and precise prompts. This approach reduced hallucinations and ensured that GPT-4o generated accurate, detailed, and contextually relevant explanations, improving the overall effectiveness of our deepfake detection and explanation system.

E. ROBUSTNESS TO UNSEEN PERTURBATIONS

We compare the performance of five models on the ExDF dataset and evaluate their robustness against unseen perturbations. To evaluate the model’s robustness, we analyze the performance of detectors under two common types of image perturbations: JPEG compression (with quality levels of 100, 65, and 30) and Gaussian blur (with σ values of 1, 2, and 3). We average the performance across all samples in the ExDF dataset to observe the overall behavior of each method. The results demonstrate that our model exhibits better stability under JPEG compression and Gaussian blur than other methods. The proposed method consistently achieves higher accuracy across varying levels of perturbation, indicating that our model is more resilient to deteriorating image quality and unseen distortions.

F. LIMITATIONS AND SOCIETAL IMPACT

In this paper, we have achieved promising results in both deepfake detection and explanation, demonstrating strong generalization capabilities. However, the current work has several limitations. First, the editing model used for dataset generation lacks sufficient robustness, which may negatively impact the quality of the manipulated data. Additionally, our approach is limited to image-based deepfakes and does not yet extend to video deepfakes. Finally, while the model effectively explains facial attribute manipulations, it provides limited support for explaining non-facial alterations. Additionally, we highlight the current deepfake threats and the irresponsible predictions made by existing methods, which harm the community and hinder effective defense against deepfakes, leading to distrust and rejection of digital content. Our approach demonstrates explainable deepfake detection, offering a ray of hope to address this flaw. However, the threat of deepfakes to society remains significant. Through this research, we aim to encourage future researchers to focus on trustable predictions, helping to build a more reliable and positive digital environment.

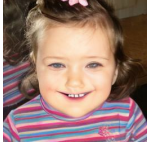



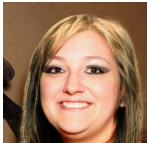

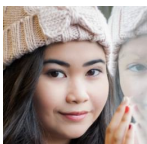
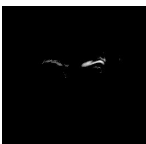
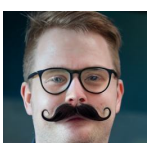

Fake Image	GT Mask	Textual Explanation
		<p>Attributes: "mouth", "eyes"</p> <p>Textual Explanation: "The image is fake. The facial attributes such as the mouth, and eyes have been altered, affecting the overall expression."</p>
		<p>Attributes: "eyes", "glasses"</p> <p>Textual Explanation: "The image is fake. The facial attributes such as eyes and glasses have been altered. Specifically, sunglasses have been added to the face."</p>
		<p>Attributes: "hair"</p> <p>Textual Explanation: "The image is fake. One specific change made to the face is the hair color, which significantly impacts the overall appearance and adds an unnatural element to the image."</p>
		<p>Attributes: "eyebrows"</p> <p>Textual Explanation: "The image is fake. The alterations made to the facial attributes have primarily focused on the eyebrows."</p>
		<p>Attributes: "mouth", "beard"</p> <p>Textual Explanation: "The image is fake. The facial attributes that have been altered include the mouth, and beard. The primary visible change is the addition or modification of the beard."</p>

Fig. 1: Uncurated samples from the ExDF dataset.

G. REFERENCES

- [1] Achiam et al., "Gpt-4 technical report," *arXiv*, 2023.
- [2] Dai et al., "Instructblip: Towards general-purpose vision-language models with instruction tuning," *NeurIPS*, 2024.







Input	Editing Instruction	Real Image	Manipulated Image	Textual Explanation	
Editing Instruction	Darken the skin tone			The image is fake. The eyes have been modified slightly	✗
Editing Instruction +Fake Image	Widen his/her mouth to look shocked			The image is fake. The mouth is too wide and the teeth are too big.	✗
Editing Instruction +Real Image +Fake Image	Mouth slightly open			The image looks fake. The person's mouth is wide, and the teeth are not aligned properly. The eyes are also slightly crossed.	✗

Fig. 2: Examples of failed cases in generating deepfake explanations.

Attribute	Editing Instruction
Mouth	"Widen his/her mouth to look shocked", "Change to smile", "Big lips" "Widen his/her mouth to smile", "Wearing lipstick" "Change to crying face", "Make his/her mouth terrifying", "Change to angry face" "Change only the lips to red", "Change to sad", "Mouth slightly open"
Eyes	"Change the eye color to green", "Change to angry face", "Close the eyes" "Dark circles under eyes", "Add long eyelashes", "Bags under eyes" "Add blue eye shadow", "Big eyes", "Narrow eyes", "Squint the eyes slightly"
Nose	"Make the nose bigger", "Decrease the overall size of the nose", "Pointy nose"
Eyebrows	"Make eyebrows bushy", "Make his/her frown" "Arched eyebrows", "Bushy eyebrows"
Accessory	"Add a beard to the face", "Add glasses", "Add a scar on the face" "Wearing sunglasses", "Double chin", "Add a mustache", "Bald", "Goatee"
Skin	"Add wrinkles to the face", "Add acne to the face", "Darken the skin tone" "Make skin paler", "Make him/her look older", "Face freckles"
Hair	"Change hair color to blonde", "Change straight hair to curly", "Gray hair" "Wavy hair", "Straight hair", "Receding hairline", "Brown hair"
Head	"Wearing a hat"
Ears	"Add earrings"

Table 1: Editing instructions for facial attributes on DMs and GANs.



Fig. 3: Qualitative analysis of deepfake explanation compared to GPT-4o and InstructBLIP. The GT Mask indicates the actual manipulated parts of the image. The bold text highlights the fake features detected by each model.

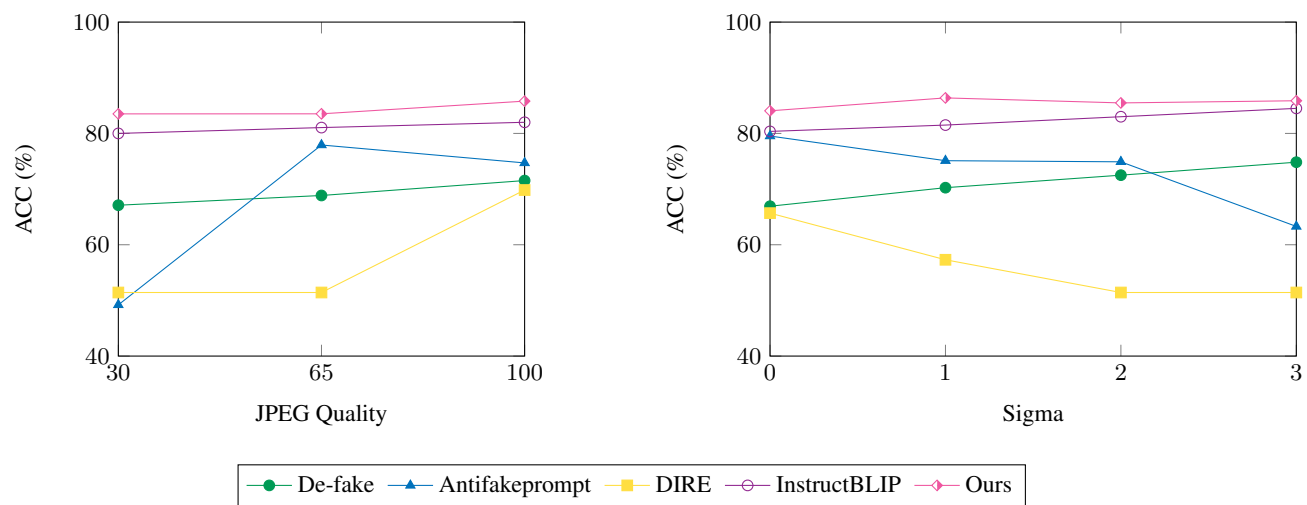


Fig. 4: Robustness against unseen perturbations. The column on the left displays robustness to JPEG compression, while the column on the right shows robustness to Gaussian blur. We report the average performance of GANs and DMs.