

Common Sense Reasoning for Deepfake Detection

–Supplementary Material–

Yue Zhang^{*1,2}, Ben Colman², Xiao Guo¹, Ali Shahriyari², and Gaurav Bharaj²

¹ Michigan State University

² Reality Defender Inc

{zhan1624, guoxia11}@msu.edu

1 DD-VQA Dataset Annotations

Annotation Tools. Annotations for DD-VQA are collected entirely by crowd workers from Amazon Mechanical Turk (AMT) ³. The dataset is collected over the course of 3 months and 3 iterations of updating annotation schemes. Approximately 9000 Human Intelligence Tasks (HITs) are launched on AMT, where each HIT involves 3-6 questions, answers, and the corresponding images. Each HIT was designed such that workers manage to earn anywhere between \$6-\$8 per hour, which follows ethical research standards on AMT [37].

Fakeness Annotations. From Tab. 2-Tab. 7, we present examples of fine-grained fake facial features and the corresponding descriptions in our dataset. We provide the annotators with fine-grained feature options and use templates to comprise the description with our templates. Some fakenesses require the annotators to provide the corresponding area, for example, “*left or right eyebrows*”. Also, for the question of which area looks unnatural brighter/darker, the answers need to include the corresponding facial areas, such as “*left/right cheeks*”, “*beside the left/right eyes*”, “*around nose*”, etc.

Challenging Annotation Cases. In Fig 2, we provide examples where at least two annotators mistakenly perceive manipulated images as real. Such cases are excluded when annotators provide inaccurate labels, as effective deception of humans requires the human face in the image to adhere to common-sense knowledge.

Uncertainty of Fakeness. There are cases where annotators express uncertainty regarding the image’s authenticity. To capture this ambiguity, we offer annotators a fakeness rating scale ranging from 0 to 5, where 0 and 1 indicate authenticity, 2 and 3 means a slight degree of fakeness, and 4 and 5 represent a high degree of fakeness. The corresponding descriptions are “real”, “a bit fake”, and “very fake”. Annotating the uncertainty of fakeness helps the model simulate human perception of fakeness, thereby enhancing its ability to generate explanations that align more accurately with human judgment.

General Questions assess the overall authenticity of an image. The format of the general question is “*Does the person in the image look fake?*”. The answers to

^{*} This work was completed during an internship at Reality Defender Inc.

³ <https://www.mturk.com/>

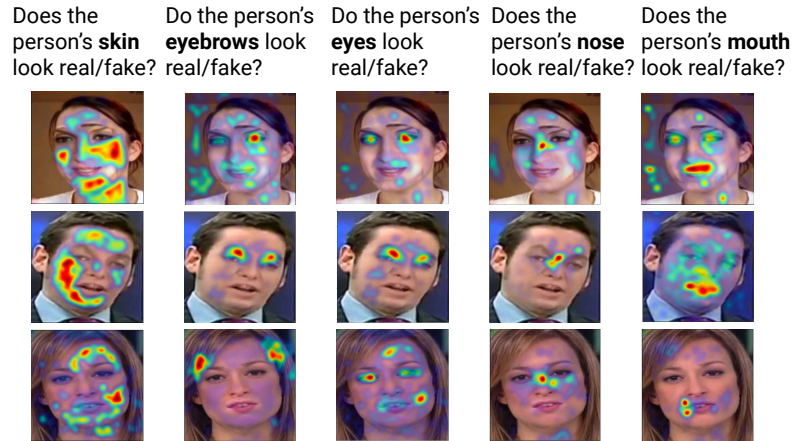


Fig. 1: Additional attention heatmap visualization of BLIP-TI.

this question cover the general reasons for authenticity or fakeness. Specifically, the general fakeness factors include “*obvious manipulated region*”, “*incomplete face feature*”, “*unrealistic texture or lighting*”, etc. Conversely, the general reasons for authenticity involve “*complete face features*”, “*face features in good shape, size, and positioning.*”, “*natural expression*”, etc.



Fig. 2: Challenging cases where annotators provide incorrect labels.

Fine-Grained Facial Feature Questions assess the authenticity of individual facial features. There are instances where specific facial components may still exhibit authenticity despite the overall image appearing fake. The detailed facial features include *eyebrows*, *skin*, *eyes*, *nose*, and *mouth*. The format of the fine-grained feature question is “*Do the person’s X look real/fake?*”, and *X* is any facial component. We show the corresponding examples in Fig. 2.

- **Eyebrows.** Humans commonly have a pair of eyebrows with a symmetrical shape, smooth hair, and a dark color. The presence of overlapping, broken and blurred eyebrows can indicate manipulated images.
- **Skin.** There is no universally “perfect” type of skin; however, generally, common skin should exhibit clarity, an even skin tone, and a smooth texture, especially at lower resolutions. Also, the presence of boundaries, discolored patches, or drastically inconsistent skin color on the face are not characteristic of a real person’s face.

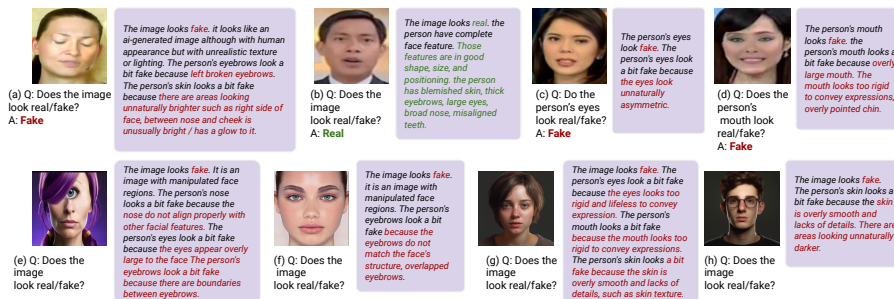


Fig. 3: Additional Qualitative Examples (a)-(d) are images from FF++. (e) is a cartoon image; (f) is a Photoshop image showing overlapped eyebrows. (g) and (h) are images from Midjourney.

- **Eyes.** Common eyes include the characteristics of symmetry, clarity, expressiveness, an appropriate size, etc. The blurred and asymmetric eyes in the manipulated image can indicate fakeness.
- **Nose.** An ideal nose should be appropriately positioned, with clear and proportionate nostrils in terms of shape and size. However, the unnaturally curved nose or nose without fine lines are obvious fake signs.
- **Mouth.** The mouth in our annotation scheme refers to mouth areas, including lips, teeth and chin. The appearance of inappropriate size and color of these areas could be used to indicate fakeness.

2 DD-VQA Enhanced Deepfake Detection

Our proposed DD-VQA generates multi-modal representations that can serve as a model-agnostic enhancement for general binary deepfake detectors. We illustrate our approach using RECCE [7] as an example. RECCE proposes a forgery detection framework that leverages the common compact representations of genuine faces based on reconstruction classification learning. Specifically, the images are fed into an encoder-decoder reconstruction network for representation learning. The encoder’s output, denoted as \mathbf{F}_1 , undergoes a multi-scale graph reasoning module to enhance better representation, denoted as \mathbf{F}_2 , which is subsequently combined with \mathbf{F}_1 . In summary, the vision representation of deepfake detection is $\mathbf{F}' = \mathbf{F}_1 + \mathbf{F}_2$. Based on this, we incorporate our DD-VQA enhanced multi-modal representation \mathbf{F} obtained from our VL model trained using the DD-VQA dataset. We first utilize a few CNN layers to transform \mathbf{F} into the same shape as \mathbf{F}' . We can obtain the final enhanced representation \mathbf{F}^{en} with $\mathbf{F}^{en} = \mathbf{F}' + \theta(\mathbf{F})$, where $\theta(*)$ represents the necessary tensor shape transformations for fusing \mathbf{F} and \mathbf{F}' .

3 Experiment Setup

Metrics We mainly use image-caption-based metrics to evaluate the quality of the generated text, as follows.

Method	DD-VQA Deepfake Detection	DD-VQA Answer Generation
	Acc \uparrow	F1 \uparrow
No Distortion	0.8749	0.9007
Resize(0.75X)	0.8621	0.8987
JPEGCompression(quality=75)	0.8593	0.8827
GaussianNoise($\sigma = 3$)	0.8434	0.8676
Color Enhancement(factor=3.0)	0.8385	0.8639

Table 1: Robustness Evaluation.

- **BLEU-4** [31] is used to evaluate the precision of the match between the generated text and reference text based on 4-grams.
- **CIDEr** [49] measures the consensus between the generated text and the referenced text, considering both word and grammar similarity and the alignment in terms of meaning and content.
- **Rouge_L** [27] evaluates the **Longest Common Subsequence** (LCS) of words between the generated text and the referenced text. Using LCS does not require consecutive matches but in-sequence matches reflecting sentence-level word order.
- **METEOR** [10] considers precision, recall, stemming, synonymy, and word order. It employs a unigram-based matching approach but extends it with additional semantic features.
- **SPICE** [3] evaluates how well a generated text can capture the specific entities present in the image, emphasizing precision, recall, and diversity.

ViT-based deepfake detection models. Efficient ViT combines a ViT with a convolutional EfficientNet B0 as the feature extractor. Convolutional Cross ViT builds upon both the Efficient ViT and the multi-scale Transformer, and enable the utilization of larger patches to achieve a broader receptive field. Although both Efficient ViT and Convolutional Cross ViT use video deepfake datasets (FF++ [35] and DFDC [12]), they extract frames from videos and use images for model training.

4 Qualitative Study

Visualization. We present additional visualization examples in Fig.1 generated by our best model BLIP-TI. The model is trained with both language modeling loss and our designed contrastive losses. These examples demonstrate that the highlighted attention areas predominantly align with the facial components mentioned in the question. We employ GradCam [38] visualization technique to show the alignments between textual tokens and the highlighted area in the image.

Robustness Evaluation. We conduct a robustness evaluation of our model, considering aspects such as resizing, compression, Gaussian noise, and color enhancement. We evaluate both detection and explanation generation performances. As shown in Tab. 1, our model appears to be robust to different variations, especially regarding the quality of generated textual explanations.

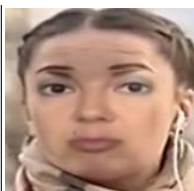

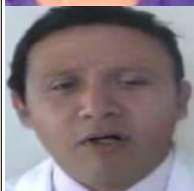
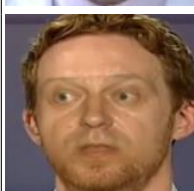
Fine-grained Features	Images	Descriptions
Overlapped eyebrows		The person's eyebrows look fake. The person's eyebrows look very fake because the person has left overlapped eyebrows and right-overlapped eyebrows.
Broken eyebrows		The person's eyebrows look fake. The person's eyebrows look very fake because the person has broken left eyebrow.
Blurry eyebrows		The person's eyebrows look fake. The person's eyebrows look very fake because the eyebrows look blurry and unclear.
Boundary between eyebrows		The person's eyebrows look fake. The person's eyebrows look fake because there is a boundary between the person's eyebrows.

Table 2: Fake Eyebrows Features.

Qualitative Examples We provide additional qualitative examples in Fig. 3. We extend our testing beyond the FF++ dataset. We evaluate our model on diverse images, including cartoon images, Photoshop images, and images generated using a diffusion model. These examples show our model can capture common-sense knowledge of human facial features well. For instance, the cartoon image of Fig. 3 (e), our model can capture the pattern of “*over large eyes*”. Also, we manipulate a real image to put another pair of eyebrows on top of the original eyebrows, as shown in Fig. 3 (f), and our model still can capture the fakeness of “*overlapped eyebrows*”. For images from Midjourney (Fig. 3 (g) and (h)), our model can capture the fakeness of “*rigid eyes and mouth*”.

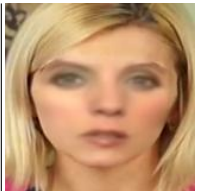

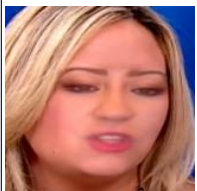
Fine-grained Features	Images	Descriptions
Blurry eyes		<p>The person's eyes look fake. The person's eyes look fake because the eyes look blurry and unclear.</p>
Unnatural asymmetric eyes		<p>The person's eyes look fake. The person's eyes look fake because the person has unnatural asymmetric eyes.</p>
Rigid Eyes		<p>The person's eyes look fake. The person's eyes look fake because the person's eyes are too rigid to convey expressions.</p>

Table 3: Fake Eyes Features.




Fine-grained Features	Images	Descriptions
Boundaries		The person's skin looks fake. The person's skin looks very fake because there are boundaries on the person's face, such as boundaries on the person's left and right cheeks.
Inconsistent skin color		The person's skin looks fake. The person's skin looks very fake because the person has inconsistent skin color.
Discolored patches		The person's skin looks fake. The person's skin looks very fake because there is a discolored path on the person's forehead.

Table 4: Fake Skin Features.


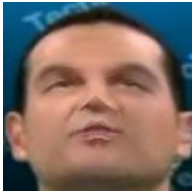
Fine-grained Features	Images	Descriptions
Unnaturally curved nose		The person's nose looks fake. The person's nose looks unnaturally curved.
nose lacks of details		The person's nose looks fake. The person's nose looks very fake because the nose lacks of pores and fine lines.

Table 5: Fake Nose Features.




Fine-grained Features	Images	Descriptions
Blurry Mouth		The person's mouth area looks fake. The person's mouth looks blurry and unclear.
Mouth with unnatural color		The person's mouth area looks fake. The person's mouth shows an unnatural white color.
unnatural coloring/blurry teeth		The person's mouth area looks fake. The person's teeth look misaligned with the rest of the mouth. The person's teeth look unnatural coloring.

Table 6: Fake Mouth Features.




Fine-grained Features	Images	Descriptions
Incomplete facial features		The image looks fake because the person has incomplete facial features.
Unclear eyeglass frame		The image looks fake because the person's eyeglass frame looks unclear.
Mustache		The image looks fake because the person's mustache does not align with other facial features.

Table 7: General Fake Features.