

A Hitchhiker's Guide to Fine-Grained Face Forgery Detection Using Common Sense Reasoning

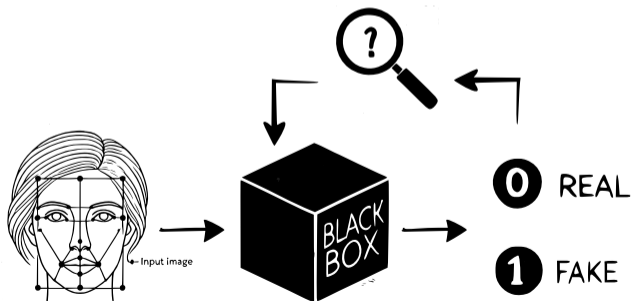
N. Foteinopoulou¹ E. Ghorbel^{1,2} D. Aouada¹

¹CVI², SnT, University of Luxembourg

²Cristal Laboratory, National School of Computer Sciences, University of Manouba

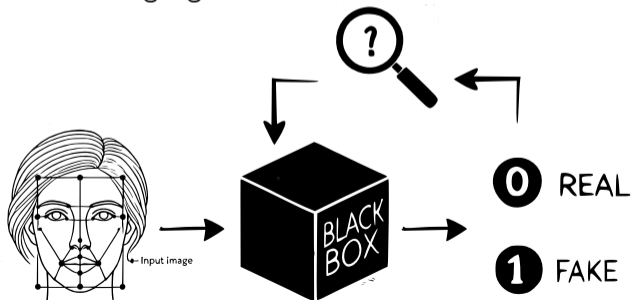
Motivation for Common Sense Reasoning in Deepfake Detection

- ▶ Current detection methods rely upon binary classifiers, with explainability and fine-grained areas as post-hoc analysis



Motivation for Common Sense Reasoning in Deepfake Detection

- ▶ Current detection methods rely upon binary classifiers, with explainability and fine-grained areas as post-hoc analysis
- ▶ VLLMs hold promise for both detecting deepfakes and explaining decisions through natural language



Motivation for Common Sense Reasoning in Deepfake Detection

Research Question 1

To what extent can existing VLLMs detect deepfake images and what rationale supports the decision?

Research Question 2

How do we fairly and comprehensively evaluate VLLMs in the fine-grained task?

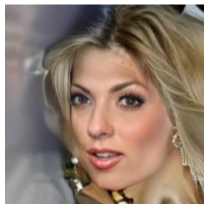
Motivation for Common Sense Reasoning in Deepfake Detection







Main Contributions:

- ▶ Transform deepfake detection into a VQA multi-label problem, enhancing explainability.
- ▶ Systematic assessment for SoTA VLLMs, promoting transparency and reproducibility.
- ▶ Comprehensive analysis of model capabilities and limitations to inform future research.

Proposed Multi-Stage Benchmark

We transform the discriminative deepfake detection task into a VQA problem.



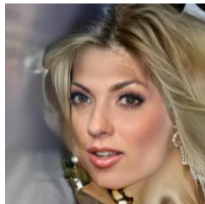
- a)  Is this image manipulated? a) Yes b) No
-  a) Yes
- b)  Of the areas in the list Nose, Eyebrows, Eyes, Lip, Hair, which ones are manipulated?
-  In the image, the nose, eye, eyebrow, lip, and hair of the blonde woman are manipulated. The other areas, such as ear, chin, and neck, are not altered.
- c)  What areas are manipulated?
-  In the image, the area that is manipulated is the woman's face. The photo has been digitally altered to enhance her features, such as her eyes, lips, and cheekbones, making her look more attractive







Stage 1: Binary task.

$\mathbf{X}_t =$ "Is this image $[s_i]$? a) Yes b) No"

where $\mathbf{s}_i \in \mathcal{S}$ is a set of standard terms used to describe deepfakes in English.

Proposed Multi-Stage Benchmark



- a)  "Is this image manipulated? a) Yes b) No"
-  a) Yes
- b)  "Of the areas in the list Nose, Eyebrows, Eyes, Lip, Hair, which ones are manipulated?"
-  "In the image, the **nose, eye, eyebrow, lip, and hair** of the blonde woman are manipulated. The other areas, such as ear, chin, and neck, are not altered."
- c)  "What areas are manipulated?"
-  "In the image, the area that is manipulated is the woman's face. The photo has been digitally altered to enhance her features, such as her **eyes, lips, and cheekbones**, making her look more attractive"

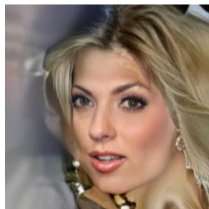
Stage 2: Multiple-choice setting for the fine-grained task.







$\mathbf{X}_t =$ "Of the areas in the list $[\mathbf{cls}_0, \dots, \mathbf{cls}_{|C|}]$

which ones are $[\mathbf{s}_i]$?"

where $\mathbf{cls}_i \in \mathcal{C}$ is the class name of the i -th class from the set of target classes \mathcal{C} .

Proposed Multi-Stage Benchmark



- a)  Is this image manipulated? a) Yes b) No
-  a) Yes
- b)  Of the areas in the list Nose, Eyebrows, Eyes, Lip, Hair, which ones are manipulated?
-  In the image, the nose, eye, eyebrow, lip, and hair of the blonde woman are manipulated. The other areas, such as ear, chin, and neck, are not altered.
- c)  What areas are manipulated?
-  In the image, the area that is manipulated is the woman's face. The photo has been digitally altered to enhance her features, such as her eyes, lips, and cheekbones, making her look more attractive

Stage 3: Open-ended VQA for the fine-grained task.

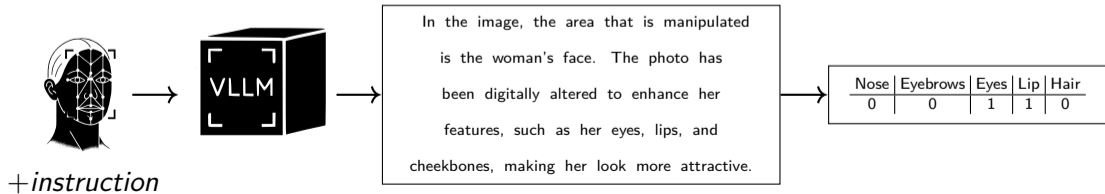
$X_t =$ "What area of this image is $[s_i]$?"

Deepfake Detection to VQA

VLLMs generate **natural language** that needs to be **transformed** for **classification** evaluation.

Matching strategy depends on the task.

- ▶ *Exact Match (EM)*: The generated sentence is exactly equal to the class name.
- ▶ *Contains*: The class name is contained in the response.
- ▶ *CLIP distance*: Sigmoid over the cosine similarity of the prediction embeddings and class name embeddings.



Fine-Grained Face Forgery Detection Using Common Sense Reasoning

VLLMs Tested

Selected Models:

- ▶ LLaVa-1.5
- ▶ BLIP2
- ▶ InstructBLIP with Flan-T5 and Flan-T5-xxl
- ▶ Ensemble of models

Baseline: CLIP

Upper Bound: GPT-4V^a

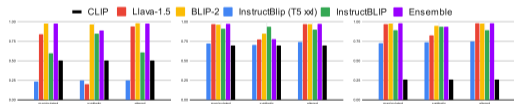
^aon a subset of each dataset

Datasets

- ▶ FF++
- ▶ DFDC
- ▶ Celeb-DF
- ▶ WildDeepFake
- ▶ StyleGAN StyleGAN2, StyleGAN3
- ▶ SeqDeepFake attributes, components
- ▶ R-splicer

Binary Detection

- Exact Match (EM) for the binary task.
- Several synonyms to test model robustness to instruction: *manipulated, deepfake, synthetic, altered, fabricated, face forgery, falsified*

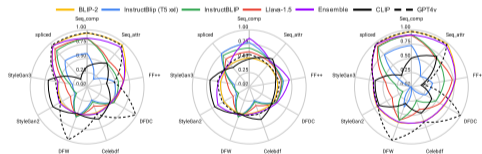


(a) SeqDF attributes

(b) SeqDF components

(c) R-splicer dataset

Figure: EM Performance of VLLMs in terms of F1-score for the top 3 synonyms



(a) Accuracy

(b) AUC

(c) F1-score

Figure: EM Performance of VLLMs on nine benchmarks

Fine-Grained VQA

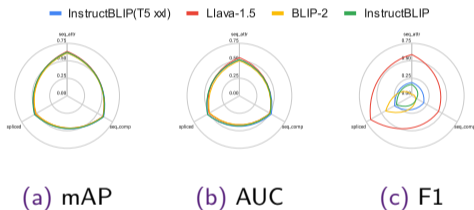


Figure: Assessment of model performance in multiple-choice settings with *contains* matching.

Multiple-choice:

- ▶ Models often mention all label names, increasing False Positives.
- ▶ Responses like "All of them" or "None of them" further complicate matching, impacting F1 scores.

Fine-Grained VQA

Table: Model performance on open-ended VQA using a) *contains* and b) *CLIP* matching

	BLIP-2			InstructBLIP			InstructBLIP-xxl			LlaVa-1.5		
	mAP	AUC	F1	mAP	AUC	F1	mAP	AUC	F1	mAP	AUC	F1
SeqDF attr.	61.8	51.0	20.4	61.3	50.4	18.3	63.1	53.6	37.5	61.7	51.1	40.0
SeqDF comp.	59.5	50.5	14.7	59.2	50.0	4.1	60.2	51.8	26.2	59.0	49.6	17.1
R-Splicer	55.8	55.6	31.3	52.3	53.2	23.5	53.8	54.0	31.1	58.7	57.5	41.6

(a) *contains* matching

	BLIP-2			InstructBLIP			InstructBLIP-xxl			LlaVa-1.5		
	mAP	AUC	F1	mAP	AUC	F1	mAP	AUC	F1	mAP	AUC	F1
SeqDF attr.	63.0	53.6	73.5	59.9	50.9	74.0	60.4	50.7	55.5	61.0	51.3	74.1
SeqDF comp.	58.8	52.7	71.0	55.5	49.0	71.7	59.9	55.7	59.8	56.1	49.6	71.7
R-Splicer	54.3	55.3	66.2	48.5	49.3	66.5	54.0	53.1	60.3	56.7	57.4	66.5

(b) *CLIP* distance

Open-ended VQA:

- ▶ *CLIP distance* matching improves recall and F1-score but slightly lowers mAP.
- ▶ May offer more reliable results for class-specific detection, reflected in F1-scores.

Qualitative Evaluation

Qualitative evaluation

The aim of the study is to understand how well the generated responses reflect the ground truth i.e. how well each model can spot the manipulated areas.

The possible manipulation areas are: entire face (facevox), mouth, nose, eyes, eyebrows. Each model is instructed to "Explain briefly what areas of this image are manipulated?".

See the image, the correct answer and the predicted answers.

Select the score that best reflects how closely the predicted answers capture the same information as the correct answer.

This means that the statements should be semantically similar but do not need to match the class names exactly.

1. completely wrong
2. mostly wrong
3. half right
4. mostly right
5. completely right

The entire evaluation should take 30-60 mins.

The areas that are manipulated are: eyebrows mouth facevox



The eyes of the man in the hat

completely wrong 1 2 3 4 5 completely correct

(a) Annotator Briefing

(b) Annotation Form

Figure: Briefing(a) and Annotation Form(b) shown to human evaluators.

Table: Open-ended qualitative evaluation with human annotators in Tab. a and BertScore in Tab. b-d

Model	Human Eval. Score
BLIP-2	0.35
InstructBLIP	<u>0.36</u>
InstructBLIP-xxl	0.33
LlaVa-1.5	0.38

(a) Human Evaluation

Model	Precision	Recall	F1
BLIP-2	79.87	79.72	79.61
InstructBLIP	81.12	<u>83.89</u>	86.90
InstructBLIP-xxl	<u>82.57</u>	81.77	81.01
LlaVa-1.5	87.40	86.37	<u>85.39</u>

(c) SeqDF comp.

Model	Precision	Recall	F1
BLIP-2	79.77	78.75	79.24
InstructBLIP	86.53	<u>83.22</u>	<u>84.81</u>
InstructBLIP-xxl	80.73	81.78	81.25
LlaVa-1.5	<u>84.86</u>	85.31	85.08

(b) SeqDF attr.

Model	Precision	Recall	F1
BLIP-2	79.55	79.76	80.04
InstructBLIP	<u>83.47</u>	<u>85.34</u>	87.39
InstructBLIP-xxl	82.53	81.87	81.23
LlaVa-1.5	85.94	86.33	<u>86.74</u>

(d) R-splicer

Key Takeaways

- ▶ **Performance of tested models:** Smaller models perform better on the binary task, but larger models show better reasoning.

Key Takeaways

- ▶ **Performance of tested models:** Smaller models perform better on the binary task, but larger models show better reasoning.
- ▶ **Zero-Shot Evaluation:** Models leverage pre-trained semantic mapping, though they lag behind task-specific models.

Key Takeaways

- ▶ **Performance of tested models:** Smaller models perform better on the binary task, but larger models show better reasoning.
- ▶ **Zero-Shot Evaluation:** Models leverage pre-trained semantic mapping, though they lag behind task-specific models.
- ▶ **Limitations:** Current datasets lack fine-grained labels and detailed descriptions.

Key Takeaways

- ▶ **Performance of tested models:** Smaller models perform better on the binary task, but larger models show better reasoning.
- ▶ **Zero-Shot Evaluation:** Models leverage pre-trained semantic mapping, though they lag behind task-specific models.
- ▶ **Limitations:** Current datasets lack fine-grained labels and detailed descriptions.
- ▶ **Future Directions:** Specialised datasets and task-specific models.

Thank you!

Scan for project page!



Scan for CVI² page!

