

# Harnessing Chain-of-Thought Reasoning in Multimodal Large Language Models for Face Anti-Spoofing

## Supplementary Material

### 1 FaceCoT dataset: additional details

#### 1.1 GPT-4o annotation details

Figure 6 illustrates our Chain-of-Thought (CoT) annotation details using GPT-4o [28]. To guide GPT-4o to generate accurate, detailed, and correctly formatted responses in a human-like reasoning style, we structure the input into five components—image, system prompt, question, hint, and label—which are concatenated and submitted to the GPT-4o API for annotation.

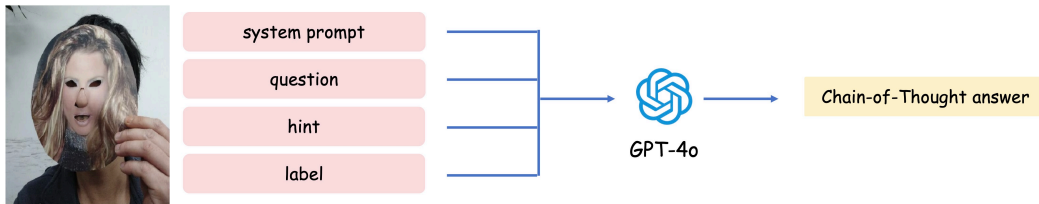


Figure 6: Details of the GPT-4o-based CoT annotation pipeline, with inputs concatenated from five components: image, system prompt, question, hint, and label.

To ensure the consistency and accuracy of our CoT annotations, we apply the following prompt design strategies. First, we explicitly define the model’s role as an “examiner” in the system prompt, thereby guiding it to engage in rigorous reasoning from an evaluator’s perspective and enhancing both the discriminative power and standardization of its outputs. Second, to enforce structural consistency, we impose an output framework composed of six core modules—caption, facial description, facial attributes, reasoning, spoofing description, and conclusion. Finally, to guarantee accuracy, we require the model’s final conclusion to match the standard answer exactly, using this alignment as a key criterion for valid automated annotations. Embedding these requirements at the prompt level effectively codifies the model’s reasoning logic and output format, significantly improving result uniformity and batch-processing scalability, while ensuring transparency at every step of the annotation workflow.

Furthermore, to prevent the model from detecting spoofing-related features without being able to establish clear decision boundaries (example of such ambiguous case is illustrated in Fig. 7), we provide tailored hints for samples of different spoofing types to guide the model toward correct reasoning and conclusions, as presented in Tab. 5. For spoofing types that the model can readily distinguish, no hints are provided.

Table 5: Hints provided for each category; other categories are straightforward and do not require additional hints

Types	Hints
Photo	Photographing a printed photo constitutes spoofing.
Newspaper	Photographing a newspaper page constitutes spoofing.
Poster	Photographing a poster constitutes spoofing.
Album	Photographing an album page constitutes spoofing.
A4	Photographing an A4-printed image constitutes spoofing.
Facial print	Photographing a facial photo constitutes spoofing.
Upper body	Photographing a printed upper-body image constitutes spoofing.
Phone	Photographing a phone screen constitutes spoofing.
Pad	Photographing a tablet display constitutes spoofing.
PC	Photographing a computer screen constitutes spoofing.
Else	-

## System Prompt

In a face anti-spoofing detection system, a person has submitted an image for inspection. Now please assume the role of the examiner and respond to the question. I need you to strictly follow the format containing six specific parts: caption, facial description, facial attributes, reasoning, spoofing description, and conclusion. It is crucial that you adhere exactly to the outlined structure, and that the final answer in conclusion matches the standard correct answer exactly. Further explanation: In caption, provide a holistic description of the image. In facial description, focus on describing the facial features of the person in the image. In facial attributes, list the facial attributes. In reasoning, perform step-by-step logical reasoning based on the image description, facial description, and facial attributes. In spoofing description, describe the spoofing method and highlight the critical spoofing characteristics in detail. In conclusion, give the final answer "Yes" or "No" that must match the standard answer exactly. Format as follows: <Caption> [Holistic description of the image] </Caption> <Facial Description> [Detailed description of the person's facial features] </Facial Description> <Facial Attributes> [Describe facial attributes: "eyes": ..., "mouth": ...etc.] </Facial Attributes> <Reasoning> [Step-by-step logical reasoning based on the above] </Reasoning> <Spoofing Description> [Details of the spoofing method: ...] </Spoofing Description> <Conclusion> [Answer "Yes" or "No" matching the standard answer exactly] </Conclusion> Please carefully apply this format to analyze the given image and answer the related question.



```
<Caption> The picture shows the content of a Japanese newspaper, with a prominent central position displaying a man in a suit and leather shoes raising a glass, and an advertising logo in the background. The article is accompanied by relevant textual explanations, mainly focusing on describing the background information related to the character's activities and alcohol. </Caption>
<Facial description> The man in the picture has a clear face, thick eyebrows, a certain hairstyle decoration, eyes staring ahead, and a natural expression on his face. The facial light is sufficient, presenting a normal and natural face. </Facial description>
<Facial Attributes> { "Eyes": "Black pupils, staring ahead", "Mouth": "Naturally closed, no obvious expression", "Nose": "Proportional coordination", "Skin": "Uniform and natural, no abnormal color difference", "Hairstyle": "Neat and deep hair color", "Facial movements": "Neutral expression, no special posture." } </Facial Attributes>
<Reasoning> The newspaper image is a clear and normal face with no traces of technical processing. Combining facial description and attribute observation, the facial expression conforms to natural light and shadow conditions, and no obvious features such as pixel abnormalities, capture errors, or post modification were detected. The use of newspapers has not introduced factors related to face swapping or facial tampering. The state of the characters in the picture is mainly presented naturally without any artificial elements, and the activity scene is consistent and reasonable with the content in the picture. </Reasoning>
<Cheat Description> No clear external intervention features </Cheat Description>
<Conclusion> No </Conclusion>
```


```
<CAPTION> This is an image from a newspaper featuring a male figure dressed in a suit. He is holding a glass containing a clear liquid, with a simple indoor setting as the background. The newspaper also includes textual content that likely describes the person or an event related to them. </CAPTION>
<FACIAL DESCRIPTION> The facial features of the person in the image are clearly visible, with well-proportioned eyes looking straight ahead, a slightly closed mouth, and a relatively serious expression. The hair is neat and slightly short-cut. </FACIAL DESCRIPTION>
<FACIAL ATTRIBUTES> { "eyes": "staring straight ahead", "mouth": "slightly closed", "nose": "clear", "skin": "smooth", "hair": "short and neat", "expression": "serious" } </FACIAL ATTRIBUTES>
<REASONING> Based on the description of the image, this appears to be a printed picture from a newspaper rather than a directly photographed portrait. The facial description indicates that the person's features and expression are normal, but considering the overall analysis of the image, the source of the person's image is from newspaper printing, not an original digital photograph. Therefore, this falls under an indirect form of image acquisition. </REASONING>
<SPOOFING DESCRIPTION> Using printed newspaper images as the submitted content instead of directly photographing the person's face. The image source is clearly from a newspaper, with a print quality that indicates a printed effect rather than a genuine photograph. This behavior falls under the category of submitting non-real face images and is considered cheating. </SPOOFING DESCRIPTION>
<CONCLUSION> Yes </CONCLUSION>
```

Figure 7: Example of a decision-boundary failure: the model correctly detects newspaper features but does not classify the image as spoofing (top). After adding a hint (bottom), the model refines its reasoning and accurately identifies the image as a spoof, demonstrating improved CoT guidance with better decision boundaries.

## 1.2 Hard case handling

In the main text, we describe that, following GPT-4o annotation, we apply a regular expression to extract the result within the "<Conclusion>...</Conclusion>" tag and cross-check it against the original label; if the match fails, this annotation is regarded as failed. Samples that could not be correctly labeled after two annotation rounds are designated as "hard cases," resulting in a total of 581 instances. These hard cases are then corrected by human experts. Figure 8

illustrates one such example: experts first verify whether the conclusion is correct, then diagnose why the reasoning and spoofing description are inaccurate; if the reasoning is flawed, they replace it with a correct step-by-step rationale; if the visual feature description is inconsistent, they refine it and iteratively update the subsequent reasoning. Through this expert review and correction process, we ultimately obtain the complete, high-quality FaceCoT-Gold100K dataset. Similarly, samples in the data expansion that fail to be correctly annotated by our caption model are also reviewed and corrected by human experts, leading to the construction of the complete FaceCoT-Silver982K dataset.



<CAPTION> A close-up portrait of a man wearing a traditional-style clothing, facing the camera against a wooden background with soft indoor lighting. </CAPTION>  
 <FACIAL DESCRIPTION> The subject has a *natural skin texture* with visible pores and slight stubble, straight black hair partly covered by a headband, well-defined straight eyebrows, almond-shaped dark eyes, a straight nasal bridge, and closed lips in a neutral expression. </FACIAL DESCRIPTION>  
 <FACIAL ATTRIBUTES> { "eyes": "almond-shaped with dark brown irises and natural highlights", "eyebrows": "straight, medium-thick, black", "nose": "straight bridge with a moderately narrow tip", "mouth": "closed lips with a neutral expression and natural contour", "skin": "*even tone with visible pores and slight stubble*" } </FACIAL ATTRIBUTES>  
 <REASONING> 1. The skin exhibits *natural microtexture and pores* rather than uniform plastic or printed artifacts. 2. Lighting is consistent across the face and background, with no specular reflections indicative of transparent overlays or masks. 3. Facial features (eyes, nose, mouth) align symmetrically and dynamically with natural depth cues. 4. *There are no visible edges, seams, or distortions* that would suggest a printed or screen-based spoof. </REASONING>  
 <SPOOFING DESCRIPTION> *No spoofing artifacts are present* </SPOOFING DESCRIPTION>  
 <CONCLUSION> No </CONCLUSION>

<CAPTION> The image shows a person standing indoors, with some decorations and items on the walls in the background. The person's face appears to have some unnatural features. </CAPTION>  
 <FACIAL DESCRIPTION> The facial features of the person in the image appear as if they are wearing a mask, seeming unnatural. There are visible boundaries at the edges of the face, indicating possible signs of an artificial face, especially around the seam between the forehead and the rest of the face. </FACIAL DESCRIPTION>  
 <FACIAL ATTRIBUTES> { "eyes": "shape is normal but lacks naturalness", "mouth": "position is normal but slightly stiff", "nose": "proportions are normal but the contour appears somewhat flat", "skin": "texture is hard, resembling plastic or an artificial surface", "forehead": "connection with the face looks like there's a boundary line", "overall expression": "stiff and unnatural" } </FACIAL ATTRIBUTES>  
 <REASONING> By observing the image as a whole, describing the face, and analyzing its detailed facial attributes, it was found that the person's face displays unnatural edge boundaries, particularly in the forehead area. This indicates the use of some prop, possibly a mask or other artificial materials, to cover the real face. Additionally, the texture of the skin and the features lack naturalness, making this face very likely not naturally formed. These characteristics clearly indicate that this image contains face spoofing. </REASONING>  
 <SPOOFING DESCRIPTION> Cheating method: Using a face mask to cover true features. Cheating characteristics: 1. Clear boundaries at the edges of the face, indicating where the mask is covering. 2. Unnatural skin texture, appearing like plastic or similar materials. 3. Stiff facial expressions lacking normal dynamic movements. </SPOOFING DESCRIPTION>  
 <CONCLUSION> Yes </CONCLUSION>

Figure 8: Illustration of hard case handling. The top shows the initial failed annotation, while the bottom presents the revised version by human experts. The subject wears a mask with a clearly visible boundary at the forehead, which is incorporated into the revised annotation.

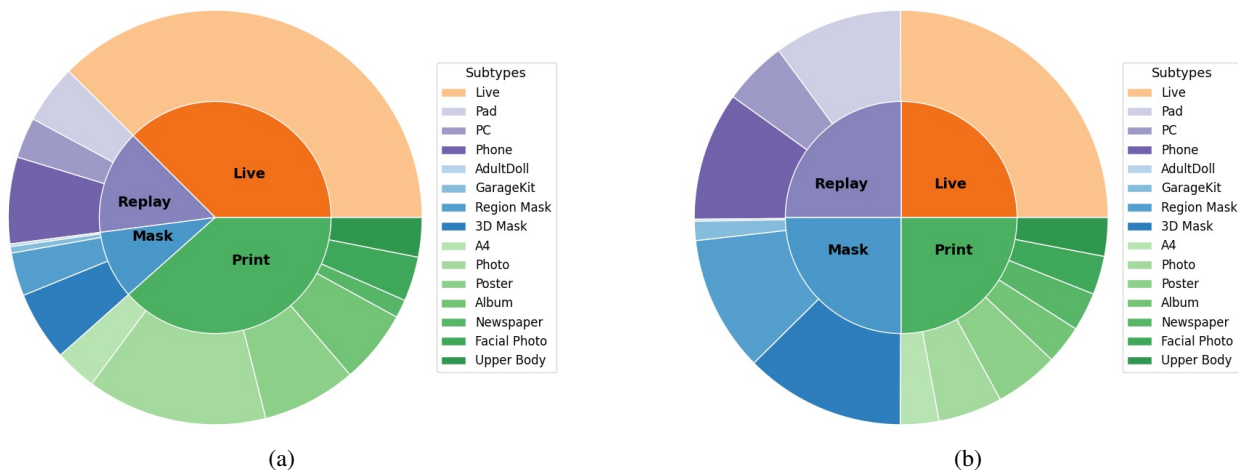


Figure 9: (a) The data types in FaceCoT-Silver982K. (b) The data types in FaceCoT-Gold100K. Both of them comprise 3 major spoofing types and 14 subtypes.

### 1.3 Statistics

Our FaceCoT dataset comprises two subsets—FaceCoT-Gold100K and FaceCoT-Silver982K—and encompasses living faces alongside 14 distinct spoofing attack types. Here, we present the relative proportions of types across the two datasets in Fig. 9 and report the exact sample counts for every category in both subsets in Tab. 6. Examples of each attack category are illustrated in Figure 10. For the annotation of FaceCoT-Gold100K, we used the GPT-4o API, incurring an average cost of approximately \$0.01 per image. To enable large-scale annotation at lower cost, we further trained a caption model via SFT and RL, which required 8 NVIDIA A100 GPUs for about one day, and subsequently employed this model to annotate FaceCoT-Silver982K at a total cost of roughly 288 GPU-hours. Finally, in the human refinement stage, six annotators manually cleaned and verified samples from both subsets over a period of three days.

Table 6: Sample counts per category in the FaceCoT-Gold100K and FaceCoT-Silver982K subsets

Types	FaceCoT-Gold100K	FaceCoT-Silver982K
Photo	5,000	138,373
Newspaper	3,000	14,425
Poster	5,000	72,079
Album	3,000	56,490
A4	3,000	31,776
Facial print	3,000	33,647
Upper body	3,000	30,167
Phone	10,000	66,434
Pad	10,000	48,516
PC	5,000	31,072
3D mask	12,768	52,637
Region mask	10,579	33,285
Garagekit	1,488	4,505
Adultdull	165	1,454
Living	25,000	367,608
Total	100,000	982,468



Figure 10: Representative examples of all 14 spoofing attack categories and living faces in the FaceCoT dataset.

## 1.4 Comparison with contemporary works

Recently, several datasets for FAS have been introduced, including PA-FAS [27] and FaceShield [36], which were developed contemporaneously with our study. Table 7 presents a quantitative comparison. Compared with these datasets, FaceCoT provides substantially larger-scale coverage, more diverse scenarios, finer-grained annotations, and a structured CoT format that supports multimodal reasoning for FAS.

Table 7: Comparison with recent FAS datasets.

Attribute	PA-FAS [27]	FaceShield [36]	FaceCoT
Subjects	800	925	<b>205,881</b>
Samples	40,000	54,959	<b>1,082,468</b>
Attack Types	11	11	<b>14</b>

## 2 Methodology details

### 2.1 Reinforcement Learning in training caption model

**Reward functions** We design a dual-reward scheme targeting both semantic accuracy and output format compliance:

- **Semantic accuracy reward:** Inspired by the "`<Conclusion>...</Conclusion>`" structure in FaceCoT, we apply a regular expression to extract the conclusion from the model’s generated output and compare it to the ground-truth label. A match yields a reward of 1; otherwise, 0.
- **Format compliance reward:** We verify whether the model’s output follows the prescribed FaceCoT template. If the structural format is correct, the reward is 1; otherwise, 0.

This dual-reward scheme simultaneously enforces correct annotation content and strict adherence to the FaceCoT formatting guidelines.

**Training strategy** We initialize the policy model with a version pre-trained via SFT. Given an input image and its associated task prompt, the policy model generates a CoT response. Each response is scored according to the dual-reward functions above, and the resulting reward signal is used to update the policy via RL. To stabilize training, we employ the SFT model as a fixed reference: we compute the KL divergence between the policy’s output distribution and that of the reference model, using it as a penalty term to prevent the policy from drifting too far from its initial semantic space. This balance preserves output reliability while enabling effective exploration.

**Training data** To enhance the caption model’s annotation capability and task adaptability on unseen data, we directly use the unlabeled images from the target annotation corpus as input during the RL stage. This construction endows the caption model with strong task-specific adaptation.

**Accuracy evaluation** We first randomly sample 2,000 instances from the dataset that have not been annotated to construct a test set for evaluation. Then, we use two models to perform automatic annotation on this test set: one trained solely with SFT, and the other further optimized with RL based on the SFT model. From the generated outputs, we extract the result within the "`<Conclusion>...</Conclusion>`" tags and compare them with the original labels of the samples. If the two labels match exactly, the annotation is considered correct; otherwise, it is considered incorrect. The final annotation accuracy is calculated using the following formula:

$$\text{Accuracy} = \frac{\text{Count}(\text{conclusion} = \text{label})}{\text{Count}(\text{conclusion} = \text{label}) + \text{Count}(\text{conclusion} \neq \text{label})} \quad (1)$$

### 2.2 Reinforcement Learning in CoT-Enhanced Progressive Learning (CEPL)

**Motivation** After the two-stage training with CEPL, the model has demonstrated remarkable anti-spoofing performance on the FAS task. Building upon the success of RL in the caption model, we further investigate its applicability in this component to boost FAS performance while preserving the model’s existing capabilities.

**Details** Specifically, after completing the two-stage training with CEPL, we introduce a third stage of RL. In this stage, we augment the original multi-task loss, which consists of CoT reasoning and classification supervision, with

an additional RL objective driven by our dual-reward functions for semantic accuracy and format compliance. The RL procedure follows the same policy-optimization paradigm described previously, with one key difference: no new data is incorporated. Instead, we directly reuse the image–text pairs employed during the second stage. This design tests whether strategic optimization of output structure and semantics, without any additional training examples, can still yield significant performance gains.

### 2.3 Details for evaluation metrics

Since standard FAS metrics such as AUC and HTER require continuous confidence scores rather than binary predictions, we adapt the output of VLMs to provide probabilistic scores. Specifically:

1. **Deterministic decoding.** To ensure output consistency and avoid randomness from beam search, we set the generation beam number to 1.
2. **Token-level logits extraction.** Instead of directly treating textual outputs (e.g., *yes* vs. *no*) as hard labels, we extract token-level logits from the first generated token. In particular, we identify the token IDs corresponding to ‘Yes’ and ‘No.’
3. **Probability computation.** We compute the softmax probability over the two logits, obtaining the confidence that a sample is *real*:

$$p_{\text{real}} = \frac{\exp(\ell_{\text{No}})}{\exp(\ell_{\text{No}}) + \exp(\ell_{\text{Yes}})}, \quad (2)$$

where  $\ell_{\text{Yes}}$  and  $\ell_{\text{No}}$  denote the logits of the “Yes” and “No” tokens, respectively.

The resulting probability  $p_{\text{real}}$  is then used to calculate AUC and HTER following standard definitions in the FAS literature. This procedure allows us to fairly evaluate LLM-based classifiers under conventional spoofing metrics.

## 3 Experiments

### 3.1 Cross-domain generalization under widely adopted protocol

In the FAS literature, a widely adopted evaluation protocol is the leave-one-out cross-domain testing on four benchmarks: OULU-NPU (O) [3], CASIA-MFSD (C) [48], Idiap Replay-Attack (I) [6], and MSU-MFSD (M) [40]. However, the performance under this protocol has already saturated (with AUC exceeding 99%), making it less discriminative for assessing fine-grained improvements. Therefore, in the main text we focus on a more challenging and generalization-oriented one-to-eleven protocol, which better highlights the advantages of our method. Nevertheless, to further demonstrate the robustness of our approach, we also report results under the widely used leave-one-out protocol. Specifically, we first apply our FAS caption model to generate CoT annotations for the training splits of the O, C, M, and I datasets. Based on these annotated datasets, we then perform four cross-dataset evaluations following the standard protocols. For example, the protocol OCI→M denotes that the model is trained on OULU-NPU, CASIA-MFSD, and Idiap Replay-Attack, and tested on MSU-MFSD. Similarly, OMI→C, OCM→I, and ICM→O are defined in the same manner. As shown in Tab 8, our method outperforms previous state-of-the-art methods, achieving the best average HTER and AUC. These results confirm the effectiveness of our approach in improving generalization in FAS.

Table 8: Cross-dataset evaluation results under widely used cross-domain protocol.

Method	O&C&I→M		O&M&I→C		O&C&M→I		I&C&M→O		Avg.	
	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)
FGHV [21]	9.17	96.92	12.47	93.47	16.29	90.11	13.58	93.55	12.88	93.51
GDA [51]	9.20	98.00	12.20	93.00	10.00	96.00	14.40	92.60	11.45	94.90
PatchNet [34]	7.10	98.46	11.33	94.58	13.40	95.67	11.82	95.07	10.91	95.95
SSAN [38]	6.67	98.75	10.00	96.67	8.88	96.79	13.72	93.63	9.82	96.46
IADG [52]	5.41	98.19	8.70	96.40	10.62	94.50	8.86	97.14	8.40	96.56
UDG-FAS [25]	5.95	98.47	9.82	96.76	5.86	98.62	10.97	95.36	8.15	97.30
TTDG [53]	4.16	98.48	7.59	98.18	9.62	98.18	10.00	96.15	7.84	97.75
SA-FAS [33]	5.95	96.55	8.78	95.37	6.58	97.54	10.00	96.23	7.83	96.42
DIVT-M [18]	2.86	99.14	8.67	96.92	3.71	99.29	13.06	94.04	7.08	97.35
GAC-FAS [15]	5.00	97.56	8.20	95.16	4.29	98.87	8.60	97.16	6.52	97.19
FLIP [32]	4.95	98.11	0.54	99.98	4.25	99.07	2.31	99.63	3.01	99.20
CFPL [20]	1.43	99.28	2.56	99.10	5.43	98.41	2.50	99.42	2.98	99.05
I-FAS [45]	<b>0.32</b>	<b>99.88</b>	0.04	99.99	3.22	98.48	<b>1.74</b>	<b>99.66</b>	1.33	99.50
<b>Ours</b>	0.42	<b>99.92</b>	<b>0.00</b>	<b>100.00</b>	<b>1.00</b>	<b>99.83</b>	2.81	99.63	<b>1.06</b>	<b>99.85</b>

### 3.2 Fine-grained analysis on spoof type robustness

To examine whether FaceCoT introduces bias toward certain spoof types, we conduct a fine-grained analysis on the Rose-Youtu [16], which contains seven representative spoofing attack types. We report the per-type detection accuracy before and after fine-tuning with FaceCoT. As shown in Tab 9, our approach achieves consistent improvements across all attack categories. These results indicate that FaceCoT provides more comprehensive semantic supervision and enhances general spoof detection capability, rather than overfitting to the dominant categories in the training set.

Table 9: Performance comparison across different spoof types in the Rose-Youtu test set before and after SFT with the proposed FaceCoT.

Cheat Type	Meaning	Number	Acc. (Zero-shot)	Acc. (After SFT)	Change
Mc	Mask: Cut eyes & mouth	202	100.00%	100.00%	–
Mf	Mask: Full face	100	74.00%	100.00%	+26.00%
Mu	Mask: Upper part cut	198	93.43%	100.00%	+6.57%
Pq	Printed paper (quivering)	200	0.00%	95.50%	+95.50%
Ps	Printed paper (still)	200	0.00%	68.00%	+68.00%
Vl	Video (Lenovo LCD)	201	0.00%	96.02%	+96.02%
Vm	Video (Mac LCD)	199	0.00%	71.36%	+71.36%

### 3.3 Comparison of zero-shot and CoT-trained models

To assess the effectiveness of our supervised CoT training, we first establish a zero-shot baseline, where large vision-language models are directly prompted with natural language to classify real versus spoof images without any fine-tuning. To further validate the robustness of our approach, we also conduct the same experiment with Qwen2.5-VL [1], a recent advanced multimodal VLM. As shown in Tab 10, both VLMs perform worse than the SOTA method I-FAS [45] under the zero-shot setting. After applying our CoT-based fine-tuning, we observe consistent and substantial gains on both models: MiniCPMV achieves a reduction of 11.61% in HTER and an improvement of 10.45% in AUC, while Qwen2.5-VL shows similar improvements (HTER reduced by 10.28% and AUC improved by 11.97%). These results demonstrate that our FaceCoT dataset and CEPL framework provide stable and significant benefits across different VLM architectures, enabling stronger discriminative ability and cross-domain generalization than relying on zero-shot reasoning alone.

Table 10: Comparison between zero-shot baselines and our supervised CoT method across different backbone VLM models considered.

Method	Average HTER (%)	Average AUC (%)
I-FAS [45]	11.31	93.71
Zero-shot(Qwen2.5-VL-7B [1])	19.60	83.75
Zero-shot(MiniCPMV-2.6-8B [42])	17.91	87.32
Ours(Qwen2.5-VL-7B [1])	9.32 (↓10.28)	95.72 (↑11.97)
Ours(MiniCPMV-2.6-8B)	6.30 (↓11.61)	97.77 (↑10.45)

### 3.4 The effect of Reinforcement Learning on CoT annotation quality

By introducing RL into the training of the caption model, the annotation accuracy is effectively improved. Although this metric demonstrates the effectiveness of RL in enhancing conclusion label accuracy, relying solely on conclusion accuracy is not sufficient to fully evaluate the semantic quality of the generated annotations. To further verify the advantages of RL-generated CoT annotations in terms of linguistic coherence and logical consistency, we design an experiment in which we treat two sets of generated CoT annotations—those produced by a model trained solely via Supervised Fine-Tuning (SFT) and those produced by an SFT-trained model further refined with RL—as separate training sets under our proposed CoT-Enhanced Progressive Learning (CEPL) framework. The comparative results, reported in Tab 11, demonstrate the tangible benefits of RL on CoT data quality: RL not only enhances annotation accuracy, but also significantly improves the consistency, coherence, and semantic reliability of the generated CoT explanations overall.

### 3.5 The effect of FaceCoT under low-resolution

In the ablation study presented in the main text, we evaluate the impact of FaceCoT data using an input resolution of  $448 \times 448$ . Given that most existing FAS methods [45, 32, 14] conduct experiments at a resolution of  $224 \times 224$ , we

Table 11: The effect of Reinforcement Learning (RL) on CoT annotation quality: Supervised Fine-Tuning (SFT) versus SFT with RL

Training methods	Results	
	HTER(%)	AUC(%)
SFT	8.00	96.97
SFT + RL	<b>6.87</b>	<b>97.27</b>

Table 12: Ablation study on CoT data at a resolution of  $224 \times 224$ , with comparison to a setup using only binary label data.

Data type	Results	
	HTER(%)	AUC(%)
Label	17.07	90.42
Label + CoT	<b>11.28</b>	<b>94.05</b>

perform an ablation study at this resolution to verify the effectiveness of our CoT data under low-resolution settings. We compare a single-stage training regime using only label classification data against a single-stage joint training regime incorporating CoT data. As shown in Tab 12, the model trained with CoT data achieves a 6.70% reduction in HTER and a 4.61% increase in AUC. These findings can be summarized as follows: (1) The relative gain at  $224 \times 224$  ( $-5.79\%$  HTER,  $+3.63\%$  AUC) is substantially larger than at  $448 \times 448$  ( $-1.42\%$  HTER,  $+1.54\%$  AUC), demonstrating that our CoT annotations help the model recover fine-grained facial cues that are otherwise lost at lower resolutions. (2) Even when applied in a simple single-stage joint training regime, the CoT-augmented model already outperforms current state-of-the-art methods, demonstrating its superior generalization and robustness conferred by CoT data training.

### 3.6 Ablation on the CoT components

To validate the effectiveness of the proposed CoT design, we conduct an ablation study by simplifying the supervision signal. Specifically, the model is trained using only the *Spoofing Description + Conclusion* components, without the intermediate steps. The results are reported in Table 13. Removing the complete CoT reasoning process leads to a noticeable performance drop across evaluation metrics. These findings support the rationale of the FaceCoT design, which follows a hierarchical global-to-local reasoning paradigm where conclusions are progressively derived from structured descriptions and intermediate reasoning steps. Such structured CoT supervision enables the model to better capture subtle spoofing cues and improves overall detection performance.

Table 13: Ablation study on the CoT components. Results are reported on OULU-NPU (O.) and the average (Avg.) over all benchmarks.

Method	O. HTER	O. AUC	Avg. HTER	Avg. AUC
Spoofing	21.59	87.17	8.59	95.88
Full CoT	5.86	97.72	7.65	96.59

## 4 Usage of LLMs

In this work, Large Language Models (LLMs) were employed as auxiliary tools in two aspects:

- **Manuscript refinement:** LLMs were used to assist in language polishing and grammar checking after the human authors had completed the technical writing. The scientific content, experiment design, and analysis were fully conducted by the authors.
- **Annotation of FaceCoT-Gold100K:** GPT-4o was used to generate Chain-of-Thought (CoT) annotations. Specifically, we carefully designed prompts to guide the model toward describing spoof-related visual cues (e.g., reflection artifacts, cutting marks), rather than allowing free-form generation. The model outputs were subsequently reviewed and refined by human experts to ensure accuracy, fairness, and domain relevance.

## 5 Limitation and feature work

We have included as many spoofing types as practically possible in FaceCoT to ensure broad coverage. Some uncommon real-world variations in devices or environments are inevitably not captured, which we leave for future extension. Furthermore, while our work primarily emphasizes the utility of FaceCoT for model training and performance gains on downstream FAS tasks, we acknowledge that its potential as a standardized benchmark for evaluation has not been fully explored. In particular, since FaceCoT provides reasoning chains, it can serve as a valuable resource to assess not only predictive accuracy but also the interpretability and coherence of model outputs. We leave a more systematic investigation of FaceCoT’s role in model evaluation and benchmarking as an important direction for future work.

## 6 Ethics statement

In designing FaceCoT, attention was paid to fairness, bias mitigation, and data privacy, ensuring that the dataset not only enhances interpretability and generalization in Face Anti-Spoofing (FAS), but also adheres to responsible research principles. To address these potential concerns, this section outlines our efforts in three dimensions: dataset fairness, language model bias mitigation, and data privacy protection.

### 6.1 Dataset bias and fairness

The original FAS datasets (e.g., CelebA-Spoof [47] and WFAS [35]) used in our work were not explicitly designed with fairness auditing or demographic balance in mind. The goal of FaceCoT, however, is to introduce a reasoning-based multimodal framework that improves interpretability and generalization in FAS models. This also provides a structured avenue for detecting and mitigating bias through Chain-of-Thought (CoT) rationales. To align FaceCoT with responsible research practices and address potential fairness concerns, we incorporated the following safeguards during our data collection and annotation process:

- **No new image data introduced:** All FaceCoT annotations are derived exclusively from publicly available FAS datasets. We did not collect or distribute any new images, and the released dataset contains only annotations indexed to existing data.
- **Bias-aware annotation pipeline:** (1) *Prompt Design:* CoT generation prompts were carefully crafted to steer the model toward spoof-specific visual cues (e.g., reflection artifacts and cutting marks), while explicitly excluding references to race, gender, age, or identity. (2) *Human-in-the-Loop Filtering:* Expert annotators reviewed all FaceCoT-Gold100K outputs and were instructed to remove any content with identity-based or stereotypical language. (3) *Random Auditing:* Manual sampling and inspection of 5,000 annotations from FaceCoT-Silver982K revealed no evidence of demographic bias or stereotype leakage. (4) *Model Validation Across Subgroups:* Models trained with FaceCoT demonstrated balanced performance across evaluation datasets, including with respect to skin tone and gender, indicating no observable subgroup bias attributable to the annotations. As shown in the outputs across various evaluation datasets in our Appendix 7, the model focused on spoofing features related to FAS rather than attributes such as age, gender, or skin tone.

### 6.2 Use of GPT-4o and mitigation of language model bias

Large foundation models such as GPT-4o may introduce bias. To mitigate such risks, we employed several safeguards:

- **Constrained use:** GPT-4o was used solely to generate CoT explanations under strict prompt constraints and was never involved in classification or decision-making tasks.
- **Annotation safeguards:** (1) FaceCoT-Gold100K annotations were reviewed and refined by human experts to eliminate any inappropriate or biased language. (2) For FaceCoT-Silver982K, we employed a reward model trained to optimize spoof-specific consistency and rule-based constraints rather than open-ended language fluency, reducing the likelihood of inherited bias.

### 6.3 Data privacy and consent

FaceCoT also respects individual privacy and consent, particularly when using publicly sourced visual data. While the datasets employed (e.g., WFAS) are released under academic or Creative Commons licenses, we designed the release to be cautious and transparent:

- **No image redistribution:** The FaceCoT release contains only annotations and metadata; no image data is redistributed or exposed.
- **Transparent documentation:** The final dataset release will include: (1) explicit documentation of all source datasets and their associated licenses; (2) clear usage guidelines requiring downstream users to adhere to original dataset terms and ethical standards.

## 7 Demonstration of result interpretability

### 7.1 Demonstration of real face

Since FAS is inherently a binary classification problem, explaining why a face is real is as important as explaining why a face is spoof. To this end, we present a representative real face case (Figure 11) to illustrate how the model perceives authenticity. Rather than relying solely on the absence of spoofing cues, the model proactively identifies positive

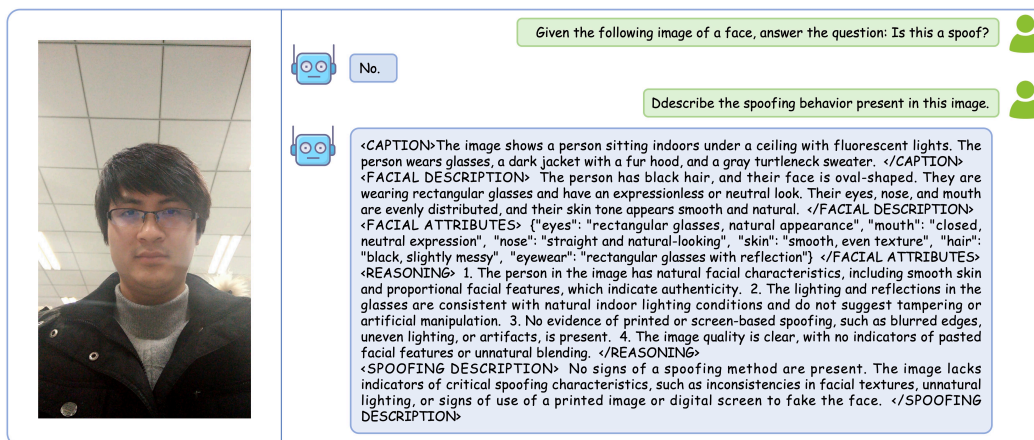
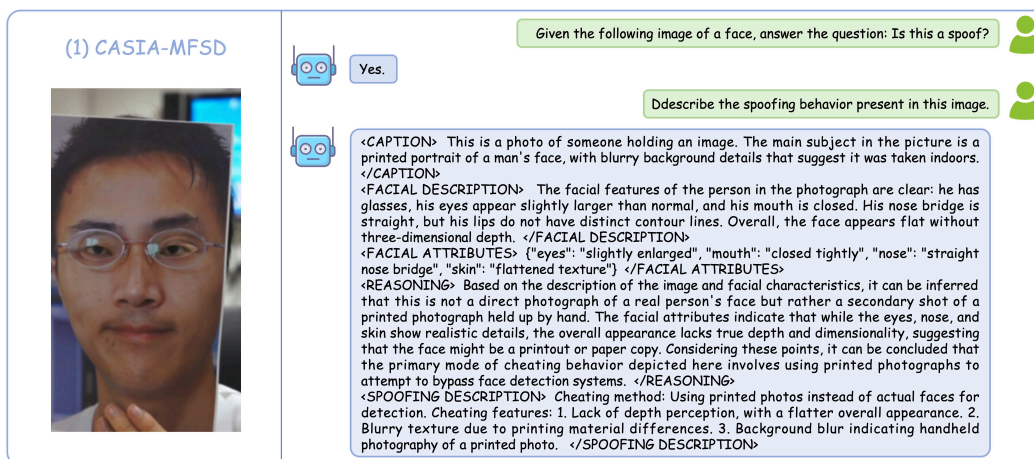


Figure 11: An interpretable result case on a real face

evidence, including natural and proportionate facial structure, realistic skin texture consistent with illumination, and coherence between facial appearance and the surrounding environment. This unified reasoning pipeline ensures that both real and spoof faces are explained under the same framework, with real cases supported by explicit authenticity cues and evidence.

## 7.2 Demonstration of eleven benchmark dataset


In this section, we present examples of our model’s interpretable outputs on eleven benchmark datasets, including MSU-MFSD [40], CASIA-MFSD [48], Idiap Replay Attack [6], OULU-NPU [3], SIW [23], Rose-Youtu [16], HKBU-MARs-V1+ [22], WMCA [10], SIW-M-V2 [12], CASIA-SURF-3DMask [43], and HiFiMask [19].



(a) An interpretable result case on CASIA-MFSD

Figure 12: Interpretable CoT outputs on eleven benchmarks

(2) CASIA-SURF-3DMask



Given the following image of a face, answer the question: Is this a spoof? ✔


Yes. 🤖

Describe the spoofing behavior present in this image. ✔

<CAPTION> The image shows an outdoor scene with a large umbrella in the background and some grass visible. In the center of the picture, there is a mannequin head wearing a white t-shirt. The face on the head appears to be superimposed onto the mannequin body. </CAPTION>  
 <FACIAL DESCRIPTION> The facial features are clearly visible but seem unnatural, particularly the skin tone and texture do not match those of the neck or shoulder areas, indicating it might be a fake face rather than a real one. </FACIAL DESCRIPTION>  
 <FACIAL ATTRIBUTES> {"eyes": "normal shape, slightly large", "mouth": "closed tightly, position normal", "nose": "average size, straight nose bridge", "skin": "color uniform but lacks realistic texture", "head": "seems to be made from other materials"} </FACIAL ATTRIBUTES>  
 <REASONING> Based on the overall description of the image, this figure does not conform to natural human characteristics as its head and shoulders appear mismatched compared to the background. Further analysis reveals that while the facial structure seems normal, the texture of the skin does not align with genuine human attributes. Specifically, the head may have been superimposed onto the mannequin's body to produce such effects. </REASONING>  
 <SPOOFING DESCRIPTION> Cheating method: Using a human face (head) spliced into an artificial mannequin body to pass off as a real person. Cheating features: 1. The facial texture does not blend well with the surrounding environment; 2. The material appearance between the head and the body looks very different, resembling two distinct objects. </SPOOFING DESCRIPTION>

(b) An interpretable result case on CASIA-SURF-3DMask

(3) HKBU-MARs-V1+



Given the following image of a face, answer the question: Is this a spoof? ✔


Yes. 🤖

Describe the spoofing behavior present in this image. ✔

<CAPTION> This is an indoor photo of a person. In the foreground, there's a person facing the camera, wearing dark clothing. The overall lighting in the picture is even. </CAPTION>  
 <FACIAL DESCRIPTION> The facial features of the person in the image are unusual. Their face appears unnatural with distinct facial contours and proportions that do not match typical human characteristics. Specifically: - The eyes are hollowed out, exposing the internal structure; - There is no nose visible, only what looks like nasal cavities; - Overall, this entire appearance seems to have been created using artificial means rather than genuine human facial features. </FACIAL DESCRIPTION>  
 <FACIAL ATTRIBUTES> {"eyes": "two eye sockets, empty holes, no pupils", "mouth": "flat, without obvious texture or luster", "nose": "absent", "skin": "normal color, but with unnatural smoothness"} </FACIAL ATTRIBUTES>  
 <REASONING> Based on the description of the image, the facial features of the subject appear unnatural, particularly the hollowed-out eyes and missing nose, which indicate that these features may not be real faces but could be artificially constructed masks. From the detailed facial descriptions, it can be observed that the eyes, nose, and skin textures all lack natural details, further suggesting they were crafted artificially. </REASONING>  
 <SPOOFING DESCRIPTION> Cheating methods: Using masks to disguise. Cheating characteristics: 1. Eyeholes present empty structures instead of actual eyes. 2. No nose exists, just the shape of the nasal cavity. 3. Mouths are plane, lacking realistic texture and gloss. 4. Overall facial features show unnatural proportions, not matching normal human appearances. </SPOOFING DESCRIPTION>

(c) An interpretable result case on HKBU-MARs-V1+

(4) HiFiMask



Given the following image of a face, answer the question: Is this a spoof? ✔

Yes. 🤖

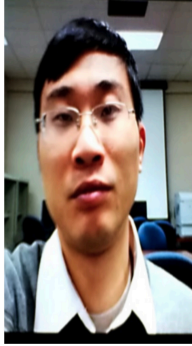
Describe the spoofing behavior present in this image. ✔

<CAPTION> This is a photograph of two people, with the main subject being a man wearing glasses and having his face covered by an unusual mask-like object. In the background, there is another person who seems to be standing in front of a wall. The overall lighting condition is even. </CAPTION>  
 <FACIAL DESCRIPTION> The male's facial features are partially obscured by a transparent plastic mask-like material that covers the entire face except for the eyes area. This cover bears clear contours resembling human skin but does not allow the true underlying appearance to be seen completely. His eyes appear through large transparent openings on this mask, while the nose area is also covered with noticeable distortion. </FACIAL DESCRIPTION>  
 <FACIAL ATTRIBUTES> {"eyes": "larger, through a transparent opening", "mouth": "mouth area covered, only partly visible", "skin": "resembles normal skin texture, but distorted due to the covering", "nose": "distorted, possibly due to reflection or refraction effects from the material"} </FACIAL ATTRIBUTES>  
 <REASONING> Based on the description of the image and facial features, it can be observed that the man's face is not entirely exposed; instead, he wears a protective mask made of transparent plastic material. While the eyes are visible, the nose area appears to have been affected by some form of distortion. Analyzing the characteristics of the facial attributes, it can be inferred that this behavior involves using non-true-to-life equipment to disguise one's identity during identification processes. </REASONING>  
 <SPOOFING DESCRIPTION> Cheating method: Wearing a transparent plastic mask-like material to conceal the true face. Cheating features: 1. Transparent material covers major facial parts, such as the nose and forehead. 2. Material reflections obscure fine details of the original face. </SPOOFING DESCRIPTION>

(d) An interpretable result case on HiFiMask

Figure 12: Interpretable CoT outputs on eleven benchmarks

(5) MSU-MFSD



Given the following image of a face, answer the question: Is this a spoof?


Yes.

Describe the spoofing behavior present in this image.

<CAPTION> This is a photo of an indoor environment, possibly an office or conference room. The person in the image is wearing glasses and appears to be taking a selfie. There are desks and chairs visible in the background, with lights mounted on the ceiling overhead. </CAPTION>  
 <FACIAL DESCRIPTION> The facial features of the person in the image include short black hair, glasses covering half of their eyes, a straight nose, lips slightly open, and overall a natural expression. The face appears clear but has some screen effects, possibly indicating it was taken from a monitor or mobile device screen. </FACIAL DESCRIPTION>  
 <FACIAL ATTRIBUTES> { "eyes": "wearing glasses, partially covered by lenses", "mouth": "slightly open", "nose": "straight", "skin": "light-colored, somewhat reflective", "hair": "short black hair", "facial effect": "screen reflection" } </FACIAL ATTRIBUTES>  
 <REASONING> Based on the description of the image, this photograph seems to have been taken through a camera system rather than directly capturing a real person's face. Further analysis of the facial features shows signs of screen reflection, particularly noticeable around the edges and lighting areas, which suggests that this may not be a direct capture of the person's actual face but instead a screen recording of a photograph or a video call. Considering these factors, it can be inferred that there might be cheating behavior present. </REASONING>  
 <SPOOFING DESCRIPTION> Cheating methods: By using a screen recording method to display an existing photograph as a live human face. Cheating characteristics: The edges of the faces show abnormal lighting reflections; the overall picture appears distorted, similar to what one would see when viewing an image on a computer screen. </SPOOFING DESCRIPTION>

(e) An interpretable result case on MSU-MFSD

(6) OULU-NPU



Given the following image of a face, answer the question: Is this a spoof?


Yes.

Describe the spoofing behavior present in this image.

<CAPTION> This is a photo of an indoor setting with office equipment and walls visible in the background. In the foreground, there is a person whose face appears to be flat without significant three-dimensional depth, resembling a printed image or a digital rendering rather than a real human face. </CAPTION>  
 <FACIAL DESCRIPTION> The facial features of the person in the image are relatively clear but appear unnatural due to their flatness. The eyes have normal size and shape, but they lack detail, giving an overall illusion of a two-dimensional effect. Overall, this face does not show any natural skin pores or reflections. </FACIAL DESCRIPTION>  
 <FACIAL ATTRIBUTES> { "eyes": "normal", "mouth": "natural", "nose": { "details": "plain surface, lacking texture" }, "skin": { "details": "flat, no textures or pores" }, "overall": { "depth": "flat", "texture": "digital or printed-like" } } </FACIAL ATTRIBUTES>  
 <REASONING> Based on the description of the image and facial features analysis, it can be determined that the face depicted in the picture lacks realistic depth, presenting as a flat image rather than a three-dimensional human face. Additionally, upon further examination, the skin details are overly smooth and devoid of genuine facial characteristics (such as pores), suggesting that this may be a photograph of a printed paper mask or a digitally generated image. Given these observations, it can be inferred that the face in the image may have been forged through some method other than genuine biological features. </REASONING>  
 <SPOOFING DESCRIPTION> Cheating methods: Using printed masks or digital generation images. Characteristics of cheating: 1. Lacks true facial depth; 2. Skin texture is overly smooth, without natural pores. </SPOOFING DESCRIPTION>

(f) An interpretable result case on OULU-NPU

(7) Replay Attack



Given the following image of a face, answer the question: Is this a spoof?

Yes.


Describe the spoofing behavior present in this image.

<CAPTION> This is a photograph of an indoor setting, featuring a male figure in the foreground with blurred background elements. The overall image quality appears to be slightly lower due to low resolution or shooting angle issues. </CAPTION>  
 <FACIAL DESCRIPTION> The man in the picture has dark curly hair, long face shape, thick eyebrows, deep-set eyes, a straight nose bridge, and sparse facial hair. His expression is natural but seems somewhat serious. </FACIAL DESCRIPTION>  
 <FACIAL ATTRIBUTES> { "eyes": "deep", "mouth": "narrow", "nose": "straight", "skin": "natural skin tone", "hair": "dark curls", "expression": "serious" } </FACIAL ATTRIBUTES>  
 <REASONING> Based on the description of the image and facial features, it can be observed that this is not a direct shot of a real person's face but rather a photo taken from another device (such as a computer screen). Considering the blurriness and unnatural lighting effects present in the image, it can be inferred that this could be a case where someone took a screenshot of a display screen for submission. Therefore, based on the above analysis, it can be determined that there may be cheating behavior involving taking screenshots of displayed content. </REASONING>  
 <SPOOFING DESCRIPTION> Cheating method: Using devices to capture screenshots of displayed images. Cheating characteristics: 1. Blurry overall image quality; 2. Lighting reflections appear as if viewed through a screen; 3. No three-dimensional depth perception; 4. Lack of realistic environment details. </SPOOFING DESCRIPTION>

(g) An interpretable result case on Replay Attack

Figure 12: Interpretable CoT outputs on eleven benchmarks

(8) Rose-Youtu



Given the following image of a face, answer the question: Is this a spoof?

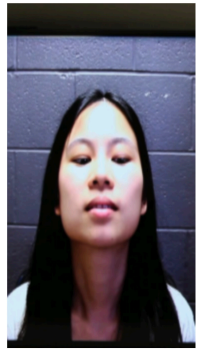
Yes.

Describe the spoofing behavior present in this image.

<CAPTION> This is an image of a desktop scene, with a person holding what appears to be a paper-cutout photo. The background includes various office items and decorations. </CAPTION>  
 <FACIAL DESCRIPTION> The figure in the picture has a clear facial expression, but it lacks depth; its features appear flat. The hair is dark and neat, the eyes are relatively large, and the mouth shape is somewhat closed. Overall, this face seems unnatural due to being a flat cut-out rather than a real three-dimensional human face. </FACIAL DESCRIPTION>  
 <FACIAL ATTRIBUTES> { "eyes": "large but lacking depth", "mouth": "closed tightly", "nose": "flat, without obvious texture", "skin": "flat appearance, no realistic skin pores or reflections", "hair": "dark color, neatly combed", "other": "overall image is two-dimensional" } </FACIAL ATTRIBUTES>  
 <REASONING> Based on the description of the image, this is not a genuine photograph of a live person's face but a printed paper-cutout of a face. From the facial description, it can be seen that while the facial details are clear, they lack the depth and realistic textures characteristic of a true human face. Further analysis confirms that all facial attributes indicate an attempt at forgery through physical means rather than utilizing actual human facial characteristics. Therefore, it can be concluded that this image involves facial cheating behavior. </REASONING>  
 <SPOOFING DESCRIPTION> Cheating method: Using printed paper-cutouts to substitute for real faces. Cheating features: 1. Lack of depth, presenting a flat effect; 2. Clear edge lines where the paper was trimmed; 3. The hand holding the paper further indicates the use of a non-living form. </SPOOFING DESCRIPTION>

(h) An interpretable result case on Rose-Youtu

(9) SiW



Given the following image of a face, answer the question: Is this a spoof?

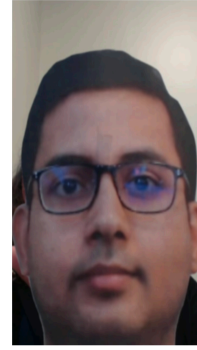
Yes.

Describe the spoofing behavior present in this image.

<CAPTION> This is a photo of an image displayed on a screen, featuring a portrait of a person against a dark wall background. </CAPTION>  
 <FACIAL DESCRIPTION> The figure in the image has long straight hair that reaches their shoulders. Their facial features are clear, with prominent eyes looking directly ahead, slightly closed lips, and well-defined cheeks. There are no noticeable obstructions or distortions on the face. </FACIAL DESCRIPTION>  
 <FACIAL ATTRIBUTES> { "eyes": "gazing forward", "mouth": "slightly closed", "nose": "straight", "skin": "evenly colored", "hair": "long and straight", "background": "dark brick wall" } </FACIAL ATTRIBUTES>  
 <REASONING> Based on the overall description of the image, it can be observed that this picture was captured through a screen rather than directly taking a photograph of the subject. Facial features appear normal but lack natural lighting effects, and there is obvious glare from the screen. From the facial attributes, while the characteristics are complete, they do not align perfectly with typical human visual presentation due to differences in reflections caused by screens. Therefore, it can be inferred that this is a screenshot taken for facial recognition verification purposes, which may pose a risk during actual identification processes. </REASONING>  
 <SPOOFING DESCRIPTION> Cheating method: Taking screenshots using mobile devices. Cheating characteristics: 1. Image originates from a display screen, showing glare and unnatural lighting conditions; 2. Lack of realistic environment illumination, such as shadows or depth cues; 3. Screen reflection leading to a non-natural perspective. </SPOOFING DESCRIPTION>

(i) An interpretable result case on SiW

(10) SiW-M-V2



Given the following image of a face, answer the question: Is this a spoof?

Yes.

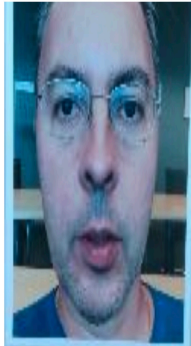
Describe the spoofing behavior present in this image.

<CAPTION> This is a photo of a facial image, with an indoor background and light source from the top. The face in the picture appears as if it's part of a flat object rather than a real human face. </CAPTION>  
 <FACIAL DESCRIPTION> The face in the picture is relatively symmetrical but lacks depth, giving it a two-dimensional appearance. It has dark hair, large black-framed glasses covering the eyes, blue irises shining through, clear nasal cavity structure, slightly closed lips, and fair skin overall. </FACIAL DESCRIPTION>  
 <FACIAL ATTRIBUTES> { "eyes": "large, wearing glasses, looking directly at the camera", "mouth": "closed tightly", "nose": "clearly visible, proportions normal", "skin": "fair, even lighting", "headwear": "no headwear or accessories other than glasses", "lighting effect": "light from above illuminates clearly" } </FACIAL ATTRIBUTES>  
 <REASONING> Based on the description of the image, this face does not possess a three-dimensional structure typical of a real person; instead, it seems like a flat representation (such as a photograph printed on paper). Further confirmation from the facial attributes indicates that while the details are relatively complete, there is no significant depth information present. Combining these observations, it can be inferred that this could be an attempt to simulate a true human face using a flat medium, indicating cheating behavior in detection systems. </REASONING>  
 <SPOOFING DESCRIPTION> Cheating methods: Using a flat printing method to create a fake human face. Cheating features: - Lack of depth perception; - Uniform surface texture, similar to paper material; - Even lighting without shadow effects. </SPOOFING DESCRIPTION>

(j) An interpretable result case on SiW-M-V2

Figure 12: Interpretable CoT outputs on eleven benchmarks

(11) WMCA



Given the following image of a face, answer the question: Is this a spoof? 👤

Yes. 👤

Ddescribe the spoofing behavior present in this image. 👤

```

<CAPTION> This is a photo of a man's face, with the background being an indoor setting featuring some blurry objects. The overall image has a certain degree of blurriness. </CAPTION>
<FACIAL DESCRIPTION> The person in the picture wears glasses and looks directly at the camera. His facial features are visible, including his eyes, nose, mouth, and hair. There is slight blur around his face, particularly noticeable on his skin texture. </FACIAL DESCRIPTION>
<FACIAL ATTRIBUTES> { "eyes": "visible but with a faint haze", "mouth": "slightly open", "nose": "normal proportion", "skin": "with traces of fuzziness", "hair": "short, light-colored", "background": "blurry interior scene" } </FACIAL ATTRIBUTES>
<REASONING> Based on the description of the image, it can be observed that the overall image quality is relatively low, possibly due to photography or display issues. From the facial attributes, it is apparent that these characteristics could have been intentionally modified or produced by non-living means (such as printing). Combining all factors, this seems to be an image generated through non-living media rather than genuine human facial data. </REASONING>
<SPOOFING DESCRIPTION> Cheating method: By uploading a printed photo instead of a real person for detection. Cheating features: 1. Image shows obvious signs of shooting from a printout, lacking dynamic features. 2. Skin texture is blurry and inconsistent with live humans' natural appearance. 3. Background exhibits reflections or glare marks commonly found on paper prints. </SPOOFING DESCRIPTION>

```

(k) An interpretable result case on WMCA

Figure 12: Interpretable CoT outputs on eleven benchmarks