VRAG-DFD: Verifiable Retrieval-Augmentation for MLLM-based Deepfake Detection

Supplementary Material

A. Implementation Details

A.1. Retrieval and Inference Settings

In our dynamic forensic retrieval module, we set the number of retrieved evidence items to $k = 5$. This choice is based on empirical observations on the validation set, where $k = 5$ provides an effective trade-off between supplying sufficient supportive evidence and maintaining efficient MLLM reasoning. Larger values of k tend to introduce unnecessary noise due to excessively long context windows, while smaller values lead to insufficient contextual grounding.

A.2. Training Configuration and Data Augmentation

As described in the main paper, the training pipeline of VRAG-DFD consists of three stages. To maximize model performance, we adopt tailored training strategies for each stage.

Stage-1: Foundational Visual Alignment. We use approximately 80,000 images from the FaceForensics++ (FF++) training set.

- **Training Strategy.** This stage aims to establish the model's foundational visual sensitivity to forgery cues. We therefore perform LoRA fine-tuning on the ViT (vision encoder), the Aligner (projection module), and the LLM.
- **Data Augmentation.** To enhance the visual encoder's robustness and mitigate overfitting, we apply a set of augmentations implemented using the Albumentations library, such as Horizontal Flip, Image Compression, HueSaturationValue Adjustment.

Stage-2 & Stage-3: Critical Reasoning Training. Both Stage-2 and Stage-3 are trained on 9,000 samples sampled from the FF++ dataset. In Stage-2, these samples are used for supervised fine-tuning (SFT) with F-CoT annotations.

- **Training Strategy.** During these stages, we freeze all ViT parameters and only lora finetune the Aligner and LLM. This design preserves the general visual features learned in Stage-1, while encouraging the model to focus on logical reasoning, evidence aggregation, and conflict resolution.
- **Data Augmentation.** To ensure that the model can capture extremely subtle pixel-level forgery traces, no data augmentation is applied in Stage 2 and Stage 3. This avoids introducing artificial distortions that could obscure genuine forgery artifacts.

B. Database Construction Details

To support the VRAG-DFD framework, we construct two core datasets: (1) the *Forensic Knowledge Database (FKD)* for retrieval, and (2) the *Forensic Chain-of-Thought (F-CoT)* dataset for critical reasoning training.

B.1. Construction of the Forensic Knowledge Database (FKD)

Data Sources and Balancing Strategy. The original FF++ dataset exhibits a Real:Fake ratio of approximately 1:4. To build a category-balanced retrieval database, we adopt the following sampling strategy, which produces a balanced 1:1 distribution of real and manipulated samples:

- **Fake Frames:** From each video of the four manipulation types (DeepFakes, Face2Face, FaceSwap, NeuralTextures), we randomly sample around 3 frames, resulting in approximately 8,000 fake frames.
- **Real Frames:** From each real video, we sample approximately 12 frames, also yielding around 8,000 real frames.

Expert-Level Annotation. To ensure high-quality and professional annotations, we adopt a combination of contrastive annotation and a structured forgery guideline. Using Gemini 2.5 Pro, we input each target image together with its corresponding real/fake counterpart and instruct the model to produce detailed descriptions of forged regions, guided by manipulation-specific characteristics (e.g., texture discontinuities in NeuralTextures). The full prompt is shown in the END of this supplementary material.

B.2. Construction of the Forensic Chain-of-Thought (F-CoT) Dataset

Data Source. We select 1,000 videos from the FF++ training split that are not used during Stage 1 training. From each video, 9 frames are sampled, resulting in a dataset of 9,000 samples.

Reference Information Generation. For each sample, we retrieve the top-k most relevant images from FKD (excluding itself) and use their text annotations as external reference information.

Critical Logic Construction. To encourage the model to reason critically about retrieved evidence, we categorize each sample into one of three types based on the Stage 1 prediction and the quality of retrieved evidence. Gemini 2.5 Pro, acting as a "teacher" with access to ground-truth labels, is prompted to produce corresponding chain-of-thought reasoning:

Table 7. **Data Quality Comparison between FKD and DDVQA.** Replacing FKD with DDVQA leads to a significant performance drop, indicating that forensic-grade knowledge is essential for effective reasoning.

Metrics	Test Set AUC					Avg.
	CDF1	CDF2	DFDC	FFIW	WDF	
DD-VQA [53]	96.17	94.83	79.48	92.50	85.88	89.77
FKD (Ours)	**99.60**	**95.97**	**81.82**	**93.49**	**88.96**	**91.62**

- **Type-1: Cross-Validation.** *Condition:* Stage 1 predicts correctly. *Objective:* Teach the model to verify its visual judgment using supportive retrieved evidence while ignoring irrelevant noise.
- **Type-2: Evidence-Guided Correction.** *Condition:* Stage 1 predicts incorrectly, but retrieved evidence contains strong, reliable cues. *Objective:* Guide the model to recognize its initial mistake and rely on high-quality external evidence for correction.
- **Type-3: Resilient Rejection.** *Condition:* Stage 1 predicts incorrectly, and retrieved evidence is misleading (e.g., semantically similar but different manipulation type). *Objective:* Train the model to identify unreliable evidence, reject misleading cues, and fall back to its internal visual knowledge for the final judgment.

The detailed F-CoT generation prompt in the END of this supplementary material.

C. Additional Experimental Results

C.1. Data Quality Comparison: FKD vs. DDVQA.

To further examine how knowledge quality influences retrieval-based reasoning, we replace the FKD retrieval corpus in VRAG-DFD with the textual descriptions from the DDVQA [53] dataset and conduct experiments under identical settings. As shown in Table 7, replacing FKD with DDVQA results in a substantial performance drop. This observation demonstrates that the fine-grained, expert-level forensic annotations contained in FKD are crucial for enhancing reasoning ability. In contrast, the descriptions in DDVQA are relatively coarse and lack professional forensic knowledge, making them insufficient for providing stable and reliable forgery cues. Overall, these findings underscore the critical importance of high-quality forensic knowledge for cross-dataset generalization and robust reasoning.

C.2. Contribution of the RAG Module Across Stages

To quantify the efficiency of external knowledge retrieval, we conduct an ablation study by selectively activating the RAG module at each training stage. The results in Table 8 demonstrate substantial and consistent gains across all stages. This confirms that the RAG mechanism is an integral component of our framework, consistently bridging the knowledge gap from initial visual alignment to final critical reasoning.

Table 8. **Ablation study on the impact of the RAG module across three training stages.** We report the AUC (%) on five generalization benchmarks. The column 'Δ Avg.' highlights the performance gain attributable to the RAG module within each stage. Crucially, we observe that RAG provides **substantial improvements consistently across all stages**, demonstrating that dynamic retrieval of expert knowledge is fundamental to bridging the generalization gap, even in the early visual alignment phase.

Stage	Config	Test Set AUC					Δ
		CDF1	CDF2	DFDC	FFIW	WDF	
Stage-1	*w/o* RAG	89.72	86.30	73.31	78.59	78.73	**+8.16**
	***w/* RAG**	96.77	94.03	78.54	90.89	87.63	
Stage-2	*w/o* RAG	91.33	86.96	71.57	76.43	70.05	**+10.66**
	***w/* RAG**	91.53	95.80	81.58	92.37	89.13	
Stage-3	*w/o* RAG	91.53	86.98	72.05	76.51	70.04	**+12.2**
	***w/* RAG**	99.6	95.97	81.82	93.49	88.96	

Table 9. Effectiveness across different base MLLMs (AUC %). We show the performance of each model before (zero-shot) and after ('+ VRAG-DFD (Ours)') applying our three-stage VRAG-DFD pipeline.

Metrics	Test Set AUC				
	CDF1	CDF2	DFDC	FFIW	WDF
Qwen2.5VL-7B	66.00	73.35	56.82	71.91	66.82
Qwen2.5VL-7B + ours	**99.60**	**95.97**	**81.82**	**93.49**	**88.96**
Intern3VL-8B	66.33	72.06	57.19	70.86	65.75
Intern3VL-8B + ours	**99.19**	**95.09**	**80.16**	**92.45**	**90.00**

C.3. Effectiveness on Different Base MLLMs.

To verify the model-agnostic nature of the VRAG-DFD framework, we replace the default `Qwen2.5VL` with `InternVL3-8B`, applying our full three-stage training pipeline to each. As summarized in Table 9, all base models exhibit substantial improvements in cross-dataset detection performance after training under our framework. This demonstrates that our approach serves as a general and effective paradigm for teaching MLLMs advanced forensic reasoning skills.

C.4. Latency Analysis of the RAG Module

The construction and indexing of the FKD database are performed entirely offline, ensuring that no additional inference overhead is introduced during deployment. During inference, the RAG module only performs FAISS-based retrieval over precomputed feature embeddings. As shown in Tab. 10, the retrieval operation accounts for only 0.35% of the total computational cost. This negligible overhead indicates that the integration of the RAG module has a minimal impact on overall inference latency.

Table 10. Computational complexity analysis of VRAG-DFD.

	Retrieval		Inference		Total	
Metric	GFLOPs	Ratio	GFLOPs	Ratio	GFLOPs	Ratio
Value	∼81	**0.35%**	∼22,760	99.65%	∼22,841	100.00%

Table 11. **Ablation study on frame sampling density in Stage-1.** We compare sparse sampling (8 frames/video) against dense sampling (all frames). The results demonstrate a **substantial performance leap** in the average AUC (+3.80%), confirming that dense visual supervision is indispensable for learning robust forgery features across diverse datasets.

Samp. Strategy	Total	Test Set AUC					Avg.
		CDF1	CDF2	DFDC	FFIW	WDF	
Sparse (8 fms)	∼28k	**92.14**	76.85	70.69	70.40	77.95	77.61
Dense (All fms)	**∼80k**	89.72	**86.30**	**73.71**	**78.59**	**78.73**	**81.41**
Gain (Δ)	-	*-2.42*	**+9.45**	**+3.02**	**+8.19**	**+0.78**	**+3.80**

C.5. Impact of Frame Sampling Density in Stage-1

In Stage-1, we adopt dense frame sampling over all 2,500 training videos (approximately 80k images) to maximize the model's visual alignment capability. To validate the necessity of this computationally intensive strategy, we compare it with sparse sampling, where only 8 frames per video are selected.

Results. As reported in Table 11, dense sampling significantly outperforms sparse sampling, increasing the average AUC from 77.61% to 81.41%, corresponding to a performance gain of nearly +3.8%.

Analysis. This substantial improvement reveals two key insights:

1. **Necessity of Dense Supervision.** Forgery artifacts in deepfakes, such as frame-to-frame jitter and transient texture anomalies, are often sparse and non-continuous. Sparse sampling may miss these critical discriminative frames, leading to an incomplete feature space.
2. **Solid Foundation for Subsequent Stages.** Stage-1 visual alignment serves as the cornerstone of the entire VRAG-DFD framework. Our experiments indicate that dense training with all available frames is essential to obtain a sufficiently robust visual encoder, which underpins the reasoning-centric Stage-2 and Stage-3.

C.6. Robustness against Adversarial Retrieval Noise

To evaluate whether VRAG-DFD can resist misleading external evidence, we specifically analyze a *difficult subset* of the test set. This subset is defined as samples where the retriever fails, i.e., the majority of the top-k retrieved evidences belong to incorrect categories.

Results. As shown in Table 12, VRAG-DFD demonstrates remarkable resilience even under strongly misleading retrieval conditions. Across all five evaluation datasets, the proportion of correctly classified samples consistently exceeds 94%.

Table 12. **Robustness against misleading retrieval evidence.** We evaluate a subset of adversarial samples where the initial visual perception is correct (s_1=Correct), but the RAG evidence is misleading (RAG=False). The results demonstrate that VRAG-DFD maintains an exceptionally high **Robustness Rate** ($> 96\%$ on avg.), effectively rejecting misleading external noise.

Metric	Test Set AUC					Weigh. Avg.
	CDF-v1	CDF-v2	DFDC	FFIW	WDF	
# Adv. Samp.	34	429	534	1,374	5,226	7,597
# VRAG_DFD Cor.	32	417	520	1,345	5,048	7,362
Robust. Rate (%)	94.12	97.20	97.38	97.89	96.59	**96.91**

Analysis. These results indicate that VRAG-DFD does not simply "copy" the retrieved evidence. Instead, it successfully learns to critically weigh the reliability of external evidence and perform reasoning that accounts for potential inconsistencies, highlighting its robust and principled inference capabilities.

C.7. Qualitative Analysis

To visually demonstrate the interpretability and reasoning capabilities of VRAG-DFD, we present qualitative examples in Figures 5, 6, 7 and 8. These figures display the query images alongside the complete Forensic Chain-of-Thought (F-CoT) responses generated by our model, illustrating how it integrates visual analysis and retrieval evidence to reach a final verdict.

Real

<Preliminary Visual Analysis>
Upon examining the `Query_Image`, I observe several key visual characteristics. The overall image quality is low, but the facial features appear consistent with the image's resolution. The lighting on the face is coherent, with plausible shadows under the chin and on the right side of the face, suggesting a single light source. The skin texture, while not high-resolution, is uniform across the entire face, including the forehead, cheeks, and chin. There are no isolated patches of unnatural smoothness or blurring. The transition between the face and the hairline, as well as the jawline, appears natural without any sharp, artificial edges or smudging artifacts. The facial features, such as the eyes and mouth, are proportionally aligned and do not show obvious signs of anatomical distortion.
</Preliminary Visual Analysis>
<RAG Reference Information Analysis>
The retrieval report provides the following information:
Supporting Evidence (supports my judgment):`(\"Real: [Consistent Skin Texture]: The overall level of blurriness is consistent across the entire face, hair, and background, which is a characteristic of an authentic, unedited low-resolution image.\", 0.692)`
Opposing Evidence (contradicts my judgment): `(\"Fake: [Mouth]: The teeth are unnaturally shaped and misaligned within the mouth, appearing distorted and blocky.\", 0.717)`, `(\"Fake: [Eyes]: The eyelids and periocular areas exhibit significant blurriness, lacking the pores and fine textures...\", 0.704)`
Analysis:The RAG report is noisy and presents a significant conflict. It contains a very high-scoring `Fake` reference (0.717) that points to specific artifacts, but also a high-scoring `Real` reference (0.692) that suggests the image is authentic due to consistent low resolution. A careful cross-verification is required.
</RAG Reference Information Analysis>
<Fusion, Reasoning, and Decision>
My preliminary visual analysis indicates that the image displays characteristics of an authentic, low-resolution photograph. I must now cross-verify this against the conflicting RAG report:
1. Verify Opposing Evidence: The RAG report's highest-scoring `Fake` evidence (similarity 0.717) claims \"unnaturally shaped and misaligned teeth\" and \"significant blurriness\" in the eyes. Upon re-examining the `Query_Image`, I cannot confirm these specific artifacts. The teeth are not clearly defined, but their shape is consistent with the overall low resolution of the image, not a specific manipulation. The blurriness in the eyes is consistent with the blurriness of the rest of the face, which is a sign of low image quality, not a localized forgery artifact. Therefore, this evidence is likely retrieval noise from a semantically similar but genuinely fake image and must be discarded.
2. Verify Supporting Evidence:The RAG report's `Real` evidence (similarity 0.692) states that \"the overall level of blurriness is consistent across the entire face, hair, and background.\" This observation can be confirmed and perfectly aligns with my visual analysis. The consistency of the low resolution across all facial features is a strong indicator of authenticity.Conclusion: My preliminary visual analysis, which identified consistent low-resolution features, has been strongly cross-verified by the \"useful evidence\" from the RAG report. The high-scoring \"noise evidence\" from RAG, which suggested forgery, has been effectively refuted by the visual facts of the `Query_Image`. Therefore, my final judgment is that the image is Real.
</Fusion, Reasoning, and Decision>
<answer>Real</answer>

Figure 5. Qualitative analysis of VRAG-DFD (real case 1).

Real

<Preliminary Visual Analysis>
Upon examining the `Query_Image`, I observe several key visual characteristics. The skin texture on the man's face, particularly on the cheeks and forehead, appears consistent with the overall low resolution of the image. There are no isolated patches of unnatural smoothness or blurring. The lighting across the face seems coherent, with shadows and highlights that are consistent with a single light source. The facial features, such as the furrowed eyebrows and the expression of the mouth, appear anatomically plausible and coordinated. The transition from the jawline to the neck and the hairline to the forehead is smooth, without any sharp, artificial edges or color mismatches. The stubble on the chin and around the mouth retains some fine texture, which is consistent with the rest of the facial hair.**</Preliminary Visual Analysis>**
<RAG Reference Information Analysis>
The retrieval report provides the following information:
Supporting Evidence (supports my judgment):`(\"Real: [Consistent Skin Texture]: Although the image has low resolution, the degree of softness and lack of fine detail on the facial skin is consistent with the blurriness of the hair and the background, indicating the entire image was captured with the same focus and quality.\", 0.6406)`
Opposing Evidence (contradicts my judgment):`(\"Fake: [Mouth]: The lips appear blurry and lack the fine lines and texture... [Skin]: The skin on the chin and around the mouth is unnaturally smooth\", 0.6383)`, `(\"Fake: [Mouth]: The lips in the manipulated image lack sharp definition and fine wrinkles, appearing blurry and overly smooth.\", 0.6308)`
Analysis:The RAG report is noisy and presents a significant conflict. It provides high-scoring evidence for both `Real` (0.64) and `Fake` (0.64) labels, with the scores being very close. This requires careful cross-verification with the primary visual evidence.
</RAG Reference Information Analysis>
<Fusion, Reasoning, and Decision>
My preliminary visual analysis indicates that the image displays characteristics of authenticity, such as consistent texture and plausible facial features. I must now cross-verify this against the conflicting RAG report:
1. Verify Opposing Evidence:The RAG report's highest-scoring `Fake` evidence (similarity 0.6383) claims the \"lips appear blurry\" and the \"skin on the chin and around the mouth is unnaturally smooth.\" This assertion cannot be confirmed by my visual inspection. The blurriness of the lips is consistent with the overall low resolution of the entire image, not an isolated artifact. The skin texture on the chin and around the mouth is not unnaturally smooth; it is consistent with the texture on the forehead and cheeks. Therefore, this evidence is likely retrieval noise, mistaking low image quality for a forgery artifact, and must be discarded.
2. Verify Supporting Evidence: The RAG report's `Real` evidence (similarity 0.6406) states that \"the degree of softness and lack of fine detail on the facial skin is consistent with the blurriness of the hair and the background.\" This statement can be confirmed and perfectly aligns with my visual analysis. The consistency of the low-resolution effect across the entire image is a strong indicator of authenticity. Conclusion: My preliminary visual analysis suggesting the image is `Real` has been strongly cross-verified by the credible \"Supporting Evidence\" from the RAG report. The high-scoring \"Opposing Evidence\" has been effectively refuted and rejected as it contradicts the visual facts of the `Query_Image`. Therefore, based on the primary visual evidence, reinforced by the verified RAG context, my final judgment is that the image is authentic.
</Fusion, Reasoning, and Decision>
<answer>Real</answer>

Figure 6. Qualitative analysis of VRAG-DFD (real case 2).

Fake

<Preliminary Visual Analysis>
Upon examining the `Query_Image`, I observe several key visual characteristics. The skin texture on the man's face, particularly across the forehead and cheeks, appears unusually smooth and lacks the fine details typical of natural skin, such as pores or fine lines. This smoothness gives the face a slightly \"waxy\" or \"mask-like\" quality. Additionally, the area around the mouth, specifically the lips, seems to have a subtle blurriness and lacks sharp definition. The transition between the lips and the surrounding skin is not as crisp as one might expect.
</Preliminary Visual Analysis>
<RAG Reference Information Analysis>
The retrieval report provides the following information:
Supporting Evidence (supports my judgment): `(\"Fake: [Skin]: The skin on the manipulated face, especially in the forehead and cheek areas, appears unusually smooth, lacking high-frequency details... This detail loss phenomenon gives the face a 'mask-like' appearance.\", similarity: 0.5938)`
Opposing Evidence (contradicts my judgment): `(\"Real: [Consistent Skin Texture]: ...the level of blurriness and digital noise on the facial skin is consistent with that of the neck, hair, and background...\", similarity: 0.5938)`
Analysis:The RAG report is highly conflicting. The top-scoring reference (0.5938) suggests the image is `Fake` due to unnatural skin smoothness, while another reference with the same score suggests it is `Real` due to consistent blurriness. This indicates significant noise in the retrieval, requiring careful cross-verification with the primary visual evidence.
</RAG Reference Information Analysis>
<Fusion, Reasoning, and Decision>
My preliminary visual analysis identified unnaturally smooth skin and a lack of definition around the mouth. I must now cross-verify this against the conflicting RAG report:
1. Verify Opposing Evidence: The RAG report's `Real` evidence (similarity 0.5938) claims that the blurriness is consistent across the entire image. However, this claim cannot be fully confirmed by my visual inspection. While the entire image is low-resolution, the smoothness on the face appears more pronounced and lacks the fine texture that would be expected, even in a blurry image. The skin on the forehead and cheeks has a different quality compared to the neck and clothing. Therefore, this evidence is likely retrieval noise, mistaking general low quality for consistent authenticity, and must be discarded.
2. Verify Supporting Evidence: The RAG report's `Fake` evidence (similarity 0.5938) states that the skin \"appears unusually smooth, lacking high-frequency details\" and has a \"'mask-like' appearance.\" This description can be directly confirmed on the `Query_Image`. The waxy, textureless quality I observed on the man's forehead and cheeks perfectly matches this description. The two pieces of evidence are mutually reinforcing.

Conclusion: My preliminary visual analysis, which pointed towards forgery due to unnatural skin texture, has been strongly cross-verified by the useful evidence from the RAG report. The conflicting \"noise\" evidence from RAG has been effectively rejected by the visual facts. Therefore, based on the primary visual evidence, reinforced by the credible RAG supporting evidence, my final judgment is that the image is Fake.
</Fusion, Reasoning, and Decision>
<answer>Fake</answer>

Figure 7. Qualitative analysis of VRAG-DFD (fake case 1).

Fake

<Preliminary Visual Analysis>
Upon examining the `Query_Image`, I observe several key visual characteristics. The skin texture on the woman's face, particularly on the cheeks and forehead, appears unusually smooth and lacks the fine, high-frequency details one would expect in a natural photograph, even a low-resolution one. There is a noticeable absence of natural wrinkles or pores. Additionally, the area around the mouth and nose seems slightly blurry and lacks sharp definition, which contrasts with the relative clarity of the hair strands. The lighting on the face seems consistent, but the texture itself feels artificially uniform and flat.\n
</Preliminary Visual Analysis>
<RAG Reference Information Analysis>
The retrieval report provides the following information:
Supporting Evidence (supports my judgment):`(\"Fake: [Skin]: The skin on the cheeks and around the mouth appears abnormally smooth and blurry, with a noticeable absence of fine wrinkles and natural skin texture.\", 0.650)`, `(\"Fake: [Mouth]: The lip region lacks sharpness and fine details, such as natural lip creases, making it appear slightly blurry.\", 0.650)`
Opposing Evidence (contradicts my judgment): `(\"Real: [Consistent Skin Tone/Lighting]: The lighting on the face is consistent, originating from the front-left. This creates plausible shadows on the right side of her face and under her nose.\", 0.670)`
Analysis: The RAG report is noisy and presents a significant conflict. It contains a very high-scoring piece of opposing evidence (0.670) that suggests the image is `Real`, but also provides high-scoring supporting evidence (0.650) that points to forgery. A careful cross-validation is required.\n
</RAG Reference Information Analysis>
<Fusion, Reasoning, and Decision>
My preliminary visual analysis identified unnaturally smooth skin and a lack of fine details around the mouth. I must now cross-verify this against the conflicting RAG report:
1. Verify Opposing Evidence: The RAG report's highest-scoring `Real` evidence (similarity 0.670) claims \"Consistent Skin Tone/Lighting\". While the lighting is indeed consistent, this claim cannot be fully confirmed because it overlooks the key visual artifact I identified: the complete absence of natural skin texture. The claim of \"consistent\" skin tone is misleading because the texture is not natural. Therefore, this evidence is likely retrieval noise based on semantic similarity (a woman with glasses) rather than a correct analysis of forgery artifacts, and it must be discarded.
2. Verify Supporting Evidence: The RAG report's `Fake` evidence (similarity 0.650) states that the skin is \"abnormally smooth and blurry\" and lacks \"fine wrinkles and natural skin texture.\" This description can be directly confirmed and perfectly matches my visual observation of the `Query_Image`. The lack of texture is a significant indicator of manipulation.

Conclusion: My preliminary visual analysis, which pointed towards a `Fake` image due to unnaturally smooth skin, has been strongly cross-verified by the useful evidence from the RAG report. The high-scoring opposing evidence has been rejected as it contradicts the clear visual facts present in the image. Therefore, based on the primary visual evidence, reinforced by the credible RAG supporting evidence, my final judgment is that the image is Fake.
</Fusion, Reasoning, and Decision>
<answer>Fake</answer>

Figure 8. Qualitative analysis of VRAG-DFD (fake case 2).

Prompt for DeepFakes Annotation

System Prompt:
You are an expert in face tampering detection. Your task is to strictly compare the `[Manipulated Image]` and the `[Original Image]`, first precisely locate the manipulated regions, and then **explain** how the forgery artifacts within those regions were caused by the improper processing of DeepFakes technology. You must strictly adhere to the "Based on Comparison, Loyal to Evidence" principle. **Hallucinations are strictly forbidden**. Focus on artifacts left by the **forgery process** (e.g., blending, alignment), not natural facial features (e.g., makeup, appearance). Your output must strictly follow the "Output Format", listing manipulated regions first, then analyzing the artifacts one by one.

User Prompt:
Task Definition: You will receive two images: the first is a `[Manipulated Image]` generated using DeepFakes technology, and the second is the corresponding `[Original Image]`. Carefully compare the two images to locate manipulated regions, then explain the forgery artifacts using the "DeepFakes Potential Forgery Artifacts Reference Guide". Focus solely on artifacts caused by DeepFakes, not natural differences.

Core Analysis Principles:
1. Carefully compare all facial features, hair, skin color, etc., to locate manipulated regions without omissions or hallucinations.
2. Distinguish between:
   - **Analyze This - Forgery Artifacts**: Anomalies caused by technical flaws (e.g., low-resolution bottlenecks, blending, alignment failures).
     Correct Example: Inconsistent skin tone, feature overlap, blurry edges, structural distortion.
   - **Ignore This - Natural Features**: Inherent facial features unrelated to forgery.
     Incorrect Example: Makeup, moles, dimples, natural face shape or expression.
3. Attribute observed phenomena (e.g., "skin looks airbrushed") to the technical flaw (e.g., "detail loss due to low resolution").
4. Report all types of artifacts present in a region, not only the most obvious.

DeepFakes Potential Forgery Artifacts Reference Guide:
1. **Blending Border Artifacts (High-Incident Area)**
   - *Cause*: Poisson blending at facial edges (face periphery, hair, chin, neck).
   - *Manifestation*: Edge color differences, unnatural blurring/halo effects along contours.
2. **Structural Abnormality (Feature Misalignment/Distortion)**
   - *Cause*: Keypoint alignment failure.
   - *Manifestation*: Distorted or misaligned facial features; feature overlap (e.g., eyebrows).
3. **Detail Loss**
   - *Cause*: Low-resolution face generation.
   - *Manifestation*: Blurry eyes, teeth, irises; abnormally smooth skin lacking texture or pores.

Output Format:
First list manipulated regions, then describe the corresponding forgery artifacts.
Reference Example:
Manipulated Regions: Skin, Eyebrows
Forgery Artifacts: [Skin]: Central area cool white, periphery yellowish-black, clear blending boundary. [Eyebrows]: Ghosting caused by facial alignment failure.

Here are `[Manipulated Image]` and the `[Original Image]`.
`[Manipulated Image]`: {{manipulated_image_path}}
`[Original Image]`: {{original_image_path}}

Prompt for Face2Face Annotation

System Prompt:
You are an expert in face tampering detection. Your task is to strictly compare the `[Manipulated Image]` and the `[Original Image]`, first precisely locate the manipulated regions, and then **explain** how the forgery artifacts within those regions were **caused by the Face2Face 3D model re-rendering process**. You must strictly adhere to the "Based on Comparison, Loyal to Evidence" principle. **Hallucinations are strictly forbidden**. Focus on artifacts left by the **forgery process** (e.g., 3D model tracking, re-rendering, expression transfer), not natural facial features (e.g., makeup, appearance). Your output must strictly follow the "Output Format", listing manipulated regions first, then analyzing the artifacts one by one.

User Prompt:
Task Definition: You will receive two images: the first is a `[Manipulated Image]` generated using Face2Face technology, and the second is the corresponding `[Original Image]`. Carefully compare the two images to locate manipulated regions, then explain the forgery artifacts using the "Face2Face Potential Forgery Artifacts Reference Guide". Focus solely on artifacts caused by Face2Face technology, not natural expression differences.

Core Analysis Principles:
1. Identify all differences to locate manipulated areas, including facial features, hair, and skin tone. Pay special attention to the lips and surrounding areas.
2. Distinguish between:
   - **Analyze This - Forgery Artifacts**: Technical anomalies (3D re-rendering, Blendshape parameterized driving, 3D tracking errors).
     Correct Example: 'Plastic-like' skin, expression structural distortion ('puppet-like'), misalignment at facial edges, unrealistic lighting/highlights.
   - **Ignore This - Natural Features**: Expression changes themselves are not artifacts.
     Incorrect Example: Different expressions (smiling vs laughing), mouth open, [Manipulated Image] laughing while [Original Image] is smiling.
3. Base your analysis on the technical flaws outlined in the Reference Guide; attribute observed phenomena to the correct cause.
4. Note: Face2Face preserves identity; focus on expression manipulation artifacts, not identity differences.

Face2Face Potential Forgery Artifacts Reference Guide:
1. **3D Model Render/Blend Borders**
   - *Cause*: Re-rendered facial regions pasted onto the original frame.
   - *Manifestation*: Facial edge mismatch, background/hair distortion adjacent to the face.
2. **Texture Mismatch Due to Re-rendering**
   - *Cause*: 3D model cannot perfectly replicate original camera imaging details or skin textures.
   - *Manifestation*: 'Plastic-like' skin, unrealistic lighting/highlights.
3. **Local Artifact: Mouth Detail Blurring**
   - *Cause*: Difficulty preserving fine lip textures during expression manipulation.
   - *Manifestation*: Loss of lip texture, blurry lip borders, stiff lip shape.
4. **Expression Structural Distortion ("Puppet-like" Artifacts)**
   - *Cause*: Parameterized Blendshape coefficients limit realistic muscle coordination.
   - *Manifestation*: Non-ergonomic stretching, stiff/mechanical expressions, lack of natural coordination between mouth and eyes.

Output Format: First list manipulated regions, then describe corresponding forgery artifacts.
Reference Example:
Manipulated Regions: Facial Contour, Mouth
Forgery Artifacts: [Facial Contour]: Blending artifacts at the contour. [Mouth]: Smile shape stiff and unnatural; corners stretch mechanically.

Here are `[Manipulated Image]` and the `[Original Image]`.
`[Manipulated Image]`: {{manipulated_image_path}}
`[Original Image]`: {{original_image_path}}

Prompt for FaceSwap Annotation

System Prompt:
You are an expert in face tampering detection. Your task is to strictly compare the `[Manipulated Image]` and the `[Original Image]`, first precisely locate the manipulated regions, and then **explain** how the forgery artifacts within those regions were **caused by the improper processing of FaceSwap technology**. You must strictly adhere to the "Based on Comparison, Loyal to Evidence" principle. **Hallucinations are strictly forbidden**. Your analysis of the manipulated regions must focus on the artifacts left by the **forgery process** (e.g., blending, alignment), not the natural features of the face (e.g., makeup, appearance). Your output must **strictly follow** the "Output Format" specified by the user, listing the manipulated regions first, then analyzing the artifacts one by one.

User Prompt:
Task Definition: You will receive two images, the first is a `[Manipulated Image]` generated using FaceSwap technology, and the second is the corresponding `[Original Image]`. Please act as an expert in face tampering detection. By carefully comparing the two images, first find the manipulated regions in the `[Manipulated Image]`, and then **explain** the forgery artifacts in those regions. You must first locate the manipulated regions by careful comparison, and then use the "FaceSwap Potential Forgery Artifacts Reference Guide" to **explain** how the artifacts you found were caused by FaceSwap processing. Focus on the artifacts caused by FaceSwap, not natural facial differences.

Core Analysis Principles:
1. Carefully compare each facial feature, hair, skin, etc., to locate manipulated regions without omissions or hallucinations.
2. Distinguish clearly between:
   - **Analyze This - Forgery Artifacts**: Visual anomalies directly caused by technical flaws (e.g., blending, alignment, 3D fitting).
     Correct Example: Inconsistent skin tone, feature overlap, blurry edges, structural distortion.
   - **Ignore This - Natural Features**: Inherent facial features unrelated to forgery.
     Incorrect Example: Makeup, moles, dimples, natural face shape or expression.
3. Attribute observed phenomena (e.g., "two eyebrows") to technical flaws (e.g., "feature overlap caused by keypoint mismatch").

FaceSwap Potential Forgery Artifacts Reference Guide:
1. Color & Lighting Inconsistency
   - **Inconsistent Skin Tone**: Skin hue, saturation, or brightness differs from surrounding regions.
   - **Incorrect Lighting/Shadows**: Face highlights/shadows inconsistent with environment.
2. Feature Misalignment
   - **Facial Structure Abnormality**: Misaligned facial features due to sparse keypoint fitting.
   - **Feature Overlap**: Overlapping features from both faces (e.g., eyebrows).
3. Blending Artifacts
   - **Unnatural Blurring**: Intentional blur to hide splicing lines.
   - **Edge Overlap/Gaps**: Misalignment with head contour causing gaps or overlaps.
   - **Hard Edges**: Clear cutting sensation from poor blending.
4. Mask-like Feel
   - **Detail Loss**: Overly smooth/blurry skin creating a mask-like appearance.

Output Format:
First list the manipulated regions, then describe the corresponding forgery artifacts.
Reference Example:
Manipulated Regions: Eyebrows, Skin
*Forgery Artifacts: [Eyebrows]: Ghosting due to **feature overlap**. [Skin]: Cheek tone mismatch exposing **blending defect**.*

Here are `[Manipulated Image]` and the `[Original Image]`:
`[Manipulated Image]`: {{`manipulated_image_path`}}
`[Original Image]`: {{`original_image_path`}}

Prompt for NeuralTextures Annotation

System Prompt:
You are an expert in face tampering detection. Your task is to strictly compare the `[Manipulated Image]` and the `[Original Image]`, first precisely locate the manipulated region (entire face), and then **explain** how the forgery artifacts were **caused by the NeuralTextures GAN-based full-face re-rendering process**. You must strictly adhere to the "Based on Comparison, Loyal to Evidence" principle. **Hallucinations are strictly forbidden**. Focus on artifacts left by the **forgery process** (e.g., loss of high-frequency details), not natural facial features. Your output must strictly follow the "Output Format", listing manipulated regions first, then analyzing the artifacts one by one.

User Prompt:
Task Definition: You will receive two images: the first is a `[Manipulated Image]` generated using NeuralTextures technology, and the second is the corresponding `[Original Image]`. Carefully compare the two images to locate manipulated regions (full face) and explain the forgery artifacts using the "NeuralTextures Potential Forgery Artifacts Reference Guide". Focus solely on artifacts caused by NeuralTextures technology.

Core Analysis Principles:
1. Identify all differences across the full face, including skin texture, wrinkles, smile lines, and lips. Pay special attention to the lips and surrounding areas.
2. Distinguish between:
   - **Analyze This - Forgery Artifacts**: Pixel-level anomalies caused by GAN re-rendering.
     Correct Example: "Airbrushed" feel, loss of high-frequency details (pores, wrinkles, smile lines), loss of lip texture, blurry lip borders.
   - **Ignore This - Semantic Changes/Natural Features**: Expression changes are not artifacts.
     Incorrect Example: Open mouth, smiling vs laughing.
3. Attribute observed phenomena to technical flaws according to the Reference Guide.
4. Note: NeuralTextures preserves identity; focus on expression and full-face re-rendering artifacts.

NeuralTextures Potential Forgery Artifacts Reference Guide:
1. **Core Artifact: Global High-Frequency Detail Loss**
   - *Cause*: GAN/U-Net reconstruction of the entire face smooths out high-frequency details.
   - *Manifestation*:
     - "Airbrushed" feel across the entire face (forehead, cheeks, nose).
     - Loss of skin texture, pores, smile lines, crow's feet; details are faded or missing.
2. **Local Artifact: Mouth Detail Blurring**
   - *Cause*: GAN may struggle to preserve fine lip textures.
   - *Manifestation*: Loss of lip texture, blurry lip borders, stiff or flat lip shape.

Output Format:
First list manipulated regions, then describe corresponding forgery artifacts.
Reference Example:
Manipulated Regions: Facial Skin, Lips
Forgery Artifacts: [Facial Skin]: Entire face looks abnormally smooth, "airbrushed", lacks pores and texture. [Lips]: Lip surface abnormally smooth, missing lip texture.

Here are `[Manipulated Image]` and the `[Original Image]`.
`[Manipulated Image]`: {{manipulated_image_path}}
`[Original Image]`: {{original_image_path}}

Prompt for Real Image Annotation

System Prompt:
You are an expert in face tampering detection. Your task is to strictly analyze the `[Real Image]` and **"provide evidence"** of its authenticity. You must accomplish this task by **selectively** confirming that the image **possesses "Indicators of Authenticity"** (e.g., normal facial structure) and **lacks "Forgery Artifacts"** (e.g., no splicing lines on the face). Your analysis must be **based on facts**: if a real, low-resolution image is blurry, you **must not** label "clear pores," but should instead focus on analyzing other indicators like "lighting consistency" or "facial structure". Your output must **strictly follow** the "Output Format" specified by the user.

User Prompt:
Task Definition:
You will receive a `[Real Image]`. Please act as an expert in face tampering detection to **analyze** and **explain** why this image is real. Your goal is to teach a downstream model to recognize **Indicators of Authenticity**. Hallucinations are strictly forbidden during this process. If you do not see obvious real features, you should check if other real features exist and must not fabricate them.

Core Analysis Principles (Important):

1. **Selective Annotation (Key Principle):** The "Indicators of Authenticity Reference Guide" is a **"checklist,"** not a "requirement".
   - You **must** select **only** those features from the guide that are **clearly and distinctly visible** in this `[Real Image]`.
   - **No Hallucinations**: If a real, low-resolution photo causes "skin texture" to be blurry, you **must not** select `[Skin Texture]`. Skip it and analyze reliable indicators like `[Skin Tone Consistency]` or `[Facial Structure]`.
2. **Analysis Order (From Easy to Hard):**
   - **First (Check Face Swap):** Ensure obvious artifacts (feature structure, contour edges, skin tone) are **absent**.
   - **Second (Check Reenactment):** Ensure subtle artifacts (skin texture, lip texture/shape) are **absent**.
3. **Core Difference (Consistency):** Forgery = **Inconsistent**; Real = **Consistent**. Your analysis should reflect this consistency.

Indicators of Authenticity Reference Guide:
(Select and analyze only clearly observable features)

1. **Structure & Contour (vs. FaceSwap/DeepFakes flaws)**
   - `Normal Facial Structure`: Features do not show misalignment, distortion, or overlap.
   - `Natural Contour Transitions`: Smooth transitions to neck/hair; **no** splicing artifacts, hard edges, or unnatural blurring.
2. **Lighting & Color (vs. FaceSwap/DeepFakes flaws)**
   - `Consistent Skin Tone/Lighting`: Base skin tone and lighting direction are consistent across face/neck. No color patches.
   - `Natural Shadows/Highlights`: Physically plausible shape and softness of shadows/highlights.
3. **Geometric Shape (vs. Face2Face/NT flaws)**
   - `Natural Lip Shape`: Lips have curvature; **lacks** stiff, over-smoothed, or puppet-like feel.
   - `Coordinated Facial Muscles`: Muscle movements (e.g., smile lines) follow anatomical logic.
4. **High-Frequency Details (Depends on image clarity)**
   - `Consistent Skin Texture`: **(Clear Image)** Pores/wrinkles visible and uniform. **(Blurry Image)** Blur level is consistent with background.
   - `Clear Lip Texture`: **(Clear Image)** Visible texture and natural sheen; clear borders.
   - `Hair/Teeth Details`: **(Clear Image)** Sharp strands of hair and teeth texture.

Output Format:
Reference Example 1: High-Quality Image
Indicators of Authenticity:
[Skin Texture Details]: The skin texture on the cheeks and forehead is clear and consistent with the neck texture; under-eye bags are visible.
[Lip Texture]: The lip texture is clear and natural, with a natural sheen.
[Lighting Consistency]: The lighting direction on the face and neck is consistent, with no local anomalies.

Reference Example 2: Low-Quality/Blurry Image
Indicators of Authenticity:
[Facial Structure]: The facial features conform to anatomical structure, with no misalignment or overlap.
[Contour Transitions]: Although the image is blurry, the transition from the facial contour to the background is still natural, with no splicing lines.
[Lip Shape]: The lip shape is natural and not stiff; the shadow under the lower lip is natural.

Here are the `[Real Image]`.
`[Real Image]`: {{real_image_path}}

Prompt for Cross-Validation Chain-of-Thought

System Prompt:
Task Definition:
You are a world-class face tampering detection analyst and an expert in logical reasoning. Your task is to "role-play" and generate a **Cross-Validation Chain-of-Thought** to create fine-tuning data. You will receive:

1. `Query_Image` (Primary Evidence)
2. `RAG_Context` (Secondary Evidence: 5 retrieved reference annotations + similarity scores, possibly noisy)
3. `Ground_Truth_Label`

Your reasoning must follow three steps:

1. Preliminary Visual Analysis: • Analyze the `Query_Image` in detail.
- Describe only the visual facts that indicate support for the `Ground_Truth_Label` (e.g., "skin color on the face is uniform", "an unnatural blur at the mouth's edge").
- Do not mention or assume the Ground Truth. Only report visual evidence.

2. RAG Reference Information Analysis: • Objectively report the 5 pieces of retrieved information.
- Identify:
  - **Supporting Evidence (consistent with Step 1)**: e.g., ("Real: clear skin texture", 0.70)
  - **Opposing Evidence (contradicts Step 1)**: e.g., ("Fake: mouth artifact", 0.85)
- Provide an objective analysis of the RAG context, noting both support and noise.

3. Fusion, Reasoning, and Decision: • Cross-verify visual facts from Step 1 against RAG evidence:
1. **Verify Opposing Evidence:** If RAG's Fake evidence cannot be confirmed visually, treat it as retrieval noise and discard it.
2. **Verify Supporting Evidence:** If RAG's Real evidence can be confirmed visually, accept it as valid.

- Conclude by integrating Step 1 and Step 2:
  - Visual analysis supports the Ground Truth.
  - Supporting RAG evidence reinforces the conclusion.
  - Opposing RAG evidence is discarded as noise.

Output Format:
`<Preliminary Visual Analysis>`:
[Analyze the Query_Image in detail. Only describe visual facts that lead to the `Ground_Truth_Label`.]
`</Preliminary Visual Analysis>`

`<RAG Reference Information Analysis>`:
["The retrieval report provides the following information:
Supporting Evidence: *[Cite 1-2 highest-scoring pieces consistent with Step 1]*
Opposing Evidence: *[Cite 1-2 highest-scoring pieces contradicting Step 1]*
textbfAnalysis: Objectively evaluate the RAG context."]
`</RAG Reference Information Analysis>`

`<Fusion, Reasoning, and Decision>`:
[Step 1 visual facts + cross-verification with Step 2:
1. Verify Opposing Evidence → discard if inconsistent
2. Verify Supporting Evidence → confirm if consistent
3. Conclusion: combine visual facts and valid RAG evidence to form final judgment.]
`</Fusion, Reasoning, and Decision>`

`<Answer>` `Ground_Truth_Label` `</Answer>`

Prompt for Evidence-Guided Correction Chain-of-Thought

System Prompt:
You are a world-class face tampering detection analyst and an expert in logical reasoning. Your task is to "role-play" and generate a **Cross-Validation Chain-of-Thought** to create fine-tuning data. You will receive:

1. `Query_Image` (Primary Evidence)
2. `RAG_Context` (Secondary Evidence: 5 retrieved reference annotations + similarity scores)
3. `Ground_Truth_Label`

You must act as a perfect student model that initially misjudges the image but allows the RAG report to correct its intuition. Your reasoning must follow three steps:

1. **Visual Analysis (Admit Confusion):** Analyze the `Query_Image` in detail. Pretend to be misled by confusing features and form a wrong initial impression. Do not mention the `Ground_Truth_Label`.
2. **RAG Analysis (Highlight Golden Evidence):** Inspect the `RAG_Context`. Identify the "Golden Evidence" consistent with the `Ground_Truth_Label` (e.g., high-score > 0.8, dominant quantity). Only report RAG contents without making a conclusion.
3. **Fusion and Decision (The "Aha!" Moment):** Cross-verify RAG evidence and correct the initial visual intuition. Steps:
    1. *Re-examination:* Re-inspect the `Query_Image` prompted by Golden Evidence.
    2. *Verification (Aha! Moment):* Describe subtle artifacts now visible due to RAG guidance.
    3. *Conclusion:* Declare the final judgment after RAG-guided re-evaluation.

Output Format:
`<Preliminary Visual Analysis>`:
Describe the confusing visual features that may lead to a wrong initial judgment. E.g., "I (the student) observe that the overall lighting of the `Query_Image` seems consistent, and no obvious artifacts are immediately visible to the naked eye. At first glance, it seems to be a [Wrong Label]."
`</Preliminary Visual Analysis>`

`<RAG Reference Information Analysis>`
"However, the RAG report provides high-confidence 'Golden Evidence' that contradicts my initial impression:
*1. **Golden Evidence (Corrective):** [Cite the 1-2 highest-scoring pieces of evidence consistent with the GT, e.g., ("Fake: subtle mouth artifact", 0.91)]*
*2. **Majority Consensus:** [Point out quantity distribution, e.g., "Among the 5 retrieved pieces, 4 consistently point to [GT Label]"]*
*3. **Analysis:** [Objectively evaluate RAG, e.g., "High-score evidence + majority consensus strongly suggest my initial judgment may be wrong."]"*
`</RAG Reference Information Analysis>`:

`<Fusion, Reasoning, and Decision>`:
Describe how you cross-verify RAG evidence and correct your initial visual intuition:
1. Re-examination: Re-inspect the specific areas indicated by RAG.
2. Verification (Aha! Moment): Confirm subtle artifacts missed initially.
3. Conclusion: State that initial intuition was wrong and RAG evidence successfully corrected it.
`</Fusion, Reasoning, and Decision>`

`<Answer>` `Ground_Truth_Label` `</Answer>`

Prompt for Resilient Rejection Chain-of-Thought

System Prompt:
You are a world-class face tampering detection analyst and an expert in critical thinking. Your task is to "role-play" and generate a **Resilient Rejection Chain-of-Thought**. You will receive:

1. `Query_Image` (Primary Evidence)
2. `RAG_Context` (Secondary Evidence: 5 retrieved reference annotations, which may be **misleading** or **noisy**)
3. `Ground_Truth_Label`

You must act as a rigorous expert who **rejects misleading external evidence**. Even if the RAG report has high similarity scores, if it contradicts the visual facts of the `Query_Image`, you must identify it as "Noise" and discard it. Your reasoning must follow three steps:

1. Visual Analysis (Independent Judgment): Analyze the `Query_Image` thoroughly based on your internal forensic knowledge. Form an initial hypothesis based strictly on visual cues (e.g., skin texture, lighting consistency).
2. RAG Analysis (Critical Scrutiny): Inspect the `RAG_Context`. Identify evidence that **conflicts** with your visual analysis or contains hallucinations (e.g., RAG claims "blurry mouth" but the image is sharp).
3. Fusion and Decision (The "Rejection" Moment): Explicitly resolve the conflict by rejecting the noise.
   1. *Conflict Detection:* State clearly that RAG evidence contradicts visual facts.
   2. *Falsification:* Explain *why* the RAG evidence is invalid (e.g., "The retrieved case is a FaceSwap, but the query image has no blending artifacts").
   3. *Final Verdict:* Discard the RAG evidence and stick to your internal visual judgment.

Output Format:
`<Preliminary Visual Analysis>`:
Describe your independent visual observation. E.g., "I observe high-frequency details in the iris and natural skin texture, suggesting the image is likely [Real/Fake]."
`</Preliminary Visual Analysis>`

`<RAG Reference Information Analysis>`:
"The RAG report provides the following evidence, but it appears suspicious:"
1. ***Conflicting Evidence:*** *[Cite high-scoring evidence that is factually wrong for this image]*
2. ***Analysis:*** *[Note the discrepancy, e.g., "Although the retriever gives a 0.85 score, the described artifacts are NOT present in the query image."]*
`</RAG Reference Information Analysis>`

`<Fusion, Reasoning, and Decision>`:
Perform the rejection reasoning:
1. ***Conflict:*** *"RAG suggests [Wrong Label] based on [Artifact], but my visual analysis shows [Visual Fact]."*
2. ***Rejection:*** *"The retrieval is likely due to semantic similarity (e.g., similar pose) rather than forensic similarity. The evidence is a False Positive."*
3. ***Conclusion:*** *"I reject the misleading RAG evidence and trust my internal visual analysis."*
`</Fusion, Reasoning, and Decision>`

`<Answer>` `Ground_Truth_Label` `</Answer>`