Explainable Deepfake Detection with Feature-robust Augmentation and Evidence-grounded Explanation Optimization

Zhu Xu
Wangxuan Institute of Computer Technology, Peking University
Beijing, China
xuzhu@stu.pku.edu.cn

Jiaqi Tang
Wangxuan Institute of Computer Technology, Peking University
Beijing, China
2601213445@stu.pku.edu.cn

Pokai Chen
Wangxuan Institute of Computer Technology, Peking University
Beijing, China
2501213417@stu.pku.edu.cn

Yuxin Peng
Wangxuan Institute of Computer Technology, Peking University
Beijing, China
pengyuxin@pku.edu.cn

Yang Liu*
Wangxuan Institute of Computer Technology, Peking University
Beijing, China
yangliu@pku.edu.cn

Abstract

Explainable deepfake detection extends binary classification by requiring models to not only predict authenticity but also provide interpretable justifications. This expanded scope is critical in practice, where users like forensic analysts need insight into the rationale behind the detection. Despite advancements, current approaches suffer from two critical deficiencies: (1)*vulnerability to image quality degradation*: detection accuracy plummets on low-quality samples, while naive augmentation strategies may induce feature drift and impair performance as diversity expands. (2) *factually flawed explanations*: explanation models may omit manipulation evidence or hallucinate irrelevant details, undermining interpretability. To address it, we propose a framework with two innovations. For robust deepfake detection, we introduce *Feature-robust Augmentation*, which comprises diversified degradation-aware augmentation strategies, and a supervised contrastive learning pattern paired with a mean-teacher architecture that stabilizes features against augmentations through consistency constraints. For explanation, we devise an *evidence-grounded preference optimization* process that guides model to prioritize genuine manipulation traces by learning from chosen-rejected explanation pairs, where rejected samples are constructed via evidence omission or irrelevant information injection. The proposed approach wins the first place in ACM Multimedia 2026 Explainable Deepfake Detection Challenge. The code is available at https://github.com/oceanflowlab/EDD.git

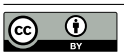
CCS Concepts

• **Computing methodologies** → **Artificial intelligence**.

Keywords

Deepfake Detection, Visual Reasoning, Reinforcement Learning

*Corresponding author



MM '26, Rio de Janeiro, Brazil.

ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3838608

ACM Reference Format:
Zhu Xu, Jiaqi Tang, Pokai Chen, Yuxin Peng, and Yang Liu. 2026. Explainable Deepfake Detection with Feature-robust Augmentation and Evidence-grounded Explanation Optimization. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3767308.3838608

1 Introduction

The rapid advancement of generative AI has made image manipulation increasingly realistic and challenging to verify, posing serious risks to digital media integrity and forensic investigations [4, 7, 11, 22, 26, 32, 34]. Although the majority of existing deepfake detection formulates this as a binary decision task, i.e., classifying an image as real or fake, such formulation fails to capture the practical demands. In practice, a label alone is seldom sufficient: fact-checkers and forensic experts must comprehend the specific visual evidence that renders an image suspect. Consequently, the field has increasingly recognized the necessity of explainable detection, wherein systems are expected to both make accurate predictions and articulate the visual cues underpinning those decisions, e.g., inconsistent object boundaries, implausible geometry, etc. By providing human-interpretable reasoning, such systems can facilitate more effective forensic scrutiny and foster greater user confidence in automated detection tools.

Despite advancements [7, 11, 13, 22] in explainable deepfake detection, existing methods face two fundamental challenges that severely undermine their real-world applicability. (1)*the brittleness of detection models under image quality degradation.* As shown in Figure 1a, we adopt BRISQUE[6] as the metric to measure the image quality of the validation set of XPlainVerse Challenge dataset, partitioning the samples into three subsets with high, medium and low image quality. We then evaluate detection performance on them. Performance on baseline(DINOv3[28] + classification head trained without any data augmentation, red bar) shows a decrease as the image quality drops, which confirms that quality degradation indeed compromises discriminative capability. A straightforward remedy, i.e., applying data augmentations, often backfires due to feature drift, where the model struggles to learn a consistent representation across augmented views. In Figure 1b,

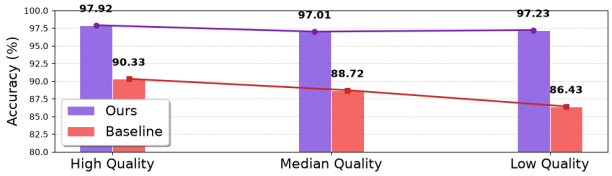
as the number of augmentation strategies for training increases, the test accuracy of baselines(red color) initially rises but subsequently declines, indicating that excessively diverse augmentations blur classification boundaries and confuse the model. (2)*inadequate evidence-verification of existing explanation mechanisms.* Prevailing approaches predominantly focus on the procedural format of reasoning, i.e., mandating multi-step thinking or sufficient response length, while neglecting the factual accuracy and completeness of the evidence presented within the rationale. We evaluate explanations of the baseline Qwen3-VL-8B-Instruct[36] on a 1k-subset of validation set, observing that two prevalent failure modes emerge: (i) *evidence omission*, where crucial manipulation traces are missing from the explanation, which exist in 68.3% samples, and (ii) *irrelevant information intrusion*, where spurious details are hallucinated that exist in 57.7% samples. These deficiencies not only mislead users but also erode the trust that explainability aims to establish.

To address these challenges, we propose a two-stage framework with two components. For detection robustness, we introduce a *Feature-robust augmentation* strategy, which first uses a degradation-aware augmentation pipeline that augments input images with diverse distortions at varying intensities. To complement this, we impose a supervised contrastive loss that pulls features of the same authenticity class(real or fake) into more compact clusters, enhancing discriminability across quality variations. However, contrastive learning operates only at the class level without preventing intra-instance feature drift. We therefore introduce a mean-teacher architecture that maintains a stable feature anchor via exponential moving average of the student's weights. The teacher aggregates information across consecutive training steps, suppressing noise and providing a consistent reference. A consistency constraint aligns all augmented student features toward this anchor, preventing intra-instance drift while preserving the benefits of augmentation diversity. For deepfake explanation, we devise an *Evidence-grounded Explanation Optimization* to guide the model via reinforcement learning. By constructing chosen-rejected explanation pairs where chosen explanations are complete and accurate, while rejected ones omit key evidence or include fabrications, we fine-tune the model to prioritize genuine manipulation evidence over superficial reasoning patterns via direct preference optimization(DPO)[25], enhancing the evidence accuracy and completeness of explanations.

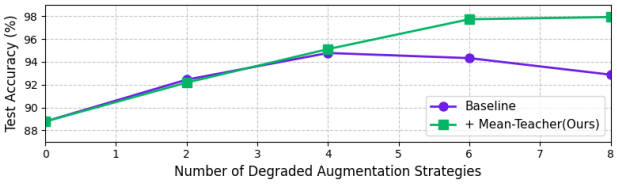
Beyond detailed evidence generation, we further develop a concise explanation model tailored for scenarios where brevity is prioritized. This model is optimized via GRPO[27] algorithm, with rewards jointly designed on semantic fidelity and conciseness, enabling it to deliver only the most critical manipulation cues. Combined with the detailed evidence explanation model, our framework flexibly adapts to diverse usage scenarios from in-depth forensic analysis to quick public fact-checking. Evaluated on the 200k XPlainVerse[19] test set, our model achieves state-of-the-art overall performance in ACM MM 2026 Explainable Deepfake Detection Challenge[20]. Specifically, we achieve the best semantic fidelity for evidence-grounded explanation, the optimal semantic fidelity and conciseness balance for concise explanation.

The contributions of this paper are summarized as follows:

- We identify and empirically validate two critical vulnerabilities in existing explainable deepfake detection systems, i.e., sensitivity to image quality and lack of evidence accuracy verification.
- We propose a framework that jointly enhances detection robustness via degradation-aware augmentation with mean-teacher stabilization, and improves explanation faithfulness through evidence-grounded preference optimization.
- Our framework shows state-of-the-art overall performance, ranks first for ACM MM 2026 Explainable Deepfake Detection Challenge[20], offers a reliable solution for explainable deepfake detection.



(a) Detection accuracy on samples with different image quality.



(b) Performance change under increasing augmentations.

Figure 1: Illustration of analysis for detection robustness. (a) Performance drops under quality degradation. (b) Mean-teacher stabilizes learning under diverse augmentations while the baseline suffers from feature drift.

2 Related Work

2.1 Robust Deepfake Detection

Recent advances have explored various strategies to improve robustness against unseen forgeries. To diminish the impact of image quality on detector performance, several works [5, 9, 16, 24, 30, 38, 41] attempt to model different degradation strategies. Though promising, such naive data augmentation may lead to feature drift, causing classification boundary to blur as augmentation diversity expands. Our method addresses such limitations by incorporating a mean-teacher architecture that maintains a stable feature anchor, enforcing consistency constraints across all augmented views to prevent feature drift while preserve the benefits of augmentation diversity.

2.2 Vision-Language Models for Explainable Deepfake Detection

Recent advancements in Vision-Language Models (VLMs) [1, 15, 43] adopted a paradigm of supervised fine-tuning followed by reinforcement learning, which has proven effective in boosting reasoning capabilities across diverse applications, such as image[2, 23, 35], video[21, 40] and 3D[17, 37, 42] understanding tasks. The rapid development of VLMs has also opened new possibilities for explainable deepfake detection. For example, ForgerySleuth [29] employs

a trace encoder to generate detailed tampering analyses. And more recent works[31] propose datasets annotated by humans to generate explanations further aligned with human preference. However, the post-training focuses on the format of reasoning, i.e, requiring multi-step thinking or sufficient response length, while neglecting the factual accuracy and completeness of evidence. Different from them, we identify the critical issue of evidence inaccuracy, termed evidence omission and irrelevant information intrusion, and address it through an evidence-grounded preference optimization process. By constructing chosen-rejected explanation pairs, we fine-tune the model to prioritize genuine manipulation evidence over superficial reasoning patterns.

3 Method

3.1 Overview

Our goal is to build a deepfake detection system that is both robust against real-world quality variations and capable of providing faithful, evidence-grounded explanations. To this end, we propose a unified framework comprising two synergistic components, as illustrated in Figure 2. The first component targets deepfake detection. Given an input image, the visual backbone extracts features and passes through a classifier for binary classification(1 for fake and 0 for real). To ensure robustness under diverse quality conditions, we incorporate a degradation-aware augmentation pipeline during training, paired with a mean-teacher architecture that anchors augmented representations toward a stable reference, preventing feature drift. We also incorporate a supervised contrastive loss to make features more distinguishable. Once a detection decision is made, the second component generates an explanation to justify its reasoning. For detailed forensic analysis, we employ an MLLM for the reasoning. We propose an evidence-grounded explanation optimization process, which explicitly teaches the model to prioritize genuine manipulation evidence over potential evidence omissions or hallucinations via DPO algorithm training, where we construct different sub-optimal responses to simulate potential problems in model reasoning. For scenarios where brevity is prioritized, we additionally provide a concise explanation model optimized via tailored reward functions. Together, these components deliver a detector that is not only accurate across diverse conditions but also transparent in its decision-making.

3.2 Robust Deepfake Detection via Feature-robust Augmentation

3.2.1 Feature Extraction and Aggregation. We employ DINOv3-7B [28] as our visual encoder. To fully exploit both global and local information, we aggregate multi-granularity tokens. Specifically, we extract CLS token $\in \mathbb{R}^{1*d}$, REG token $\in \mathbb{R}^{4*d}$ with global semantics, AVG patch tokens $\in \mathbb{R}^{1*d}$ preserve fine-grained details, where d is feature dimension. These tokens are concatenated and subsequently passed through an MLP head for classification, supervised by a binary cross-entropy loss $\mathcal{L}_{\text{BCE}}$. This design ensures that the detector leverages both high-level contextual cues and fine-grained manipulation traces simultaneously.

3.2.2 Degradation-Aware Augmentation Pipeline. To build robustness against quality degradations, we design a augmentation pipeline that exposes model to diverse distortion types. Table 1 summarizes our strategies and intensity sampling. Distractors (random text or images) are additionally applied to simulate real-world overlays. All degradations are applied in random probability and order, ensuring that the model encounters a wide spectrum of quality conditions during training.

Table 1: Summary of degradation strategies and intensities.

Degradation	Sampling Intensities and Conditions
Smoothing	3 kernel types: anisotropic Gaussian ($\sigma \sim \mathcal{U}(0, 30)$), fspecial Gaussian ($\sigma \sim \mathcal{U}(0, 30)$), uniform square ($w \sim \mathcal{U}(3, 30)$)
Resize	Upsample $\sim \mathcal{U}(1, 2)$, downsample $\sim \mathcal{U}(0.5, 1)$, interpolation: linear/cubic/area
Gaussian Noise	3 types: per-channel, grayscale, correlated ; two regimes: $\sigma \sim \mathcal{U}(2, 100)$ and $\sigma \sim \mathcal{U}(80, 100)$
Non-Gaussian Noise	Speckle ($\sigma \sim \mathcal{U}(2, 25)$)/ Poisson
JPEG Compression	Quality factor $\sim \mathcal{U}(10, 95)$
Color Enhance	Brightness or contrast factor $\sim \mathcal{U}(0.5, 1.5)$
Distractors	Random text overlay or image patch overlay

To enhance robustness against degradation variations and explicitly enforce the features more authenticity-aware, i.e., discriminative for real and fake prediction, we impose a supervised contrastive loss, as naively applying augmentations to expand training diversity can paradoxically induce feature drift, causing representations of augmented samples to deviate from their original distributions and impair discriminative performance. Supervised contrastive loss handles this by pulling features of the same authenticity class into compact clusters regardless of their degradation condition, while pushing features of different class apart. For a given image x_i with label y_i, we generate a degraded view $x_i^d = t(x_i)$ where t is sampled from our augmentation pipeline. The loss pulls features of the same class together while pushing features of different classes apart:

$$\mathcal{L}_{\text{supcon}} = \sum_{i \in \mathcal{B}} \frac{-1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \log \frac{\exp(\mathbf{x}_i \cdot \mathbf{x}_p / \tau)}{\sum_{a \in \mathcal{A}(i)} \exp(\mathbf{x}_i \cdot \mathbf{x}_a / \tau)}, \quad (1)$$

where $\mathcal{B}$ is the current batch, $\mathcal{P}(i)$ denotes the set of positive samples sharing the same label as i (including both original and degraded views), $\mathcal{A}(i)$ is the set of all anchors excluding i, and τ is a temperature parameter.

3.2.3 Mean-Teacher Stabilization. While supervised contrastive loss effectively enforces authenticity-aware feature clustering, it operates solely at the class level to ensure real and fake samples are more distinguishable without constraining how different augmented views of the same sample relate to each other beyond pulling them toward the same class centroid. As degradation diversity increases, this leaves room for intra-instance feature drift: different views of the same image may scatter widely within the same class cluster, undermining prediction stability and consistency. A potential remedy would be to impose a consistency loss between different augmented views of the same sample. However, this approach lacks a stable reference: when both views are independently perturbed, the model may simply enforce agreement between two noisy predictions without anchoring to a meaningful representation, causing features to drift arbitrarily. To complement this, we introduce a mean-teacher architecture that directly constrains each sample's feature across augmentations toward a stable reference. As proved by prior works[18, 33], exponential moving average (EMA) aggregates information across consecutive model states, producing

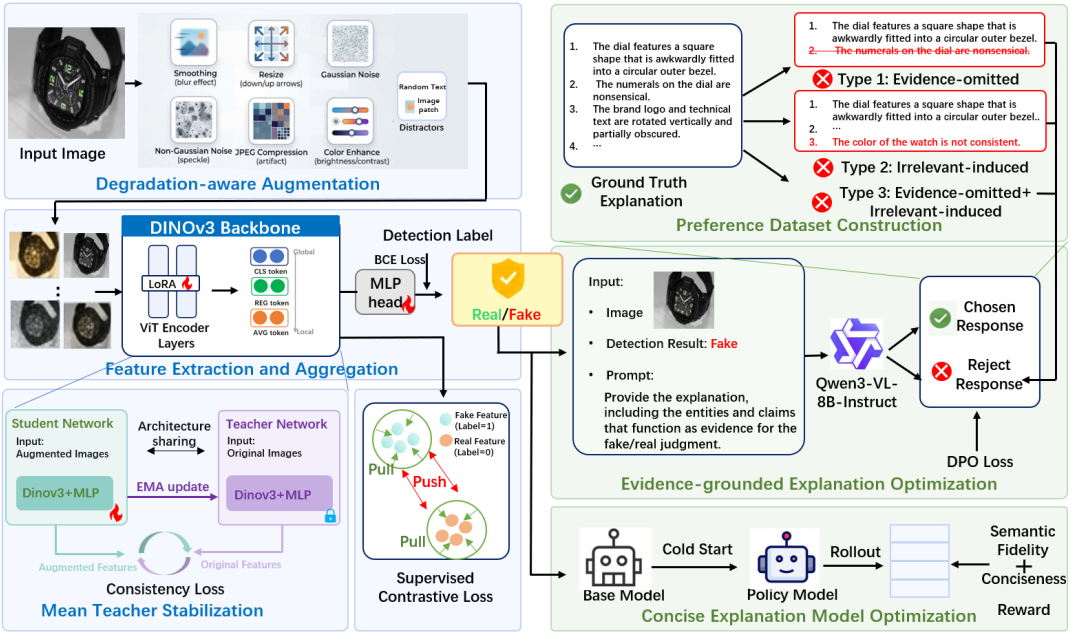


Figure 2: Overview of our proposed framework. Given an input image, the visual backbone extracts features and fed into a classifier for prediction. During training, the detection branch incorporates degradation-aware augmentations, supervised contrastive learning, and mean-teacher consistency regularization to enhance robustness against quality degradation and feature drift. For explanation generation, the image and deepfake label are sent to the MLLM for generation. To enhance evidence accuracy and completeness, we propose an evidence-grounded optimization process, which generates evidence-omitted and irrelevant information induced explanation to construct rejected samples, guiding the model via the DPO algorithm. We also provide a concise explanation model that is trained via GRPO with tailored semantic fidelity and conciseness rewards.

more accurate and consistent predictions by canceling out noise while preserving the signal [14]. The teacher network θ_{tea}, updated as an EMA of the student network θ_{stu}, inherently suppresses stochastic degradation noises to provide stable sample features. For each image x_i, we feed clean view and all degraded views into student network, while the teacher receives only clean view. A consistency loss enforces alignment between the student's features from degraded views and the teacher's feature from the clean view:

$$\mathcal{L}_{\text{consis}} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \|\mathbf{z}_{\text{stu}}(t(x_i)) - \mathbf{z}_{\text{tea}}(x_i)\|_2^2. \tag{2}$$

where $\mathcal{T}$ is the set of all augmentations. This alignment constraint anchors all augmented representations toward a stable reference, preventing feature drift while preserving the benefits of augmentation diversity. The overall loss is:

$$\mathcal{L}_{\text{det}} = \mathcal{L}_{\text{BCE}} + \lambda_1 \mathcal{L}_{\text{supcon}} + \lambda_2 \mathcal{L}_{\text{consis}}. \tag{3}$$

where λ_1 and λ_2 are coefficients to balance the loss terms.

3.3 Evidence-Grounded Explanation Optimization

Beyond accurate detection, our system is required to provide rationales that faithfully reflect the model's decision process. Our empirical findings reveal that existing explanation models frequently suffer from two critical issues: evidence omission, i.e., failing to mention crucial manipulation traces, and irrelevant information intrusion, i.e., hallucinating spurious details. While supervised fine-tuning (SFT) enables the MLLM to generate plausible explanations by mimicking ground-truth rationales, it does not explicitly discourage such factual errors. To bridge this gap, we propose an evidence-grounded preference optimization based on reinforcement learning that explicitly teaches the model to prioritize genuine manipulation evidence over superficial reasoning patterns.

3.3.1 Preference Dataset Construction. We construct a preference dataset $\mathcal{D}_{\text{pref}}$ consisting of chosen-rejected explanation pairs. For each training sample with ground-truth explanation e^* that accurately describes the specific manipulation traces present in the image, we use advanced MLLM Qwen3.5-27B[36] to construct three types of rejected explanations based on the ground-truth: (1) *Evidence-Omitted*: We instruct MLLM to remove key manipulation traces from e^*, producing an explanation that lacks certain critical evidence. (2) *Irrelevant-induced*: We inject content that exists within image but is irrelevant to the deepfake judgement into e^*, simulating hallucinated evidence. (3) *Combined above:* We further construct rejected explanations that not only lack critical evidence, but also injected with redundant information.

For each sample with the e^* as the chosen explanation, we pair a rejected explanation e^-, forming preference pairs (e^+, e^-) where e^+ indicates the preferred explanation. Then using $\mathcal{D}_{\text{pref}}$, we further fine-tune the SFT model using Direct Preference Optimization

(DPO) [25], which optimizes the policy π_θ to favor chosen explanations over rejected ones. Given a prompt q (sample image with a prompt asking for detection rationale), the DPO loss is:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(q,e^+,e^-)\sim\mathcal{D}_{\text{pref}}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(e^+\mid q)}{\pi_{\text{ref}}(e^+\mid q)} - \beta\log\frac{\pi_\theta(e^-\mid q)}{\pi_{\text{ref}}(e^-\mid q)}\right)\right] \quad (4)$$

where π_{ref} is a reference model (the SFT checkpoint), and β controls the deviation from the reference. This objective implicitly rewards explanations that are factually complete and penalizes those that omit evidence or include hallucinations, effectively aligning the model's reasoning with evidence-grounded manipulation traces.

3.3.2 Concise Explanation Model. While detailed evidence explanations are valuable for forensic analysis, many real-world scenarios, such as social media fact-checking or real-time alerts, require brief, digestible outputs. To accommodate such use cases, we additionally develop a concise explanation model that describes the most critical manipulation cues into compact statements. This model is optimized via Group Relative Policy Optimization (GRPO) [27], with reward functions regarding: (1) *semantic fidelity*, which adopts BERTScore[39] to measure the semantic similarity between the rollout and reference explanation; and (2) *conciseness*, which adopts SLEscore[3] to penalize verbosity. The GRPO training encourages the model to find an optimal balance between overall reward, producing explanations that are both faithful and concise.

4 Dataset Overview.

The challenge uses XPlainVerse dataset[19], which is sourced from MultifakeVerse[8] dataset. XPlainVerse is the largest Explainable Deepfake Detection Dataset containing 1 million images, including both authentic and manipulated images paired with reference explanations that ground the evaluation of model-generated justifications. This challenge contained a subset of XPlainVerse, separately contains 450k, 110k, 200k data in train, validation and test set.

5 Experiments

Evaluation Metrics. Deepfake detection is evaluated by detection accuracy and macro F1 score. Complex explanations are evaluated by BERTscore[12], along with Entity and Claim F1 that measures the semantic faithfulness in explanations. Simple explanations are evaluated with BERTscore along with SLE score[3], which measures the simplicity of prediction. The overall score is the weighted sum of all metrics. Refer to Challenge paper[20] for more metric details.

5.1 Our experimental setup

For the detection, we employ DINOv3 [28] as the visual encoder, fine-tuned via LoRA [10](rank=16, α=16). We train it for 10 epochs with a batch size of 64 and an initial learning rate of 1×10^{-4}. λ_1/λ_2 is 0.5/0.2. τ is 0.2, β is 0.5, EMA rate is 0.99. For explanation model, we adopt Qwen3-VL-8B-Instruct [36] as base model and fine-tuned with LoRA (rank=32, α=32). The model first undergoes supervised fine-tuning on the ground-truth explanation annotations for 3 epochs, followed by reinforcement learning with DPO or GRPO algorithm for the evidence-grounded and concise explanation models, respectively. All experiments are conducted on 8 HUAWEI Ascend 910B NPUs.

Table 2: Public leaderboard Results of Top-5 Teams. Qwen3-VL-8B/InternVL3.5-13B are challenge[20] baselines.

Team name	Overall Score	Detection F1	Complex BERT	Simple Overall	Explanation Score	Rank
Pixel Slueth(Ours)	**0.841617**	0.942399	**0.706168**	**0.775501**	**0.740834**	**1**
Team Antvengers	0.838154	**0.947941**	0.696179	0.760555	0.728367	2
MSUteam	0.8312	0.933982	0.70113	0.755707	0.728419	3
HIT VIRLAB	0.822242	0.926326	0.698713	0.737602	0.718157	4
Team1	0.814503	0.916683	0.696807	0.727839	0.712323	5
Qwen3-VL-8B[20]	0.642776	0.634234	0.664774	0.637863	0.651318	-
InternVL3.5-13B[20]	0.637566	0.625662	0.661781	0.637161	0.649471	-

Table 3: Final hidden-test results for Top-5 team.

Team	Detection F1	Complex Bert	Simple Bert	SLE Score	Entity F1	Claim F1	Explain Score	Overall Score	Rank
Pixel Sleuth(Ours)	0.9424	**0.7063**	**0.6819**	**0.9782**	**0.5280**	**0.4205**	**0.5800**	**0.7612**	**1**
Antvengers	**0.9479**	0.6958	0.6699	0.9720	0.5078	0.3940	0.5618	0.7549	2
MSUteam	0.9340	0.7004	0.6509	0.9726	0.4987	0.3932	0.5571	0.7456	3
Team1	0.9167	0.6959	0.6321	0.9335	0.5155	0.4021	0.5590	0.7378	4
HIT VIRLAB	0.9263	0.6988	0.6493	0.9287	0.4467	0.3436	0.5235	0.7249	5

5.2 Challenge Results

We present the public leaderboard results[1] on the Challenge test set in Table 2, and the final hidden-test results(where Entity and Claim F1 Score are reported on a 10k hidden evaluation subset) in Table.3, comparing with other top-tier teams. Our method ranks first for the overall score among 138 participating teams with 414 submission entries in total. We yield the highest score for all explanation metrics, showcasing that we own (1) most accurate complex explanation, thanks to our evidence-grounded learning to guide the model to focus on the accurate forensic traces for more accurate and complete reasoning; (2) most concise and semantic correct simple explanation, which validates that our curated rewards can find an optimal balance between semantic fidelity and expression conciseness. As a result, our method yields the best overall score, showing detection robustness, explanation interpretability.

5.3 Qualitative Results.

Robustness on Deepfake Detection. As shown in Figure 4, the baseline model exhibits error predictions when confronted with quality-degraded samples. In contrast, our model maintains accurate predictions across different degradation types. The degradation-aware contrastive learning exposes the model to diverse degradation patterns during training, while the mean-teacher architecture ensures that augmented features are aligned toward a stable feature anchor, preventing feature drift. As a result, our model learns quality-invariant representations for deepfake detection.

Evidence-accuracy on Explanation. As shown in Figure 3, the baseline's explanation is dominated by irrelevant information, indicating that without proper guidance, the MLLM tends to generate superficially plausible but factually ungrounded rationales. The SFT-only model, while benefiting from ground-truth supervision, achieves only partial correctness, as it correctly identifies some manipulation traces (e.g., unnatural text on the banner and floating contact) but still omits critical evidence such as the unnatural high-contrast of ball, and occasionally hallucinates non-existent artifacts like loss of detail on player's limbs. This confirms our empirical finding that SFT alone is insufficient to ensure complete and

[1]The challenge website is https://explainable-deepfake-detection.github.io, and Public leaderboard is shown in https://www.codabench.org/competitions/16461/#/results-tab.

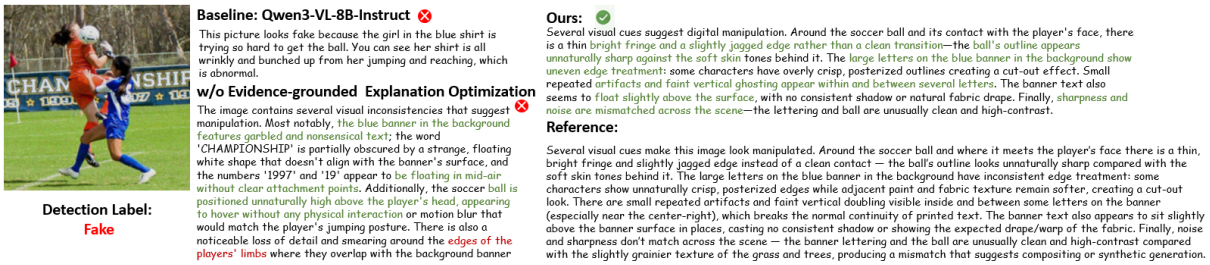


Figure 3: Visualizations for deepfake explanation. Result of baseline[36] shows irrelevant information, fails to identify valid evidence. SFT-only model captures partial evidence (marked in green) but still suffers irrelevant details(marked in red). Our model generates complete and accurate explanation(marked by green) without introducing spurious information.

Figure 4: Qualitative comparison of detection robustness under samples with various image degradations. The baseline model misclassifies degraded samples, while our model shows correct predictions across different degradation types.

faithful evidence grounding. In contrast, our model produces a concise yet comprehensive explanation that accurately enumerates all detectable manipulation traces without fabricating spurious details, demonstrating the effectiveness of evidence-grounded preference optimization in aligning model reasoning with factual evidence.

5.4 Ablation Studies

Since the ground-truth annotations for the Challenge test set are not publicly available, we instead conduct ablation studies on the training and validation sets, splitting them into 420k training samples and 140K testing samples to systematically validate the effectiveness of our designs.

Ablation on Detection Designs. We choose the model of naive Dinov3 with MLP head as baseline and incorporate our designs on top of it. Results are shown in Table 4: (1)Degradation-aware Augmentation("DA") brings a substantial improvement, confirming it effectively enhances its robustness to real-world quality variations. (2)Introducing supervised contrastive loss("SC") yields further gain by pulling real and fake samples apart while aligning degraded versions of the same sample. (3)Incorporating mean-teacher stabilization("MT") further boosts performance by anchoring augmented representations toward a stable feature reference to prevent potential feature drift caused by varied augmentations and ensures training stability. Finally, the combination of above designs yields SOTA performance.

Ablation on Deepfake Explanation Designs. We use Qwen3-VL-8B-Instruct[36] as baseline, and incorporate different training strategies on top for evidence-grounded explanation. Since the official evaluation code for Fact and Claim F1 takes a considerable amount of time, we randomly sample a 1k subset from our test set for evaluation. Table 5 reports the performance. (1) SFT improves all metrics by enabling the model to learn the basic pattern of evidence description. (2) Applying DPO with evidence-omitted rejected responses further improves performance by explicitly teaching the model to avoid missing critical manipulation traces, and using irrelevant-augmented samples enhances the model's ability to suppress hallucinated details. (3) Combining all three types(irrelevant-induced, evidence-omitted and combined above) of rejected samples yields the best results, confirming that evidence omission and irrelevant hallucination are complementary failures we addressed.

Table 4: Ablation on detection components.

Configuration	Detection F1	Detection Accuracy
Baseline	0.8824	0.8877
+ DA	0.9402	0.9411
+ DA& SC	0.9588	0.9623
+ DA& MT	0.9593	0.9627
Full(+ DA& MT & SC)	**0.9724**	**0.9793**

Table 5: Ablation on complex explanation model designs.

Training Strategy	Fact F1	Claim F1	BERTScore
Baseline	0.4402	0.3362	0.6038
+ SFT	0.5173	0.4218	0.6673
+ DPO (Omit Only)	0.5666	0.4860	0.6912
+ DPO (Irrelevant Only)	0.5743	0.5042	0.6856
+ DPO (Irrelevant + Omit)	0.5942	0.5144	0.7131
Full(DPO with all three types)	**0.5969**	**0.5227**	**0.7184**

6 Conclusion

In this work, we propose a framework for the explainable deepfake detection task. For robust detection, we introduced degradation-aware contrastive learning with a mean-teacher architecture that prevents feature drift while preserving augmentation benefits. For faithful explanation, we devised evidence-grounded preference optimization via DPO that explicitly teaches the model to prioritize genuine manipulation evidence, complemented by a concise explanation model optimized through GRPO with tailored reward for brevity-sensitive scenarios. Our method ranks first for ACM Multimedia 2026 Explainable Deepfake Detection Challenge, providing a promising method for future research toward more reliable and transparent forensic AI systems.

Acknowledgements. This work was supported by the grants from the National Natural Science Foundation of China (62372014, 62525201, 62132001, 62432001), Beijing Nova Program and Beijing Natural Science Foundation (4252040, L247006).

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923* (2025).

[2] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24185–24198.

[3] Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. Simplicity Level Estimate (SLE): A Learned Reference-Less Metric for Sentence Simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 12053–12059. doi:10.18653/v1/2023.emnlp-main.739

[4] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*.

[5] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. 2020. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*. PMLR, 3247–3258.

[6] Rehan Guha. 2024. *BRISQUE: Blind/Referenceless Image Spatial Quality Evaluator*. doi:10.5281/zenodo.11104461

[7] Xiao Guo, Xiufeng Song, Yue Zhang, Xiaohong Liu, and Xiaoming Liu. 2025. Rethinking Vision-Language Model in Face Forensics: Multi-Modal Interpretable Forged Face Detector. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 105–116.

[8] Parul Gupta, Shreya Ghosh, Tom Gedeon, Thanh-Toan Do, and Abhinav Dhall. 2025. Multiverse Through Deepfakes: The MultiFakeVerse Dataset of Person-Centric Visual and Conceptual Manipulations. arXiv:2506.00868 [cs.MM] https://arxiv.org/abs/2506.00868

[9] Benedikt Hopf and Radu Timofte. 2025. Practical Manipulation Model for Robust Deepfake Detection. arXiv:2506.05119 [cs.CV] https://arxiv.org/abs/2506.05119

[10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[11] Zhengchao Huang, Bin Xia, Zicheng Lin, Zhun Mou, and Wenming Yang. 2024. Ffaa: Multimodal large language model based explainable open-world face forgery analysis assistant. *arXiv preprint arXiv:2408.10072* (2024).

[12] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. 4171–4186.

[13] Kartik Kuckreja, Parul Gupta, Muhammad Haris Khan, and Abhinav Dhall. 2026. Pixels Don't Lie (But Your Detector Might): Bootstrapping MLLM-as-a-Judge for Trustworthy Deepfake Detection and Reasoning Supervision. arXiv:2602.19715 [cs.CV] https://arxiv.org/abs/2602.19715

[14] Samuli Laine and Timo Aila. 2017. Temporal Ensembling for Semi-Supervised Learning. arXiv:1610.02242 [cs.NE] https://arxiv.org/abs/1610.02242

[15] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems* 36 (2023), 34892–34916.

[16] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. 2021. Spatial-Phase Shallow Learning: Rethinking Face Forgery Detection in Frequency Domain. arXiv:2103.01856 [cs.CV] https://arxiv.org/abs/2103.01856

[17] Wentao Mo and Yang Liu. 2026. Distilling Neuro-Symbolic Programs into 3D Multi-modal LLMs. arXiv:2606.01215 [cs.CV] https://arxiv.org/abs/2606.01215

[18] Daniel Morales-Brotons, Thijs Vogels, and Hadrien Hendrikx. 2024. Exponential Moving Average of Weights in Deep Learning: Dynamics and Benefits. arXiv:2411.18704 [cs.LG] https://arxiv.org/abs/2411.18704

[19] Abhijeet Narang, Kartik Kuckreja, Shreya Ghosh, Muhammad Haris Khan, Jianfei Cai, and Abhinav Dhall. 2026. XPlainVerse: A Million-Scale Benchmark for Explainable Deepfake Detection. arXiv:2607.03562 [cs.CV] https://arxiv.org/abs/2607.03562

[20] Abhijeet Narang, Kartik Kuckreja, Shreya Ghosh, Muhammad Haris Khan, Usman Tariq, Jianfei Cai, and Abhinav Dhall. 2026. Explainable Deepfake Detection Challenge. arXiv:2607.21007 [cs.CV] https://arxiv.org/abs/2607.21007

[21] Jinyoung Park, Jeehye Na, Jinyoung Kim, and Hyunwoo J. Kim. 2026. DeepVideo-R1: Video Reinforcement Fine-Tuning via Difficulty-aware Regressive GRPO. arXiv:2506.07464 [cs.CV] https://arxiv.org/abs/2506.07464

[22] Siran Peng, Zipei Wang, Li Gao, Xiangyu Zhu, Tianshuo Zhang, Ajian Liu, Haoyuan Zhang, and Zhen Lei. 2025. MLLM-Enhanced Face Forgery Detection: A Vision-Language Fusion Solution. *arXiv preprint arXiv:2505.02013* (2025).

[23] Yuxin Peng, Zishuo Wang, Geng Li, Xiangtian Zheng, Sibo Yin, and Hulingxiao He. 2026. A Survey on Fine-Grained Multimodal Large Language Models. *Chinese Journal of Electronics* 35, 2 (2026), 771–803. doi:10.23919/cje.2025.00.336

[24] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. 2020. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *Proceedings of the European Conference on Computer Vision*. Springer, 86–103.

[25] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[26] Simiao Ren, Yao Yao, Kidus Zewde, Zisheng Liang, Ning-Yau Cheng, Xiaoou Zhan, Qinzhe Liu, Yifei Chen, Hengwei Xu, et al. 2025. Can Multi-modal (reasoning) LLMs work as deepfake detectors? *arXiv preprint arXiv:2503.20084* (2025).

[27] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junyang Song, et al. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning. *arXiv preprint arXiv:2402.03300* (2024).

[28] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. 2025. DINOv3. arXiv:2508.10104 [cs.CV] https://arxiv.org/abs/2508.10104

[29] Zhihao Sun, Haoran Jiang, Haoran Chen, Yixin Cao, Xipeng Qiu, Zuxuan Wu, and Yu-Gang Jiang. 2024. ForgerySleuth: Empowering Multimodal Large Language Models for Image Manipulation Detection. *arXiv preprint arXiv:2411.19466* (2024).

[30] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. 2024. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 5052–5060.

[31] Hao Tan, Jun Lan, Zichang Tan, Ajian Liu, Chuanbiao Song, Senyuan Shi, Huijia Zhu, Weiqiang Wang, Jun Wan, and Zhen Lei. 2026. Veritas: Generalizable Deepfake Detection via Pattern-Aware Reasoning. arXiv:2508.21048 [cs.CV] https://arxiv.org/abs/2508.21048

[32] Shahroz Tariq, David Nguyen, MAP Chamikara, Tingmin Wu, Alsharif Abuadbba, and Kristen Moore. 2025. LLMs Are Not Yet Ready for Deepfake Image Detection. *arXiv preprint arXiv:2506.10474* (2025).

[33] Antti Tarvainen and Harri Valpola. 2018. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. arXiv:1703.01780 [cs.NE] https://arxiv.org/abs/1703.01780

[34] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. 2024. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905* (2024).

[35] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2024. CogVLM: Visual Expert for Pretrained Language Models. arXiv:2311.03079 [cs.CV] https://arxiv.org/abs/2311.03079

[36] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388* (2025).

[37] Zhiwen Yang and Yuxin Peng. 2026. GaLa-2.5D: Global-Local Alignment with 2.5D Semantic Guidance for Camera-Based 3D Semantic Scene Completion in Autonomous Driving. *Chinese Journal of Electronics* (2026). https://api.semanticscholar.org/CorpusID:289116718

[38] Dengyong Zhang, Feifan Qi, Jiahao Chen, Jiaxin Chen, Rongrong Gong, Yuehong Tian, and Lebing Zhang. 2025. Fake Face Detection Based on Fusion of Spatial Texture and High-Frequency Noise. *Chinese Journal of Electronics* 34, 1 (2025), 212–221. doi:10.23919/cje.2023.00.342

[39] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations (ICLR)*.

[40] Minghang Zheng, Zihao Yin, Yi Yang, Yuxin Peng, and Yang Liu. 2026. Temporal-Aware Reasoning Optimization for Video Temporal Grounding. arXiv:2606.09248 [cs.CV] https://arxiv.org/abs/2606.09248

[41] Jiaran Zhou, Yuezun Li, Baoyuan Wu, Bin Li, Junyu Dong, et al. 2024. Freqblender: Enhancing deepfake detection by blending frequency knowledge. *Advances in Neural Information Processing Systems* 37 (2024), 44965–44988.

[42] Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. 2025. LLaVA-3D: A Simple yet Effective Pathway to Empowering LMMs with 3D-awareness. arXiv:2409.18125 [cs.CV] https://arxiv.org/abs/2409.18125

[43] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479* (2025).