Anchored-prompt Zero-Training Framework for Deepfake Detection, Localization and Explanation

Wenfei Xu[1*], **Xibo Fan**[1*], **Junbo Gao**[1*] and **Ye Zhu**[1†]

[1]School of Artificial Intelligence, Hebei University of Technology

1318464155@qq.com, 2965659239@qq.com, 1416339629@qq.com, zhuye@hebut.edu.cn

Abstract

With the rapid development of generative models, forged content has become increasingly realistic, making it essential to accurately determine the authenticity of visual content and provide reliable evidence for corresponding decisions. Most existing deepfake forensic methods primarily focus on detection or localization, making it difficult to jointly perform detection, localization, and explanation grounded in visual evidence. To address this limitation, we propose an anchored-prompt zero-training framework that simultaneously performs deepfake detection, manipulated-region localization, and visible forgery-trace explanation. First, we construct a schema anchored prompt that integrates task instructions, output-format constraints, and forensic evidence categories into a unified instruction. A frozen Qwen2-VL-7B-Instruct model then performs evidence-guided forensic reasoning. Finally, a consistency filtering module normalizes the generated JSON outputs, validates the predicted bounding boxes, and enforces semantic consistency between classification results and spatial evidence. Extensive experiments on the DDL-X challenge benchmark demonstrate the effectiveness of the proposed framework across deepfake detection, localization, and explanation tasks. Our final solution ranked among the top ten teams in DDL-X Track 3, indicating its strong overall performance and practical potential for real-world deepfake forensics.

1 Introduction

In recent years, with the rapid development of deep learning, especially generative adversarial networks (GANs) [Goodfellow *et al.*, 2014], diffusion models (DMs) [Ho *et al.*, 2020], and large vision-language models, highly realistic forged images can now be easily generated by ordinary users, while such forged images are becoming increasingly difficult to identify. If maliciously disseminated, these forged contents may cause serious negative impacts on personal privacy, social stability, the credibility of news media, and even the judicial system. To address these risks, both academia and industry have proposed a large number of deepfake detection methods and released a series of related datasets. As shown in Figure 1, early methods mostly formulate deepfake detection as a binary classification task. However, they usually only output image-level real/fake labels or confidence scores, lacking precise localization of manipulated regions and making it difficult to explain the reasons behind model predictions to users. Therefore, relying solely on binary classification results is often insufficient for real-world applications such as forensic investigation, content moderation, and trustworthy media analysis. To improve the verifiability and interpretability of deepfake analysis systems, the task of Deepfake Detection and Localization (DDL) has received increasing attention. Unlike traditional detection tasks, DDL requires models not only to determine whether an image has been manipulated, but also to localize the specific forged regions. The localization results can provide intuitive evidence for the detection conclusion, making model predictions more transparent. However, most existing DDL methods still remain at the level of visual localization, usually presenting suspicious regions in the form of masks, bounding boxes, or heatmaps, while

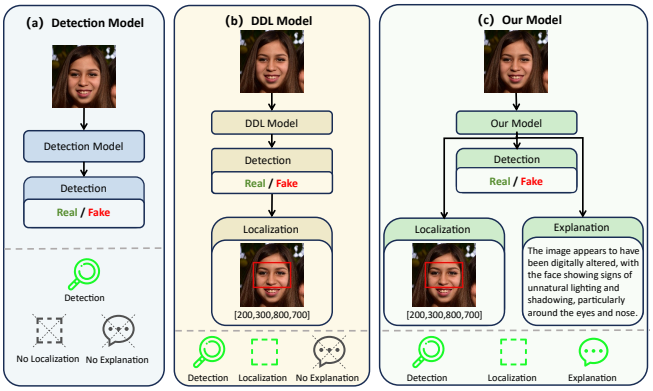

Figure 1: Comparison of framework capabilities. Existing deepfake methods are often limited to image-level detection or detection with localization. In contrast, the proposed framework provides a unified solution that simultaneously supports detection, manipulated-region localization, and natural-language explanation.

*These authors contributed equally.

†Corresponding author.

lacking natural-language explanations for users. As a result, non-expert users may still find it difficult to understand why a model considers a certain region to contain forgery traces. Track 3: Deepfake Detection, Localization, and Explainability Challenge at the IJCAI-ECAI 2026 Workshop on AIGC Safety further extends the traditional DDL task. This track focuses on deepfake image analysis in real-world scenarios, requiring participating systems not only to determine whether an input image is real or fake, but also to provide the corresponding forged regions and generate explanatory text consistent with visible visual evidence. Compared with methods that only output detection scores or localization boxes, this task places greater emphasis on the consistency among detection results, spatial localization, and textual explanations, thereby posing higher requirements on the comprehensive understanding ability and explainable reasoning capability of models. Meanwhile, the challenge released DDL-X [Miao *et al.*, 2025a; Miao *et al.*, 2025b], a dataset that leverages Qwen3-VL to enrich the DDL-I corpus. This dataset supports the development of deepfake detection technology and improves the explainability of detection results.

This paper presents our submitted system for DDL-X Track 3. Different from supervised detectors that require task-specific training or private checkpoints, our method formulates the challenge as a prompt-based multimodal forensic reasoning problem. We use a public instruction-following vision-language model as the backbone and guide it with an evidence-oriented forensic prompt. The generated response is further processed by a structured validation module, which recovers JSON outputs, normalizes fields, validates coordinates, and enforces consistency among classification, localization, and explanation. Our main contributions are summarized as follows:

- We design a prompt-echo schema anchoring module that combines the task objective, output schema, and forensic evidence checklist into a unified instruction. It constrains the generation format while directing the model toward image-specific forensic evidence.
- We formulate deepfake analysis as zero-training multimodal forensic reasoning. Using a frozen vision-language model, the proposed method jointly predicts image authenticity, manipulated-region bounding boxes, and natural-language explanations without task-specific training or additional detector heads.
- We develop a consistency filtering module that performs JSON recovery, field normalization, label validation, coordinate clipping, bounding-box verification, and cross-output consistency checking. This module prevents malformed or contradictory generations from directly becoming final predictions.
- Extensive backbone comparisons, component-wise ablations, and qualitative analyses demonstrate the effectiveness of the three modules for deepfake detection, localization, and explanation. Our system ranked among the top ten teams in DDL-X Track 3.

2 Related Work

2.1 Deepfake Detection

FaceForensics++ provided a standard benchmark dataset for face manipulation detection [Rossler *et al.*, 2019], while Face X-ray improved model generalization by modeling blending boundaries [Li *et al.*, 2020]. ViT learns visual representations through global attention over sequences of image patches [Dosovitskiy *et al.*, 2020], Swin Transformer supports both high-resolution modeling and dense prediction through hierarchical shifted windows [Liu *et al.*, 2021], and MAE provides a scalable self-supervised pretraining paradigm based on masked image reconstruction [He *et al.*, 2022]. Building on these advances, C2P-CLIP injects category-common prompts into the text branch of CLIP, using semantic concepts to assist in identifying forged images [Tan *et al.*, 2025]. Generalization has become a core issue in deepfake detection: methods such as AIDE, DiffusionFake, GM-DF, IAPL, and OMAT specifically investigate transferable forgery features across generators, data domains, and latent priors [Yan *et al.*, 2025; Sun *et al.*, 2024; Lai *et al.*, 2025; Li *et al.*, 2026; Zhou *et al.*, 2026]. Some other studies focus on practical deployment by exploring weakly supervised schemes, data ecosystem construction, and fairness constraints [Qiao *et al.*, 2024; Lai *et al.*, 2026; Lin *et al.*, 2024; Ding *et al.*, 2026]. However, most of the above detection algorithms mainly focus on producing image-level authenticity labels or confidence scores, and cannot simultaneously provide complete manipulated-region localization and interpretable textual explanations.

2.2 Deepfake Localization

Unlike detectors that only output classification scores, localization models are required to provide spatial evidence, such as pixel-level masks, manipulation boundaries, region boxes, or text tokens. ManTra-Net learns generic manipulation trace representations for pixel-level tracing of unknown editing operations [Wu *et al.*, 2019]. MVSS-Net adopts multi-view and multi-scale supervision, jointly constraining image-level classification and pixel-level segmentation so that the two tasks can mutually benefit each other [Chen *et al.*, 2021]. EAN explicitly formulates image manipulation localization as a segmentation problem between internal and external boundary regions, leveraging neighboring feature interactions to enhance both explicit and implicit edges while alleviating false positives caused by feature redundancy in large models [Chen *et al.*, 2024]. Cross-modal studies further extend the targets of localization. DGM4 unifies authenticity judgment, manipulation type recognition, image-region grounding, and text-token grounding for image-text media. HAMMER first performs manipulation-aware contrastive learning and then models image-text interactions through cross-attention [Shao *et al.*, 2024]. SIDA targets social media images and simultaneously outputs real/synthetic/manipulated categories, manipulation masks, and textual explanations [Huang *et al.*, 2025]. MDSM/AMD considers semantically consistent deceptive narratives generated by multimodal large models, weakening the traditional shortcut of image-text mismatch, and evaluates models with joint metrics for detection, type recognition, and localization [Zhang *et al.*, 2026].

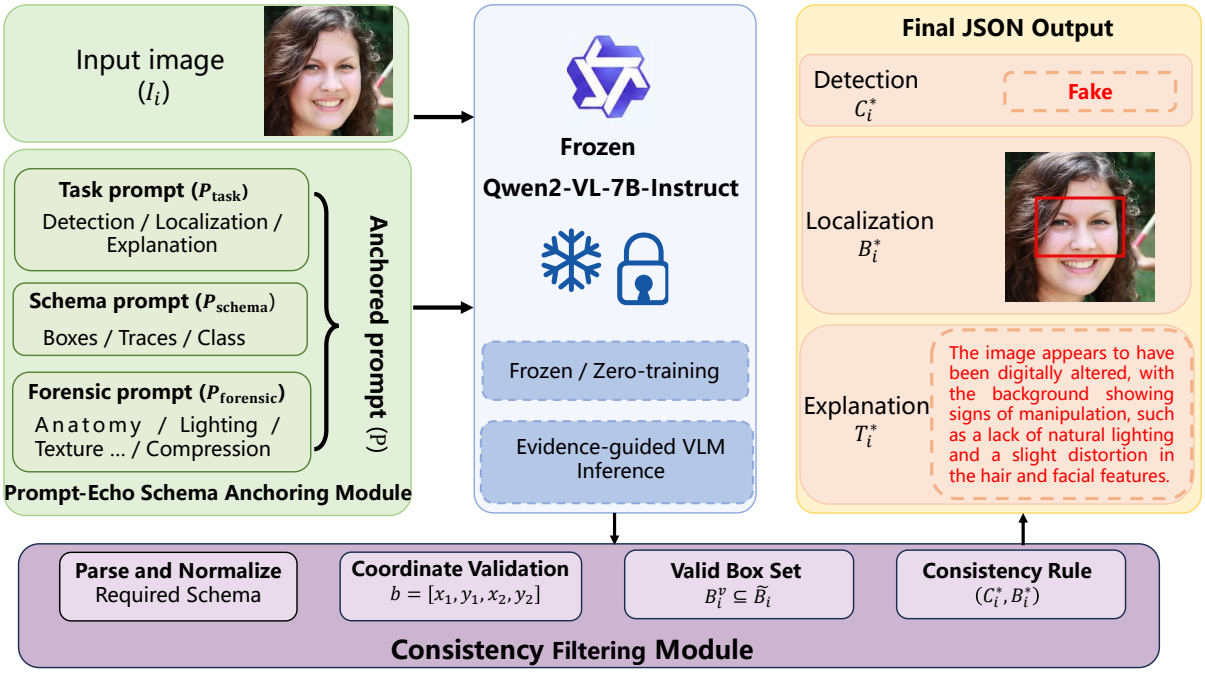

Figure 2: Overview of the proposed zero-training multimodal forensic reasoning framework. The input image is combined with an evidence-oriented anchored prompt and processed by a frozen Qwen2-VL-7B-Instruct model to generate a raw response containing an initial prediction. The raw response is then parsed, normalized, and refined by semantic-logical consistency filtering to produce the final JSON result.

2.3 Deepfake Explainability

M2F2-Det combines forgery-oriented CLIP prompt learning with an adapter connected to a large language model, enabling the model to output both forgery scores and textual reasons [Guo *et al.*, 2025]. However, fluent natural language does not necessarily correspond to real visual evidence. Especially for high-quality forged samples, models may produce seemingly plausible but incorrect descriptions of manipulated regions or artifacts. To reduce such explanation hallucinations, FFTG first constructs manipulation masks and region/type cues from real–fake differences, and then uses these verifiable signals to constrain the textual explanations generated by multimodal large models [Sun *et al.*, 2025]. This idea emphasizes that explanations should be anchored in verifiable region and type evidence rather than freely generated by an independent language model attached after a classifier. The image-box and token grounding in DGM4 [Shao *et al.*, 2024], the integrated category–mask–reason output in SIDA [Huang *et al.*, 2025], and the manipulation-oriented reasoning in AMD [Zhang *et al.*, 2026] all reflect a trend from merely "readable descriptions" toward "evidence-constrained explanations."

The DDL-X task integrates the above three requirements, requiring models to simultaneously perform authenticity classification, spatial localization, and textual explainability generation within a unified framework. This task setting encourages algorithms to generate case-specific explanations grounded in real evidence, rather than generic descriptions. Following this design principle, our method generates textual explanations entirely based on the suspicious regions identified by the detector. As a result, the output of the explanation module is strictly constrained to visible forensic features, such as unnatural edge transitions, texture misalignment, distorted facial details, and local synthetic artifacts, and avoids producing broad descriptions without visual evidence.

3 Method

3.1 Framework Overview

We propose a zero-training multimodal forensic reasoning framework for DDL-X Track 3. The task requires the model to produce a structured prediction for each input image, including image-level classification, manipulated-region localization, and visible-trace explanation. Different from conventional deepfake detectors that mainly focus on binary real/fake classification, our framework formulates the task as a unified visual forensic reasoning problem, where detection, localization, and explanation are generated in a single inference pipeline.

As shown in Figure 2, the proposed framework consists of three main components. First, the prompt-echo schema anchoring module constructs an evidence-oriented instruction that explicitly specifies the required JSON schema and forensic evidence categories. This module guides the model to preserve the target output fields and to ground its explanation in observable visual traces. Second, the frozen Vision-Language Model (VLM) performs evidence-guided zero-training inference. Given the input image and the anchored prompt, Qwen2-VL-7B-Instruct generates a raw response containing an initial prediction for bounding boxes,

visible forgery traces, and a real/fake classification result. Third, the consistency filtering module parses and normalizes the raw response, removes invalid bounding boxes, and enforces semantic-logical consistency between classification and localization.

The whole framework is inference-only. We do not introduce additional training data, task-specific fine-tuning, private model weights, ensemble models, or extra detector heads. In this way, the proposed framework converts deepfake detection from a label-only decision problem into a structured forensic reasoning pipeline. The prompt anchoring module improves output controllability, the frozen VLM provides multimodal reasoning ability without training, and the consistency filtering module improves the reliability of the final structured submission.

3.2 Prompt-Echo Schema Anchoring Module

Although VLM can follow natural-language instructions, their outputs may still suffer from schema drift and semantic drift. Schema drift means that the model may generate free-form text instead of the required JSON structure. Semantic drift means that the explanation may become generic and not directly related to visible forensic evidence. To reduce these problems, a prompt-echo schema anchoring module is designed with task, schema and forensic prompt. The anchored prompt P is composed as:

$$P = P_{\text{task}} \oplus P_{\text{schema}} \oplus P_{\text{forensic}}, \tag{1}$$

where P_{task} defines the forensic analysis objective, P_{schema} specifies the required output format, P_{forensic} describes the visual evidence dimensions to be checked, and $\oplus$ denotes prompt concatenation.

The three prompt components play different roles. The task prompt defines the model identity and analysis objective, making the generation process focus on deepfake forensic inspection. The schema prompt explicitly anchors the required JSON fields, including Bounding boxes, Visible forgery traces, and Classification result, which can specify the exact output fields and reduces format uncertainty during generation. The forensic prompt anchors the visual reasoning dimensions, such as lighting consistency, texture artifacts, boundary traces, compression artifacts, and color coherence, which provides visual evidence categories and guides the model to ground its explanation in observable image regions. By combining these components, the prompt not only asks for a prediction, but also constrains how the prediction should be structured and justified.

3.3 Evidence-guided Zero-training VLM Inference

Since DDL-X Track 3 requires classification, localization, and explanation simultaneously, a training-based solution usually needs task-specific annotations and additional trainable components. In contrast, our framework adopts a zero-training inference strategy. We use a frozen Qwen2-VL-7B-Instruct model as the multimodal reasoning backbone and guide it with the anchored forensic prompt introduced above. No task-specific training, fine-tuning, private checkpoint, ensemble model, or extra detection head is used.

Given an input image I_i and the fixed anchored prompt P, the frozen VLM first generates a raw textual response:

$$R_i = F_\theta(I_i, P), \tag{2}$$

where θ denotes the frozen parameters of Qwen2-VL-7B-Instruct. The response R_i is expected to contain an initial prediction for bounding boxes, visible forgery traces, and the real/fake classification result.

The raw response is not directly used as the final output because VLM generations may contain malformed fields, invalid boxes, or inconsistencies between classification and localization. Therefore, R_i is further processed by the consistency filtering module to obtain a structurally valid and logically consistent final prediction.

3.4 Consistency Filtering Module

Given the raw response R_i generated by the frozen VLM, the consistency filtering module first converts it into an initial structured prediction and then refines it into the final structured output. Even with prompt-echo schema anchoring, the raw output may still contain malformed fields, invalid bounding boxes, or contradictions between classification and localization. For example, the model may predict `fake` but fail to provide a valid manipulated region, or predict `real` while still producing suspicious bounding boxes. Such inconsistencies are harmful because the task requires classification and localization outputs to be semantically aligned. The proposed consistency filtering module improves the semantic-logical consistency between classification and localization.

Before applying the consistency rule, we first parse and normalize the raw response. The parser extracts the JSON-like content from R_i and maps possible field variations to the required schema. After parsing and normalization, the raw response is converted into an initial structured prediction:

$$\tilde{Y}_i = \mathrm{Norm}(\mathrm{Parse}(R_i)) = \{\tilde{B}_i, \tilde{T}_i, \tilde{C}_i\}, \tag{3}$$

where $\mathrm{Parse}(\cdot)$ denotes JSON-like content extraction and $\mathrm{Norm}(\cdot)$ denotes field normalization. The classification field is normalized to either `real` or `fake`. The explanation field is converted into a plain textual description. For localization, the parser checks whether the predicted bounding boxes can be represented as numerical coordinate tuples. Non-numerical, incomplete, or malformed boxes are discarded before coordinate validation. For each predicted bounding box $b = [x_1, y_1, x_2, y_2]$, we keep only boxes whose coordinates lie within the normalized range and satisfy the valid corner order:

$$B_i^v = \left\{ b \in \tilde{B}_i \,\middle|\, \begin{array}{l} 1 \le x_1 < x_2 \le 1000, \\ 1 \le y_1 < y_2 \le 1000 \end{array} \right\}. \tag{4}$$

Boxes that do not satisfy these constraints are discarded before the final consistency decision.

We then correct the classification and localization results according to the relationship between the predicted label and the valid spatial evidence:

$$(C_i^*, B_i^*) = \begin{cases} (\texttt{fake}, B_i^v), & \tilde{C}_i = \texttt{fake} \text{ and } |B_i^v| > 0, \\ (\texttt{real}, \texttt{None}), & \text{otherwise.} \end{cases} \tag{5}$$

Algorithm 1 Semantic-logical consistency filtering

Input: Raw response R_i
Output: Final prediction $Y_i^* = \{B_i^*, T_i^*, C_i^*\}$
Extract JSON-like content from R_i
Map field variations to the required schema
Obtain $\tilde{Y}_i = \{\tilde{B}_i, \tilde{T}_i, \tilde{C}_i\}$ by parsing and normalization
Initialize valid box set $B_i^v \leftarrow \emptyset$
for each predicted box $b \in \tilde{B}_i$ **do**
  if $1 \le x_1 < x_2 \le 1000$ and $1 \le y_1 < y_2 \le 1000$ **then**
    Add b to B_i^v
  end if
end for
if $\tilde{C}_i =$ `fake` and $|B_i^v| > 0$ **then**
  $C_i^* \leftarrow$ `fake`
  $B_i^* \leftarrow B_i^v$
else
  $C_i^* \leftarrow$ `real`
  $B_i^* \leftarrow$ `None`
end if
Normalize explanation T_i^* from $\tilde{T}_i$
Save $Y_i^* = \{B_i^*, T_i^*, C_i^*\}$ as the final JSON output
return Y_i^*

where C_i^* and B_i^* denote the final classification and localization outputs. The explanation field is normalized from $\tilde{T}_i$ and saved as T_i^*.

This rule follows a conservative forensic principle: an image should not be predicted as `fake` unless the model can provide effective spatial evidence. If the initial prediction is `fake` but no valid bounding box is available, the prediction is corrected to `real`. If the final prediction is `real`, the bounding box field is set to `None`.

The consistency rule mainly enforces the alignment between the final classification label and the localization output. The explanation field is normalized from $\tilde{T}_i$ and retained as the textual evidence description. Therefore, the filtering module should be understood as reducing structural and classification-localization conflicts, rather than fully verifying the semantic faithfulness of the generated explanation. The detailed filtering process is summarized in Algorithm 1.

4 Experiments

4.1 Experimental Setup

Dataset

We conduct experiments on the DDL-X dataset released for Track 3: Deepfake Detection, Localization, and Explainability Challenge at the IJCAI-ECAI 2026 Workshop on AIGC Safety. Different from conventional deepfake benchmarks that mainly focus on image-level binary classification, DDL-X evaluates three complementary capabilities: real/fake detection, manipulated-region localization, and natural-language explanation generation. Therefore, it provides a suitable benchmark for evaluating whether a forensic system can produce not only a decision, but also spatial and textual evidence supporting the decision.

During the competition, the organizers provided development data for format checking and method development, while the final test annotations were hidden from participants and evaluated automatically by the Codabench platform. In our experiments, we follow the official data usage rules and do not use any private training data or hidden labels. The reported analysis is based on the predictions generated by our released inference pipeline.

Evaluation Metrics

The competition comprehensively evaluates participating systems from three aspects: deepfake detection, manipulated-region localization, and textual explainability. For the deepfake detection task, the official evaluation adopts Accuracy (ACC) as the evaluation metric, which measures the correctness of real/fake classification over all input images. ACC is defined as:

$$\mathrm{ACC} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{y}_i = y_i), \tag{6}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. It equals 1 when the prediction matches the ground-truth label, and 0 otherwise. Here, N denotes the number of images in the test set, y_i represents the ground-truth label of the i-th image, and $\hat{y}_i$ represents the corresponding predicted label.

For the manipulated-region localization task, the official evaluation adopts Intersection over Union (IoU) as the evaluation metric, which measures the spatial overlap between the predicted manipulated region and the ground-truth manipulated region. For a single forged image, IoU is defined as:

$$\mathrm{IoU}_i = \frac{|B_i^{\mathrm{pred}} \cap B_i^{\mathrm{gt}}|}{|B_i^{\mathrm{pred}} \cup B_i^{\mathrm{gt}}|}. \tag{7}$$

where B_i^{pred} denotes the predicted manipulated region, and B_i^{gt} denotes the ground-truth manipulated region. The overall mean IoU (mIoU) is formulated as:

$$\mathrm{mIoU} = \frac{1}{|\mathcal{F}|}\sum_{i\in\mathcal{F}}\mathrm{IoU}_i. \tag{8}$$

where $\mathcal{F}$ denotes the set of forged images, and $|\mathcal{F}|$ denotes the number of forged images.

For the textual explainability task, the official evaluation adopts BERTScore [Zhang *et al.*, 2019] to measure the semantic consistency between the generated explanation and the reference explanation. BERTScore extracts contextual semantic representations of the candidate text and the reference text using a pretrained language model, and computes their matching degree based on token-level maximum cosine similarity.

Implementation Details

We use the frozen `Qwen/Qwen2-VL-7B-Instruct` checkpoint [Qwen Team, 2024] at revision `eed13092ef`, without task-specific training or fine-tuning. It is loaded in `bfloat16` with Accelerate automatic device mapping and run on one NVIDIA A30 (24 GB) using PyTorch 2.5.1, CUDA 12.1, and Transformers 4.45.0. Images are processed

Table 1: Ablation Results of Each Core Component. All metrics take our complete pipeline as the baseline.

Setting	ACC	mIoU	BERTScore
w/o Anchored prompt	53.32	3.37	46.97
w/o coordinate validation	56.17	2.45	82.99
w/o Consistency filtering	55.16	5.09	84.43
Ours	58.59	7.09	86.42

Table 2: VLM Backbones comparison under the same forensic prompt and inference setting.

Backbone	ACC (%)	mIoU (%)	BERTScore (%)
LLaVA-1.5-7B	48.37	0.00	74.18
InternVL2-8B	43.97	0.00	69.53
Qwen2-VL-7B (Ours)	58.59	7.09	86.42

independently, converted to RGB, and resized with Lanczos while preserving aspect ratio to fit within 1920×1080. We use the checkpoint-provided single-turn chat template, with one image and the forensic prompt in the user message. Decoding uses `max_new_tokens=200`, `do_sample=False`, `use_cache=True`, and `repetition_penalty=1.0`; temperature, top-p, and top-k are unset, and no random seed is set because sampling is disabled. The 103,585-image development set requires approximately 96 GPU-hours; resumable writing handles interruptions, and the 2.3% out-of-memory cases follow the fallback rule described above.

4.2 Ablation Analysis

Component Ablation

We conduct one-at-a-time ablations of the anchored prompt, coordinate validation, and consistency filtering. Table 1 shows that removing any component degrades all three metrics. The prompt provides the largest overall gain, coordinate validation most strongly affects localization IoU, and consistency filtering yields moderate but consistent improvements, confirming their complementary roles.

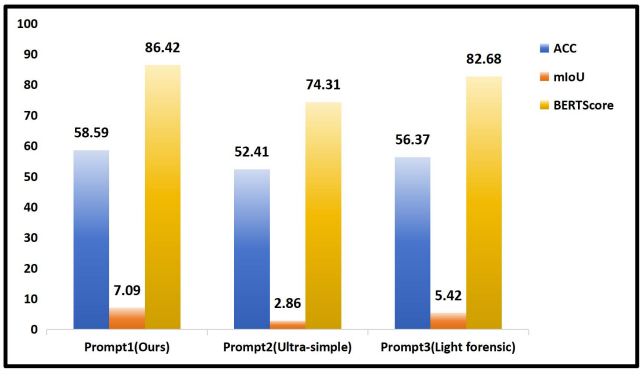

Figure 3: Quantitative comparison of three prompt variants across three core evaluation metrics: overall classification ACC, fake-region IoU, and BERTScore for forensic text explanation. The tested prompts include the original detailed forensic prompt, an ultra-simplified basic detection prompt, and a lightweight condensed forensic prompt.

Ablation on VLM Backbones

Under identical prompts and inference settings, Table 2 shows that Qwen2-VL-7B performs best on classification, localization, and explanation. LLaVA-1.5-7B and InternVL2-8B obtain zero IoU, indicating weaker spatial grounding for this zero-training forensic task.

Ablation on Prompt

We compare an ultra-simple detection prompt, a lightweight forensic prompt, and the full anchored prompt. Figure 3 shows that the ultra-simple version degrades all metrics, particularly IoU, while the lightweight version approaches the full prompt. The detailed forensic guidance achieves the best overall classification, localization, and explanation performance.

4.3 Qualitative Results

To further analyze the comprehensive capability of the proposed method in deepfake detection, tampered-region localization, and textual explanation, we present a qualitative visualization analysis on several representative examples.

As shown in Figure 4, the left panel illustrates the forensic analysis prompt and the corresponding input images used during inference. The right panel further presents a comparison between the ground-truth annotations and the model outputs, where the green dashed region denotes the ground-truth results and the red dashed region denotes the predicted results. It should be noted that the explanation shown in the ground-truth section is only a partial excerpt of the original explanation, intended to highlight the key forensic cues related to the tampered region. From the detection results, the model can effectively distinguish forged images from authentic ones. For the first three examples containing facial manipulations, the model correctly predicts the label fake. For the fourth authentic image, the model correctly predicts real and does not generate an additional localization box. This indicates that the proposed method is capable of identifying face-manipulated images with clear local anomalies while, to some extent, avoiding false alarms on real images. From the localization results, the predicted red bounding boxes show a high degree of consistency with the ground-truth green bounding boxes. From the explanation results, the generated natural-language descriptions are semantically consistent with the localized regions.

As shown in Figure 5, the proposed method still has two main limitations. First, for complex forged images containing multiple manipulated regions, the current system tends to capture only the most salient local anomaly and has difficulty covering all ground-truth manipulated regions simultaneously. Meanwhile, the single-bounding-box setting also causes the explanation text to focus only on the detected local region, resulting in incomplete explanation coverage. Second, although the classification-localization consistency filtering strategy adopted in this paper can reduce contradictory outputs where an image is classified as fake without valid spatial evidence and improve the structural reliability of the final

Figure 4: Qualitative visualization results. The left panel shows the prompt template and the input images used to guide the model for deepfake forensic analysis, where the specific JSON output format is omitted for brevity. The right panel presents a comparison between the ground-truth annotations and the model predictions, where the green dashed box denotes the ground-truth results and the red dashed box denotes the model outputs. Green boxes indicate the ground-truth tampered regions, while red boxes indicate the predicted tampered regions. Each example includes the image-level detection result, the tampered-region localization box, and the corresponding natural-language explanation. Note that the explanation text in the ground-truth annotations is only a partial excerpt of the original explanation.

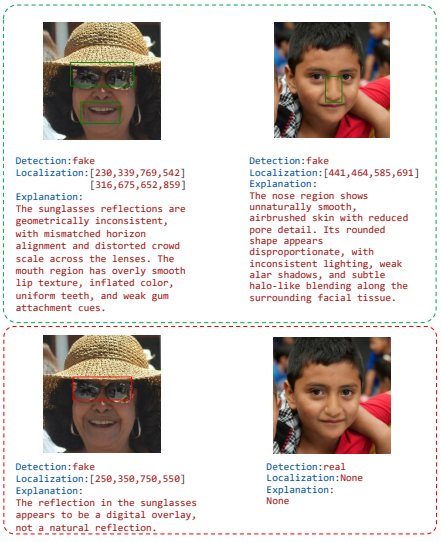

Figure 5: Failure examples. The green dashed boxes denote the ground-truth results, while the red dashed boxes denote the model outputs. The green boxes indicate the ground-truth manipulated regions, and the red boxes indicate the predicted manipulated regions.

submission, it may also introduce a certain risk of missed detections.

5 Conclusion

This paper presents our submitted system for Track 3 of the IJCAI-ECAI 2026 Workshop on AIGC Safety. We formulate deepfake detection, localization, and explanation as a prompt-based multimodal forensic reasoning problem and build a zero-training pipeline with a public VLM, evidence-oriented prompting, and classification-localization consistency constraints.

Experiments show that the framework can generate complete challenge-format predictions without supervised training, private data, or private model weights. Although the Zero-Training VLM still has limited recall and explanation faithfulness, conservative validation improves the reliability of positive fake predictions and prevents malformed generations from producing invalid submissions. Future work will improve explanation grounding and reduce prompt echoing through stronger VLMs, robust structured decoding, and evidence-aware verification.

Acknowledgements

The work was supported by National Natural Science Foundation of China under the Grant 62276088, Natural Science Foundation of Hebei Province under the Grant F2024202017, Basic Research Project of Hebei Universities in Shijiazhuang under the Grant 241790817A, and the Central Guidance Fund for Local Science and Technology Development Projects under Grant 246Z0106G.

References

[Chen *et al.*, 2021] Xinru Chen, Chengbo Dong, Jiaqi Ji, Juan Cao, and Xirong Li. Image manipulation detection by multi-view multi-scale supervision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14185–14193, 2021.

[Chen *et al.*, 2024] Yun Chen, Hang Cheng, Haichou Wang, Ximeng Liu, Fei Chen, Fengyong Li, Xinpeng Zhang, and Meiqing Wang. Ean: Edge-aware network for image manipulation localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(2):1591–1601, 2024.

[Ding *et al.*, 2026] Feng Ding, Wenhui Yi, Yunpeng Zhou, Xinan He, Hong Rao, and Shu Hu. Decoupling bias, aligning distributions: Synergistic fairness optimization for deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 32704–32713, 2026.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Goodfellow *et al.*, 2014] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[Guo *et al.*, 2025] Xiao Guo, Xiufeng Song, Yue Zhang, Xiaohong Liu, and Xiaoming Liu. Rethinking vision-language model in face forensics: Multi-modal interpretable forged face detector. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 105–116, 2025.

[He *et al.*, 2022] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Huang *et al.*, 2025] Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. Sida: Social media image deepfake detection, localization and explanation with large multimodal model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28831–28841, 2025.

[Lai *et al.*, 2025] Yingxin Lai, Hongyang Wang, Jing Yang, Xiangui Kang, Bin Li, Linlin Shen, and Zitong Yu. Gmdf: Generalized multi-scenario deepfake detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 4300–4309, 2025.

[Lai *et al.*, 2026] Yingxin Lai, Zitong Yu, Jun Wang, Linlin Shen, Yong Xu, and Xiaochun Cao. Agent4faceforgery: Multi-agent llm framework for realistic face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14073–14083, 2026.

[Li *et al.*, 2020] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5001–5010, 2020.

[Li *et al.*, 2026] Yiheng Li, Zichang Tan, Guoqing Xu, Zhen Lei, Xu Zhou, and Yang Yang. Towards generalizable ai-generated image detection via image-adaptive prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21262–21272, 2026.

[Lin *et al.*, 2024] Li Lin, Xinan He, Yan Ju, Xin Wang, Feng Ding, and Shu Hu. Preserving fairness generalization in deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16815–16825, 2024.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[Miao *et al.*, 2025a] Changtao Miao, Yi Zhang, Weize Gao, Zhiya Tan, Weiwei Feng, Man Luo, Jianshu Li, Ajian Liu, Yunfeng Diao, Qi Chu, et al. Ddl: A large-scale datasets for deepfake detection and localization in diversified real-world scenarios. *arXiv preprint arXiv:2506.23292*, 2025.

[Miao *et al.*, 2025b] Changtao Miao, Yi Zhang, Man Luo, Weiwei Feng, Kaiyuan Zheng, Qi Chu, Tao Gong, Jianshu Li, Yunfeng Diao, Wei Zhou, et al. Mffi: Multi-dimensional face forgery image dataset for real-world scenarios. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 13235–13242, 2025.

[Qiao *et al.*, 2024] Tong Qiao, Shichuang Xie, Yanli Chen, Florent Retraint, and Xiangyang Luo. Fully unsupervised deepfake video detection via enhanced contrastive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):4654–4668, 2024.

[Qwen Team, 2024] Qwen Team. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct, 2024.

[Rossler *et al.*, 2019] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019.

[Shao *et al.*, 2024] Rui Shao, Tianxing Wu, Jianlong Wu, Liqiang Nie, and Ziwei Liu. Detecting and ground-

ing multi-modal media manipulation and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5556–5574, 2024.

[Sun *et al.*, 2024] Ke Sun, Shen Chen, Taiping Yao, Hong Liu, Xiaoshuai Sun, Shouhong Ding, and Rongrong Ji. Diffusionfake: Enhancing generalization in deepfake detection via guided stable diffusion. *Advances in Neural Information Processing Systems*, 37:101474–101497, 2024.

[Sun *et al.*, 2025] Ke Sun, Shen Chen, Taiping Yao, Ziyin Zhou, Jiayi Ji, Xiaoshuai Sun, Chia-Wen Lin, and Rongrong Ji. Towards general visual-linguistic face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19576–19586, 2025.

[Tan *et al.*, 2025] Chuangchuang Tan, Renshuai Tao, Huan Liu, Guanghua Gu, Baoyuan Wu, Yao Zhao, and Yunchao Wei. C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7184–7192, 2025.

[Wu *et al.*, 2019] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9543–9552, 2019.

[Yan *et al.*, 2025] Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. In *International Conference on Learning Representations*, volume 2025, pages 70702–70720, 2025.

[Zhang *et al.*, 2019] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[Zhang *et al.*, 2026] Yuchen Zhang, Yaxiong Wang, Yujiao Wu, Lianwei Wu, Li Zhu, and Zhedong Zheng. The coherence trap: When mllm-crafted narratives exploit manipulated visual contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8760–8769, 2026.

[Zhou *et al.*, 2026] Yue Zhou, Xinan He, KaiQing Lin, Bing Fan, Feng Ding, and Bin Li. Breaking latent prior bias in detectors for generalizable aigc image detection. *Advances in Neural Information Processing Systems*, 38:30649–30679, 2026.