# A Reproducible Detection, Localization, and Explanation System for DDL-X Track 3

**Yihang Lu**[1], **Tianshuo Zhang**[1], **Siran Peng**[2], **Haoyuan Zhang**[1], **Zhen Lei**[2]
[1]University of Chinese Academy of Sciences
[2]Institute of Automation, Chinese Academy of Sciences
luyihang24@mails.ucas.ac.cn, tianshuo.zhang@nlpr.ia.ac.cn, pengsiran2023@ia.ac.cn,
zhanghaoyuan2023@ia.ac.cn, zhen.lei@ia.ac.cn

## Abstract

DDL-X Track 3 requires an integrated system to classify face images, localize manipulated regions, and explain visible forgery traces. We present an image-to-JSON pipeline that coordinates three learned modules: a ConvNeXt-B face classifier, a three-detector localization ensemble with weighted box fusion (WBF), and Qwen2.5-VL-3B-Instruct adapted with LoRA for explanation generation. Fixed post-processing converts face scores into image labels, fuses and selects detector boxes, and conditions the explanation model on the image, label, and localized regions. On an internal hold-out split, the classification branch reached $92.17\%$ image-level accuracy. On the 24,000-image internal development selection set, $76.0\%$ of fake images achieved a best-region IoU above $0.7$ under WBF. On a balanced 512-example explanation holdout, the LoRA-adapted model improved mean BERTScore F1 from $0.8354$ for a condition-specific template to $0.8877$. On the final DDL-X Track 3 leaderboard, our team ranked second with an official score of $0.7025$. We release the inference code, fixed configurations, and model weights for rerunning the pipeline from images to JSON.

## 1 Introduction

DDL-X Track 3 couples three outputs: a binary image decision, bounding boxes for visible manipulated regions, and an English explanation of visible forgery traces. The challenge builds on recent datasets for face forgery under diversified real-world conditions, including MFFI [Miao *et al.*, 2025b] and DDL [Miao *et al.*, 2025a]. Its central systems problem is not only predicting each field, but also coordinating outputs so that localization supports explanation while classification controls the required output schema.

Recent multimodal forensic systems have approached related goals from different directions. M2F2-Det couples face-forgery scoring with textual explanations [Guo *et al.*, 2025]; SIDA jointly studies detection, localization, and explanation [Huang *et al.*, 2025]; and ForgeryGPT and FakeVLM connect authenticity decisions with interpretable artifact descriptions [Zhang *et al.*, 2024; Wen *et al.*, 2025]. In contrast to methods centered on one large multimodal model, our challenge system assigns specialized learned modules to the three outputs and connects them through fixed interfaces.

Our contributions are threefold. First, we develop a coordinated image-to-JSON pipeline in which all required fields are produced automatically by learned modules and fixed post-processing, without manual intervention at inference time. Second, we report held-out component diagnostics for classification, localization, and explanation quality. Third, we release one integrated inference entrypoint with the weights and configurations required to regenerate submission-format JSON files from images, and we report the official final leaderboard result for the complete system.

## 2 Method

Figure 1 summarizes the complete inference path. The classifier provides the image label, the localization ensemble provides boxes, and both outputs condition explanation generation. The three learned modules are connected by deterministic aggregation, selection, prompt construction, and JSON serialization.

### 2.1 Classification

Input images are preprocessed using MTCNN [Zhang *et al.*, 2016] from `facenet-pytorch` version 2.5.3. We retain all detected faces using cascade thresholds $0.6/0.7/0.7$ and a minimum face size of 20 pixels, and expand each detected box by a factor of $1.4$ before cropping. The first learned module is a ConvNeXt-B [Liu *et al.*, 2022] binary classifier applied to these face crops. The image fake probability is the maximum probability among accepted face crops, and the fixed image-level threshold is $0.20$. This OR-like aggregation matches the task semantics that an image is fake when any visible face is manipulated and preserves sensitivity when only one of several faces is forged. Because taking a maximum can increase false positives as the number of faces grows, face acceptance is fixed before aggregation and the operating threshold is calibrated on image-level rather than face-level outputs; the calibration therefore includes the effect of multi-face aggregation. The threshold was selected with the fixed localization pipeline from candidates between $0.20$ and $0.60$ on the calibration split using the end-to-end development proxy. On the untouched holdout, $0.20$ also produced

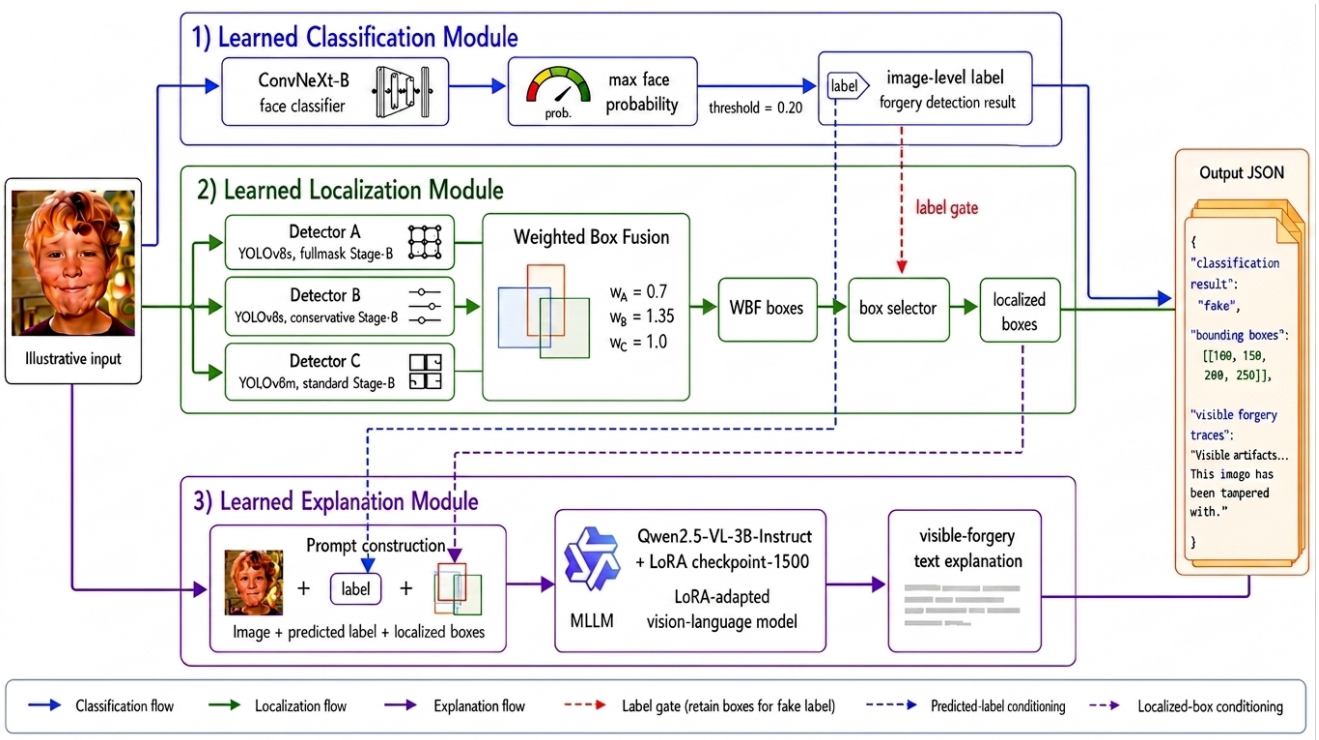


Figure 1: Main DDL-X Track 3 image-to-JSON inference path. Three coordinated learned modules produce the image-level classification, localized regions, and visible-forgery explanation. The classification result gates the retention of fused boxes without entering WBF, while the image, predicted label, and localized boxes condition Qwen2.5-VL explanation generation. The confidence-triggered landmark fallback described in Section 2.2 is omitted for visual clarity; the input thumbnail and JSON values are illustrative.

the highest accuracy, recall, and F1 among the evaluated candidates (Table 1), whereas larger thresholds traded recall for a smaller number of false positives.

### 2.2 Localization

Detector A and detector B use YOLOv8s backbones, whereas detector C uses YOLOv8m. Their Stage-B variants are full-mask for detector A, conservative for detector B, and standard for detector C. WBF [Solovyev *et al.*, 2021] uses weights $0.7/1.35/1.0$, pre-fusion confidence $0.125$, IoU $0.35$, and post-fusion confidence $0.175$. At most three boxes are retained, and each fused box requires support from at least two detectors. The landmark fallback is activated only when the maximum face fake probability reaches the image-level threshold $0.20$ and no detector box survives the WBF confidence and support filters. In that case, deterministic boxes centered on the MTCNN-predicted nose, left eye, right eye, and mouth landmarks provide coarse challenge-specific regions; they are not treated as detector-supported manipulation evidence.

### 2.3 Explanation

Qwen2.5-VL-3B-Instruct receives the image and a programmatically constructed prompt containing the predicted label and retained WBF or fallback boxes. For a fake prediction, the prompt requests visually supportable manipulation evidence and a concise tampering conclusion. For a real prediction, it requests evidence of visual consistency and a non-tampering conclusion. The generated response replaces only the explanation field; the model is not asked to emit JSON or coordinates.

## 3 Training Strategy

**Classifier.** The initial ConvNeXt-B classifier was initialized from ImageNet-pretrained weights and trained on $2{,}069{,}511$ face crops, comprising $1{,}065{,}678$ real and $1{,}003{,}833$ fake faces. The corresponding validation manifest contained $230{,}749$ crops. Training used eight GPUs, per-GPU batch size $128$ (global batch $1024$), AdamW with learning rate $10^{-4}$ and weight decay $10^{-4}$, mixed precision, and a positive-class weight computed from the training labels. Crops were resized to $224 \times 224$, with horizontal flipping and mild color jitter. The resumed run was configured for ten epochs; its preserved log covers epochs 2–8, with the epoch-8 checkpoint reaching $99.70\%$ accuracy and $0.9999$ ROC AUC on the source-distribution validation manifest. For final development adaptation, all parameters except the classifier head and last ConvNeXt feature stage were frozen. The adaptation set contained $100{,}000$ face crops ($60{,}000$ real and $40{,}000$ fake), with $6{,}830$ internal validation crops. The model was adapted for one epoch on eight GPUs with per-GPU batch size $96$, AdamW, head learning rate $10^{-4}$, last-stage learning rate $10^{-5}$, weight decay $10^{-4}$, and mixed precision.

**Detectors.** All three detectors use Ultralytics YOLOv8 [Jocher *et al.*, 2026] and share a two-stage schedule. Detectors A and B use YOLOv8s, whereas detector C uses YOLOv8m. Stage A performs one fullmask-to-box training epoch on $2{,}069{,}472$ records using six GPUs, global batch size $96$, image size $512$, and AdamW. Recovery checkpoints are written every $4{,}000$ steps. Stage B adapts each detector to the development distribution for three epochs. Detector A uses learning rate $2 \times 10^{-4}$; detector B repeats the continuation with a conservative rate of $10^{-4}$; detector C starts from a COCO-pretrained YOLOv8m checkpoint and uses $2 \times 10^{-4}$.

**Explanation model.** We fine-tuned Qwen2.5-VL [Bai *et al.*, 2025] with LoRA [Hu *et al.*, 2022] while freezing the vision encoder and modality aligner. The training JSONL contains $67{,}240$ unique images and paired reference explanations: $33{,}255$ real and $33{,}985$ fake examples. Each record contains one image, a condition-specific user prompt, and an assistant reference response. LoRA was applied to the language-model linear projections with rank $16$, scaling $32$, and dropout $0.05$. Training used eight GPUs, per-device batch size $1$, four-step gradient accumulation, learning rate $5 \times 10^{-5}$, cosine decay, warmup ratio $0.03$, weight decay $0.1$, FP16, and maximum sequence length $4096$. Checkpoint-1500, reached after $0.714$ of the scheduled epoch, was retained after internal evaluation.

# 4 Experiments

## 4.1 Evaluation Protocol

We report internal development evidence because the three output types require different diagnostics. Classification metrics are computed after maximum face-score aggregation, treating fake as the positive class. We report accuracy, precision, recall, and F1 at threshold $0.20$, together with a threshold sweep on the untouched holdout split. Localization is evaluated on the 24,000-image internal development selection set using the number of fake images whose best region IoU exceeds $0.7$. We additionally report the development proxy used during model selection; this proxy assumes fixed text quality and is not an official leaderboard score or a complete detection metric such as mAP or mean IoU.

The Qwen training split contains $67{,}240$ unique images, while the tune, calibration, and holdout pools each contain $12{,}000$ unique images, with zero image overlap among the four pools. For evaluation, we independently sampled a balanced $512$-example subset ($256$ fake and $256$ real images) from each pool using seeds 20260531, 20260601, and 20260602, respectively. Candidate explanations are generated using classifier-predicted labels and the final localization outputs after WBF and, when triggered, landmark fallback, matching the deployed inference conditions; no ground-truth labels or boxes are supplied to the generation model. Candidates are compared with the paired reference explanations stored in the challenge-derived JSONL records. We use BERTScore [Zhang *et al.*, 2020] with `roberta-large`, layer 17, no IDF weighting, and no baseline rescaling. Whitespace is normalized before scoring. The classifier source and adaptation train/validation manifests likewise have no image-path overlap. We verify image-path disjointness for the reported splits; stricter separation by identity, source video, manipulation method, or near-duplicate group was not available from the challenge-derived manifests.

Table 1: Image-level threshold sweep on the 12,000-image holdout split. Fake is the positive class; all values are percentages.

| Threshold | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| **0.20** | **92.17** | 97.09 | **90.98** | **93.93** |
| 0.22 | 92.07 | 97.40 | 90.51 | 93.83 |
| 0.25 | 91.68 | 97.70 | 89.64 | 93.49 |
| 0.28 | 91.34 | 97.96 | 88.86 | 93.19 |
| 0.30 | 91.02 | 98.11 | 88.23 | 92.91 |
| 0.35 | 90.52 | 98.47 | 87.13 | 92.45 |
| 0.40 | 89.80 | 98.71 | 85.83 | 91.82 |
| 0.45 | 89.13 | 98.92 | 84.63 | 91.22 |
| 0.50 | 88.54 | 99.13 | 83.55 | 90.67 |
| 0.60 | 87.04 | **99.43** | 81.03 | 89.29 |

Table 2: Localization result on the 24,000-image internal development selection set. The proxy is used only for internal model selection.

| Method | Fake IoU $> 0.7$ | Rate | Dev proxy |
|---|---|---|---|
| A+B+C WBF | 12,158 | 75.99% | 0.7677 |

## 4.2 Classification and Localization

The source-distribution classifier reached $99.70\%$ face-level accuracy and $0.9999$ ROC AUC before development adaptation. After adaptation to the different development distribution, the final checkpoint reached $80.02\%$ face-level accuracy and $0.9006$ ROC AUC on its $6{,}830$-crop internal validation set. After face-score aggregation, image-level accuracy was $92.17\%$ on the held-out development split. At threshold $0.20$, the image-level classifier obtained $97.09\%$ precision, $90.98\%$ recall, and $93.93\%$ F1. Table 1 shows that increasing the threshold monotonically improved precision but reduced recall; the default $0.50$ threshold lowered accuracy by $3.63$ percentage points and F1 by $3.26$ points relative to $0.20$.

With the selected WBF configuration, $12{,}158$ of $16{,}000$ fake images ($75.99\%$) achieved a best-region IoU above $0.7$ on the internal development selection set. We report this result as a component diagnostic rather than a standalone methodological contribution.

## 4.3 Explanation Evaluation

On the calibration set, maximum generation lengths of $384$, $512$, and $640$ tokens produced mean F1 values of $0.8866$, $0.8867$, and $0.8867$, respectively. We selected $512$ tokens because it achieved the highest unrounded score without the cost of the $640$-token setting. Table 3 reports the untouched holdout comparison. The LoRA-adapted model primarily improves recall, reflecting greater coverage of the reference explanation than the short template.

The released rerun command uses greedy decoding with an output budget of $2048$ new tokens; this deployment setting is separate from the $512$-token checkpoint-selection protocol.

Table 3: BERTScore on the balanced 512-example explanation holdout.

| Method | Precision | Recall | F1 |
|---|---|---|---|
| Fixed template | 0.8868 | 0.7896 | 0.8354 |
| Qwen2.5-VL + LoRA | **0.8891** | **0.8864** | **0.8877** |

### 4.4 Official Challenge Result

The submitted system ranked second on the official leaderboard, with an official final score of 0.7025. The official evaluation measures the complete image-to-JSON submission, including classification, localization, and explanation fields. This result complements the internal component diagnostics above and confirms that the modular pipeline was competitive under the challenge's end-to-end evaluation.

## 5 Reproducibility

The public reproduction path starts from the official test image directory and runs the fixed submitted pipeline. The wrapper `run_end_to_end_from_images.sh` takes an image directory, model-root directory, output directory, and GPU selector. It writes per-image JSON files, a submission zip, and run summaries. The JSON schema contains `Classification result`, `Bounding boxes`, and `Visible forgery traces`. For real images, boxes are `null`; for fake images, coordinates are integers on the DDL-X 1-1000 scale.

The release includes the classifier, three detector checkpoints, Qwen base model, LoRA adapter, fixed configuration, training and evaluation scripts, and SHA256 manifests. Package validation produced 100,000 JSON files: 63,989 fake and 36,011 real predictions, all with a complete final explanation statement. The code is available in the GitHub repository, and the model assets are available in the Hugging Face model repository. The original Qwen2.5-VL-3B-Instruct source is linked from the official Qwen model page. Both public repositories were accessible when this manuscript was finalized.

## 6 Limitations

This paper reports a challenge-specific system and does not claim cross-dataset generalization. The development proxy and BERTScore analysis are internal selection diagnostics rather than official challenge metrics, and the localization proxy is not a replacement for standard detection metrics such as mAP or mean IoU. BERTScore is sensitive to response length and semantic overlap and cannot replace human assessment of forensic correctness. Maximum face-score aggregation improves sensitivity but may increase false positives for images with many detected faces. When confidence-gated WBF yields no retained box for a fake prediction, the landmark fallback supplies geometry-based facial regions rather than detector-supported manipulation evidence and may bias explanations toward fixed facial areas. The available image-level records support thresholded precision–recall analysis, but not a complete score-level ROC or PR curve for the final checkpoint; retaining continuous image scores would enable threshold-independent AUC reporting. Future work could further isolate detector combinations, WBF hyperparameters, and explanation conditioning through a full factorial ablation. Qwen reruns may produce byte-different explanations across software and hardware environments. Small box differences may also arise from preprocessing or detector-library versions.

## 7 Conclusion

We presented an integrated DDL-X Track 3 pipeline with specialized learned modules for classification, localization, and explanation. Under the selected WBF configuration, 76.0% of fake development images achieved a best-region IoU above 0.7, while LoRA adaptation achieved higher reference-overlap BERTScore than condition-specific templates. The final system ranked second on the official DDL-X Track 3 leaderboard with a final score of 0.7025, demonstrating the competitiveness of the complete image-to-JSON system. The released image-to-JSON implementation makes the coordination logic, model weights, and output construction independently inspectable.

## References


[Bai *et al.*, 2025] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 105–116, 2025.

[Hu *et al.*, 2022] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28831–28841, 2025.

[Jocher *et al.*, 2026] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLOv8. Ultralytics software, version 8.4.53, 2026.

[Liu *et al.*, 2022] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

[Miao *et al.*, 2025a] Changtao Miao, Yi Zhang, Weize Gao, Zhiya Tan, Weiwei Feng, Man Luo, Jianshu Li, Ajian Liu, Yunfeng Diao, Qi Chu, Tao Gong, Zhe Li, Weibin Yao, and Joey Tianyi Zhou. DDL: A large-scale datasets for deepfake detection and localization in diversified real-world scenarios, 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, Joey Tianyi Zhou, and Xiaoshuai Hao. MFFI: Multi-dimensional face forgery image dataset for real-world scenarios. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pages 13235–13242, Dublin, Ireland, 2025. Association for Computing Machinery.

[Solovyev *et al.*, 2021] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107:104117, 2021.

[Wen *et al.*, 2025] Siwei Wen, Junyan Ye, Peilin Feng, Hengrui Kang, Zichen Wen, Yize Chen, Jiang Wu, Wenjun Wu, Conghui He, and Weijia Li. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[Zhang *et al.*, 2016] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[Zhang *et al.*, 2020] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020.

[Zhang *et al.*, 2024] Fanrui Zhang, Jiawei Liu, Jiaying Zhu, Esther Sun, Dong Li, Qiang Zhang, and Zheng-Jun Zha. ForgeryGPT: A multimodal LLM for interpretable image forgery detection and localization, 2024.