Deepfake Detection, Localization, and Trace Explanation with Multi-Task Visual Modeling

Xing Fan[1] , **Jiayue Yu**[2]
[1]Independent Researcher
[2]Shanghai Jiao Tong University

Abstract

Deepfake analysis in realistic settings requires more than binary classification: a useful system should also localize manipulated regions and explain the visual evidence behind its decision. In this paper, we present a multi-task framework for the DDL-X Track 3 task, which jointly evaluates image-level detection, forgery localization, and visible-trace explanation. Our method uses a shared visual encoder with task-specific heads for classification and localization, and optionally extends this architecture with a vision-language module for textual trace generation. To better exploit the available annotations, we convert bounding-box supervision into dense mask targets and train the localization branch with a combination of binary cross-entropy and Dice loss. The overall system supports reproducible data splitting, distributed multi-GPU training, and direct export to the official challenge JSON format. The resulting framework provides a practical foundation for studying unified detection, localization, and explanation on real-world deepfake benchmarks. The code is publicly available at https://github.com/fanxing2585/deepfake_detection.

1 Introduction

Recent deepfake detection and localization methods increasingly depend on strong visual feature extraction and multi-scale representation learning to capture subtle forgery artifacts and improve robustness [Guan *et al.*, 2023; Lai *et al.*, 2023]. Standard supervised backbones such as ResNet [He *et al.*, 2015] and Swin Transformer [Liu *et al.*, 2021] remain effective choices for this purpose, offering complementary strengths in local pattern modeling and global contextual reasoning. More generally, recent self-supervised learning research has emphasized the importance of balancing invariant and equivariant information in learned representations. For large scale unlabeld data, self-supervised learning methods [Dangovski *et al.*, 2021; Devillers and Lefort, 2022; Garrido *et al.*, 2023; Wang *et al.*, 2024; Wang *et al.*, 2026] preserve transformation-aware structure instead of enforcing pure invariance, which can also be beneficial for downstream tasks that rely on fine-grained spatial cues.

Recent benchmarks such as MFFI and DDL highlight that modern deepfake analysis must handle diverse manipulation types, compression artifacts, varying image quality, and complex real-world content distributions [Miao *et al.*, 2025b; Miao *et al.*, 2025a]. These challenges make the problem more difficult than standard image classification. A useful deepfake analysis system should capture both global semantic cues, such as overall facial realism, and local visual inconsistencies, such as abnormal boundaries, blending artifacts, or unnatural texture transitions. In addition, when textual explanation is required, the model must connect linguistic output to visually grounded evidence rather than produce generic descriptions.

The DDL-X Track 3 task [DDL 2.0 Workshop, 2026] formalizes this broader setting by evaluating three capabilities simultaneously: image-level detection, manipulation localization, and visible-trace explanation. These three objectives are closely related. A model that correctly identifies forged content should ideally rely on visual evidence that is spatially localized, and a meaningful explanation should describe the same suspicious regions that contribute to the detection result. Motivated by this observation, we treat Track 3 as a unified multi-task learning problem rather than a collection of independent subtasks.

In this paper, we present a practical multi-task framework for deepfake analysis. Our approach uses a shared visual encoder with task-specific heads for classification and localization, together with an optional vision-language module for explanation generation. To better exploit the available supervision, we convert bounding-box annotations into dense mask targets, which allow the localization branch to be trained with standard segmentation losses while still remaining compatible with the challenge output format. The resulting system is designed not only to support model development, but also to provide a complete pipeline for reproducible data splitting, multi-GPU training, checkpoint selection, and challenge-style JSON export.

Our workshop contribution is threefold. First, we formulate DDL-X Track 3 as a joint modeling problem in which detection, localization, and explanation are learned from a shared visual representation. Second, we show how coarse bounding-box annotations can be transformed into dense spa-

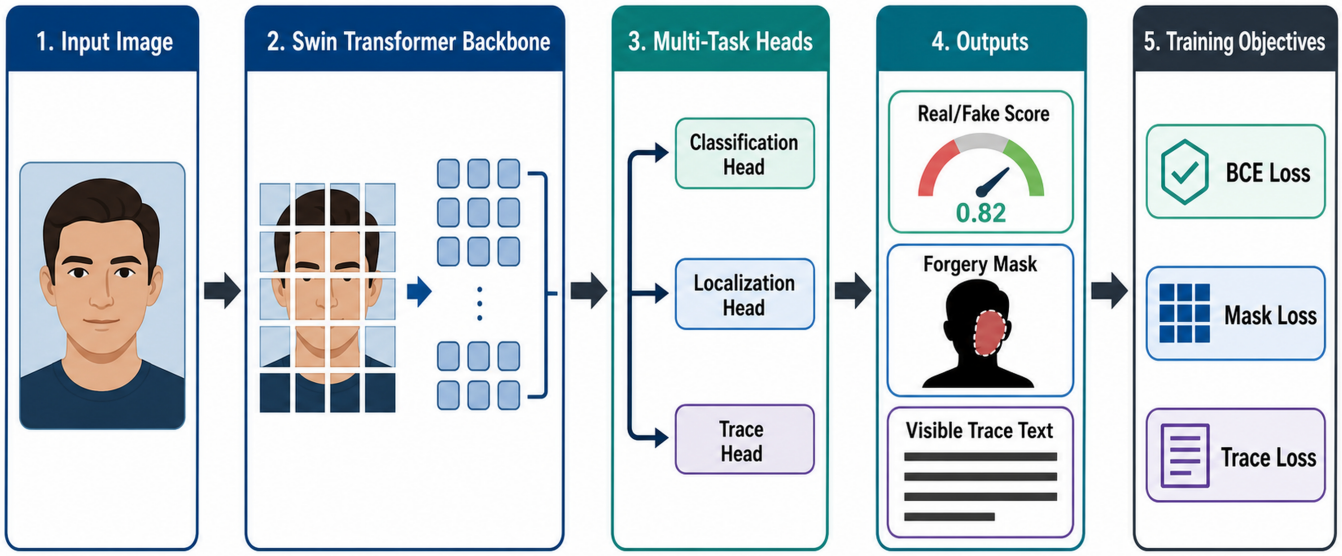

Figure 1: Overview of the proposed multi-task framework. A Swin Transformer backbone extracts shared visual features from the input image, which are then consumed by classification, localization, and trace-generation heads. The model produces a real/fake score, a forgery mask, and visible trace text, and is optimized with joint classification, mask, and trace losses.

tial supervision for forgery localization. Third, we provide an end-to-end implementation framework that supports practical experimentation on the private competition dataset and enables direct evaluation under the official submission protocol.

2 Related Work

Deepfake detection benchmarks. The progress of deepfake detection has been closely tied to the development of large-scale benchmarks. Early influential datasets such as FaceForensics++ standardized evaluation for manipulated facial images and videos and showed that forgery detection performance is highly sensitive to compression and visual quality [Rössler *et al.*, 2019]. Celeb-DF further increased realism and reduced obvious synthesis artifacts, making generalization more challenging for existing detectors [Li *et al.*, 2020]. The DFDC benchmark expanded the scale and diversity of face-swapping data and highlighted the difficulty of robust detection in realistic unconstrained settings [Dolhansky *et al.*, 2020]. More recently, DDL and MFFI shifted attention toward more realistic and diversified forgery scenarios, while also emphasizing localization and interpretability beyond binary classification [Miao *et al.*, 2025a; Miao *et al.*, 2025b].

Deepfake detection and localization. Most early methods focused primarily on image- or video-level classification, often exploiting visual artifacts, frequency inconsistencies, or mismatches between manipulated regions and surrounding context [Nirkin *et al.*, 2022]. However, binary predictions alone provide limited interpretability. To address this limitation, several works explored joint detection and manipulation localization. Nguyen et al. proposed a multi-task framework for simultaneous classification and segmentation of manipulated content [Nguyen *et al.*, 2019]. Dang et al. incorporated attention-based localization to improve both prediction and visual interpretability [Dang *et al.*, 2020]. More recent methods further improved fine-grained localization through collaborative feature learning and segmentation-oriented designs [Guan *et al.*, 2023; Lai *et al.*, 2023].

Explainability in deepfake analysis. Explainability has recently emerged as an important direction in deepfake forensics. Instead of only predicting whether an image or video is forged, explainable systems aim to indicate where the evidence is located and how it can be described in human-understandable terms. Recent works such as ExDDV introduced datasets and benchmarks for explainable deepfake detection with both textual and spatial supervision [Hondru *et al.*, 2025], while newer multimodal approaches have explored integrating visual evidence with language-based explanations [ALBarqawi *et al.*, 2025]. Our work is aligned with this trend, but focuses on a practical multi-task formulation for the DDL-X Track 3 setting, where detection, localization, and visible-trace explanation are learned within a shared visual pipeline.

Vision backbones for joint visual reasoning. Modern deepfake detectors increasingly rely on stronger visual backbones and multi-scale feature learning to capture subtle manipulation artifacts and improve robustness in forgery detection and localization [Guan *et al.*, 2023; Lai *et al.*, 2023]. In practice, both convolutional backbones such as ResNet [He *et al.*, 2015] and hierarchical Transformer backbones such as Swin Transformer [Liu *et al.*, 2021] are widely used choices. ResNet provides a strong and efficient convolutional baseline for visual recognition, while Swin Transformer offers a useful compromise between global context modeling and dense spatial prediction, making it particularly attractive for joint classification and localization tasks. change the word I want

to cite the papers.

3 Method

Figure 1 summarizes the overall architecture of the proposed system. Given an input face image x, the model predicts three outputs: an image-level real/fake score, a pixel-level manipulation mask, and a textual description of visible forgery traces. The central design choice is to share one visual encoder across the three tasks so that global semantic cues and local artifact cues are learned in a common feature space. This is particularly useful in the DDL-X setting, where image-level classification, region localization, and explanation are strongly correlated.

From an optimization perspective, the architecture can be viewed as a multi-objective model with one shared backbone and several lightweight task-specific heads. During training, the classification and localization heads are optimized jointly on all labeled samples, while the explanation component is trained only when trace annotations are available. During inference, the same shared visual representation supports all three outputs in a single forward pass, which makes the pipeline simple to deploy for challenge-style JSON generation.

3.1 Visual Encoder

The detector is implemented with `timm` backbones. The baseline configuration uses ResNet-50, while the stronger submitted variants use Swin-Base. In both cases, the encoder maps the input image to a hierarchy of feature maps with progressively lower spatial resolution and stronger semantic abstraction. We expose intermediate stages rather than only the deepest output, because localization requires spatial detail that would be weakened if the model relied exclusively on the last stage.

The backbone is therefore used in two complementary ways. First, the deepest feature tensor provides a compact global representation for image-level classification. Second, a set of multi-scale feature tensors is passed to the localization branch, which preserves information about suspicious boundaries, texture inconsistencies, and local appearance anomalies. This shared-feature design keeps the model parameter-efficient while still supporting different levels of reasoning for global and spatial tasks.

3.2 Detection Head

The detection branch applies global average pooling to the deepest visual feature map and predicts one binary logit. Concretely, let $F^{(L)} \in \mathrm{R}^{C\times H\times W}$ denote the last-stage feature map. Global average pooling reduces $F^{(L)}$ to a C-dimensional vector, which is then passed through a linear classifier to obtain the image-level logit z_{cls}. Training uses binary cross entropy with logits:

$$L_{\text{cls}} = \mathrm{BCEWithLogits}(z_{\text{cls}}, y), \tag{1}$$

where $y \in \{0, 1\}$ is the ground-truth real/fake label. At inference time, the sigmoid probability is interpreted as the fake confidence. A fixed decision threshold converts this probability into the final label `real` or `fake`. This branch is intentionally simple: most representational capacity is delegated to the shared backbone, while the head itself acts as a lightweight classifier over learned visual evidence.

3.3 Localization Head

The localization branch uses a lightweight feature pyramid. Multi-scale encoder features are projected to a common channel size, merged top-down, and decoded into a single-channel mask logit map. More specifically, each selected backbone feature is first passed through a channel projection layer so that features from different resolutions can be fused consistently. A top-down pathway then upsamples coarser features and combines them with finer-resolution features, producing a spatially richer representation for manipulation localization. The final decoder predicts a one-channel mask logit map, which is upsampled to the input image resolution.

Ground-truth localization masks are generated from the annotated bounding boxes. For real images or images with `Bounding boxes = None`, the target mask is all zeros. The localization loss combines binary cross entropy and soft Dice loss. The BCE term supervises each pixel independently, while the Dice term encourages region-level overlap and reduces sensitivity to the strong foreground-background imbalance typical in forged-region localization:

$$L_{\text{mask}} = 0.5 \cdot \mathrm{BCE}(M, T) + 0.5 \cdot \mathrm{DiceLoss}(M, T), \tag{2}$$

The full detector loss is:

$$L = L_{\text{cls}} + \lambda_{\text{mask}} \cdot L_{\text{mask}}, \tag{3}$$

where M denotes the predicted mask logits, T denotes the target mask, and λ_{mask} is set by the training configuration. After thresholding the predicted mask, connected activated regions are converted back into challenge-format bounding boxes on the 1–1000 coordinate grid. This design lets the model train with dense supervision while still producing the structured box output expected by the benchmark.

3.4 Trace Explanation

To support visible forgery-trace prediction, we attach a retrieval-based trace head to the shared visual encoder and train it jointly with the detection and localization branches. Instead of generating free-form text autoregressively, the trace head maps the visual representation of an input image into a trace semantic space and retrieves the most relevant trace description from a predefined candidate set constructed from the training annotations.

Concretely, the shared encoder extracts visual features that are reused across all three tasks. The trace head takes these features as input and predicts matching scores over candidate visible-trace descriptions. The final textual output is obtained by selecting the highest-scoring trace entry, or a small set of top-ranked entries, from the trace vocabulary. This design is simpler and more stable than free-form text generation, while still allowing the model to produce semantically meaningful explanations grounded in visual evidence.

Because the retrieval head is trained jointly with the classification and localization objectives, the selected trace text

Table 1: Implemented model variants in the current system.

Model	Backbone	Image Size	Notes
Detector baseline	ResNet-50	384	Classification with bbox-to-mask localization
Strong detector	Swin-Base	384	Stronger visual backbone with longer training schedule
Multi-task VLM	Swin-Base + trace head	384	Joint detection, localization, and trace generation

Table 2: Dataset split statistics.

Split	Images	Fake Images
Train	93,226	52,297
Validation	10,359	5,811
Total paired samples	103,585	58,108

is encouraged to reflect the same suspicious cues that support the detector, such as abnormal facial boundaries, blending artifacts, or unnatural local textures. No additional post-training stage is required, and the entire framework remains end-to-end trainable within a unified multi-task setting.

4 Experimental Setup

4.1 Dataset and Split

The local dataset is organized as paired image and JSON files:

```
DDL_X/
  image/
    <id>.jpg or <id>.png
  json/
    <id>.json
```

Each JSON annotation contains the image-level label, bounding boxes, and a natural-language visible forgery trace. The field `Classification result` is mapped to a binary target, where `real = 0` and `fake = 1`. The field `Bounding boxes` is either `None` or a list of boxes in `[x1, y1, x2, y2]` format. Following the challenge annotation convention, box coordinates are interpreted on a 1–1000 normalized grid and converted to image-size binary masks during training. This setup follows the annotation style used in DDL, which provides both detection labels and localization supervision for real-world deepfake analysis [Miao *et al.*, 2025a].

We use a reproducible stratified 9:1 train/validation split.

4.2 Evaluation Protocol

The official Track 3 page describes three evaluation dimensions:

- **Detection:** accuracy.
- **Localization:** intersection-over-union, computed only on fake ground-truth images.
- **Explainability:** BERTScore and phase-2 subjective evaluation.

Our validation code follows the same detection and localization definitions. Detection accuracy is computed from the binary class prediction. Localization IoU is computed by thresholding the predicted mask and comparing it against the bounding-box-derived ground-truth mask, restricted to fake images.

For submission, predicted masks are converted back into bounding boxes in the 1–1000 coordinate space. The inference scripts write one JSON file per image with the fields `Confidence`, `Bounding boxes`, `Visible forgery traces`, and `Classification result`.

4.3 Training Details

All models are trained in PyTorch using distributed data-parallel training and mixed precision. Input images are resized to 384×384 and normalized with ImageNet statistics. We optimize the models with AdamW, weight decay 10^{-4}, linear warmup, and cosine learning-rate decay. The main Swin-Base variants are trained for 100 epochs with a per-GPU batch size of 32. Pretrained Swin-Base models use an initial learning rate of 1×10^{-4}, while baseline runs use 3×10^{-4}.

The training objective combines classification, localization, and trace-prediction losses:

$$\mathcal{L} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{trace}}\mathcal{L}_{\text{trace}}.$$

For the trace-head model, we use $\lambda_{\text{cls}} = 0.2$, $\lambda_{\text{mask}} = 0.5$, and $\lambda_{\text{trace}} = 0.3$. The localization loss combines binary cross entropy and soft Dice loss over bounding-box-derived masks, while the trace loss is multi-label binary cross entropy over frequent visible-trace terms. Best checkpoints are selected using validation performance.

5 Experimental results

5.1 Competition Leaderboard Result

Our final submission was evaluated on the official competition leaderboard and ranked **5th** overall under the team name **XQ999**, which is shown in Figure 2. The submission achieved a **final score of 0.5770**, with task-specific scores of **ACC = 0.1601**, **IoU = 0.1985**, **BERTScore = 0.0585**, and **Rubrics = 0.1599**. This result indicates that the proposed multi-task framework is competitive under the official evaluation protocol, while still leaving room for improvement in both localization quality and explanation generation.

Compared with our internal validation analysis, the leaderboard result provides an external assessment of the full pipeline, including classification, manipulation localization, and visible-trace explanation. We view this ranking as evidence that jointly modeling the three tasks is a practical direction for real-world deepfake analysis.

Rank	Team	Final Score	ACC	IoU	BERTScore	Rubrics
1	[redacted]	0.8940	0.1995	0.3079	0.0943	0.2923
2	[redacted]	0.7025	0.1776	0.2495	0.0813	0.1941
3	[redacted]	0.6857	0.1562	0.2363	0.0851	0.2081
4	[redacted]	0.6181	0.1601	0.2200	0.0769	0.1610
5	XQ999	0.5770	0.1601	0.1985	0.0585	0.1599
6	[redacted]	0.3697	0.1593	0.1529	0.0454	0.0121
7	[redacted]	0.3505	0.1590	0.1818	0.0000	0.0097
8	[redacted]	0.2553	0.1090	0.1201	0.0160	0.0102
9	[redacted]	0.2522	0.0600	0.1612	0.0000	0.0031
10	[redacted]	0.1630	0.0655	0.0031	0.0313	0.0630

Figure 2: Official final leaderboard of the competition. Our team, **XQ999**, ranked 5th overall with a final score of 0.5770. For privacy, other team names are masked.

5.2 Learning Curves

Figure 3 illustrates the loss decomposition for the Swin-Base model with the trace head. The classification loss decreases rapidly and remains substantially smaller than the other objectives throughout training, suggesting that image-level real/fake discrimination becomes relatively easy once strong visual features are learned. In contrast, the mask and trace losses remain noticeably larger, indicating that manipulated-region localization and visible forgery-trace prediction are more challenging tasks.

The validation curves for classification and mask prediction remain relatively stable, whereas the trace-related loss begins to increase after approximately 20 epochs. This behavior suggests that the addition of the trace head does not significantly destabilize the shared visual encoder, but the trace-generation branch itself is harder to optimize and may be more prone to overfitting or optimization instability. These observations indicate that further improvements may require a stronger trace head design, more careful loss balancing, or a more sensitive analysis of the relative weights assigned to the different training objectives.

Figure 4 compares the total-loss learning curves of the representative model variants over the first 80 epochs. All curves are computed from the same training and validation split. The ResNet-50 baseline converges steadily but remains higher than the Swin-based variants, indicating that the stronger transformer backbone provides a better optimization trajectory for the multi-task objective.

Swin-Base reduces both training and validation loss more effectively than ResNet-50. Adding the trace head changes the total-loss scale because the model is optimized with an additional trace objective, but the curve remains stable throughout training. This suggests that the trace supervision can be jointly optimized with classification and localization without causing training instability. Together with the ablation results, the learning curves support the use of a Swin-Base backbone and multi-task trace supervision for the final system.

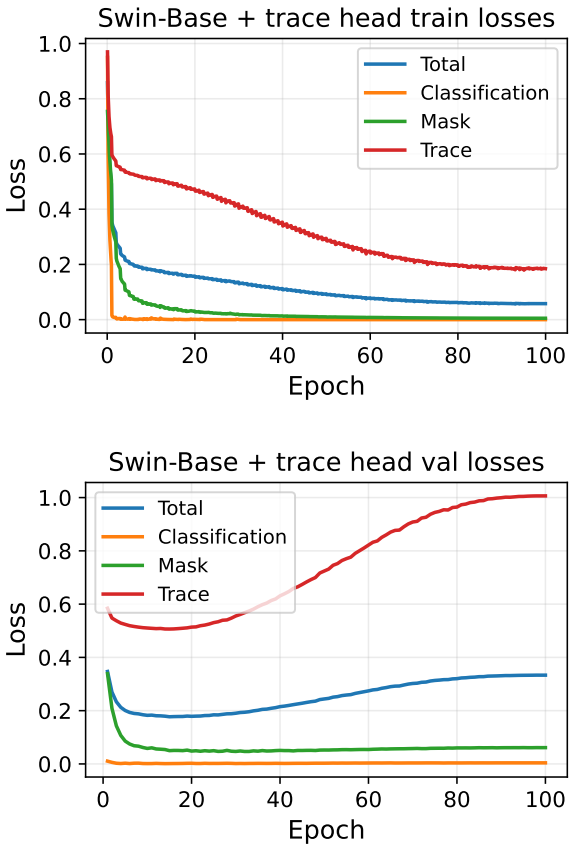

Figure 3: Loss decomposition of the Swin-Base model with the trace head. Top: training losses. Bottom: validation losses. The total loss is optimized jointly with classification, mask, and trace objectives.

5.3 Ablation study

Since the official test partition and labels are not publicly released, we report all quantitative results on the held-out validation split. The validation set is generated with a fixed random seed and preserves the real/fake class distribution of the full dataset. Therefore, the reported numbers should be interpreted as validation-set performance rather than official leaderboard results.

Table 3 evaluates the effect of adding the generative trace head while keeping the detector backbone fixed. The detection-localization model already achieves strong image-level accuracy and fake-only localization IoU, showing that bounding-box-derived mask supervision is effective for learning spatial forgery evidence. After introducing the generative trace head, localization IoU further improves, suggesting that trace-oriented supervision can encourage the shared representation to focus more strongly on manipulation-related visual cues. At the same time, the generative trace module substantially improves BERTScore while preserving nearly identical detection performance. These results indicate that generative trace prediction is a practical way to improve explanation quality without degrading the core detection and localization tasks.

Table 3: Effect of the generative trace head on the validation split.

Variant	Mask	Trace	Acc.	Fake IoU	BERTScore
Detection + Localization	Yes	Fixed	**0.9999**	0.9280	0.8066
+ Generative trace head	Yes	Generation	0.9998	**0.9436**	**0.8607**

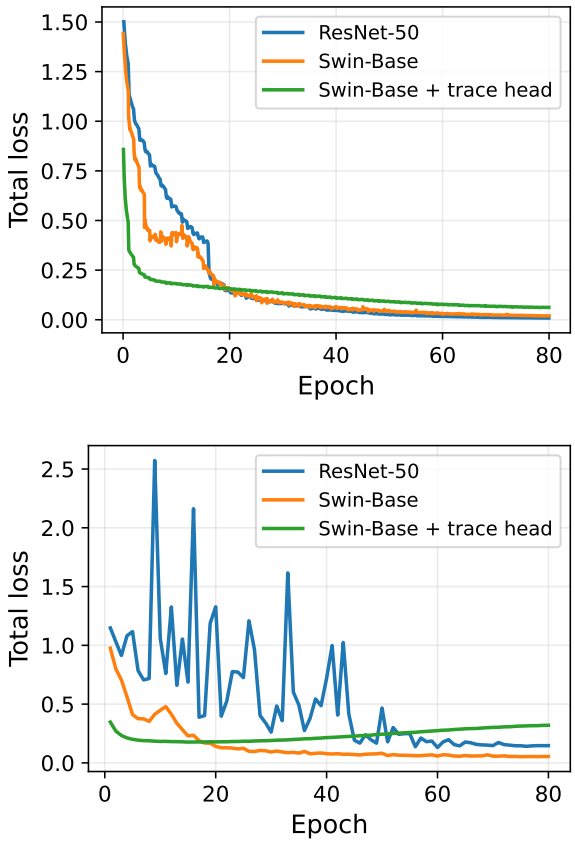

Figure 4: Training and validation total-loss curves. We compare ResNet-50, Swin-Base, and Swin-Base with the trace head under the same validation split.

These artifacts demonstrate that the implementation already supports end-to-end challenge submission. In the future work, we plan to replace Table 1 with a quantitative comparison covering detection accuracy, fake-only IoU, and trace-generation quality.

6 Discussion

The main advantage of the method is that it treats detection and localization as coupled visual tasks. A single shared encoder encourages the detector to learn visual evidence that is also spatially meaningful. The bounding-box-to-mask conversion gives dense supervision while remaining faithful to the available labels.

The approach also has limitations. Bounding boxes are coarse and may cover both manipulated and unmanipulated pixels, so mask supervision is approximate. The explanation component depends on the quality and style consistency of the training annotations. Finally, converting predicted masks to one bounding box can underrepresent multiple manipulated regions; a connected-component post-processing step may improve this for images with several separated artifacts. More broadly, the present manuscript is strongest as a workshop system paper: it introduces the modeling pipeline, design rationale, and implementation choices, but it still needs a more complete quantitative study to support stronger empirical claims.

Table 4: Backbone ablation on the validation split. All variants use the same data split, image size, localization head, and evaluation protocol.

Backbone	Pretrained	Acc.	Fake IoU	Score
ResNet-50	No	0.9875	0.8586	0.6268
Swin-Base	No	0.9986	0.8636	0.6315
Swin-Base	Yes	0.9999	0.9276	0.6638

7 Conclusion

We presented a practical multi-task framework for DDL-X deepfake analysis that jointly addresses image-level detection, manipulated-region localization, and visible-trace explanation within a shared visual architecture. By combining a common backbone with task-specific heads, the proposed approach provides a unified way to model the three core objectives of Track 3 while remaining compatible with the official challenge output format. Experimental results on the held-out validation split show that the framework achieves strong detection and localization performance, and that retrieval-based trace prediction further improves explanation quality without degrading the core visual tasks.

Overall, the results suggest that joint modeling is a promising direction for real-world deepfake analysis. Future work will focus on stronger trace representations, more precise localization of multiple manipulated regions, and a more comprehensive study of the trade-offs between detection accuracy, localization fidelity, and explanation quality.

References

[ALBarqawi *et al.*, 2025] Ahmad ALBarqawi, Mahmoud Nazzal, Issa Khalil, Abdallah Khreishah, and NhatHai Phan. Vigtext: Deepfake image detection with vision-language model explanations and graph neural networks, 2025.

[Dang *et al.*, 2020] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K. Jain. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[Dangovski *et al.*, 2021] Rumen Dangovski, Li Jing, Charlotte Loh, Seungwook Han, Akash Srivastava, Brian Cheung, Pulkit Agrawal, and Marin Soljacic. Equivariant contrastive learning, 2021.

[DDL 2.0 Workshop, 2026] DDL 2.0 Workshop. Track 3: Deepfake detection, localization, and explainability. https://ai-safety-workshop-ijcai2026.github.io/Track3.html, 2026.

[Devillers and Lefort, 2022] Alexandre Devillers and Mathieu Lefort. Equimod: An equivariance module to improve self-supervised learning, 2022.

[Dolhansky *et al.*, 2020] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton-Ferrer. The deepfake detection challenge dataset. *CoRR*, abs/2006.07397, 2020.

[Garrido *et al.*, 2023] Quentin Garrido, Laurent Najman, and Yann LeCun. Self-supervised learning of split invariant equivariant representations. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[Guan *et al.*, 2023] Weinan Guan, Wei Wang, Jing Dong, Bo Peng, and Tieniu Tan. Collaborative feature learning for fine-grained facial forgery detection and segmentation, 2023.

[He *et al.*, 2015] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015.

[Hondru *et al.*, 2025] Vlad Hondru, Eduard Hogea, Darian Onchis, and Radu Tudor Ionescu. Exddv: A new dataset for explainable deepfake detection in video, 2025.

[Lai *et al.*, 2023] Yingxin Lai, Zhiming Luo, and Zitong Yu. Detect any deepfakes: Segment anything meets face forgery detection and localization, 2023.

[Li *et al.*, 2020] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3204–3213, 2020.

[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*, 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, 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, page 13235–13242, New York, NY, USA, 2025. Association for Computing Machinery.

[Nguyen *et al.*, 2019] Huy H. Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. In *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2019.

[Nirkin *et al.*, 2022] Yuval Nirkin, Lior Wolf, Yosi Keller, and Tal Hassner. Deepfake detection based on discrepancies between faces and their context. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6111–6121, 2022.

[Rössler *et al.*, 2019] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *International Conference on Computer Vision (ICCV)*, 2019.

[Wang *et al.*, 2024] Qin Wang, Kai Krajsek, and Hanno Scharr. Equivariant representation learning for augmentation-based self-supervised learning via image reconstruction. In *NeurIPS 2024 Workshop: Self-Supervised Learning - Theory and Practice*, 2024.

[Wang *et al.*, 2026] Qin Wang, Alessio Quercia, Benjamin Bruns, Abigail Morrison, Hanno Scharr, and Kai Krajsek. Self-supervised learning based on transformed image reconstruction for equivariance-coherent feature representation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(31):26425–26434, Mar. 2026.