IMFD: End-to-end Multi-Face Forgery Detection through Instruction-based Large Vision-Language Models

Dasom Choi[1], **Sangjun Moon**[1], **Hyeongchan Im**[1], **Jaeeon Park**[3], *__Jingun Kwon__[1], **Hidetaka Kamigaito**[2], **Taro Watanabe**[2], **and Manabu Okumura**[3]
[1]Chungnam National University
[2]Nara Institute of Science and Technology (NAIST), [3]Institute of Science Tokyo
{dasomchoi, sangjunmoon, hyeongchanim}@o.cnu.ac.kr
jingun.kwon@cnu.ac.kr, {kamigaito.h, taro}@is.naist.jp
park.j.an@m.titech.ac.jp, oku@pi.titech.ac.jp

Abstract

The rapid increase of deepfakes has raised significant concerns due to their spread on social media. Traditional multi-face forgery detectors crop and verify each face independently, ignoring background context and inter-face relationships, which often yields suboptimal performance. To overcome these limitations, we leverage instruction-based Large Vision–Language Models (LVLMs), which can interpret entire images and follow complex textual instructions. We propose a simple yet effective single-stage multi-face forgery detector, called IMFD (Instruction-based Multi-face Forgery Detector), which is trained end-to-end to jointly localize faces and predict per-face forgery labels. Rather than treating face box prediction only as a joint objective, IMFD explicitly integrates predicted face bounding boxes into the instruction as visual cues that enhance instruction grounding and forgery detection. To support the training and evaluation of IMFD, we convert existing multi-face forgery datasets into an instruction-based format. Experimental results and analyses show that IMFD improves multi-face forgery detection by integrating face bounding boxes into the instruction, and consistently outperforms various state-of-the-art methods.

1 Introduction

The rapid advancement of diffusion-based generative models has enabled the synthesis of highly realistic images, audio, and videos from simple text instructions (Dhariwal and Nichol, 2021). While beneficial in industries such as film and advertising (Xu et al., 2024; Wang et al., 2025), these models have also been increasingly abused to produce harmful content, specifically deepfake faces, undermining trust in digital media (Bond, 2022; Li et al., 2022). Therefore, effective and reliable face forgery detection methods are needed.

Traditional approaches typically frame face forgery detection as binary classification and fall into three groups: methods using biological clues such as eye blinking, head poses, and skin texture (Liu et al., 2020), signal-level artifacts to capture synthesis boundaries (He et al., 2021), and data-driven methods that train neural networks on labeled datasets (Ju et al., 2024). Despite strong results in controlled settings, most methods adopt a two-stage pipeline: first cropping each face and then classifying each cropped face independently. This ignores global background context and inter-face relationships, leading to error propagation in scenes with many faces at varying scales and poses (Le et al., 2021).

To address these limitations, the OpenForensics dataset was introduced, enabling end-to-end multi-face forgery detection models (Le et al., 2021). To better distinguish real and forged faces, incorporating bi-grained contrastive learning into the detection backbone has been studied (Zhang et al., 2024). However, these detectors remain object-centric and often miss subtle semantic manipulations. Recently, Large Vision–Language Models (LVLMs) combining large-scale vision encoders and language models have shown strong cross-modal reasoning (OpenAI, 2024). They process whole images and complex prompts, reasoning over both global context and fine-grained details.

Motivated by these advancements, we propose an instruction-based multi-face forgery detector, named **IMFD**. IMFD takes an image and a textual instruction as input, and is trained end-to-end with a joint objective that integrates two subtasks: instruction grounding with forgery detection and face bounding-box prediction. This single-stage framework teaches the model both where to look and what decision to make for each face by injecting predicted face coordinates into the instruction

* Corresponding author.
Our code is available at https://github.com/somcandy08/IMFD.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 3539–3546
November 6-10, 2026 ©2026 Association for Computational Linguistics

Instruction

```
<image>\nThere are {total number of people} people in
the original image and its dimensions are {width in
pixels} pixels in width and {height in pixels} pixels in
height. The coordinates of people's faces at {all face
bounding box coordinates} assign labels sequentially
from left to right starting at 1 only to individuals
whose faces are identifiable. In this case, what are the
numbers of the fake faces and real faces?
```

Response

```
the fake faces are {fake face labels and coordinates}
and the real faces are {real face labels and
coordinates}.
```

Figure 1: Instruction format for multi-face forgery detection using an ordered labeling output schema. "<image>" is a placeholder for the given image.

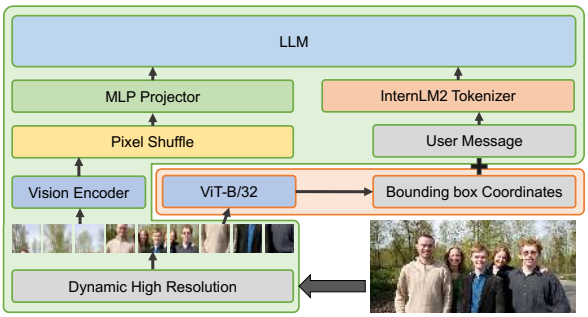


Figure 2: Overview of IMFD. The green block is a standard LVLM architecture such as InternVL. The red block predicts face boxes that are injected into the instruction.

as visual cues. To improve robustness, we incorporate task-relevant details (Misra et al., 2020), such as the number of people, image resolution, and predicted face coordinates, directly into the instruction to emphasize information critical for detecting forged faces without external resources. For analysis, we also compare IMFD with a two-stage setting that uses ground-truth face coordinates. To support the training and evaluation of IMFD, we transform existing multi-face forgery benchmarks into an instruction-based format.

We conduct experiments on the large-scale OpenForensics dataset (Le et al., 2021) and perform in-depth analyses to validate our method. Experimental results show that IMFD consistently outperforms various state-of-the-art baselines. The code and instruction-formatted dataset are publicly available at https://github.com/somcandy08/IMFD.

2 Instruction-based Forgery Dataset

Since no instruction-based datasets exist, we convert existing multi-face forgery datasets into an instruction-based format to train and evaluate IMFD in an instruction-based setting. Each sample pairs an image with an instruction that specifies how to index faces and how to format the output, ensuring consistent prompts and automatic evaluation. The LVLM is expected to identify and label any forged faces. To reduce typical LVLM errors even after fine-tuning (Chen et al., 2024a), we augment the instruction with auxiliary details (Misra et al., 2020).

First, the total number of people is provided for the model to accurately verify face counts. Second, the input image resolution is included to help the model detect potential signs of forgery. Third, the face coordinate information explicitly directs the model's attention to specific facial regions to improve alignment between a given image and textual instructions. Figure 1 shows the instruction format for multi-face forgery detection. With these details incorporated, the LVLM more effectively captures the correspondence between faces and the accompanying instruction. We assign labels to faces from left to right, starting with the numeric number one based on the center coordinates of each face. We then embed these labels into the instruction format. The corresponding binary labels (real or fake) are inherited from the original dataset annotations. We include face bounding boxes in the output as well, enabling the model to better learn spatial grounding.

3 Our End-to-end IMFD

We introduce our end-to-end IMFD. Figure 2 shows an overview of the proposed model, which leverages LVLMs by jointly learning instruction grounding with forged face detection and bounding box prediction. We define $\mathbf{X}_{img}$ as the input from the image modality, and $\mathbf{X}_{txt} = \{[\text{img}], w_1, w_2, ..., w_n\}$ as the input from the textual modality, where $[\text{img}]$ is a special token that will later be replaced by the image embedding.

Vision Encoder. Given $\mathbf{X}_{img}$, we obtain patch embeddings:

$$X_0 = [\mathbf{x}_{\text{cls}}; \text{PatchEmbed}(X_{\text{img}})] + E_{\text{pos}}, \quad (1)$$

$$X'_l = \text{MHA}(\text{LN}(X_{l-1})) + X_{l-1}, (l = 1, ..., L) \quad (2)$$

$$X_l = \text{MLP}_1(\text{LN}(X'_l)) + X'_l, \quad (3)$$

where $\mathbf{x}_{\text{cls}}$ is the special token, and E_{pos} denotes the positional embedding. LN, MHA, and MLP denote layer normalization, multi-head attention, and a feed-forward layer.

Face Bounding Box Prediction. Face bounding boxes are predicted from a single text query,

"A photo of face", without explicit class prediction. Given an image $\mathbf{X}_{img}$ and anchor coordinates $Bboxes$, the procedure is as follows.

Text–Anchor Similarity. A pre-trained CLIP ViT-B/32 encodes the query to a feature F_{query}, which we L_2-normalize (Radford et al., 2021). For each anchor feature $F_{\text{anchor},i}$, we compute the following:

$$S_i = F_{\text{anchor},i} \cdot F_{\text{query}}^{\top}, \quad S = [S_1, \ldots, S_N] \in \mathbb{R}^N, \tag{4}$$

$$F_{\text{anchor},i} = \frac{\text{CLIP}_{\text{img}}(c_i)}{\|\text{CLIP}_{\text{img}}(c_i)\|_2}, \quad (i = 1, \ldots, N) \tag{5}$$

where F_{anchor} is the normalized feature for the i-th anchor, computed by passing the anchor crop c_i from the input image $\mathbf{X}_{img}$ through the pre-trained Vision Transformer CLIP image encoder architecture. Here, N denotes the number of anchors. To obtain the anchor coordinates c_i, we first divide the image into a grid of rectangular regions and record each cell's coordinates as an anchor. For each grid cell, we generate additional anchor boxes by applying various width and height coefficients to scale the region around the cell's center (Radford et al., 2021).

Anchor Selection. Scores are min–max-normalized and anchors whose scores exceed the skip threshold τ_{skip} are retained: $\mathcal{I} = \{\, i \mid S_i \geq \tau_{\text{skip}} \,\}$. If no anchor satisfies the threshold (i.e., $\mathcal{I} = \varnothing$), we select the anchor with the highest score only if the image–text similarity $s_{\text{img}} = F_{\text{img}} \cdot F_{\text{text}}^{\top}$ exceeds the true-positive threshold τ_{tp}, where F_{img} is obtained by passing $\mathbf{X}_{img}$ through the CLIP image encoder. If s_{img} falls below the false-positive threshold τ_{fp}, all boxes are discarded.

Face Box Selection and Refinement. The selected anchors and their scores are refined using Weighted Boxes Fusion (WBF) to remove overlapping anchors: $B_{pred} = \text{WBF}(Bboxes[\mathcal{I}], S;\ \tau_{\text{iou}})$, where τ_{iou} is the IoU threshold used to determine which overlapping boxes are fused together.

Integrating LLM and Vision Encoder. We map image features to the LLM embedding space and replace the `[img]` slot with the projected vector. Specifically,

$$F_{\text{img}} = \text{MLP}_2(X_L[:, 1:, :]), \tag{6}$$

$$X_0 = \text{Embed}(X_{\text{txt}}) + E_{\text{pos}}, \tag{7}$$

$$X_0[i] = F_{\text{img}}, \quad Y_{\text{pred}} = \text{LLM}(X_0), \tag{8}$$

where MLP_2 projects the image embeddings to the same dimensionality as the word embedding space in the LLM. $Embed(X_{txt})$ denotes the token embeddings of the input text, and i is the position of the special `[img]` token appended to the text input. The image representation F_{img} replaces the embedding at this position, allowing the LLM to incorporate visual information during generation.

Objective Function. We optimize the average of text and box losses:

$$\mathcal{L}_{\text{gen}} = \text{CrossEntropy}(Y_{\text{pred}}, Y), \tag{9}$$

$$\mathcal{L}_{\text{bbox}} = \text{L1Loss}(B_{\text{pred}}, B), \tag{10}$$

$$\mathcal{L} = (\mathcal{L}_{\text{gen}} + \mathcal{L}_{\text{bbox}})/2. \tag{11}$$

where Y and B denote the ground-truth text and boxes, respectively.

During training, we use ground-truth face bounding-box coordinates. At test time, we use the predicted face bounding-box coordinates and derive the total number of people, which are then injected into the instruction. Image resolution is read directly from the input image.

4 Experiments

4.1 Experimental Settings

Datasets. We used OpenForensics, a large-scale multi-face forgery detection dataset (Le et al., 2021). The training, validation, and test-development datasets consist of 44,122, 7,308, and 18,895 pairs, respectively. For the test dataset used in the evaluation (ALL), the fake face ratio was 0.58. To investigate the multi-face forgery detection performance of the LVLM on images with a large number of faces, we also conducted evaluations on only 1,522 images (LARGE) that contain more than the average number of faces (2.63) in the test-development dataset.

Implementation Details. We employed InternVL2-2B (Chen et al., 2024b) as the backbone model and fine-tuned it with LoRA (Hu et al., 2022). Detailed settings are provided in Appendix A.

Compared Models. Following the settings in (Yan et al., 2023b; Zhang et al., 2024), we compared our end-to-end **IMFD** with two-stage state-of-the-art (SOTA) baselines including Capsule (Nguyen et al., 2019), UCF (Yan et al., 2023a), EfficientNet-B4 (Tan and Le, 2019), F3Net (Qian et al., 2020), and the single-stage SOTA method, COMICS (Zhang et al., 2024). All baselines were trained on the standard OpenForensics splits, whereas IMFD was trained on our instruction-based variant. We used ground-truth face bounding

Model	ALL $F1_c$	ALL $F1_a$	ALL Acc	LARGE $F1_c$	LARGE $F1_a$	LARGE Acc	ALL Lat. ↓	ALL Max Lat. ↓	ALL Thpt. ↑
Two-Stage Methods									
EfficientNet	0.69	0.29	0.54	0.90	0.46	0.36	**8.0**	363.7	**128.8**
Capsule	0.70	0.29	0.55	0.90	0.47	0.56	14.9	1166.8	67.1
UCF	0.23	0.14	0.06	0.48	0.30	0.09	31.5	5394.2	31.9
F3Net	0.65	0.31	0.35	0.81	0.45	0.44	43.8	939.1	22.8
IMFD	**0.98**†	**0.98**†	**0.94**†	**0.97**†	**0.98**†	**0.87**†	162.8	**263.7**†	6.3
Single-Stage Methods									
COMICS	0.81	0.74	0.40	**0.81**	0.74	0.33	**457.7**	**605.3**	**4.4**
IMFD	**0.90**†	**0.84**†	**0.77**†	0.75	**0.80**†	**0.40**†	565.3	875.2	3.6

Table 1: Comparison with SOTA baselines. Latency (Lat.) and maximum latency (Max Lat.) are reported in milliseconds, and throughput (Thpt.) is the number of images processed per second. † indicates that the improvement is significant (p<0.01) compared with the underlined score (typically the best baseline), as determined by paired bootstrap resampling (Koehn, 2004).

Model	# Peop.	Image Rel.	Coordi.	$F1_c$	$F1_a$	Acc
IMFD	✗	✗	✗	0.91	0.61	0.83
	✓	✗	✗	0.94	0.81	0.87
	✗	✓	✗	0.73	0.56	0.32
	✗	✗	✓	0.96	0.66	0.92
	✗	✓	✓	**0.98**	0.93	**0.96**
	✓	✓	✓	**0.98**	**0.98**	0.94

Table 2: Ablation studies on ALL regarding the effectiveness of integrating auxiliary details into instructions.

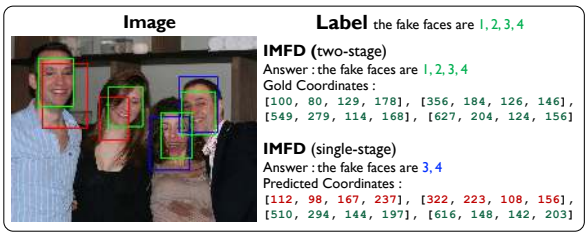


Figure 3: Face bounding-box prediction and forgery detection by IMFD in two-stage and single-stage settings.

boxes for the two-stage setting, while predicted face bounding boxes were used for the single-stage setting. Additional upper-bound results obtained by injecting ground-truth face bounding boxes coordinates into the instruction, along with comparisons across instruction variants, are provided in Appendix B.

Evaluation Metrics. For IMFD, we evaluate micro F1 ($F1_c$) for overall performance, macro F1 ($F1_a$) to balance rare labels, and exact matching (Acc), which counts an image as correct only when the set of predicted fake faces exactly matches the ground truth. For two-stage models, we collected the independently detected results to calculate $F1_c$, $F1_a$, and Acc. For the single-stage COMICS model, to evaluate detection results as a classification problem, we assign a fake face when a predicted bounding box overlaps with a ground-truth box by an IoU threshold of 0.5 or higher.

Results. As shown by the results in Table 1, IMFD outperforms the other models in both two- and single-stage settings. This demonstrates the effectiveness of IMFD and the instruction-based dataset. While there is a trade-off between computational cost and detection performance, IMFD achieves significantly lower maximum latency compared to all two-stage methods in the two-stage setting. This is because, when a large number of faces are present in a single image, IMFD avoids the costly and sequential steps of face cropping and per-face detection required by traditional methods. However, the single-stage IMFD, which uses predicted face bounding boxes, shows a noticeable drop in performance compared to the two-stage IMFD that leverages ground-truth bounding boxes.

Error Analysis. We calculated COCO-style Average Precision (AP) for face bounding box prediction (Lin et al., 2014). The AP score of 81.9 for the ALL dataset highlights the importance of high-quality face localization for the single-stage IMFD in Table 1. We also confirmed that our method can predict face bounding boxes for a large number of faces in a single image, achieving an AP of 83.6 for the LARGE dataset. Figure 3 provides a qualitative comparison between the two-stage and single-stage settings. The two-stage IMFD accurately detects all fake faces using the ground-truth face coordinates, while the single-stage IMFD only identifies a subset of fake faces based on the predicted bounding boxes. This highlights that errors in bounding box prediction for single-stage IMFD can directly affect the accuracy of fake face detection.

Impact of Auxiliary Details. We analyzed the effectiveness of auxiliary details in instructions for multi-face forgery detection. Table 2 presents the

results obtained using IMFD. Even without auxiliary details, IMFD ($F1_c$) substantially outperforms the best two-stage baselines in Table 1. Solely injecting image resolution was not effective. Resolution alone lacks spatial correspondence and thus cannot satisfy the left-to-right labeling required by instruction-based detection. A more detailed analysis of the effect of input image resolution is provided in Appendix C. Additional Grad-CAM case studies are provided in Appendix D.

5 Conclusion and Discussion

We proposed IMFD, a single-stage end-to-end instruction-based detector that jointly localizes faces and reasons about their authenticity using LVLMs. We convert existing multi-face forgery datasets into an instruction-based format with auxiliary details. On OpenForensics, IMFD consistently outperforms strong two-stage and single-stage baselines. In the future, we will further investigate various LVLM backbones, improve bounding-box quality, and extend the framework to video.

Limitations

Our study has several limitations. First, experiments are limited to OpenForensics because other multi-face forgery benchmarks are not publicly available. Second, the single-stage setting is sensitive to face localization quality, and errors in predicted bounding boxes directly affect forgery detection performance. Third, we focus only on image-based forgery detection and do not address temporal consistency in videos. In addition, the relatively small number of zero-fake samples may limit robustness in such cases. Future work includes evaluation on more diverse benchmarks, improving box prediction quality, and extending the framework to video.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-2026-25569031, Development of an AIBOM-Based Platform for Strengthening the Reliability and Security of the AI Supply Chain). This study was financially supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (RS-2026-25580044). This work was supported by the Korea Basic Science Institute (KBSI) Young Researcher Infrastructure Support Program funded by the Ministry of Science and ICT (MSIT) (Grant No. RS-2026-25473607). Following are results of a study on the "Convergence and Open Sharing System" Project, supported by the Ministry of Education and National Research Foundation of Korea.

References

Shannon Bond. 2022. The latest marketing tactic on LinkedIn: AI-generated faces. NPR.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. 2024a. DRESS: Instructing large vision-language models to align and interact with humans via natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14239–14250.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024b. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34.

Yang He, Ning Yu, Margret Keuper, and Mario Fritz. 2021. Beyond the spectrum: Detecting deepfakes via re-synthesis. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 2534–2541. International Joint Conferences on Artificial Intelligence Organization.

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

Yan Ju, Shan Jia, Jialing Cai, Haiying Guan, and Siwei Lyu. 2024. GLFF: Global and local feature fusion for AI-synthesized image detection. *IEEE Transactions on Multimedia*, 26:4073–4085.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Trung-Nghia Le, Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. 2021. OpenForensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10117–10127.

Changjiang Li, Li Wang, Shouling Ji, Xuhong Zhang, Zhaohan Xi, Shanqing Guo, and Ting Wang. 2022. Seeing is living? rethinking the security of facial liveness verification in the deepfake era. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 2673–2690, Boston, MA. USENIX Association.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755. Springer International Publishing.

Zhengzhe Liu, Xiaojuan Qi, and Philip H. S. Torr. 2020. Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8060–8069.

Kanishka Misra, Allyson Ettinger, and Julia Rayz. 2020. Exploring BERT's sensitivity to lexical cues using tests from semantic priming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4625–4635, Online. Association for Computational Linguistics.

Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. 2019. Capsule-Forensics: Using capsule networks to detect forged images and videos. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2307–2311.

OpenAI. 2024. GPT-4 technical report. *Preprint*, arXiv:2303.08774. Version 6.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. 2020. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *Computer Vision – ECCV 2020*, pages 86–103. Springer International Publishing.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626.

Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR.

Haohan Wang, Wei Feng, Yaoyu Li, Zheng Zhang, Jingjing Lv, Junjie Shen, Zhangang Lin, and Jingping Shao. 2025. Generate e-commerce product background by integrating category commonality and personalized style. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Meng Xu, Tong Zhang, Fuyun Wang, Yi Lei, Xin Liu, and Zhen Cui. 2024. MPDS: A movie posters dataset for image generation with diffusion model. *Preprint*, arXiv:2410.16840.

Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. 2023a. UCF: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22412–22423.

Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. 2023b. DeepfakeBench: A comprehensive benchmark of deepfake detection. In *Advances in Neural Information Processing Systems*, volume 36.

Cong Zhang, Honggang Qi, Shuhui Wang, Yuezun Li, and Siwei Lyu. 2024. COMICS: End-to-end bi-grained contrastive learning for multi-face forgery detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10):10223–10236.

A Implementation Details

We employed InternVL2-2B (Chen et al., 2024b) as the backbone model and fine-tuned it with LoRA (Hu et al., 2022). Table 5 summarizes the LVLM backbones used in our experiments, including their parameter sizes and underlying LLMs. Experiments were conducted using the Transformers library with a random seed of 42 on an NVIDIA A6000 GPU. The input image resolution was fixed at 448×448. We set τ_{iou} and τ_{skip} to 0.01 and 0.8, respectively. We set τ_{tp} and τ_{fp} to 0.0 and -2.0, respectively.

Figure 4 shows the three instruction settings used in the two-stage IMFD analysis. Setting (1) uses the instruction without auxiliary details. Setting (2) adds only the total number of people. Setting (3) incorporates the total number of people, image resolution, and face coordinates. These settings are used in the Grad-CAM case study to analyze how auxiliary details affect instruction grounding and multi-face forgery detection.

B Upper-bound Analysis

To analyze the upper bound of instruction-based multi-face forgery detection independently of face localization errors, we additionally evaluate several LVLM backbones using ground-truth face

bounding-box coordinates injected into the instruction. This setting removes the effect of bounding-box prediction noise and allows us to compare instruction variants under controlled conditions. We consider three instruction variants: (1) the base instruction without auxiliary details, (2) the instruction augmented with only the total number of people, and (3) the instruction augmented with the total number of people, image resolution, and face coordinates.

Table 3 shows that augmenting the instruction with auxiliary details consistently improves performance across LVLM backbones. In particular, instruction variant (3), which includes both global and spatial cues, achieves the best overall performance in most cases. This confirms that the benefit of auxiliary details is not limited to a specific model family. We also observe that increasing the LLM size does not always lead to better performance. For example, InternVL2-2B remains competitive with, and in some cases outperforms, larger variants such as InternVL2-4B and InternVL2-8B. These results suggest that performance in this task depends more on how effectively the model aligns visual faces with the instruction than on model scale alone.

Model	Inst.	ALL			LARGE		
		$F1_c$	$F1_a$	Acc	$F1_c$	$F1_a$	Acc
LLaVA-1.5 -Vicuna	(1)	0.77	0.41	0.56	0.84	0.52	0.43
	(2)	0.73	0.56	0.32	0.93	**0.81**	**0.64**
	(3)	**0.91**†	**0.92**†	**0.78**†	**0.94**	**0.81**†	0.55
LLaVA-1.6 -Vicuna	(1)	0.86	0.52	0.74	0.89	0.57	0.55
	(2)	0.73	0.56	0.32	0.93	**0.81**	0.64
	(3)	**0.96**†	**0.94**†	**0.90**†	**0.94**†	**0.81**†	**0.76**†
InternVL2 -2B	(1)	0.91	0.61	0.83	0.94	0.65	0.74
	(2)	0.94	0.81	0.87	0.96	0.90	0.79
	(3)	**0.98**†	**0.98**†	**0.94**†	**0.97**†	**0.98**†	**0.87**†
InternVL2 -4B	(1)	0.91	0.66	0.84	0.94	0.77	0.73
	(2)	0.95	0.78	0.88	0.95	0.78	0.76
	(3)	**0.96**†	**0.91**†	**0.89**†	**0.95**	**0.96**†	**0.80**†
InternVL2 -8B	(1)	0.91	0.65	0.84	0.94	0.76	0.76
	(2)	0.94	0.77	0.86	**0.95**	0.90	0.77
	(3)	**0.96**†	**0.98**†	**0.90**†	**0.95**	**0.97**†	**0.80**†
MiniCPM -Llama3-V-v2.5	(1)	0.53	0.15	0.25	0.50	0.18	0.00
	(2)	0.73	0.56	0.32	0.92	0.79	0.59
	(3)	**0.90**†	**0.95**†	**0.77**†	**0.95**	**0.97**†	**0.80**†
MiniCPM -V-v2.6	(1)	0.49	0.11	0.26	0.43	0.13	0.00
	(2)	0.72	0.46	0.29	**0.93**	0.81	0.64
	(3)	**0.91**†	**0.96**†	**0.77**†	0.92	**0.95**†	**0.70**†

Table 3: Upper-bound results on **OpenForensics** obtained by injecting ground-truth face bounding-box coordinates into the instruction. (1), (2), and (3) denote the three instruction variants described in the main text. **ALL** and **LARGE** denote evaluation on the full test set and on the subset of images containing more than the average number of faces (2.63), respectively. † indicates that the improvement is significant ($p<0.01$) compared with the underlined score, as determined by paired bootstrap resampling (Koehn, 2004).

C Impact of Input Image Resolution

To provide a more detailed analysis of image resolution, we evaluate IMFD under different input resolutions using InternVL2-2B as the backbone and instructions augmented with ground-truth face coordinates. Table 4 reports the results on ALL and LARGE. As the input resolution increases from 112×112 to 672×672, the performance improves consistently across all evaluation metrics.

In particular, on ALL, the $F1_c$ and $F1_a$ scores reach 0.981 and 0.985, respectively, at 672×672. A similar trend is observed on LARGE. These results suggest that higher-resolution inputs provide more discriminative visual details for instruction-based multi-face forgery detection.

Image Rel.	ALL			LARGE		
	$F1_c$	$F1_a$	Acc	$F1_c$	$F1_a$	Acc
112x112	0.9173	0.9414	0.6538	0.8894	0.9401	0.6538
224x224	0.9414	0.9492	0.7424	0.9260	0.9566	0.7424
448x448	0.9762	0.9846	0.9359	0.9744	0.9838	0.8739
672x672	**0.9810**	**0.9849**	**0.9484**	**0.9808**	**0.9841**	**0.9008**

Table 4: Impact of input image resolution on IMFD using InternVL2-2B.

D Grad-CAM Case Study

We evaluated whether instruction-based LVLMs can parse multi-face images with left-to-right numeric labels by visualizing IMFD Grad-CAM heat maps on test samples (Selvaraju et al., 2017). Figure 5 shows the results. (1) uses the original instructions without auxiliary details, (2) adds only the total number of people, and (3) incorporates the number of people, image resolution, and face coordinates. In example (A), the model's attention aligns with its predictions: with (1) and (2), it focuses on and labels only the first two forged faces, while with (3), attention extends to all three faces. Example (B) further shows that, when using (3), the LVLM with auxiliary details can localize and detect fake faces even in images with many faces.

Model	Params	Backbone LLM	HuggingFace Repository
LLaVA-1.5	7B	vicuna-7b-v1.5	https://huggingface.co/liuhaotian/llava-v1.5-7b
LLaVA-1.6	7B	vicuna-7b-v1.5	https://huggingface.co/liuhaotian/llava-v1.6-vicuna-7b
InternVL2-2B	2.2B	internlm2-chat-1_8b	https://huggingface.co/OpenGVLab/InternVL2-2B
InternVL2-4B	4.2B	Phi-3-mini-128k-instruct	https://huggingface.co/OpenGVLab/InternVL2-4B
InternVL2-8B	8.1B	internlm2_5-7b-chat	https://huggingface.co/OpenGVLab/InternVL2-8B
MiniCPM-Llama3-V-2.5	8.1B	Llama3-8B-Instruct	https://huggingface.co/openbmb/MiniCPM-Llama3-V-2_5
MiniCPM-V-2.6	9.5B	Qwen2-7B-Instruct	https://huggingface.co/openbmb/MiniCPM-V-2_6

Table 5: LVLM backbones used in this study, including parameter sizes, backbone LLMs, and model repositories for reproducibility.

Image	Inst.	Prompt
	(1)	**Instruction :** Label the faces in the given photo as follows: Assign labels to the faces from left to right, starting with 1. Do not assign labels to the backs of heads; only label visible faces. If there's no forgery face, just only give answer like this: None. If there is a forgery face, just only give answer like this: The fake face is {label numbers of the photo with Forgery faces}.
		Response : the fake faces are 4, 5 and the real faces are 1, 2, 3, 6.
	(2)	**Instruction :** There are 6 people in total. Assign labels sequentially from left to right starting at 1 only to individuals whose faces are identifiable, what are the numbers of the fake faces and real faces?
		Response : the fake faces are 4, 5 and the real faces are 1, 2, 3, 6.
	(3)	**Instruction :** There are 6 people in the original image and its dimensions are 1024 pixels in width and 768 pixels in height. The coordinates of people's faces at [121, 303, 75, 94], [235, 207, 63, 90], [362, 237, 48, 64], [724, 252, 66, 80], [456, 294, 45, 45], [588, 255, 42, 46], assign labels sequentially from left to right starting at 1 only to individuals whose faces are identifiable. In this case, what are the numbers of the fake faces and real faces?
		Response : the fake faces are 4[456, 294, 45, 45], 5[588, 255, 42, 46] and the real faces are 1[121, 303, 75, 94], 2[235, 207, 63, 90], 3[362, 237, 48, 64], 6[724, 252, 66, 80].

Figure 4: Face forgery detection instructions and responses.

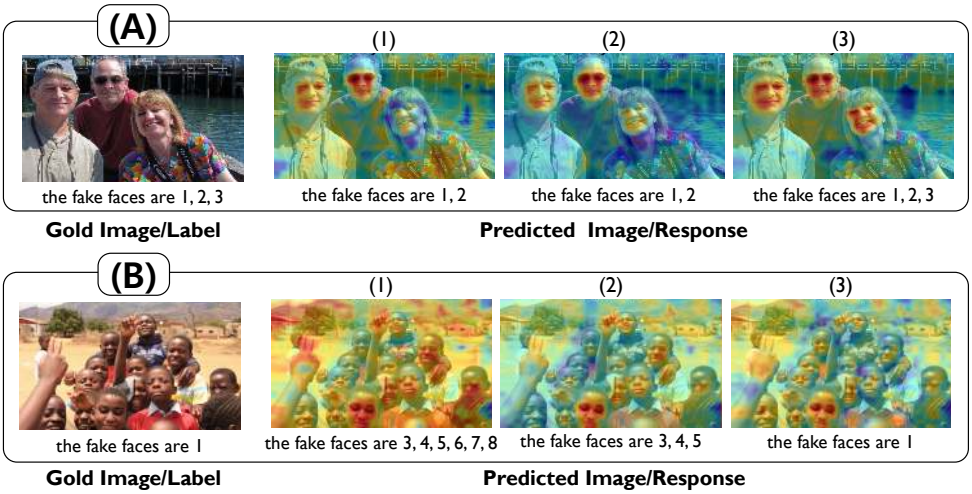


Figure 5: Original and Grad-CAM visualizations for the 18th layer on images containing forged faces.