Deepfake-Agent: Aggregating Semantic Forgery Clues for Generalizable Detection

Supplementary Material

1. Implementation Details

Our Deepfake-Agent framework integrates three specialized MLLM-based forensic experts. For the Image Editing Expert ($\mathcal{E}_{\text{edit}}$) and Face-Swap Expert ($\mathcal{E}_{\text{face}}$), we emply FakeShield [6] and M2F2-Det [2], respectively. These two experts are used without modifications to ensure fair comparison with baseline methods. Also, we build up our semantic forgery expert and aggregator on LLaVA-7b [4] and LLaVA-34b [4], respectively. Both of these two models are fine-tuned using our proposed reinforcement learning objectives as described in Sec. **??**. To assess the large model's capability in aggregating different forgery experts, we also evaluate the performance of LLaVA-110B [3] in Tab. **??**. All these pre-trained model weights are summarized as below:

- FakeShield huggingface.co/zhipeixu/fakeshield-v1-22b
- M2F2-Det huggingface.co/CHELSEA234/llava-v1.5-7b-M2F2-Det
- LLaVA-7b huggingface.co/llava-hf/llava-1.5-7b-hf
- LLaVA-34b huggingface.co/liuhaotian/llava-v1.6-34b
- LLaVA-110b huggingface.co/lmms-lab/llava-next-110b

The Semantic-Forgery Expert ($\mathcal{E}_{\text{sem}}$) comprises two MLLM proposers: GPT-4o [1] and Gemini-2.0 Pro [5]. We employ their respective API endpoints with the same configurations: we set the temperature parameter to 0.1 to ensure deterministic and focused responses while maintaining some reasoning flexibility. The maximum output tokens are configured at 512. Also, we use standardized prompts for these two models to ensure the format is consistent, which will be further illustrated in Sec . 2.

All experiments are conducted on a cluster of 8 NVIDIA A6000 GPUs with 48GB VRAM each. Training the semantic-forgery expert requires approximately 242 hours for 4 epochs, while the forgery aggregator training takes approximately 48 hours for 3 epochs. Inference throughput averages 2.3 seconds per image during evaluation, with batch size set to 1 to accommodate the large memory usage caused by multiple MLLMs. Our implementation is based on PyTorch 2.1.0 with CUDA 11.8 support. We utilize the Transformers library (version 4.35.0) for all MLLM components. For reinforcement learning optimization, we extend the GRPO implementation with custom reward functions tailored for forensic tasks. Additional dependencies include OpenCV 4.8.0 for image processing and NumPy 1.24.0 for numerical operations.

2. Prompt Engineering

Each component in our Deepfake-Agent framework employs specialized prompts. For the MLLM proposers in the semantic-forgery expert, we employ a standardized prompt that systematically guides the analysis of semantic inconsistencies. This structured prompt ensures consistent output formats while systematically guiding the models to evaluate multiple dimensions of semantic integrity, allowing each MLLM to leverage its unique reasoning capabilities and world knowledge. The exact prompt structure used for both GPT-4o and Gemini-2.0 Pro is:

Forgery Analysis Prompt

```
Analyze this image for potential digital
forgeries by examining aspects such as
Do objects follow physical laws?  Does
the scene make real-world sense?  Are
there contradictions between different
objects?
Provide analysis in exact format:
Decision:  [real/fake]
Reasoning:  [2-3 sentence explanation]

Now analyze the provided image:
```

Adjudicator's Initial Prompt

```
You are a forensic adjudicator
synthesizing multiple proposers'
opinions:
Expert Analyses:
Proposer 1 Analysis:
Decision:  {T_GPT-4o_decision}
Reasoning:  {T_GPT-4o_reasoning}
Proposer 2 Analysis:
Decision:  {T_Gemini_decision}
Reasoning:  {T_Gemini_reasoning}
Your task:
• Compare and evaluate both analyses
• Identify the most compelling evidence
  and possible different opinions
• Provide unified decision with bounding
  box
Output in exact format:
Decision:  [real/fake]
Reasoning:  [detailed synthesis]
Bounding Box:  [x1,y1,x2,y2]

Now synthesize the expert opinions:
```

The adjudicator in the semantic-forgery expert employs a multi-stage prompting strategy that integrates outputs from both proposers. The initial adjudication prompt combines visual input with textual hypotheses from the proposers. The subsequent causal self-verification phase uses a counterfactual prompt to validate the forgery evidence:

Adjudicator's Self-Verification Prompt

```
Previous Analysis:
Decision: {previous_decision}
Reasoning: {previous_reasoning}
Bounding Box: {previous_bbox}
Counterfactual Assessment:
Temporarily ignore the region specified
by the bounding box and re-evaluate the
image.
Critical questions:
• Does the image appear authentic
  without this region?
• Is this evidence causally necessary
  for your decision?
Provide analysis in exact format:
Decision: [real/fake]
Reasoning: [causal explanation]
Verification: [confirmed/refuted]

Now perform the counterfactual
assessment:
```

The forgery aggregator utilizes a system prompt that structures the fusion of multiple expert opinions, ensuring that the aggregator produces well-structured outputs with clear reasoning traces. The prompt is as follows:

Aggregator Prompt

```
You are a forensic aggregator analyzing
outputs from three experts:
Expert Outputs:
[PS Expert]: {T_ps}
[Face Expert]: {T_face}
[Semantic Expert]: {T_semantic}
Analyze their arguments and provide your
final assessment:
• Compare the expert reasoning and
  identify the most credible analysis
• Resolve any conflicts between expert
  opinions
• Provide your final verdict with clear
  justification
Output format:
Final Decision: [real/fake]
Integrated Reasoning: [detailed
explanation]

Now synthesize the expert analyses:
```

Table 1. Single proposer in the semantic-forgery expert. Performance measured on Deepfake-Eval-2024 test set. [Key: Acc.: acurracy, F1: F1 score; Pre.: Precision; Rec.: Recall.]

Configuration	F1	Prec.	Rec.	Acc.
GPT-4o	79.3	83.5	75.6	76.2
Gemini-2.0 Pro	77.8	81.2	74.8	74.9
Both proposers	**82.6**	**88.1**	**79.2**	**79.6**

Table 2. Comparison of different aggregation strategies on the multi-domain forgery test set. [Key: Acc.: acurracy, F1: F1 score; Pre.: Precision; Rec.: Recall.]

Aggregation	Acc.	F1	Pre.	Rec.
Majority voting	72.5	82.1	70.1	77.4
Confidence weighting	75.1	82.2	77.4	67.2
RL-based (ours)	**84.2**	**90.9**	**90.4**	**91.6**

3. Additional Ablation Studies

We conduct additional ablation experiments to further validate our design choices and understand the contribution of each component.

Semantic Expert with a Single Proposer. We feed the trained aggregator in the proposed semantic-forgery expert with responses from only one proposer. As shown in Tab. 1, using both GPT-4o and Gemini-2.0 Pro as proposers achieves the best performance (82.6 F1-score), outperforming single-proposer configurations. The GPT-4o only configuration achieves 79.3 F1-score, while Gemini-2.0 Pro alone reaches 77.8 F1-score, demonstrating the complementary nature of different MLLMs' world knowledge and reasoning styles.

Alternative Aggregation Strategies. We compare against several baselines: majority voting, confidence-weighted averaging, and a transformer-based fusion network. As shown in Tab. 2, our RL-based aggregator achieves superior performance (84.2% accuracy, 90.9 F1-score) compared to these alternatives.

Self-Verification Iteration Number. We experiment with maximum iterations ranging from 1 to 5 and find that 3 iterations provide the optimal balance between localization accuracy and computational efficiency. Beyond 3 iterations, we observe diminishing returns with minimal improvement in evidence localization quality while inference time increases linearly. This configuration achieves 87% human-evaluated localization quality while maintaining reasonable inference latency of 2.1 seconds per image.

4. Additional Visualizations

Self-Verification. Fig. 1 illustrates additional examples where the red bounding boxes gradually localize the forgery evidence. It is worth mentioning that in the last example,

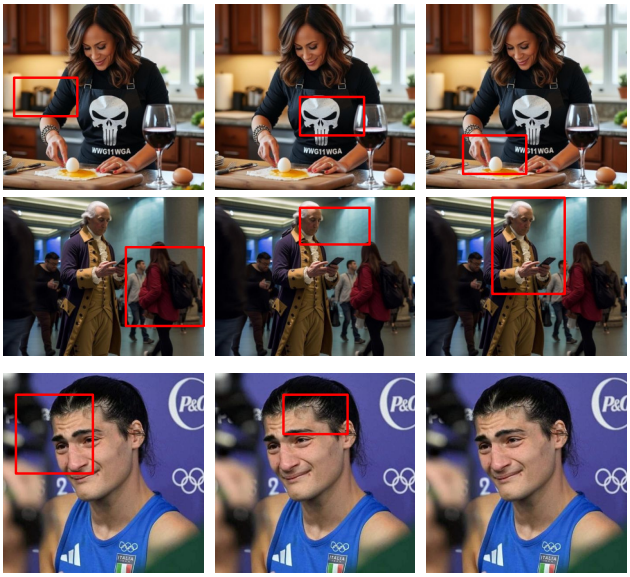

Figure 1. Additional self-verification localization examples.

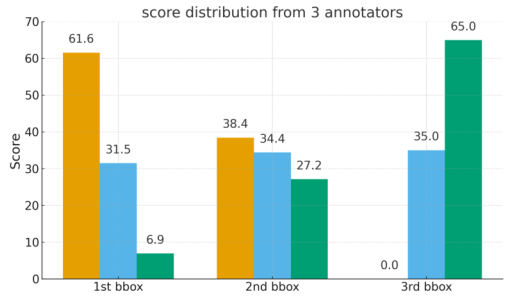

Figure 2. Human assessment scores distribution. The x-axis denotes 3 different iterations.

our proposed method claims forgery in the first two rounds, but in the third round, it changes its decision back to the real and returns the image without a bounding box.

Human Assessment. Tab. 1 in the main paper shows each round's annotation qualities, and Fig. 2 illustrates the detailed human preference scores regarding each iteration of forgery evidence. We observe that the annotator's opinion diverges at the first and second rounds since the forgery localization quality can be similar sometimes, while all the third round's bounding boxes are regarded as the best in describing the forgery evidence.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[2] 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. 1

[3] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[4] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024. 1

[5] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 1

[6] Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. *arXiv preprint arXiv:2410.02761*, 2024. 1