

Omni-Fake: Benchmarking Unified Multimodal Social Media Deepfake Detection

Supplementary Material

Overview of Supplementary Material

In this supplementary document, we provide:

- **S1. Extended Experimental Results:** extra quantitative results, including explanation quality, replay ratio ablation, and RL reward design.
- **S2. Dataset statistics:** core statistics of the OMNI-FAKE benchmark across modalities and splits.
- **S3. Implementation Settings:** key details of the training setup for SFT and GSPO-based RL.
- **S4. Case Studies and Representative Samples:** qualitative examples with visualizations of masks, intervals, textual explanations, and representative samples from OMNI-FAKE-SET and OMNI-FAKE-OOD across all modalities.

S1. Extended Experimental Results

We report additional experiments on explanation quality, the effect of replay ratios during multimodal training, and the RL reward used for alignment.

S1.1 Explanation Study

We evaluate explanations using ROUGE-L (longest common subsequence F-measure), cosine semantic similarity (CSS) between sentence embeddings, and human expert ratings on a 1–5 scale for factual correctness and usefulness. ROUGE-L reflects lexical and structural overlap, CSS captures semantic similarity, and human scores provide a direct assessment of explanation quality.

Table 1 reports results for images, audio, generic videos, and audio–visual talking-head clips. CSS is high across modalities, ROUGE-L is moderate due to paraphrasing, and human scores are above 4 on average, indicating that explanations are generally accurate and informative.

Table 2. Replay ratio ablation on the Audio→Image two-modality training setup.

Replay Ratio	Avg. ACC (↑)	Avg. IoU (↑)
0%	70.19	31.72
5%	75.35	33.42
10%	78.14	36.07
15%	80.45	39.49
30%	79.86	37.31

Table 1. Explanation quality across modalities.

Modality	Automatic metrics		Human experts
	ROUGE-L (↑)	CSS (↑)	Mean score (↑)
Image	0.41	0.79	4.39
Audio	0.32	0.70	4.27
Video	0.37	0.72	4.21
AV-TH	0.45	0.86	4.52

S1.2 Replay Ratio Ablation

We study replay in a two-modality curriculum (Audio → Image). The model is first trained on audio only, then on images while replaying a proportion $p \in \{0\%, 5\%, 10\%, 15\%, 30\%\}$ of audio data. We evaluate the final model with the average detection ACC and average localization IoU over both modalities.

Table 2 shows that very small replay (0–5%) slows but does not prevent forgetting; a large ratio (30%) protects early modalities but harms learning of later ones; ratios around 10–15% give the best trade-off. We therefore use 15% replay in all experiments.

S1.3 RL Reward Design

In the RL alignment stage, we optimize a composite reward

$$r(x, y) = \lambda_{\text{fmt}} r_{\text{fmt}} + \lambda_{\text{acc}} r_{\text{acc}} + \lambda_{\text{bbox}} r_{\text{bbox}} + \lambda_{\text{int}} r_{\text{int}}, \quad (1)$$

where:

- r_{fmt} checks output format `<think>` and `<answer>` tags and field validity.
- r_{acc} measures global classification correctness for REAL / TAMPERED / FULLY SYNTHETIC or REAL / FAKE.
- r_{bbox} scores spatial localization via box IoU.
- r_{int} scores temporal localization via interval IoU.

We set

$$\lambda_{\text{fmt}} = 0.3, \quad \lambda_{\text{acc}} = 1.0, \quad \lambda_{\text{bbox}} = 1.0, \quad \lambda_{\text{int}} = 1.0,$$

balancing structural correctness and global decisions, while putting stronger weight on localization quality. This configuration yields stable RL training and consistent gains in detection and localization.

S2. Dataset Statistics

We summarize core statistics of OMNI-FAKE across four modalities (images, audio, videos, and audio–visual talking-head clips) and three label types (REAL, FULLY SYNTHETIC, TAMPERED), for both the in-distribution OMNI-FAKE-SET and out-of-distribution OMNI-FAKE-OOD splits.

These statistics show that OMNI-FAKE is large-scale, spans multiple modalities and manipulation types, and includes a substantial OOD split, making it suitable for evaluating unified multimodal deepfake detectors under distribution shifts.

S3. Implementation Settings

All experiments are conducted on a single node with 4×NVIDIA H20 96GB GPUs using PyTorch, DeepSpeed ZeRO-2 and FlashAttention-2. Our base model is Qwen/Qwen2.5-Omni-7B, which is first fine-tuned with LoRA rank 16, $\alpha = 32$, dropout 0.05 on the merged Omni-Fake SFT dataset. We then apply GSPO-based reinforcement learning on the RL-formatted multimodal data, using the composite reward described above. Hyperparameters follow standard large-model training practice and emphasize stability rather than aggressive tuning. The complete training process requires approximately 100 GPU-hours on the H20 system.

S4. Case Studies

We present qualitative case studies across all modalities to illustrate how our unified multimodal detector reasons about real, fully synthetic, and tampered media. The examples cover challenging boundary cases and highlight the model’s strengths in fine-grained spatial localization, temporal interval detection, and detailed natural-language explanations. Across modalities, our method consistently identifies subtle inconsistencies such as texture misalignment, unnatural temporal dynamics, or cross-modal desynchronization while avoiding false alarms on high-quality real content.

For images and videos, our detector produces accurate bounding boxes on small manipulated regions and explains the visual cues behind each decision. Audio and AV-talking-head cases demonstrate the model’s ability to detect synthetic speech artifacts, temporal editing, and audio–visual mismatch. These examples show that the model not only outputs correct labels but also provides grounded, interpretable reasoning aligned with human perception. Such qualitative evidence complements our quantitative results and demonstrates the robustness and transparency of our unified approach.

In addition, we also present representative samples from OMNI-FAKE-SET and OMNI-FAKE-OOD to illustrate the visual and distributional diversity of the benchmark, as well as the high quality of the underlying data.

Limitation and future work. While Omni-Fake covers four major modalities, it does not yet include some emerging formats such as 3D avatars or multilingual speech synthesis, which may become increasingly relevant as generative models advance. In addition, although the benchmark incorporates diverse manipulation types, the landscape of generative technologies evolves rapidly, and newly emerging manipulation styles may still fall outside its current scope. We view these points as natural directions for future expansion to keep Omni-Fake aligned with the growing diversity of real-world multimodal deepfakes.

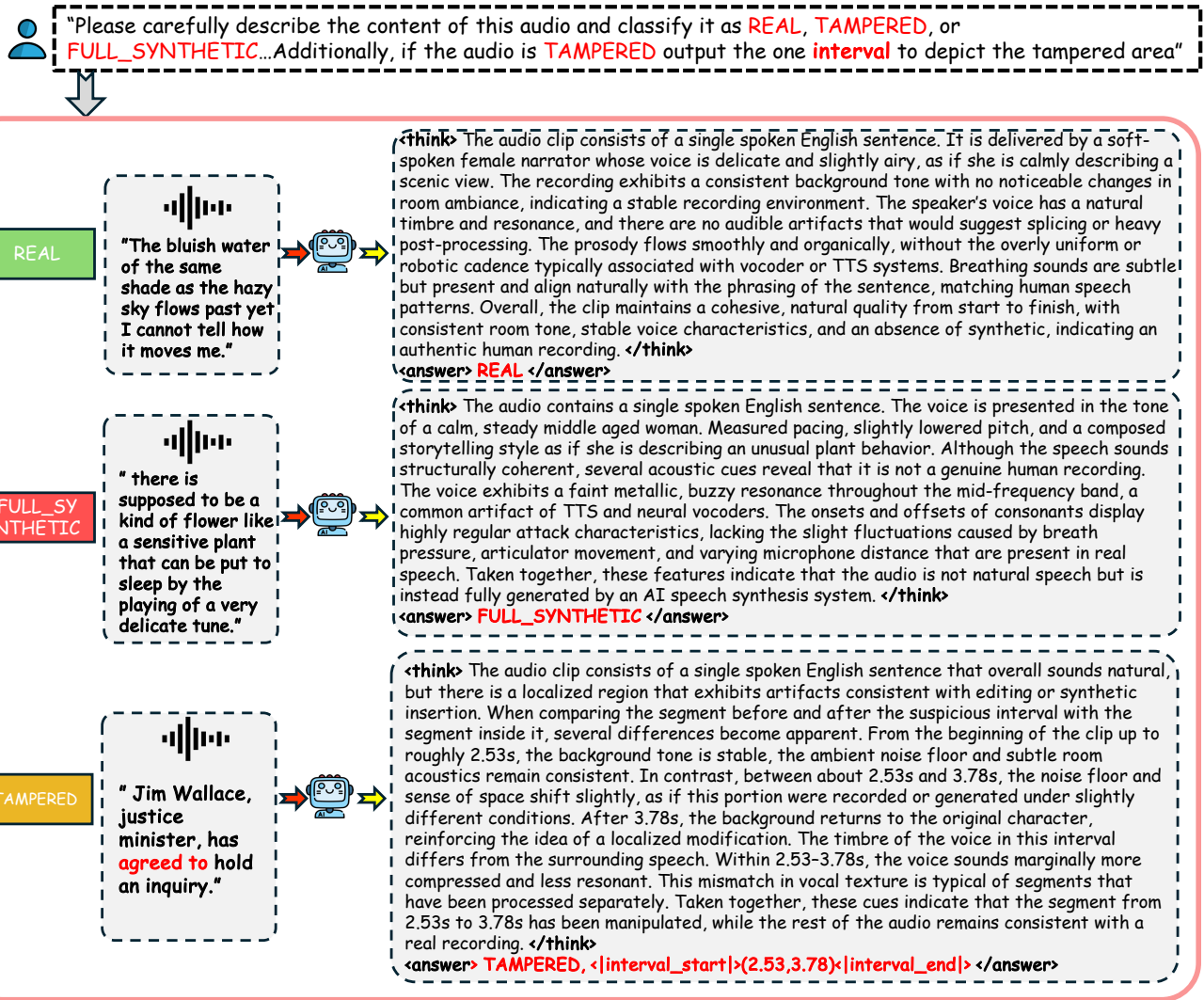


Figure 3. Audio case studies with REAL, TAMPERED, and FULLY SYNTHETIC examples. Forged temporal intervals are highlighted for TAMPERED audio, together with predictions and explanations.

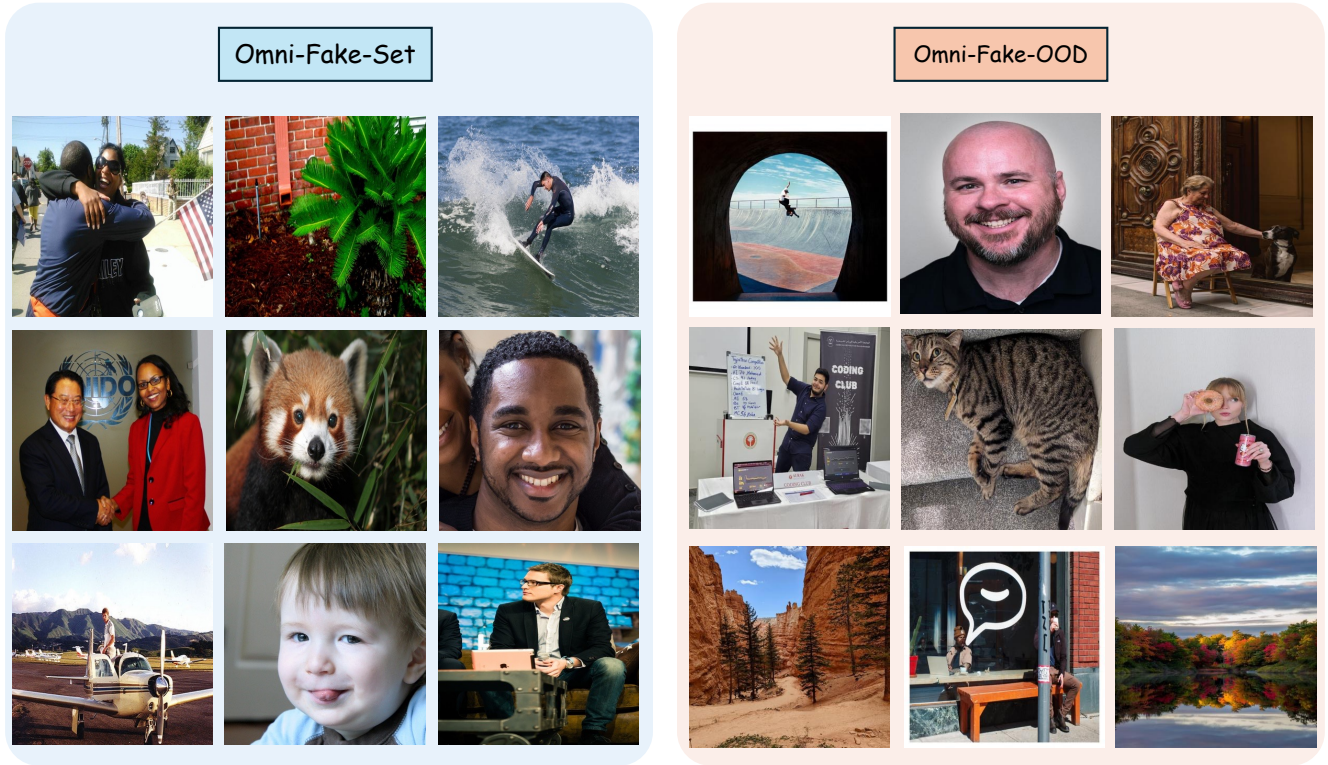


Figure 5. Representative REAL sample from the OMNI-FAKE dataset.



Figure 6. Representative FULL_SYNTHETIC sample generated by modern diffusion-based models in OMNI-FAKE.

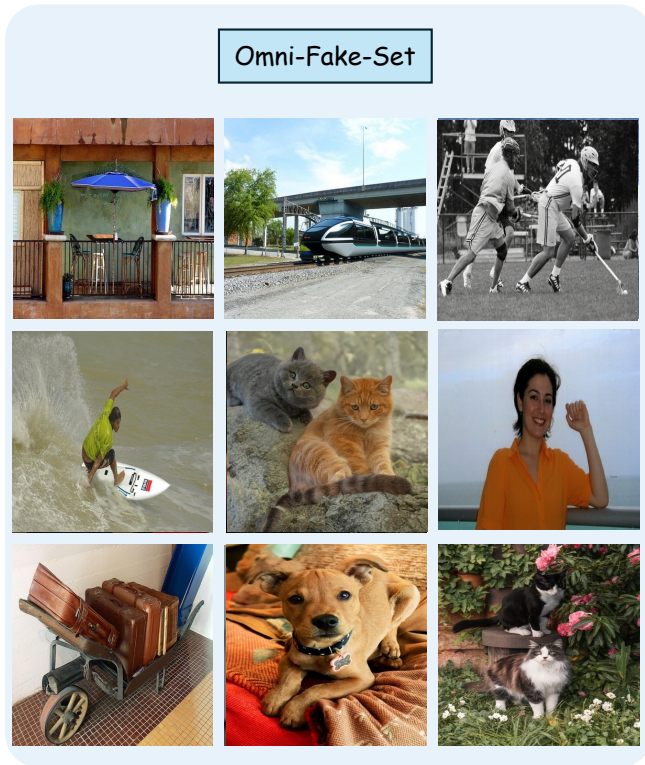


Figure 7. Representative TAMPERED sample containing localized manipulations in OMNI-FAKE.

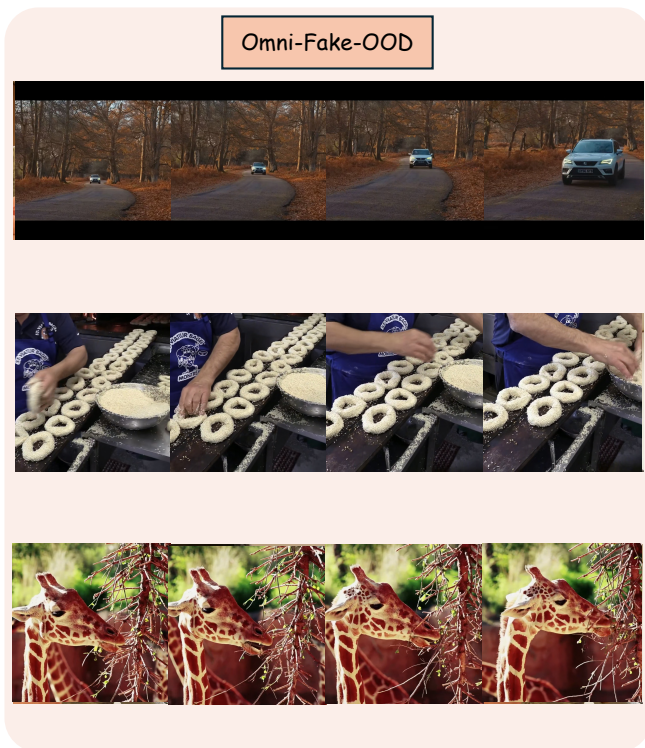
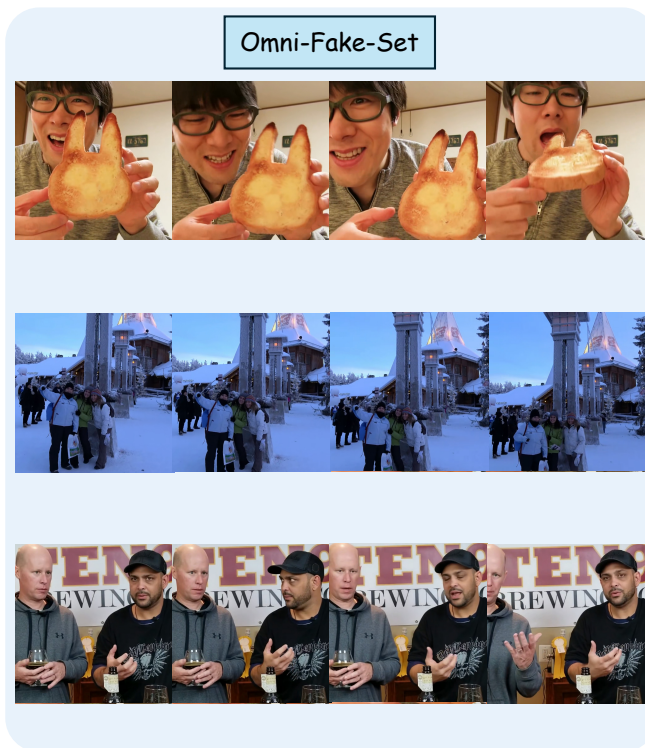


Figure 8. Representative REAL video sample from OMNI-FAKE. The frames illustrate high-quality authentic motion and natural temporal dynamics.

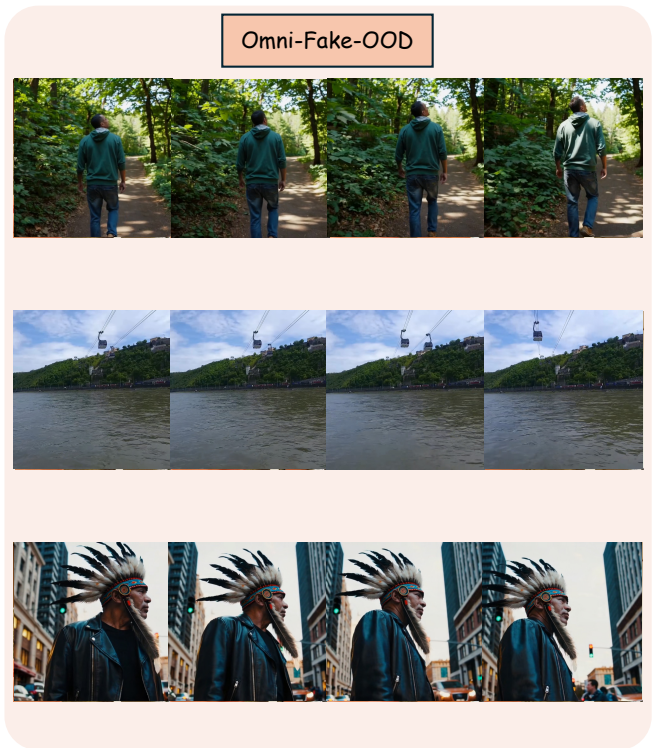
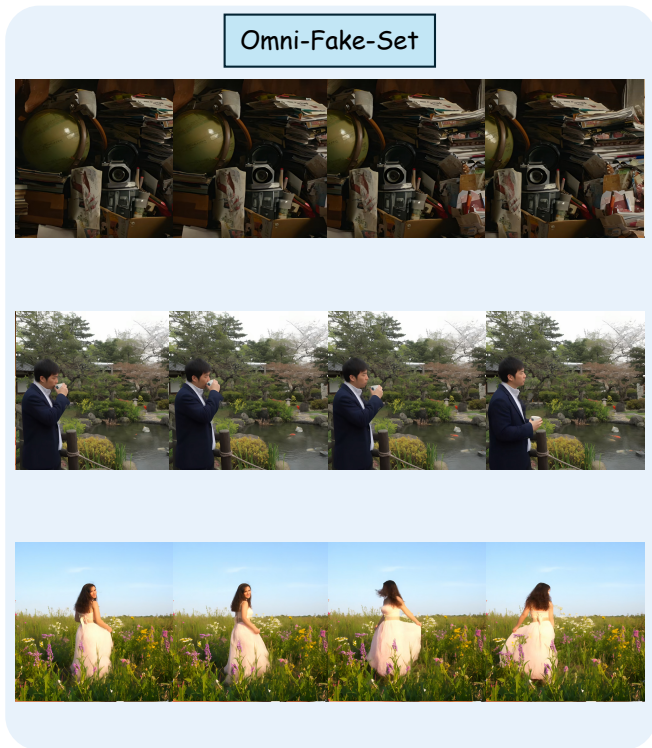


Figure 9. Representative FULLY SYNTHETIC video sample from OMNI-FAKE. This example reflects typical AI-generated motion patterns and texture consistency.

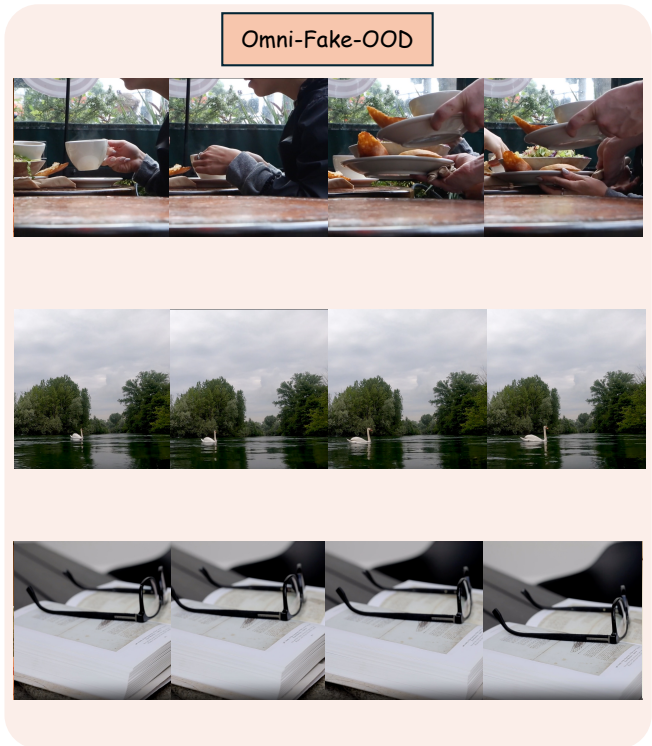


Figure 10. Representative TAMPERED video sample from OMNI-FAKE. Only part of the temporal sequence is manipulated while the rest remains authentic.

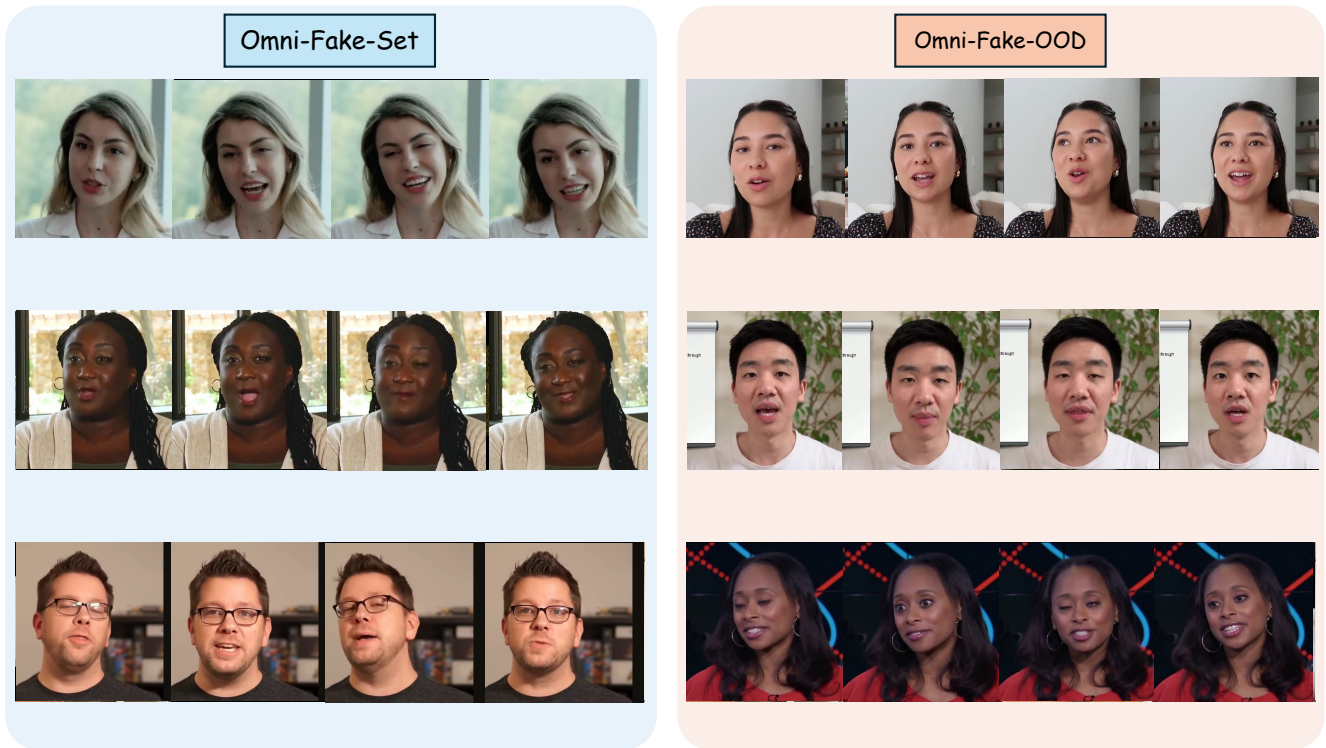


Figure 11. Representative **REAL** audio-visual talking-head samples from OMNI-FAKE-SET (left) and OMNI-FAKE-OOD (right). Samples show high visual clarity, diverse recording environments, and consistent lip-audio synchronization.

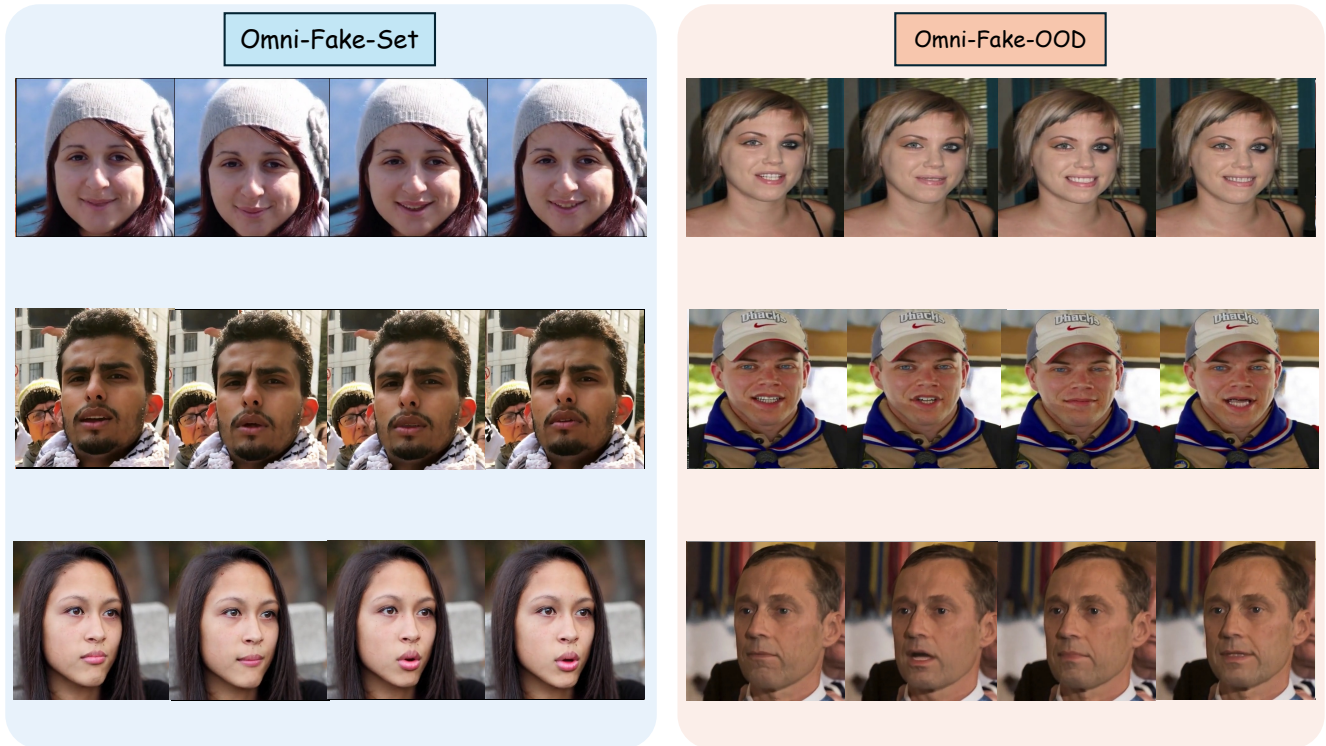


Figure 12. Representative **FULLY SYNTHETIC** audio-visual talking-head samples from OMNI-FAKE-SET (left) and OMNI-FAKE-OOD (right). Synthetic samples exhibit high realism across identity appearance.