Agent4FaceForgery: Multi-Agent LLM Framework for Realistic Face Forgery Detection

Yingxin Lai[1], Zitong Yu[1*], Jun Wang[1*], Linlin Shen[2], Yong Xu[3], Xiaochun Cao[4]
[1]Great Bay University
[2]Shenzhen University
[3]Harbin Institute of Technology
[4]School of Cyber Science and Technology, Sun Yat-sen University
[*]Corresponding authors.

Abstract

Face forgery detection faces a critical challenge: a persistent gap between offline benchmarks and real-world efficacy, which we attribute to the ecological invalidity of training data. This work introduces Agent4FaceForgery to address two fundamental problems: (1) how to capture the diverse intents and iterative processes of human forgery creation, and (2) how to model the complex, often adversarial, text-image interactions that accompany forgeries in social media. To solve this, we propose a multi-agent framework where LLM-powered agents, equipped with profile and memory modules, simulate the forgery creation process. Crucially, these agents interact in a simulated social environment to generate samples labeled for nuanced text-image consistency, moving beyond simple binary classification. An Adaptive Rejection Sampling (ARS) mechanism ensures data quality and diversity. Extensive experiments validate that the data generated by our simulation-driven approach brings significant performance gains to detectors of multiple architectures, fully demonstrating the effectiveness and value of our framework.

Introduction

The rise of generative technologies (Rombach et al. 2022; Li et al. 2024) enables the creation of hyper-realistic face forgeries. While useful for creative purposes, these forgeries present serious risks like misinformation and fraud (Yan et al. 2023c; Wang et al. 2024b). Consequently, there is an urgent need for robust, accurate detection methods that can generalize to new forgery techniques in the wild.

Despite progress in detection algorithms, a critical bottleneck remains: the persistent gap between performance on benchmark datasets and efficacy in real-world online environments. We posit that this gap stems fundamentally from the ecological invalidity of current training data and evaluation paradigms. Existing datasets (Rossler et al. 2019a; Dolhansky et al. 2020, 2019), even multimodal ones (Huang et al. 2024a,b; Khan and Dang-Nguyen 2024), typically consist of curated, static examples that fail to capture the complex, dynamic lifecycle of face forgeries in the wild. Specifically, they fail to represent the intent and process of human-driven forgery, overlooking how real forgeries are created by humans with diverse motivations, skills, and stylistic preferences. Current datasets rarely reflect this diversity of intent or the iterative process of creation. Furthermore, they lack adequate representation of social context and multimodal interaction. Forgeries online are not consumed in isolation. This data realism gap is the central barrier to developing generalizable and deployable forgery detectors.

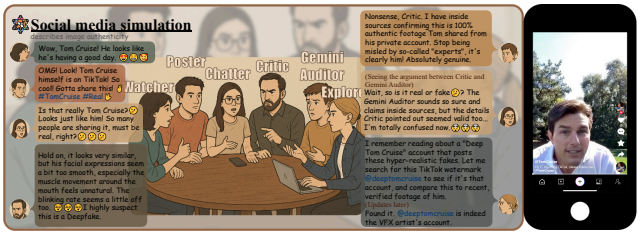

Figure 1: The illustrative example of our proposed agent-based social simulation. Diverse agents engage in a human-like deliberation on the image's authenticity.

To directly address this critical need for ecologically valid training data, we introduce Agent4FaceForgery, a framework using an LLM-powered multi-agent system to simulate the entire forgery lifecycle. The framework's Multi-Agent Simulation Core first captures the nuance of human forgery: each agent uses a Profile to define its intent, a Memory for iterative learning, and an Action module to perform visual edits and generate text, thereby simulating the adaptive process of human creators. To ensure the quality of this output, Adaptive Rejection Sampling (ARS) acts as a dynamic filter, scoring forgeries using both agent self-assessments and external detectors. Its difficulty threshold tightens over time to progressively select higher-quality, more challenging examples. Finally, a Multi-Role Social Simulation generates realistic context by having agents with different roles (e.g., Critic, Auditor) interact with the forgeries, as shown in Fig. 1. Here, agents with different roles (e.g., Creator, Critic, Auditor) interact with a forged image, generating realistic comments and claims. This process creates a stream of context-aware training data reflecting real-world social dynamics. Critically, it enables us to construct training samples based on text-image consistency, not just image authenticity, providing the challenging data needed for robust multimodal detectors. Our main contributions include:

- We introduce a novel multi-agent LLM framework specifically designed to simulate the face forgery lifecycle, generating realistic multimodal training data that captures human intent, process, and social context.
- We propose specific mechanisms (Agent profiles/memo-

ry/actions) to address the ecological invalidity of current forgery datasets, bridging the gap between offline training and real-world detection challenges.

- We experimentally validate that training detectors (CLIP and MLLMs) on data generated by Agent4FaceForgery significantly improves their generalization, and interpretability.

Related Work

Face Forgery Detection. Early deepfake detection works often uses CNNs like Xception (Rossler et al. 2019b) and ResNet (He et al. 2016) for binary classification based on visual features (Yang, Li, and Lyu 2019; Li, Chang, and Lyu 2018). While successful within datasets, these methods struggle with cross-domain generalization due to overfitting specific generation artifacts. To improve robustness, research shifted towards universal traces, particularly in the frequency domain. SPSL (Liu et al. 2021) learning from DCT representations, and M2TR (Wang et al. 2022) using frequency decomposition with cross-attention. However, newer generative models with less pronounced frequency artifacts challenge these approaches. Consequently, recent strategies explore reconstruction-based methods like RECCE (Cao et al. 2022a), which identify inconsistencies during image reconstruction, and multimodal approaches like SIAF (Huang et al. 2024a), integrating LLMs and textual context for better robustness and interpretability. This evolution reflects the ongoing adaptation required to counter advancing forgery techniques.

LLM-Based Simulation . Intelligent agents, which perceive, act autonomously, and learn, are increasingly capable thanks to LLMs enabling human-like behaviors. Applications are widespread, including social simulation (e.g., AI Town (Park et al. 2023)), task automation (e.g., AutoGen (Wu et al. 2023)), collaborative task decomposition (MetaGPT (Hong et al. 2023)), linguistic style replication (SV (Yang et al. 2024)), recommendation simulation (Agent4Rec (Zhang et al. 2024)), and value modeling (ALI-Agent (Wang et al. 2024a)). However, limitations persist. In face forgery detection, using single LLMs (like GPT-4 (Achiam et al. 2023)) for annotation (Huang et al. 2024a; Wang et al. 2019) is prone to hallucination, compromising reliability for realistic fakes. Critically, existing approaches often lack collaborative multi-agent structures necessary for intricate tasks like forgery annotation. Advanced multi-agent systems are key to achieving greater accuracy and interpretability in practice.

Method

Problem Formulation

The core challenge hindering the real-world efficacy of forgery detectors is the scarcity of high-quality, multimodal training data that mirrors real-world complexity. Our goal is to synthesize such data through simulation. Formally, given an unaltered real face image $\mathbf{x}$ and an optional text description $\mathbf{c}$, we aim to generate a large-scale, comprehensive multimodal forgery dataset $\mathcal{D}$. This dataset is composed of two sample types:

- **Real Samples**: $\{(\mathbf{x}_i, \mathbf{c}_i, y_i = 0, \delta_i = 1)\}_{i=1}^{M}$
- **Forged Samples**: $\{(\mathbf{x}'_j, \mathbf{c}'_j, y_j = 1, \delta'_j)\}_{j=1}^{N}$

where M and N are the total numbers of real and forged samples. For each sample, $\mathbf{x}$ is the image, $\mathbf{c}$ is the textual description, $y \in \{0, 1\}$ is the image authenticity label ($y = 1$ for forged), and $\delta \in \{0, 1\}$ is the text-image consistency label. A consistency label of $\delta = 1$ confirms an accurate correspondence, while $\delta = 0$ flags a mismatch (e.g., a forged image with a misleading description). Critically, our framework generates forged samples with both matching ($\delta'_j = 1$) and mismatching ($\delta'_j = 0$) text descriptions. The final output is a dataset $\mathcal{D}$ of context-aware samples, mitigating the bottleneck of scarce ecologically valid training data.

Agent Architecture

Generating high-quality multimodal data with complex social context is challenging because: 1) a single error or "hallucination accumulation" in intermediate steps can invalidate the entire sample, and 2) the content of each interaction depends on the forgery details and the creator's original intent, making it hard to maintain these complex dependencies consistently. The core insight of our approach is to separate the task generation process into two distinct phases: first creating a realistic "forgery blueprint" of the task (Phase 1), and then using this blueprint for the generation of naturalistic multi-turn conversations that fill in the conversational details (Phase 2). This separation allows us to ensure both the correctness of the underlying task structure and the naturalness of the resulting conversations.

Phase 1: Generating Forged Blueprints

The goal of this phase is to generate a "forged blueprint" for each simulated forgery. A blueprint consists of two core elements: a high-quality forged image $\mathbf{x}'$ and an initial, creator-generated textual description $\mathbf{c}'$. Together, these form the basis for the subsequent social simulation in Phase 2. To generate these blueprints, we design autonomous agents with a sophisticated cognitive architecture. Each agent's behavior is governed by three interconnected modules, with GPT-4V acting as the unified cognitive core.

Profile Module. The agent's profile is a cornerstone for emulating the nuanced behaviors of real-world human creators. To ground our simulation, we initialize each agent's profile $\mathbf{p}_k$ by analyzing the FF++ benchmark dataset. The profile consists of two components: a vector of quantifiable traits $\mathbf{v}_k$ and a natural language description of a creator's stylistic taste $\mathbf{c}_k$.

Quantifiable Traits ($\mathbf{v}_k$). To formalize an agent's behavioral tendencies, we define the set of forgeries created by a specific creator k in the dataset as forgeries_k. The traits are:

- **Forgery Frequency (T_k^{freq}):** An agent's overall productivity, defined as the total count of their works: $T_k^{\text{freq}} = |\text{forgeries}_k|$.
- **Methodological Diversity (T_k^{div}):** The variety of manipulation techniques an agent uses. This is the count of unique methods $\{\text{method}_i\}$ found across all of the

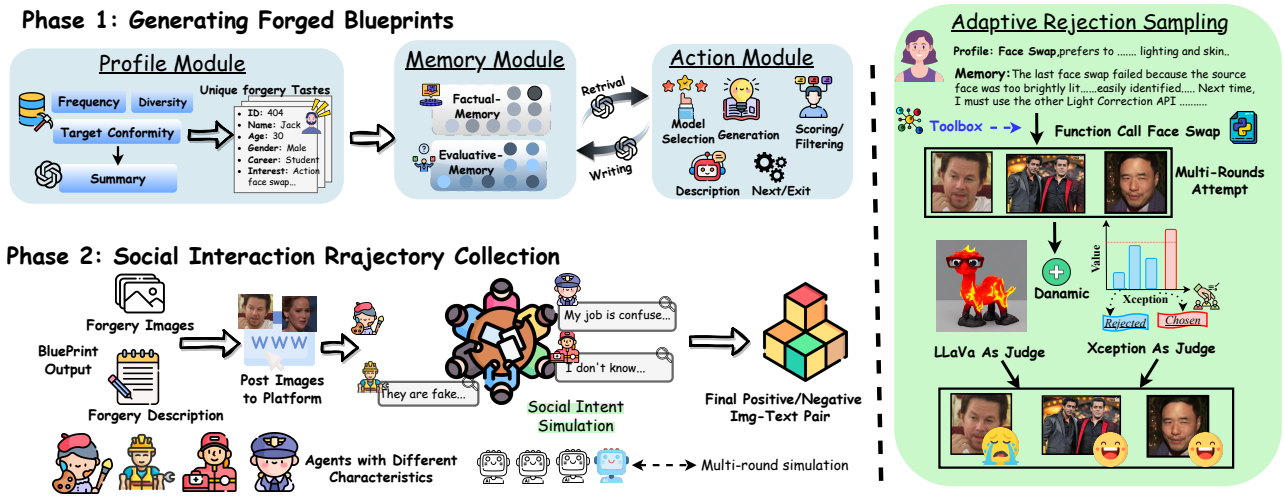

Figure 2: The overall framework of Agent4FaceForgery. Our simulator consists of two core facets: LLM-empowered Generative Agents and a Media Presentation Environment. Agent profiles are initialized using datasets characterizing real media. Agents, enhanced with specialized memory and action modules tailored for media evaluation and authenticity assessment scenarios, simulate a wide range of behaviors including viewing content, evaluating authenticity, flagging suspicious items, sharing media, ignoring content, and potentially participating in discussion.

agent's forgeries $i \in \text{forgeries}_k$:

$$T_k^{\text{div}} = \left| \bigcup_{i \in \text{forgeries}_k} \{\text{method}_i\} \right|. \quad (1)$$

- **Target Conformity (T_k^{conf}):** An agent's inclination to select popular targets. For each forgery i, target_i is the targeted identity. We define a popularity function $\text{Pop}(\cdot)$ as the total number of times a target is manipulated in the entire dataset. The conformity is the average popularity of an agent's chosen targets:

$$T_k^{\text{conf}} = \frac{1}{|\text{forgeries}_k|} \sum_{i \in \text{forgeries}_k} \text{Pop}(\text{target}_i). \quad (2)$$

These three metrics form the numerical vector $\mathbf{v}_k$. For the qualitative component $\mathbf{c}_k$, we prompt GPT-4V to generate a stylistic preference description by analyzing a sample of L forgeries from a creator's work, where L is a predefined hyperparameter. Together, they form the agent's complete "forgery gene," $\mathbf{p}_k = (\mathbf{v}_k, \mathbf{c}_k)$, which drives its decision-making.

Memory Module. To enable continuous refinement, the memory module maintains two memory types: *factual memory* for objective details of past edits, and *evaluative memory* for subjective assessments (e.g., seam visibility). Memories are logged in structured JSON, containing not only operational data but also cognitive states like high-level plans. This supports three key operations: memory retrieval, writing, and an LLM-driven reflection process to analyze outcomes and guide future actions.

Action Module. The Action Module translates intent into action. An agent's action at time t, denoted $\text{Action}_k^{(t)}$, is a pair $(\text{Edit}(\cdot), \text{Desc}(\cdot))$ consisting of a visual edit and a textual description. The visual edit is a sequential application of operators:

$$\text{Edit}(\mathbf{x}; \mathbf{p}_k, \mathcal{M}_k) = O_n(\dots O_1(\mathbf{x}; \theta_1) \dots; \theta_n), \quad (3)$$

where each operator O_i is chosen from a toolbox $\mathcal{O}_{ops}$ (including methods like Flux Pro and Deepfake APIs), and its parameters θ_i are determined by the agent's profile $\mathbf{p}_k$ and its memory $\mathcal{M}_k$. The textual description $\text{Desc}(\cdot)$ can be either an accurate caption or an intentionally misleading statement.

Adaptive Rejection Sampling (ARS). The agents leverage their cognitive architecture within an iterative algorithm to synthesize and curate the final blueprints. To ensure the generated data is both diverse and challenging, we use an ARS mechanism as a quality control gate. A candidate blueprint $(\mathbf{x}_i', \mathbf{c}_i')$ is scored using a fused metric s_i:

$$s_i = \lambda\, s_i^{\text{LLM}} + (1 - \lambda)\, s_i^{\text{disc}}, \quad (4)$$

where s_i^{disc} is the score from an external forgery detector, s_i^{LLM} is the agent's internal quality assessment based on its memory, and λ is a weighting hyperparameter.A sample is accepted if its score s_i exceeds an adaptive threshold τ. To ensure the mechanism is logically sound and robust, the threshold determination includes an initial warm-up phase. For the first N_warmup samples, we employ a fixed, lenient threshold, τ_warmup, to build a diverse initial pool of accepted samples. After the warm-up phase, the threshold τ becomes fully data-driven and is periodically updated to the q-th quantile of the scores of all previously accepted samples:

$$\tau = \text{Quantile}(\{s_j \mid j \in \text{Accepted Samples}\}, q), \quad (5)$$

where $\{s_j\}$ is the set of scores from all accepted samples and $q \in [0, 1]$ is the single hyperparameter defining the rejection rate.

Data Generation Pipeline. The generation process unfolds in five structured steps:

1. **Multi-round Forgery.** Given a set of real face images $\mathcal{X}_{\text{real}}$, the agent's *Action Module* generates a candidate by first formulating and then executing a multi-step forgery plan. For each random selected image $\mathbf{x} \in \mathcal{X}_{\text{real}}$, the agent constructs a sequence of operators $(\mathcal{O}_1, \ldots, \mathcal{O}_n)$ by sampling from a comprehensive toolbox $\mathcal{O}_{\text{ops}}$. This toolbox is populated with diverse forgery instruments, categorized into:
   - **Identity Manipulation:** Face-swapping methods (e.g., based on DeepFaceLab, FaceSwap).
   - **Attribute & Expression Editing:** GAN-based editors for expressions or attributes like age and gender (e.g., StarGAN, AttGAN).
   - **Style-Based Synthesis:** GAN-inversion techniques that allow for fine-grained stylistic edits (e.g., SBI).

   The selection of this operator chain is a probabilistic process, directly governed by the agent's *Profile Module* ($\mathbf{p}_k$). The agent maintains a probability distribution over the tools in $\mathcal{O}_{\text{ops}}$, where the likelihood of choosing a specific tool is weighted by its stylistic taste ($\mathbf{c}_k$) and methodological diversity trait (T_k^{div}). For instance, an agent profiled for high-realism forgeries might have a higher probability of first selecting a face-swap operator, and then selecting a blending operator as a second step.

   The execution of this plan, $\mathbf{x}'_i = \mathcal{O}_n(\ldots \mathcal{O}_1(\mathbf{x}) \ldots)$, produces the final forged image. Following this, a textual corresponding forgery description $\mathbf{c}'_i = \text{Desc}(\cdot)$ is generated, and the agent assigns the image-text consistency label $\delta'_i \in \{0, 1\}$ based on its strategic intent.
2. **Scoring and Filtering.** Each candidate is evaluated by two models pre-trained on the FF++ dataset. Xception and LLaVA, computes the agent's internal quality and consistency score $s_i^{\text{LLM}} = f_{\text{agent}}(\mathbf{x}'_i, \mathbf{c}'_i, \mathcal{M}_k)$. These scores are fused into a single metric s_i.Determine whether they are challenge samples based on the confidence score and the probability value of the LLaVA text output.
3. **Memory Update.** Both successful and rejected forgery attempts are logged into the *Memory Module*. After a fixed number of iterations, the agent initiates a reflection phase, analyzing past outcomes to refine its forgery techniques, e.g., blending or style adjustments, for subsequent rounds.
4. **Output Data.** The retained forged samples, denoted as $\{(\mathbf{x}'_i, \mathbf{c}'_i, y_i = 1, \delta'_i)\}$, are combined with the real face samples from $\mathcal{X}_{\text{real}}$, which are formatted as $\{(\mathbf{x}_j, \mathbf{c}_j, y_j = 0, \delta_j = 1)\}$, to form the final multimodal dataset $\mathcal{D}$.

Phase 2: Social Interaction Trajectory Collection

Positive-Negative Sample Construction. Traditional approaches to data construction for forgery detection frequently depend on binary "real" or "fake" labels assigned at the image level, often disregarding the pivotal role of social discourse and user reactions in the dissemination of deepfakes. To address this shortcoming, we propose an innovative framework that simulates multi-user interactions within a social media environment, thereby aligning with the realistic dynamics of the "social interaction–deepfake" interplay.

Socal Intent Simulation. Drawing inspiration from observable behaviors on social media platforms, we introduce five distinct user roles, each powered by MLLMs. These roles engage with forged images through a variety of actions, such as viewing, commenting, sharing, and labeling, thereby emulating a wide range of user interactions:

- **Watcher**: Frequently designates content as "liked" or "interesting" but rarely investigates its authenticity.
- **Explorer**: Compares multiple posts related to the same event, enhancing the probability of detecting forgery artifacts through comparative analysis.
- **Critic**: Emphasizes quality and credibility, often highlighting suspicious forgeries in comments or reports.
- **Chatter**: Susceptible to misinformation due to social influences, yet capable of correction through group discussions.
- **Poster**: Reposts or re-edits content, amplifying the propagation of forged images across platforms.

Collectively, these roles simulate a rich and varied set of user interactions, yielding textual responses that closely resemble those observed in authentic social media contexts.

Hard Negative Generation : To amplify the complexity of text-image inconsistencies, we introduce the *Gemini Auditor*, a specialized role engineered to produce intentionally deceptive statements. For example, the Gemini auditor might designate an evidently spliced image as "100% authentic" or manipulate attributes such as gender or identity, thereby injecting ambiguity into the deepfake scenario. These adversarial assertions facilitate the creation of robust negative samples by inducing pronounced conflicts between text and image content, compelling detection models to confront deceptive pairings and bolstering their resilience.

Social Environment Labeling. The interactions among the Watcher, Explorer, Critic, Chatter, Poster, and Gemini roles generate a diverse collection of comments, labels, and edits for each forged image. When integrated with ground-truth labels (e.g., real or fake), these interactions enable the automated construction of positive and negative sample pairs based on the consistency between text and image content. Negative samples arise in cases such as a forged image ($y = 1$) paired with a claim asserting it is "perfectly real," or a real image ($y = 0$) accompanied by a declaration of "obvious forgery." Conversely, positive samples emerge when the text accurately reflects the image's authenticity or when social feedback, such as a Critic identifying artifacts, rectifies an initial mislabeling.We formalize this labeling process with the function $\delta(x', c')$:

$$\begin{cases} 1, & \text{if } y = 1 \text{ and } c' \text{ claims "perfectly real"}, \\ 1, & \text{if } y = 0 \text{ and } c' \text{ claims "obvious forgery"}, \\ 0, & \text{otherwise}, \end{cases} \tag{6}$$

where $\delta = 1$ denotes a negative sample indicative of a text–image mismatch, and $\delta = 0$ signifies alignment or a corrected positive sample. This methodology capitalizes on so-

Table 1: **Frame-level** cross-database evaluation from FF++(HQ) to DFD, DFDC-P, Wild Deepfake, and Celeb-DF in terms of AUC (%) and EER (%). The FF++ results represent intra-domain performance, while others represent generalization to unseen domains. The **best** results are indicated in bold, and the second-best results are underlined.

Method	*FF++*		DFD		DFDC-P		Wild Deepfake		Celeb-DF	
	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER
Xception (Chollet 2017)	99.09	3.77	87.86	21.04	69.80	35.41	66.17	40.14	65.27	38.77
EN-b4 (Tan and Le 2019b)	99.22	3.36	87.37	21.99	70.12	34.54	61.04	45.34	68.52	35.61
Face X-ray (Tan and Le 2019b)	87.40	-	85.60	-	70.00	-	-	-	74.20	-
F3-Net (Qian et al. 2020a)	98.10	3.58	86.10	26.17	72.88	33.38	67.71	40.17	71.21	34.03
MAT (Nguyen et al. 2019)	99.27	3.35	87.58	21.73	67.34	38.31	70.15	36.53	70.65	35.83
GFF (Luo et al. 2021)	98.36	3.85	85.51	25.64	71.58	34.77	66.51	41.52	75.31	32.48
LTW (Sun et al. 2021)	99.17	3.32	88.56	20.57	74.58	33.81	67.12	39.22	77.14	29.34
LRL (Chen et al. 2021)	99.46	3.01	89.24	20.32	76.53	32.41	68.76	37.50	78.26	29.67
DCL (Sun et al. 2022)	99.30	3.26	91.66	16.63	76.71	31.97	71.14	36.17	82.30	26.53
PCL+I2G (Zhao et al. 2021)	99.11	-	-	-	-	-	-	-	81.80	-
SBI (Shiohara and Yamasaki 2022)	88.33	20.47	88.13	17.25	76.53	30.22	68.22	38.11	80.76	26.97
UIA-ViT (Zhuang et al. 2022)	-	-	94.68	-	75.80	-	-	-	82.41	-
RECCE (Cao et al. 2022b)	99.32	3.38	89.91	19.95	75.88	32.41	67.93	39.82	70.50	35.34
UCF (Yan et al. 2023b)	97.05	-	80.74	-	75.94	-	-	-	75.27	-
FFTG (Sun et al. 2025)	99.23	3.12	**94.79**	15.31	84.74	23.43	83.55	24.40	84.80	22.73
Ours	**99.50**	**2.97**	93.25	**13.04**	**88.10**	**19.19**	**86.50**	**21.87**	**87.10**	**20.12**

cial interactions to produce nuanced, context-sensitive labels that mirror the intricacies of real-world scenarios.

Experiments

Experimental Settings

Datasets. Following standard practice (Qian et al. 2020b; Yan et al. 2024a; Cao et al. 2022a), we built our model based on FF++ (Rossler et al. 2019a). We evaluate performance on several benchmarks, including Celeb-DF (V2) (Li et al. 2020), DFDC (Dolhansky et al. 2020), WildDeepfake (Zi et al. 2020), DFD (Dolhansky et al. 2020), and DFDC-P (Dolhansky et al. 2019). Robustness is further assessed using the DF40 protocol (Yan et al. 2024a; vlf 2025).

Implementation Details. We use LLava (Liu et al. 2024), as the backbone model, and RetinaFace (Deng et al. 2020) to detect facial areas and scaled the face image to 224×224 with a patch size of 16. We trained the model using the Adam optimizer with the learning rate set to 3e-6. The frame setting is consistent with common practices in (Cao et al. 2022a; Qian et al. 2020b; Yan et al. 2023b). *In our experiments, we generated approximately 25k image-text pairs.*

Comparison with State-of-the-Art Methods

Cross-Dataset Generalization. We first evaluated the detector's ability to generalize to completely unseen datasets. As shown in Table 1, all models were trained on FF++ (HQ) and tested on datasets including DFD, DFDC-P, WildDeepfake, and Celeb-DF. Our model achieved top-tier or second-best performance across all cross-dataset scenarios. For instance, our model reached an AUC of 87.10% on highly challenging Celeb-DF, and it achieved an AUC of 86.50% on WildDeepfake. This superior generalization performance stems from the higher ecological validity of our simulated data. By modeling diverse human intents through the Profile module and complex social interactions via the PNS module, our data better captures the fundamental characteristics of real-world forgeries, allowing the model to generalize beyond overfitting to specific dataset artifacts.

Robustness to Diverse Manipulations. Next, we assessed the model's robustness against a variety of unseen forgery algorithms using the DF40 protocol. As shown in Table 2, our model achieved an average AUC of 93.9% when tested against six advanced forgery techniques like uniface, e4s, and simswap, significantly outperforming all existing SOTA methods. This exceptional robustness further demonstrates the value of our data generation framework. Through the ARS mechanism that filters for high-difficulty samples and the diverse agent Profiles, our framework generates training data that covers a broader spectrum of forgery traces. This process enables the detector to learn more essential and universal forgery features, rather than simply memorizing the patterns of a few specific known forgery types.

Ablation Study

Effectiveness of Agent-Generated Annotations. A core claim of our work is that simulating the forgery lifecycle yields higher-quality multimodal annotations. We validated this through a multi-faceted comparison of data from Agent4FaceForgery against human annotations (DD-VQA) and direct MLLM annotations (GPT-4o). As shown in Table 3, our approach excels on multiple fronts. First, in terms of annotation quality, our data achieved a Precision of 94.41% and an F1-score of 69.06%, far surpassing other methods. This demonstrates that our agent simulation (especially the 'Memory' and 'PNS' modules) generates text that is better aligned with visual evidence and less prone to hallucination. Second, on downstream tasks, models trained with our data performed best, whether training a standard CLIP model (achieving an AVG-AUC of 91.23%) or fine-tuning an MLLM (LLaVA), which reached an accuracy of 77.98% on Celeb-DF. These combined results confirm that our simulation process produces superior multimodal training data, leading to tangible improvements in detector performance, generalization, and interpretability.

Sequential Training with A4FF Data. Agent4FaceForgery is designed as a data generation and augmentation framework. Our standard training strategy is sequential: pre-training a model on a base dataset (FF++) and then fine-tuning it with our generated data to enhance performance. The effectiveness of this strategy is clearly demonstrated

Table 2: Assessing detector robustness to diverse manipulation algorithms within the FF++ dataset. We report frame-level AUC (%) against six techniques specified in DF40 (Yan et al. 2024b).

Method	**Venue**	**uniface**	**e4s**	**facedancer**	**fsgan**	**inswap**	**simswap**	**Avg.**
RECCE (Cao et al. 2022b)	CVPR 2022	84.2	65.2	78.3	88.4	79.5	73.0	78.1
SBI (Shiohara and Yamasaki 2022)	CVPR 2022	64.4	69.0	44.7	87.9	63.3	56.8	64.4
CORE (Ni et al. 2022)	CVPRW 2022	81.7	63.4	71.7	91.1	79.4	69.3	76.1
IID (Huang et al. 2023)	CVPR 2023	79.5	71.0	79.0	86.4	74.4	64.0	75.7
UCF (Yan et al. 2023b)	ICCV 2023	78.7	69.2	80.0	88.1	76.8	64.9	76.3
LSDA (Yan et al. 2023a)	CVPR 2024	85.4	68.4	75.9	83.2	81.0	72.7	77.8
CDFA (Lin et al. 2024)	ECCV 2024	76.5	67.4	75.4	84.8	72.0	76.1	75.4
ProgressiveDet (Cheng et al. 2024)	NeurIPS 2024	84.5	71.0	73.6	86.5	78.8	77.8	78.7
Ours	-	**96.3**	**92.4**	**92.9**	**94.8**	**92.4**	**94.6**	**93.9**

Table 3: Comparison of different annotation approaches. We report precision, recall and F1-score for annotation quality evaluation, AUC (%) and EER for CLIP-based forgery detection and ACC (%) and explanation quality (Precision/Recall) for MLLMs evaluation on FF++ and Celeb-DF (CDF) datasets.

Method	Annotation Evaluation			CLIP Evaluation		MLLM Evaluation			
	Precision	Recall	F1	AVG-AUC	AVG-EER	FF++-ACC	CDF-ACC	Precision	Recall
w/o Text	-	-	-	84.36	20.64	50.13	65.30	10.41	8.10
DD-VQA (Human)	62.46	51.52	52.06	88.25	18.04	73.54	65.60	62.94	53.62
GPT-4o-mini	61.27	44.00	47.18	87.56	19.21	94.84	73.98	58.26	41.85
Ours	**94.41**	**60.04**	**69.06**	**91.23**	**16.35**	**96.35**	**77.98**	**89.02**	**59.02**

Table 4: Ablation on the sequential training strategy. Models are first pre-trained on a base dataset and then fine-tuned with data generated by Agent4FaceForgery (A4FF). The results demonstrate significant performance improvements across multiple backbone models, confirming the effectiveness of our generated data for augmentation.

Model	**A4FF Data**	**AUC(%)** ↑	**EER(%)** ↓
Phi-3.5	×	81.5	25.3
	✓	**90.4**	**19.0**
Qwen-VL 2.5	×	82.7	26.1
	✓	**91.7**	**18.4**
LLaVA	×	83.2	24.8
	✓	**92.2**	**16.8**

Table 5: Results of different backbone models with and without Agent4FaceForgery on different datasets.

Backbone	CLIP	Agent	WDF		DFDC-P	
			AUC	EER	AUC	EER
Xception (Rossler et al. 2019b)	×	×	66.17	40.14	69.80	35.41
	×	✓	**73.14**	**35.12**	**77.64**	**28.09**
EN-B4 (Tan and Le 2019a)	×	×	61.04	45.34	70.12	34.54
	×	✓	**73.78**	**34.57**	**80.04**	**27.62**
ViT-L (Radford et al. 2021)	×	×	65.39	38.11	68.03	36.06
	×	✓	76.12	30.07	77.54	28.71
	✓	✓	**79.32**	**26.51**	**79.44**	**28.79**
ViT-B (Radford et al. 2021)	×	×	69.92	36.20	71.33	33.76
	×	✓	73.55	30.12	81.79	26.31
	✓	✓	**86.50**	**21.87**	**88.10**	**19.19**

in Table 4. After augmentation with Agent4FaceForgery (A4FF) data, the performance of various models, including Phi-3.5, Qwen-VL 2.5, and LLaVA, improved significantly. For instance, the LLaVA model's AUC improved from 83.2% to 92.2%, and its EER dropped from 24.8% to 16.8%. This strongly validates the value of our framework as a powerful data augmentation tool.

Impact on Different Backbone Architectures. To demonstrate the universal applicability of our generated data, we evaluated its effectiveness across four different backbone architectures: Xception, EfficientNet-B4, ViT-B, and ViT-L.

Table 6: Ablation study regarding the effectiveness of each proposed module via cross-dataset evaluation. The results show an incremental benefit in each module.

Method	**Modules**			**Metrics (AUC(%) | AP(%) | EER(%))**		
	FT	**ARS**	**PNS**	**CDF**	**DFD**	**DFDC**
LLaVA	-	-	-	51.8 | 68.0 | 49.3	69.3 | 94.9 | 37.2	57.4 | 60.4 | 45.3
Only FT	✓	-	-	83.2 | 90.8 | 24.8	91.5 | 98.6 | 16.3	82.5 | 90.5 | 22.1
Only ARS	-	✓	-	88.0 | 93.7 | 21.0	92.1 | 98.8 | 16.8	84.2 | 86.5 | 22.3
Only PNS	-	-	✓	91.0 | 95.0 | 17.5	93.8 | 99.0 | 16.2	85.5 | 87.2 | 20.5
Ours	✓	✓	✓	92.2 | 95.6 | 16.8	94.9 | 99.2 | 15.7	86.7 | 88.0 | 19.5

Table 7: Ablation on the number of agents in the social simulation. Performance is evaluated on the DFD and Celeb-DF datasets. The results show diminishing returns as the number of agents increases, with 12 agents providing only marginal gains over 6, which justifies our agent count configuration and balances performance with computational cost.

Configuration	**DFD**		**Celeb-DF**		**Time(h)**
	AUC(%) ↑	EER(%) ↓	AUC(%) ↑	EER(%) ↓	
Baseline (No Social Sim)	88.1	20.0	74.5	31.0	3.8
2 Agents	89.8	18.1	77.9	27.8	4.5
4 Agents	91.3	16.5	81.5	24.6	5.3
6 Agents	92.8	14.9	85.3	22.0	6.1
12 Agents	**93.0**	**14.5**	**85.8**	**21.6**	7.5

The results in Table 5 show that all architectures exhibited significant performance gains when trained with data from Agent4FaceForgery compared to using only standard FF++ data. It is particularly noteworthy that combining CLIP pre-training with our agent-generated data on a ViT-B backbone yielded an AUC of 86.50% on WDF and 88.10% on DFDC-P. This powerful synergy shows that our framework provides effective ecological signals for a wide range of models, greatly enhancing their generalization capabilities.

Effectiveness of Core Modules. We further conducted an ablation study to isolate the contribution of our framework's core components: Forgery Tree simulation (FT), ARS, and PNS. As detailed in Table 6, we started with a baseline LLaVA model trained only on FF++ (achieving just 51.8%

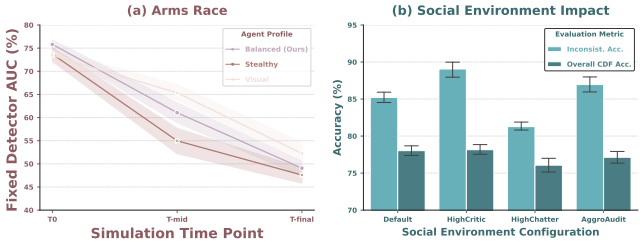

Figure 3: Left (a): Comparison of forgery evasion capability (AUC, lower is better) evolution over simulation time for Agents with different profiles. Right (b): Comparison of MLLM detection accuracy (Inconsistency vs. Overall) when trained on data generated under different Social Environment configurations.

AUC on CDF). Adding data from the 'FT' module alone boosts performance to 83.2%, validating the benefit of simulating diverse forgery intents. The ARS and PNS modules also provided substantial gains individually, with the PNS module being particularly impactful (reaching 91.0% AUC), highlighting the importance of simulating social context and generating challenging text-image consistency samples. The full system, with all modules working in concert, achieved the best performance (92.2%AUC).

Impact of Agent Number in Social Simulation. To justify our choice of agent count for the social simulation, we experimented with varying numbers of agents. Table 7 shows a clear trend of diminishing marginal returns. While performance on DFD and Celeb-DF consistently improves as the number of agents increases from 2 to 6, the gain from 6 to 12 agents is minimal. For instance, the Celeb-DF AUC increases from 85.3% to only 85.8%, at a significant additional time cost. Therefore, we selected 6 agents as our default configuration, striking a reasonable balance between performance and computational efficiency.

Agent Profile Modulates Forgery Evolution. We investigate the impact of agent profiles on simulated forgery evolution against a fixed detector in Fig. 3(a). Measuring evasion capability via AUC (lower is better), all profiles (Balanced, Stealthy, Visual) demonstrate adaptive generation, progressively reducing detectability from T0 to T-final, validating the efficacy of the Memory/Reflection and ARS modules. Crucially, the Stealthy profile achieves superior evasion, reaching the lowest final AUC (≈48.5%), while the 'Visual' profile lags. This confirms that the agent's profile, representing forgery intent, significantly modulates the adaptive generation process within the simulated arms race.

Social Simulation Enhances Detector Robustness. Fig. 3(b) evaluates how the simulated social environment during data generation impacts the robustness of a trained MLLM detector. We assess MLLM accuracy on text-image inconsistency detection (Inconsist. Acc.) and overall generalization (Overall CDF Acc.). Training with data from the 'HighCritic' environment yields the most robust detector (≈88.7% Inconsist. Acc., ≈78.6% CDF Acc.), demonstrating the value of incorporating simulated critical feedback. Conversely, the HighChatter environment degrades performance. The AggroAudit setting, featuring adversarial text generation, notably enhances robustness

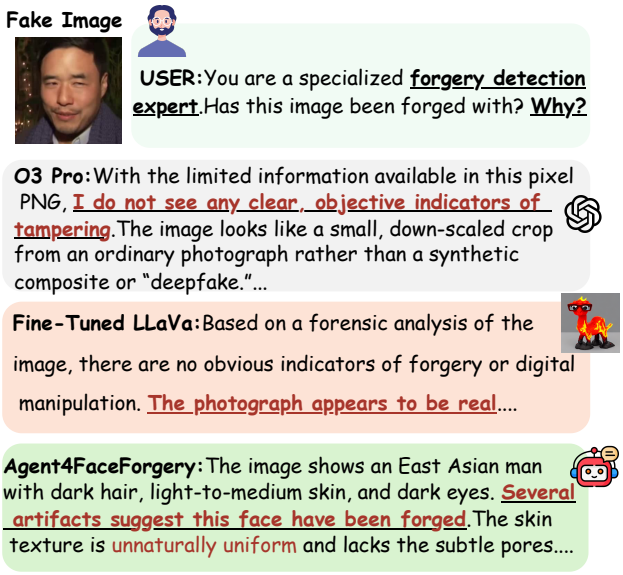

Figure 4: Qualitative examples in challenge scenarios.

Figure 5: Qualitative examples of Agent-Generated Images.

against inconsistencies (≈87.1% vs. ≈85.3% baseline). These results underscore the importance of social simulation and context-aware PNS construction. By generating ecologically valid training data reflecting diverse social interactions and potential adversarial text, our framework demonstrably improves detector robustness to complex multimodal manipulations.

Qualitative Analysis. As shown in Fig. 4 showcases a qualitative analysis where our Agent4FaceForgery demonstrates superior forgery detection over other models. Presented with a forged image, baseline models O3 Pro and a fine-tuned LLaVA fail to identify any manipulation, stating a lack of "clear, objective indicators" and deeming the photograph "to be real." Conversely, Agent4FaceForgery accurately flags the image as forged, offering a succinct explanation: "The skin texture is unnaturally uniform...". This highlights Agent4FaceForgery's enhanced ability to discern subtle forgery cues missed by other approaches. In addition, Fig. 5 provides qualitative examples of the high-fidelity and stylistically diverse portraits generated by our framework.

Conclusion

We introduce Agent4FaceForgery, a multi-agent framework that simulates human behavior to generate realistic multimodal data for face forgery detection. Refined via adaptive rejection sampling, our synthesized data significantly boosts detector performance and generalization across challenging cross-domain benchmarks. The framework's extensible design offers insights into forgery tactics and supports future work in self-supervised learning to counter evolving threats.

References

2025. Towards General Visual-Linguistic Face Forgery Detection. In *CVPR*.

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Cao, J.; Ma, C.; Yao, T.; Chen, S.; Ding, S.; and Yang, X. 2022a. End-to-End Reconstruction-Classification Learning for Face Forgery Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4113–4122.

Cao, J.; Ma, C.; Yao, T.; Chen, S.; Ding, S.; and Yang, X. 2022b. End-to-End Reconstruction-Classification Learning for Face Forgery Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4113–4122.

Chen, S.; Yao, T.; Chen, Y.; Ding, S.; Li, J.; and Ji, R. 2021. Local relation learning for face forgery detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Cheng, J.; Yan, Z.; Zhang, Y.; Luo, Y.; Wang, Z.; and Li, C. 2024. Can We Leave Deepfake Data Behind in Training Deepfake Detector? *arXiv preprint arXiv:2408.17052*.

Chollet, F. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1251–1258.

Deng, J.; Guo, J.; Ververas, E.; Kotsia, I.; and Zafeiriou, S. 2020. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5203–5212.

Dolhansky, B.; Bitton, J.; Pflaum, B.; Lu, J.; Howes, R.; Wang, M.; and Ferrer, C. C. 2020. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*.

Dolhansky, B.; Howes, R.; Pflaum, B.; Baram, N.; and Ferrer, C. 2019. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hong, S.; Zheng, X.; Chen, J.; Cheng, Y.; Zhang, C.; Wang, Z.; Yau, S. K. S.; Lin, Z.; Zhou, L.; Ran, C.; et al. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*.

Huang, B.; Wang, Z.; Yang, J.; Ai, J.; Zou, Q.; Wang, Q.; and Ye, D. 2023. Implicit Identity Driven Deepfake Face Swapping Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4490–4499.

Huang, Z.; Hu, J.; Li, X.; He, Y.; Zhao, X.; Peng, B.; Wu, B.; Huang, X.; and Cheng, G. 2024a. SIDA: Social Media Image Deepfake Detection, Localization and Explanation with Large Multimodal Model. *arXiv preprint arXiv:2412.04292*.

Huang, Z.; Xia, B.; Lin, Z.; Mou, Z.; and Yang, W. 2024b. FFAA: Multimodal Large Language Model based Explainable Open-World Face Forgery Analysis Assistant. *arXiv preprint arXiv:2408.10072*.

Khan, S. A.; and Dang-Nguyen, D.-T. 2024. CLIPping the Deception: Adapting Vision-Language Models for Universal Deepfake Detection. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*.

Li, T.; Tian, Y.; Li, H.; Deng, M.; and He, K. 2024. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37: 56424–56445.

Li, Y.; Chang, M.-C.; and Lyu, S. 2018. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *IEEE International workshop on information forensics and security (WIFS)*. Ieee.

Li, Y.; Yang, X.; Sun, P.; Qi, H.; and Lyu, S. 2020. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3207–3216.

Lin, Y.; Song, W.; Li, B.; Li, Y.; Ni, J.; Chen, H.; and Li, Q. 2024. Fake It till You Make It: Curricular Dynamic Forgery Augmentations towards General Deepfake Detection. arXiv:2409.14444.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Liu, H.; Li, X.; Zhou, W.; Chen, Y.; He, Y.; Xue, H.; Zhang, W.; and Yu, N. 2021. Spatial-Phase Shallow Learning: Rethinking Face Forgery Detection in Frequency Domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Luo, Y.; Zhang, Y.; Yan, J.; and Liu, W. 2021. Generalizing Face Forgery Detection with High-frequency Features. In *CVPR*, 16317–16326.

Nguyen, H. H.; Fang, F.; Yamagishi, J.; and Echizen, I. 2019. 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)*, 1–8. IEEE.

Ni, Y.; Meng, D.; Yu, C.; Quan, C.; Ren, D.; and Zhao, Y. 2022. CORE: Consistent Representation Learning for Face Forgery Detection. arXiv:2206.02749.

Park, J. S.; O'Brien, J. C.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*.

Qian, Y.; Yin, G.; Sheng, L.; Chen, Z.; and Shao, J. 2020a. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European Conference on Computer Vision*, 86–103. Springer.

Qian, Y.; Yin, G.; Sheng, L.; Chen, Z.; and Shao, J. 2020b. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, 86–103. Springer.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Niessner, M. 2019a. FaceForensics++: Learning to Detect Manipulated Facial Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Nießner, M. 2019b. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1–11.

Shiohara, K.; and Yamasaki, T. 2022. Detecting Deepfakes with Self-Blended Images. In *CVPR*, 18720–18729.

Sun, K.; Chen, S.; Yao, T.; Zhou, Z.; Ji, J.; Sun, X.; Lin, C.-W.; and Ji, R. 2025. Towards General Visual-Linguistic Face Forgery Detection (V2). *arXiv preprint arXiv:2502.20698*.

Sun, K.; Liu, H.; Ye, Q.; Liu, J.; Gao, Y.; Shao, L.; and Ji, R. 2021. Domain General Face Forgery Detection by Learning to Weight. In *AAAI*, volume 35, 2638–2646.

Sun, K.; Yao, T.; Chen, S.; Ding, S.; Ji, R.; et al. 2022. Dual Contrastive Learning for General Face Forgery Detection. In *AAAI*.

Tan, M.; and Le, Q. 2019a. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*. PMLR.

Tan, M.; and Le, Q. V. 2019b. Efficientnet: Rethinking model scaling for convolutional neural networks. *ICML*.

Wang, H.; Zhang, A.; Duy Tai, N.; Sun, J.; Chua, T.-S.; et al. 2024a. ALI-Agent: Assessing LLMs' Alignment with Human Values via Agent-based Evaluation. *Advances in Neural Information Processing Systems*, 37: 99040–99088.

Wang, J.; Wu, Z.; Ouyang, W.; Han, X.; Chen, J.; Jiang, Y.-G.; and Li, S.-N. 2022. M2tr: Multi-modal multi-scale transformers for deepfake detection. In *Proceedings of the 2022 international conference on multimedia retrieval*, 615–623.

Wang, T.; Liao, X.; Chow, K. P.; Lin, X.; and Wang, Y. 2024b. Deepfake Detection: A Comprehensive Survey from the Reliability Perspective. *ACM Comput. Surv.*, 57(3).

Wang, X.; Cai, Z.; Gao, D.; and Vasconcelos, N. 2019. Towards universal object detection by domain attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7289–7298.

Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Zhang, S.; Zhu, E.; Li, B.; Jiang, L.; Zhang, X.; and Wang, C. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*, 3(4).

Yan, Z.; Luo, Y.; Lyu, S.; Liu, Q.; and Wu, B. 2023a. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. *arXiv preprint arXiv:2311.11278*.

Yan, Z.; Luo, Y.; Lyu, S.; Liu, Q.; and Wu, B. 2024a. Transcending Forgery Specificity with Latent Space Augmentation for Generalizable Deepfake Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yan, Z.; Yao, T.; Chen, S.; Zhao, Y.; Fu, X.; Zhu, J.; Luo, D.; Yuan, L.; Wang, C.; Ding, S.; et al. 2024b. DF40: Toward Next-Generation Deepfake Detection. *arXiv preprint arXiv:2406.13495*.

Yan, Z.; Zhang, Y.; Fan, Y.; and Wu, B. 2023b. UCF: Uncovering Common Features for Generalizable Deepfake Detection. *arXiv preprint arXiv:2304.13949*.

Yan, Z.; Zhang, Y.; Yuan, X.; Lyu, S.; and Wu, B. 2023c. DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS)*.

Yang, X.; Li, Y.; and Lyu, S. 2019. Exposing deep fakes using inconsistent head poses. In *ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Yang, Y.; Achananuparp, P.; Huang, H.; Jiang, J.; and Lim, E.-P. 2024. Speaker Verification in Agent-generated Conversations. *arXiv preprint arXiv:2405.10150*.

Zhang, A.; Chen, Y.; Sheng, L.; Wang, X.; and Chua, T.-S. 2024. On generative agents in recommendation. In *Proceedings of the 47th international ACM SIGIR conference on research and development in Information Retrieval*, 1807–1817.

Zhao, T.; Xu, X.; Xu, M.; Ding, H.; Xiong, Y.; and Xia, W. 2021. Learning self-consistency for deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 15023–15033.

Zhuang, W.; Chu, Q.; Tan, Z.; Liu, Q.; Yuan, H.; Miao, C.; Luo, Z.; and Yu, N. 2022. UIA-ViT: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection. In *European Conference on Computer Vision*, 391–407. Springer.

Zi, B.; Chang, M.; Chen, J.; Ma, X.; and Jiang, Y.-G. 2020. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM international conference on multimedia*, 2382–2390.