TRIDENT: Tri-modal Deepfake Perception, Detection, and Hallucination Grand Challenge

Abstract

The rapid evolution of generative AI has ushered in an era of hyper-realistic, tri-modal forgeries spanning images, video, and audio. While detection performance has reached high numerical accuracy, modern forensic systems remain "black boxes," often achieving results through stochastic shortcuts or dataset biases rather than grounded reasoning. The lack of interpretability leads to the Hallucination Dilemma, where models justify correct classifications with non-existent artifacts, resulting in a critical failure mode in high-stakes legal and journalistic environments. We propose TRIDENT , Tri-modal Deepfake Perception, Detection, and Hallucination Grand Challenge. TRIDENT is a novel competition designed to shift the community toward accountable and explainable forensics. Built upon the large-scale TriDF benchmark (https://j1anglin.github.io/TriDF/), the challenge requires participants to move beyond binary classification. Models are evaluated across three interdependent dimensions: Perception (the ability to localize fine-grained artifacts), Detection (decision robustness across 16 forgery families), and Hallucination (the logical consistency between perceived evidence and final labels). By providing a standardized probing protocol involving Structured VQA and Open-Ended Reasoning tasks, TRIDENT establishes a rigorous framework for evaluating the next generation of accountable and explainable Deepfake detectors. The proposed challenge invites the multimedia research community to bridge the gap between human-centric evidence and AI-driven forensics, ensuring that the future of digital media authentication is as interpretable as it is accurate.

Keywords

Multimodal Deepfake Forensics, Interpretable Deepfake Detection, Explainable Forensic Reasoning, Detection Hallucination, Deepfake Evidence Grounding, Standardized Forensic Probing

ACM Reference Format:

. 2018. TRIDENT: Tri-modal Deepfake Perception, Detection, and Hallucination Grand Challenge. In *Proceedings of MM (Conference MM'26).* ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Motivation and Rationale

The rapid maturation of generative artificial intelligence has fundamentally altered the digital ecosystem, moving beyond simple image synthesis toward a sophisticated, multi-modal reality. While these advancements drive creative innovation, they have simultaneously given rise to a *Synthetic Media Paradox*: as deepfakes become visually and acoustically indistinguishable from authentic recordings, our detection systems are becoming increasingly powerful but alarmingly less transparent. In a 2026 landscape, where disinformation campaigns leverage high-fidelity images, temporal video sequences, and cloned audio in tandem, traditional binary detection that labels content as simply Real or Fake is no longer a sufficient defense. The TRIDENT Grand Challenge is born from the urgent need to transition from black-box detection toward a white-box forensic paradigm that prioritizes accountability, interpretability, and explanability.

The foundational philosophy of TRIDENT is rooted in the *Forensic Triad*: the belief that a trustworthy AI must not only reach the correct binary conclusion but must do so with the correct reasons. Previous challenges often overlook the *Interpretability Gap*, where models achieve decent accuracy by potentially exploiting dataset biases or non-semantic shortcuts rather than identifying actual generative artifacts, resulting in the *Hallucination Dilemma*, a critical failure mode where a detector justifies a correct classification through fabricated evidence and points to phantom artifacts that do not exist. In legal, journalistic, or security contexts, such correct but ungrounded decisions are inadmissible and dangerous. TRIDENT mitigates the gap by requiring the model to demonstrate a triad-based Deepfake examination: 1) Perception: *Can the model identify and localize the fine-grained manipulation artifacts across image, video, and audio?* 2) Detection: *Does the model maintain high classification accuracy across diverse forgery families?* 3) Hallucination: *Is the model's explanation grounded in reality, or is it fabricating evidence?*

As the premier venue for multimedia research, ACMMM is uniquely positioned to address these challenges: Deepfake detection is inherently a multimedia problem that intersects forensics, system security, and responsible AI. This challenge directly aligns with the ACMMM mission of advancing trustworthy multimedia analysis by forcing a holistic evaluation of how models perceive, classify, and explain synthetic content across the full spectrum of modern media.

2 Task Definition

To facilitate broad participation while maintaining rigorous standards, the TRIDENT Challenge is partitioned into independent modality tracks: **Image**, **Video**, and **Audio**. To recognize specialized expertise across the multimedia community, each track will be evaluated and ranked independently. Participants may compete in one or more tracks. For each track, participants must demonstrate a holistic forensic capability and subject their models to a standardized probing protocol consisting of three question types. The performance is then mapped onto the three forensic dimensions of the triad: **Perception**, **Detection**, and **Hallucination**. Within

Conference MM'26, Brazil

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/XXXXXXX.XXXXXXX

each modality track, participants are ranked based on a composite performance across three forensic dimensions.

2.1 Input Prompts and Submission Protocols

To encourage broad participation while maintaining strict evaluation standards, the TRIDENT challenge uses a flexible input protocol with a strict output standard. Participants will receive a "Starter Kit" with prompt templates and validation scripts. All predictions must be submitted in a standardized JSON format for compatibility with the automated evaluation pipeline, which uses three query types that require specific output structures:

- **Structured VQA (TFQ & MCQ):** The model must resolve automated queries regarding the existence and location of a series of artifacts. The queries are categorized into True-False Questions (TFQ) and Multiple-Choice Questions (MCQ). Outputs must be strict boolean strings ("True" or "False") for TFQs and selected options for MCQs.
- **Type-A Open-Ended Questions (Type-A OEQ):** Informed that the presented sample is a DeepFake, the model must generate a structured text paragraph itemizing observable artifacts.
- **Type-B Open-Ended Questions (Type-B OEQ):** Presented with unknown media, the model must provide a structured response containing: 1) a binary label indicating the authenticity ("Likely Authentic" or "Likely Manipulated"), and 2) a concise reasoning paragraph justifying the decision.

2.2 Triad Assessment and Sub-Task Mapping

(I) Perception: Evidence Recognition & Localization.
This dimension assesses the model's ability to recognize and localize specific flaws and artifacts that are semantically or visually identifiable to humans. Drawing primarily from Structured VQA and Type-A OEQ responses, this assessment is conducted exclusively on manipulated samples to isolate the model's "sight" and "hearing" capabilities from classification bias.
(II) Detection: Forensic Decision Making
The second dimension evaluates the fundamental capability to distinguish authentic media from high-fidelity forgeries within a realistic, uninformed forensic setting. This metric is derived exclusively from Type-B OEQ responses, where the model is presented with a balanced distribution of real and fake samples without prior knowledge of their status. Evaluation focuses on traditional robust metrics derived from the binary decisions and associated confidence values. It provides a clear measure of the model's discriminative power across diverse forgery families.
(III) Hallucination: Reliability and Grounding.
The novel aspect of the TRIDENT challenge is the evaluation of detection hallucination. This dimension measures the logical alignment between the binary classification decision and its perceived evidence. By analyzing responses from both Type-A and Type-B OEQs, we penalize systems that report non-existent artifacts, ensuring that forensic explanations are faithful to observable ground truth.

3 Data Overview

The TRIDENT challenge utilizes the **Tri-Perspective DeepFake Detection Benchmark (TriDF)** [13], a comprehensive resource specifically designed to evaluate MLLMs. TriDF comprises over 76K multimodal instruction-tuning pairs across *image*, *video*, and *audio* modalities. By incorporating 16 different DeepFake techniques, ranging from GANs to modern Diffusion Transformers, the benchmark presents significant challenges for generalization. To analyze the capabilities of the model, the questions are divided into 23K TFQ, 24K MCQ, and 29K OEQ, allowing for a detailed assessment of interpretability, perceptual reasoning, and hallucination tendencies.

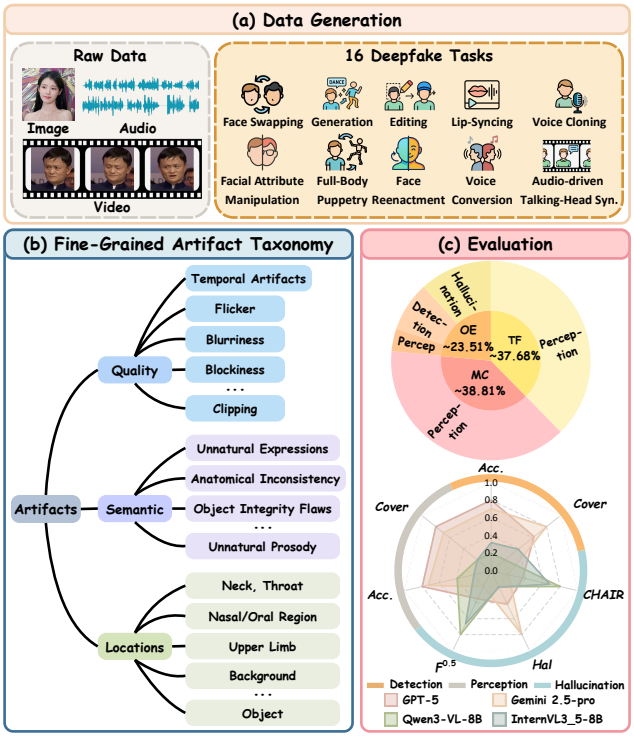

Figure 1: An overview of TriDF. TriDF is a comprehensive benchmark tailored to interpretable DeepFake detection, containing 5K high-quality samples generated by 16 DeepFake techniques across three modalities, equipped with a hierarchical taxonomy of fine-grained artifacts that decomposes perception, detection, and hallucination tendencies into artifact-wise analyses.

3.1 Multimodal Generation Pipeline

To ensure rigorous diversity, TriDF source authentic references from over 30 public datasets [5, 7, 15–17, 20, 22, 33, 34]. The generation pipeline aggregates over 50 distinct models, evolving from GANs [30] to modern Stable Diffusion editors [6, 36], state-of-the-art Diffusion Transformers [1, 4], and proprietary commercial systems [9, 10, 21]. Crucially, TriDF selects real samples exclusively from test sets and generates corresponding forgeries to establish strict one-to-one real-fake pairings, ensuring high fidelity via automated quality control.

3.2 Task Taxonomy

To comprehensively cover the threat landscape, TriDF organizes 16 specific DeepFake tasks into two primary categories based on the manipulation mechanism. The first category, **Partially Manipulated**, targets the alteration of specific attributes within authentic media, encompassing tasks such as *Face Swapping*, *Facial Attribute Manipulation*, *Lip-Syncing*, *Face Reenactment*, and *Subject-Driven Editing*. In contrast, the **Fully Synthetic** category focuses on creating entirely new content from noise or text descriptions, including *Audio-Driven Talking Heads*, *Text-to-Video/Image Generation*, *Human-Scene Synthesis*, and *Voice Cloning*.

3.3 Hierarchical Artifact Annotation

To ensure diagnostic precision and mitigate the self-preference biases often found in automated MLLM evaluations [3, 24], TriDF moves beyond unstructured text labels by implementing a standardized, hierarchical taxonomy that grounds annotations in observable evidence. Specifically, TriDF distinguishes between **Quality Artifacts**, defined as localized signal degradations such as blurriness, noise, or jitter that are spatially grounded to specific regions to test localization acuity, and **Semantic Artifacts**, which encompass high-level logical inconsistencies including anatomical errors, object integrity flaws, or unnatural audio-visual prosody requiring commonsense reasoning. This structured framework provides the necessary ground truth to strictly differentiate between accurate *perception* and plausibility-driven *hallucination*, a capability lacking in binary-labeled datasets.

3.4 Dataset Splits and Generalization Strategy

We partition the dataset into three subsets: a Training Set (∼60%), a Public Validation Set (∼20%), and a Private Test Set (∼20%), utilizing a generator-disjoint strategy. The Public Validation Set is provided for model validation and preliminary leaderboard feedback, while the Private Test Set is withheld exclusively for the final leaderboard ranking. This structure ensures that the final evaluation measures true forensic generalization against unknown threats, preventing overfitting to the validation data.

4 Evaluation Metrics and Ranking

To comprehensively assess the performance of interpretable DeepFake detection systems, we adopt a tri-perspective evaluation protocol covering Perception, Detection, and Hallucination Robustness. All metrics are computed per sample and aggregated (macro-averaged) over the test set within each modality track.

4.1 Metric Definitions

(I) Perception. Perception evaluates the model's ability to localize and identify artifacts *exclusively on manipulated samples*. This capability is assessed through both structured choices (TFQ, MCQ) and open-ended generation (Type-A OEQ). For open-ended responses, we employ an **LLM-based evaluator** to parse the generated text and map it to a discrete set of reported artifacts (R_{art}) for comparison against the ground truth (Y_{art}).

- **TFQ Accuracy (Acc_{TFQ}):** The accuracy of binary verification questions regarding the presence of specific artifacts or location cues.
- **MCQ Score ($\text{Score}_{\text{MCQ}}$):** A balanced score for multi-label selection designed to neutralize random guessing. For a question with correct options set C (where $K = |C|$) and total options M, the score for the selected set S is:

$$\text{Score}_{\text{MCQ}}(S, C) = \max\left(0, \sum_{i \in S \cap C} \frac{1}{K} - \sum_{i \in S \setminus C} \frac{1}{M - K}\right). \tag{1}$$

- **Explanatory Coverage (Cover):** Using the artifacts (R_{art}) extracted by the LLM, this metric calculates the proportion of ground-truth artifacts (Y_{art}) correctly recovered:

$$\text{Cover}(R) = \frac{|R_{\text{art}} \cap Y_{\text{art}}|}{\max(1, |Y_{\text{art}}|)} \tag{2}$$

(II) Detection. Detection assesses the holistic capability to distinguish real from fake, evaluated on *both authentic and manipulated samples* using Type-B OEQs.

- **Detection Accuracy (Acc_{Det}):** The standard accuracy of the binary authenticity decision (*Real* vs. *Fake*).

(III) Hallucination Robustness. We evaluate the faithfulness of explanations to ensure reasoning reliability. This dimension focuses on the Precision of the generated content (as parsed by the LLM evaluator) and its trade-off with perception recall.

- **Hallucinated Artifact Rate (CHAIR):** Measures the proportion of reported artifacts (R_{art}) that are incorrect (hallucinated), serving as a proxy for the *false discovery rate*:

$$\text{CHAIR}(R) = 1 - \frac{|R_{\text{art}} \cap Y_{\text{art}}|}{|R_{\text{art}}|} \tag{3}$$

- **Balanced Interpretability Score ($F^{0.5}$):** To strictly evaluate the reliability of the generated explanations, we adopt the $F^{0.5}$ score. By conceptually mapping the *non-hallucination rate* (1 − CHAIR) to **Precision** and the *explanatory coverage* (Cover) to **Recall**, we employ the F^{β} formulation with $\beta = 0.5$. This design mathematically prioritizes precision, ensuring that a system is penalized more heavily for fabricating non-existent artifacts than for missing subtle ones:

$$F^{0.5}(R) = \frac{(1 + 0.5^2) \cdot (1 - \text{CHAIR}(R)) \cdot \text{Cover}(R)}{0.5^2 \cdot (1 - \text{CHAIR}(R)) + \text{Cover}(R)} \tag{4}$$

(Empty responses imply total unreliability, setting CHAIR = 1 *and* $F^{0.5} = 0$.*)*

4.2 Official Ranking: Tri-Metric Composite Score

To determine the final leaderboard ranking, we compute a **Tri-Metric Composite Score (TCS)**. We define normalized scores ($S \in [0, 100]$) using representative metrics for each dimension: standard accuracy for Detection (S_{Det}); the average of TFQ and MCQ for Perception (S_{Perc}) to balance verification with selection; and the precision-weighted $F^{0.5}$ for Hallucination Robustness (S_{Hal}):

$$S_{\text{Det}} = 100 \times \text{Acc}_{\text{Det}} \tag{5}$$

$$S_{\text{Perc}} = 100 \times \left(0.5 \cdot \text{Acc}_{\text{TFQ}} + 0.5 \cdot \text{Score}_{\text{MCQ}}\right) \tag{6}$$

$$S_{\text{Hal}} = 100 \times F^{0.5} \tag{7}$$

Composite Score Calculation. The final TCS is a weighted sum designed to balance the necessity of accurate detection with the requirement for reliable and precise interpretation:

$$\text{TCS} = w_{\text{Det}} \cdot S_{\text{Det}} + w_{\text{Hal}} \cdot S_{\text{Hal}} + w_{\text{Perc}} \cdot S_{\text{Perc}} \tag{8}$$

We assign weights to align closely with the priorities of the challenge. The highest priority is given to the fundamental ability to distinguish authenticity, which is weighed at $w_{\text{Det}} = 0.4$. To ensure the reliability of generated explanations, we assign a substantial weight of $w_{\text{Hal}} = 0.3$ to Hallucination Robustness. Using the $F^{0.5}$ score, this dimension serves as a rigorous filter that rewards the retrieval of true artifacts (Cover) while imposing heavy penalties for the fabrication of false ones (CHAIR). Lastly, we allocate an equal weight of $w_{\text{Perc}} = 0.3$ to Atomic Perception. This component enhances the generative evaluation by focusing on fine-grained sensitivity to specific manipulation cues through structured TFQ and MCQ.

Winner Determination. The submission with the highest TCS wins the track. In the event of a tie, the tie-breaking order is: (1) Higher S_{Det}, (2) Higher S_{Hal}, and (3) Higher S_{Perc}.

5 Area Review

The field of DeepFake analysis is shifting from binary classification toward interpretable, multimodal reasoning. This section reviews this evolution and highlights the critical gap in evaluating hallucination and perceptual grounding.

5.1 From Binary Detection to MLLM Reasoning

Conventional DeepFake detectors, typically formulated as binary classifiers, often suffer from overfitting to dataset-specific cues [2, 26, 35]. To improve generalization, recent approaches incorporate explicit forensic priors, such as frequency artifacts and cross-view inconsistencies [19, 25, 31], or leverage physiological and temporal cues in video [11, 23]. Despite these forensic advances, binary verdicts remain opaque and do not provide the transparent reasoning essential for user trust. Consequently, the field is shifting toward MLLMs that unify detection with linguistic reasoning. Methods like FakeShield [28] and SIDA [12] utilize knowledge-guided learning to explain their decisions. However, this introduces a new vulnerability: *hallucination*. MLLMs may generate linguistically plausible but factually incorrect rationales [14, 38]. While recent works attempt to ground explanations via mask-guided localization [24] or temporal consistency [29], rigorous mechanisms to verify whether these explanations are based on observable artifacts remain unexplored.

5.2 Evolution of DeepFake Benchmarks

Benchmarking has paralleled this methodological shift. Foundational datasets like FaceForensics++ [22] and DFDC [8] focused on binary accuracy, while later suites like GenImage [37] addressed cross-generator transferability. Recent efforts have operationalized explainability. Datasets such as MMTD-Set [28] and FakeClue [27] provide rationales in natural-language, and benchmarks such as FakeBench [18] and LOKI [32] evaluate detection along with reasoning across multiple modalities.

Why This Challenge? Critically, existing benchmarks primarily evaluate the *plausibility* of model outputs rather than the *faithfulness* of their perception. They lack strict metrics to confirm whether a model genuinely perceives low-level visual artifacts or merely hallucinates high-level inconsistencies. Our challenge addresses this by establishing an interdependent evaluation triad. We evaluate Perception (evidence identification) as a necessary foundation for Detection (decision-making), ensuring that diagnostic claims are supported by observable reality. By introducing the Hallucination dimension to penalize ungrounded reasoning, we steer the community toward models that are not just accurate, but diagnostically reliable.

6 Organizer Team

- **Wen-Huang Cheng** is a University Distinguished Chair Professor in the Department of Computer Science and Information Engineering at National Taiwan University and a Visiting Professor at the Korea Advanced Institute of Science and Technology (KAIST). His current research interests include multimedia, computer vision, and machine learning. He has actively participated in international events and played significant leadership roles in prestigious journals, conferences, and professional organizations. These roles include serving as Editor-in-Chief for IEEE CTSoc News on Consumer Technology, Senior Editor for IEEE Consumer Electronics Magazine (CEM), Associate Editor for IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) and IEEE Transactions on Multimedia (TMM), General Chair for ACM MMAsia (2023), IEEE ICME (2022), and ACM ICMR (2021), Technical Program Chair for ACM MM (2025), ACM ICMR (2022), IEEE ICME (2020), IEEE VCIP (2018), Chair for IEEE CASS Multimedia Systems and Applications (MSA) technical committee, and governing board member for IAPR. He has received numerous research and service awards, including the NVIDIA Academic Grant Program Award (2025), the 2024 Best Paper Award of IEEE Consumer Electronics Magazine, the Best Paper Award at the 2021 IEEE ICME and the Outstanding Associate Editor Award of IEEE TMM (2021 and 2020, twice). He is an IEEE Fellow, IET Fellow, and ACM Distinguished Member. [Google Scholar] [Email]
- **Hong-Han Shuai** is a Professor at National Yang Ming Chiao Tung University (NYCU), where he is at the forefront of research in multimedia processing, deep learning, computer vision, and data mining. His research has been prominently featured at leading Artificial Intelligence/Data Mining conferences such as MM, NeurIPS, CVPR, ICCV, ECCV, ACL, EMNLP, NAACL, AAAI, KDD, WWW, ICDM, CIKM, and VLDB, and in top-tier journals including TKDE, TKDD, TMM, TNNLS, TACL, and JIOT. He has played significant roles in the academic community, serving as Associate Editor for both IEEE Transactions on Multimedia (TMM) and IEICE Transactions on Information and Systems, Guest Editor for

ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), and as TPC Co-Chair for conferences like IEEE ICCE-TW, International Conference on Technologies and Applications of Artificial Intelligence (TAAI) and IPPR Conference on Computer Vision, Graphics, and Image Processing (CVGIP). He also contributes to the technical committee for IEEE Multimedia Systems and Applications (MSA). He has been honored with several research awards, notably the Best Paper Awards at the 2021 IEEE International Conference on Multimedia and Expo, 2025 World Congress on Medical and Health Informatics (MedInfo). [Google Scholar] [Email]

- **Khoa D. Doan** is currently an Assistant Professor of Computer Science in the College of Engineering and Computer Science at VinUniversity, Vietnam. Before his current appointment, he worked as an AI Researcher at Baidu Research, USA. He received his PhD in Computer Science at Virginia Tech, and MS in Computer Science at the University of Maryland, College Park. His research focuses on developing computational frameworks that enable complex machine learning models to be more suitable and secured/trustworthy for practical uses in various domains such as computational advertising, computer vision, and natural language processing. He has chaired multiple workshops on Trustworthy and Secured Machine Learning, such as BUGS@NeurIPS'23 and DIG-BUGS@ICML'25, conferences such as ACML'24, and served as an Area Chair at conferences such as NeurIPS, AISTATS, ICML, ICLR, and ACML. Currently, he serves on the Editorial Board of Discover Data and ACM AI Letters. [Google Scholar] [Email]
- **Hongxia Xie** is an Associate Professor (tenure-track) in the College of Computer Science and Technology at Jilin University, China, and serve as the principal investigator of the Affective Vision and Computing (AVC) Lab. Her research interests include computer vision, affective computing, and vision-language models. [Google Scholar] [Email]
- **Ling Lo [Main Contact]** is a postdoctoral researcher at National Yang Ming Chiao Tung University, Taiwan. She received her Ph.D. and B.S. in Electronics Engineering from National Yang Ming Chiao Tung University. Her research lies at the intersection of computer vision and machine learning, with a focus on affective computing, generative AI for multimedia, and trustworthy visual content creation. Her work has been published in top-tier venues, including ACM MM, CVPR, ICCV, AAAI, ACM TOMM, IEEE TMM, and IEEE TAFFC, and she received the 2021 ICME Best Paper Award. [Google Scholar] [Email]
- **Jian-Yu Jiang-Lin [Main Contact]** is currently pursuing the Ph.D. degree with the Communications and Multimedia Laboratory, National Taiwan University, Taipei, Taiwan. He received the B.S. degree in electronics engineering and the M.S. degree in artificial intelligence from National Yang Ming Chiao Tung University, Hsinchu, Taiwan, in 2022 and 2024, respectively. His research interests include trustworthy Multimodal Large Language Models, with a particular focus on safety alignment, explainability, deepfake detection, and embodied intelligence. [Email]
- **Kang-Yang Huang** received the M.S. degree in Computer Science from National Taiwan University (NTU), Taiwan, in 2025, and he earned the B.S. degree from the Undergraduate Honors Program in Electrical Engineering and Computer Science at National Yang Ming Chiao Tung University (NYCU), Hsinchu, Taiwan. His research focuses on the intersection of computer vision and deep learning, encompassing generative AI and visual language models. He has received the 2024 Best Paper Award from IEEE Consumer Electronics Magazine. [Email]
- **Ling Zou** is currently pursuing the Ph.D. degree in Computer Science at National Taiwan University (NTU), Taipei, Taiwan. She received the M.S. degree in computer science and information engineering from NTU in 2025, and the B.S. degree in computer science and information engineering from Ming Chuan University (MCU), Taiwan. Her research focuses on the intersection of computer vision and deep learning, with particular interests in the acceleration of multimodal large language models (MLLMs) and generative AI. [Email]

7 Commitment

The organizing team of TRIDENT 2026 is fully committed to the long-term success and scientific integrity of this challenge. We recognize that the value of a Grand Challenge lies not only in the competition itself but in the sustained availability of its resources for the global research community. To this end, the organizers guarantee the maintenance of a dedicated challenge website for a minimum of three years following the conclusion of ACM Multimedia 2026. This portal will serve as a central repository for the TRIDENT challenge, standardized evaluation scripts, and a permanent archive of the winning methodologies. Beyond technical maintenance, we are committed to providing a robust support infrastructure during the active phase of the competition, including a comprehensive "Starter Kit" consisting of a prompt template for the VQA and OEQ tasks and the evaluation scripts, ensuring a low barrier to entry for diverse research groups. Our team will actively manage a community support channel. Furthermore, we commit to coordinating closely with the ACM Multimedia conference chairs to publicize the challenge across major academic and industrial networks, fostering a diverse participant pool. Finally, we pledge to ensure a transparent and objective judging process, culminating in a dedicated session at the conference where top-performing teams will be invited to publish their insights, further contributing to the collective understanding of trustworthy multimedia AI.

References

[1] Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. 2025. FLUX. 1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. *arXiv preprint arXiv:2506.15742* (2025).

[2] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. 2022. End-to-end reconstruction-classification learning for face forgery detection. In *CVPR*.

[3] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *ICML*.

[4] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. 2024. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *ECCV*.

[5] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. 2024. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *CVPR*.

[6] Ha-Yeong Choi, Sang-Hoon Lee, and Seong-Whan Lee. 2023. Diff-HierVC: Diffusion-based Hierarchical Voice Conversion with Robust Pitch Generation and Masked Prior for Zero-shot Speaker Adaptation. In *Interspeech*.

[7] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. Voxceleb2: Deep speaker recognition. In *Interspeech*.

[8] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. 2020. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397* (2020).

[9] Google DeepMind. 2025. Gemini 2.5 Flash Image (Nano Banana). https://deepmind.google/models/gemini-image/flash/

[10] Google DeepMind. 2025. Veo 3. https://deepmind.google/models/veo/

[11] Yue-Hua Han, Tai-Ming Huang, Kai-Lung Hua, and Jun-Cheng Chen. 2025. Towards More General Video-based Deepfake Detection through Facial Component Guided Adaptation for Foundation Model. In *CVPR*.

[12] Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. 2025. Sida: Social media image deepfake detection, localization and explanation with large multimodal model. In *CVPR*.

[13] Jian-Yu Jiang-Lin, Kang-Yang Huang, Ling Zou, Ling Lo, Sheng-Ping Yang, Yu-Wen Tseng, Kun-Hsiang Lin, Chia-Ling Chen, Yu-Ting Ta, Yan-Tsung Wang, et al. 2025. TriDF: Evaluating Perception, Detection, and Hallucination for Interpretable DeepFake Detection. *arXiv preprint arXiv:2512.10652* (2025).

[14] Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2025. Why Language Models Hallucinate. *arXiv preprint arXiv:2509.04664* (2025).

[15] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *ICLR*.

[16] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *CVPR*.

[17] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. 2020. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*.

[18] Yixuan Li, Xuelin Liu, Xiaoyang Wang, Shiqi Wang, and Weisi Lin. 2025. FakeBench: Uncover the Achilles' Heels of Fake Images with Large Multimodal Models. *IEEE TIFS* (2025).

[19] Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Yao Zhao, and Jingdong Wang. 2024. Forgery-aware Adaptive Transformer for Generalizable Synthetic Image Detection. In *CVPR*.

[20] Kun Liu, Qi Liu, Xinchen Liu, Jie Li, Yongdong Zhang, Jiebo Luo, Xiaodong He, and Wu Liu. 2025. Hoigen-1m: A large-scale dataset for human-object interaction video generation. In *CVPR*.

[21] OpenAI. 2025. GPT-4o Image Generation. https://openai.com/index/hello-gpt-4o/

[22] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. Faceforensics++: Learning to detect manipulated facial images. In *ICCV*.

[23] Stefan Smeu, Dragos-Alexandru Boldisor, Dan Oneata, and Elisabeta Oneata. 2025. Circumventing shortcuts in audio-visual deepfake detection datasets with unsupervised learning. In *CVPR*.

[24] Ke Sun, Shen Chen, Taiping Yao, Ziyin Zhou, Jiayi Ji, Xiaoshuai Sun, Chia-Wen Lin, and Rongrong Ji. 2025. Towards general visual-linguistic face forgery detection. In *CVPR*.

[25] Chuangchuang Tan, Huan Liu, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. 2024. Rethinking the Up-Sampling Operations in CNN-based Generative Network for Generalizable Deepfake Detection. In *CVPR*.

[26] Chengrui Wang and Weihong Deng. 2021. Representative forgery mining for fake face detection. In *CVPR*.

[27] Siwei Wen, Junyan Ye, Peilin Feng, Hengrui Kang, Zichen Wen, Yize Chen, Jiang Wu, Wenjun Wu, Conghui He, and Weijia Li. 2025. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. In *NeurIPS*.

[28] Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. 2025. FakeShield: Explainable Image Forgery Detection and Localization via Multi-Modal Large Language Models. In *ICLR*.

[29] Zhipei Xu, Xuanyu Zhang, Xing Zhou, and Jian Zhang. 2025. AvatarShield: Visual Reinforcement Learning for Human-Centric Video Forgery Detection. *arXiv preprint arXiv:2505.15173* (2025).

[30] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. 2023. Styleganex: Stylegan-based manipulation beyond cropped aligned faces. In *ICCV*.

[31] Yongqi Yang, Zhihao Qian, Ye Zhu, Olga Russakovsky, and Yu Wu. 2025. D^3: Scaling Up Deepfake Detection by Learning from Discrepancy. In *CVPR*.

[32] Junyan Ye, Baichuan Zhou, Zilong Huang, Junan Zhang, Tianyi Bai, Hengrui Kang, Jun He, Honglin Lin, Zihao Wang, Tong Wu, Zhizheng Wu, Yiping Chen, Dahua Lin, Conghui He, and Weijia Li. 2025. LOKI: A Comprehensive Synthetic Data Detection Benchmark using Large Multimodal Models. In *ICLR*.

[33] Jianhui Yu, Hao Zhu, Liming Jiang, Chen Change Loy, Weidong Cai, and Wayne Wu. 2023. Celebv-text: A large-scale facial text-video dataset. In *CVPR*.

[34] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. In *Interspeech*.

[35] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. 2021. Multi-attentional deepfake detection. In *CVPR*.

[36] Wenliang Zhao, Yongming Rao, Weikang Shi, Zuyan Liu, Jie Zhou, and Jiwen Lu. 2023. Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion. In *CVPR*.

[37] Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. 2023. Genimage: A million-scale benchmark for detecting ai-generated image. In *NeurIPS*.

[38] Yueying Zou, Peipei Li, Zekun Li, Huaibo Huang, Xing Cui, Xuannan Liu, Chenghanyu Zhang, and Ran He. 2025. Survey on ai-generated media detection: From non-mllm to mllm. *arXiv preprint arXiv:2502.05240* (2025).