NeurIPS 2024 Competition: Erasing the Invisible: A Stress-Test Challenge for Image Watermarks

Mucong Ding* Tahseen Rabbani* Bang An* Souradip Chakraborty
Chenghao Deng Mehrdad Saberi Yuxin Wen Xuandong Zhao Mo Zhou
Anirudh Satheesh Lei Li Yu-Xiang Wang Vishal M. Patel* Soheil Feizi*
Tom Goldstein* Furong Huang*

wavesbench@googlegroups.com

Abstract

"Erasing the Invisible" is a pioneering competition designed to rigorously stress-test image watermarks, aiming to enhance their robustness significantly. Its standout feature is the introduction of dual tracks for black-box and beige-box attacks, providing a nuanced approach to validate the reliability and robustness of watermarks under varied conditions of visibility and knowledge. The competition spans from July 18 to October 31, inviting individuals and teams to register and participate in a dynamic challenge. Throughout the competition, employing a dataset of 10k images accessed through the Hugging Face API, competitors will receive updated evaluation results on a rolling basis and submit their refined techniques for the final evaluation, which will be conducted on an extensive dataset of 50k images. The evaluation process of this competition not only emphasizes the effectiveness of watermark removal but also highlights the critical importance of maintaining image quality, with results reflected on a continuously updated leaderboard. "Erasing the Invisible" promises to elevate watermarking technology to new heights of resilience, setting a precedent for future research and application in digital content security and safeguarding against unauthorized use and misinformation in the digital age.

Keywords

Responsible AI, Generative AI, Image Watermarks, AI-generated Content Detection, Red Teaming.

1 Competition description

1.1 Background and impact

The advent of generative AI has ushered in a new era of digital media creation, yielding highly realistic images, videos, and audio that blur the lines between reality and synthetic production. This technological advancement brings with it significant challenges, notably in ensuring digital content authenticity. Deepfakes, a notorious byproduct of generative AI, pose a significant threat by enabling the creation of synthetic media that can misrepresent individuals, spread misinformation, and influence public perception, particularly during sensitive times such as political campaigns and conflicts.

In response to these challenges, our proposed NeurIPS competition focuses on stress-testing and validating the robustness of robust image watermarking — a critical field that straddles digital rights

*Competition coordinators.

Preprint. Under review.

management, content authentication, and secure communication. Watermarking techniques involve embedding information within an image in a manner that is imperceptible during normal viewing but can be detected or extracted to confirm the content's authenticity and ownership (Saberi et al., 2024). This is increasingly vital in today's digital landscape, where content is easily distributed and susceptible to unauthorized use and manipulation.

The competition is designed as the ultimate stress-test to challenge and elevate the state of watermark robustness in the digital era. It seeks to inspire innovation in the art of embedding undetectable and resilient watermarks within images, utilizing cutting-edge advancements in digital image processing, cryptography, and machine learning. Participants are invited to craft sophisticated attacks aimed at bypassing or removing watermarks without compromising image integrity, particularly focusing on those hidden within AI-generated images. This initiative underscores the critical importance of reliable content verification techniques in an age dominated by generative AI technologies, pushing the boundaries of what is currently achievable in digital watermarking.

Impact. Developing strong watermarking solutions has wide-reaching benefits: it protects intellectual property, ensures fair content distribution, and cuts down on copyright losses in industries like media and digital art. The competition drives AI innovation in detecting watermarks and improving digital security and authentication. Moreover, strong watermarking is key to fighting deepfakes and misinformation, helping verify the authenticity of content, which supports democracy and helps resolve conflicts by separating real content from manipulated ones.

Relevance and Participation. The inclusion of generative AI challenges amplifies the competition's relevance to the NeurIPS community, appealing to a wide audience interested in AI ethics, digital forensics, and security. It represents an intersection ripe for interdisciplinary research, bridging machine learning, digital watermarking, and generative AI. Given the pressing issues of deepfakes and misinformation, we anticipate strong global interest from AI and digital security experts, political scientists, journalists, and beyond. The urgent need for effective solutions guarantees a broad appeal, engaging not just domain experts but also policymakers and the public in the quest for digital content integrity.

Applications and Real-Life Scenarios. The competition will simulate real-world scenarios where participants identify watermarked content in digital media streams, reflecting applications in news platforms, social media, and content creation tools. These scenarios highlight watermarking's pivotal role in ensuring digital media credibility, especially in contexts like elections, international relations, and public health communication.

1.2 Novelty

This competition introduces a pioneering challenge in the realm of digital watermarking, focusing on the resilience of watermarks against AI-driven manipulations, particularly in AI-generated content like deepfakes. This is the first of its kind at NeurIPS, designed to bridge the gap in current watermarking techniques by addressing the increasingly sophisticated landscape of digital media manipulation. By incorporating a diverse and challenging dataset, including the latest in generative AI technology, the competition aims to spur innovation in watermark detection and removal without sacrificing image quality. This unique approach sets it apart, aiming to advance the field of digital watermarking into new territories of effectiveness and resilience.

1.3 Data

We will generate watermarked images with prompts from **DiffusionDB** (Wang et al., 2022) with a specific, self-trained diffusion model. The dataset has 1.8 million unique prompts collected from real users of the Stable Diffusion service. This dataset provides a rich and diverse range of contexts, reflecting real scenarios of how people create AI images. The DiffusionDB dataset is available under the CC0 1.0 License.

For rolling-leaderboard evaluation, we sample 10k prompts each time and generate non-watermarked images and watermarked images using these prompts. For the final evaluation, a ten-times larger set of prompts will be used. Prompts used by the competition will be randomly sampled and confidential. Only the watermarked images will be released to participants for attacking method development.

The enormous 1.8 million size of DiffusionDB minimizes the risk of cheating or overfitting to this publicly available dataset.

1.4 Tasks and application scenarios

Competition tracks. This competition contains two tracks, black-box attack, and beige-box attack tracks, with identical evaluation metrics and image data source (i.e., randomly sampled from the same pool of generated images), but in beige-box attack tracks, we additionally provide the attribute information.

This competition is structured around the dual tasks of embedding and removing digital watermarks in a manner that balances robustness with imperceptibility. The embedding challenge requires the organizing team to develop algorithms that can integrate watermarks into digital images subtly yet securely, ensuring they remain detectable even after sophisticated manipulation attempts. Conversely, the removal task by the participants focuses on algorithms capable of erasing watermarks without compromising the original media's integrity, a test of finesse and precision in image processing.

Participants in the competition will have the opportunity to engage with two distinct tracks of watermark attacks, reflecting varying degrees of access to underlying watermarking methodologies.

The **black-box attack track** simulates a real-world scenario where attackers must operate without knowledge of the watermarking technique employed. In this track, participants are provided with 10k watermarked images, yet the specific watermark method applied to each remains undisclosed. This track reflects a common situation in an industry where watermark methods are kept confidential to protect against unauthorized removal or tampering. The participants' challenge is to develop techniques that can effectively disrupt the watermark without any insights into the methods used to embed them, thereby testing the watermark's resilience in conditions that mimic actual adversarial attacks.

On the other hand, the **beige-box attack track** offers a more transparent approach by providing participants with images alongside labels indicating the watermarking methodology used. Although participants will not have access to the model used to generate the (watermarked) image – paralleling proprietary industrial models – the disclosure of the watermark methodology invites more innovative and intellectually stimulating solutions to break the watermark system. This track balances opacity and transparency to create a unique scenario that can elicit creative and scientifically interesting attack strategies. The beige-box track is designed to encourage a deeper understanding of watermark robustness and the development of novel attack vectors that could potentially counteract watermarking even when some information about the process is known.

Both tracks serve to test and improve watermarking technology against varying levels of attack complexity, thereby fostering the advancement of more resilient watermarking solutions that can withstand real-world challenges.

These tasks directly reflect critical issues in digital rights management and the fight against misinformation, applicable across various sectors, including media, entertainment, and secure communications. For instance, in the media industry, robust watermarking can deter piracy by tracing leaked content back to the source. In the context of misinformation, watermarks can help verify the authenticity of content circulating online, an essential tool for maintaining the integrity of public discourse.

The scientific and technical challenges presented by these tasks are considerable, requiring advancements in digital image processing, cryptography, and machine learning. However, they are grounded in real-world capabilities and are within reach of current technology, making them appropriate challenges for the NeurIPS community. The competition leverages a diverse pool of watermarking methods and a database of prompts of AI image generation, illustrating the universal relevance of robust watermarking. This broad applicability underscores the competition's significance, addressing both immediate and emerging needs in safeguarding digital content integrity.

1.5 Metrics

For the comprehensive evaluation of watermark robustness and attack potency, we utilize a multi-dimensional metric system that considers both watermark detection accuracy and image quality post-attack. We highlight the important aspects of metrics and evaluation design below.

- We apply the True Positive Rate at a stringent 0.1% False Positive Rate (**TPR@0.1%FPR**) as the watermark detection performance metric.
- We implement **five** diverse types of watermark methods, and the average detection TPR@0.1%FPR will be used as the overall watermark detection accuracy.
- We implement **eight** diverse types of image quality metrics and combine them as the aggregated quality degradation score.
- For the top submissions, for breaking ties when the aggregated image quality degradation scores are close, we apply **GPT4-vision** and **human judges** to evaluate the subtal image quality degradation.

All metrics used can be categorized as either watermark detection performance metrics or image quality depredation metrics. In the reminder of this subsection, we elaborate the detailed design respectively, and conclude by how we rank the submissions.

Watermark detection performance. For detection performance, specifically, we focus on the True Positive Rate at a stringent 0.1% False Positive Rate (**TPR@0.1%FPR**), acknowledging its criticality in maintaining low false positives in real-world applications. TPR@0.1%FPR not only addresses the inadequacies of alternative metrics such as the p-value and Area Under the Receiver Operating Characteristic (AUROC), but also is more challenging to be attacked and thus a good replacement of metrics such as detection and bit accuracies.

For each watermark method, we calculate the TPR@0.1%FPR metric on the union of two equally-sized sets of images: one is the non-watermarked images post-attack, and the other is the watermarked images post-attacked correspondingly. The ground-truth labels for detection of the non-watermarked images are 0 and are 1 for the watermarked counterparts. We require the competition participants to use not only the watermarked images but also an equal amount of non-watermarked images; however, they will not know which image is watermarked; see section 2.1 for details.

We consider five diverse types of watermark methods; see section 1.6 for details. A rich set of watermark methods prevents the easy success of attacks to overfit and tailor to one specific type of watermark. There is no one watermark method that is robust to every type of attack, and on the other side, it is very challenging to attack a large and diverse set of watermark methods successfully.

Image quality degradation. For image quality degradation, we argue that it is necessary to consider a diverse type of image quality metrics because it could be easy to overfit and fool a specific metric. The eight types of image quality metrics can be categorized into four groups.

- *Image similarities*, including Peak-Signal to Noise Ratio (**PSNR**) (Hore & Ziou, 2010), Structural Similarity Index (**SSIM**), and Normalized Mutual Information (**NMI**), which assess the pixel-wise accuracy after attack or distortions.
- *Distribution distances* such as Frechet Inception Distance (**FID**) (Heusel et al., 2017) and a variant based on CLIP feature space (**CLIP-FID**) (Kynkäänniemi et al., 2019).
- *Perception-based metrics*: Learned Perceptual Image Patch Similarity (**LPIPS**) (Zhang et al., 2018).
- *Image quality assessments*: **aesthetics** and **artifacts** scores (Xu et al., 2023), which quantify the changes in aesthetic and artifact features. Changes in aesthetics and artifact scores after the attack are used to evaluate the quality degradation.

Each of the eight quality metrics under consideration has unique ranges and scales. To develop an overarching image quality metric that synthesizes these diverse inputs, normalization into a common interval is crucial. We aim to estimate an appropriate interval using our "starting kit" of 26 types of attacks (see section 1.6 for details). We define the normalized scale for each image quality metric by assigning the 10% quantile value of all attacked images (within the "starting kit") as the 0.1 point and the 90% quantile as the 0.9 point. Quality metrics are always ranked in ascending order of image degradation. After normalizing these eight image quality metrics, we aggregate them by averaging, defining this as the normalized (and aggregated) quality metric.

Ranking the submissions. We have described the design of the two aggregated scores for each submission on watermark detection performance and image quality degradation, respectively. As the evaluation outcome, we plot the submission as a "dot" in the 2D plot of combined detection performance vs. quality degradation and show this plot to the competition participants. To rank the

submissions, we calculate the minimal distance of the submitted "dots" to the origin. Since both aggregated detection performance and quality degradation metrics are appropriately normalized to the $[0, 1]$ range, this distance is meaningful despite the performance and quality being on different units. The submission ("dot") with the smallest distance to the original is the winner, which means that it attacks the watermark detection performance significantly while not downgrading the image quality severely.

Other measurements In the case of ties and to further differentiate the top teams, we will make use of judges.

GPT4-vision has been used as a judge for identifying the effect of watermarks on text (Tu et al., 2023). Due to its multi-modal nature, we can also use it to assess image similarity. In particular, after we narrow down our entrants to the top 5 per the quality-versus-detection plot (described above), the entrants' images are sent to GPT4-vision to evaluate the similarity of the team's attacked image to the ground truth. The image quality scores of GPT4-vision judges will be calculated from the ranking of the teams based on their average ranking.

Humans , namely, a selection of organizers will also rank the image similarity of each team's attacked set to unattacked images without knowledge of the teams' identities. The image quality scores of human judges will be calculated based on the attack method's average ranking from this blind review process.

Other evaluation details. The error bars are not necessary since metrics like TPR@0.1%FPR are sensitive, and we are also using many types of image quality metrics. We will use a large enough test set of 1,000 images per watermark method (see section 2.1 for details) for intermediate evaluations and a ten times larger test set (and even larger if needed) for final evaluation to break ties between submissions with closed final score. The submissions, which are the attacked images, are not required to be reproduced by the organizer.

1.6 Baselines, code, and material provided

We will release a *"starting kit"* at least two weeks before the competition begins, encompassing the diverse set of baseline watermark attacks spanning three distinct categories. In the meantime, we will also provide the data loading and submission tools, the viewer of evaluation results, and the interactive leaderboard on *Hugging Face*.

Baseline attacks. Baselines include a rich set of **26 baseline attacks** (An et al., 2024) showcasing common vulnerabilities of image watermarks. We consider three categories of attacks, where each attack is applied with an appropriate range of four to six attack strengths.

- **Distortion attacks**, including (1) *Geometric distortions*: rotation, resized-crop, and erasing; (2) *Photometric distortions*: adjustments in brightness and contrast; (3) *Degradation distortions*: Gaussian blur, Gaussian noise, and JPEG compression; and (4) *Combo distortions*: combinations of geometric, photometric, and degradation distortions, both individually and collectively.
- **Image regeneration attacks**, including (1) attacks that use surrogate diffusion models to noise/denoise an image and (2) attacks that use surrogate VAEs to encode and decode images.
- **Adversarial attacks**, including (1) embedding attacks that perturb the image in the feature space while maintaining imperceptible changes in the image space and (2) surrogate detector attacks that train a surrogate watermark classifier and conduct adversarial perturbations based on it.

Watermark methods. We implement a rich set of **five** watermark methods to be attacked by competition participants, each representing unique techniques for embedding an invisible signature.

- **Stable Signature** (Fernandez et al., 2023): in-processing via model modification.
- **Tree-Ring** (Wen et al., 2023): in-processing via random seed modification.
- **StegaStamp** (Tancik et al., 2020): post-processing watermark.
- **DWT-DCT** (Al-Haj, 2007): watermark in frequency embedding.
- **RivaGAN** (Zhang et al., 2019): video watermark applied to images.

We will not publicize the implementation details of prominent watermarking algorithms nor open-source the actual code during the competition period. However, under the beige box attack track, watermarked images will be attributed with the name of the watermark method used, and competition participants are allowed to use the official paper and implementation of the watermark methods, see section 2.1 for details.

Other code and materials. Before the competition begins, we will generate a gigantic set of non-watermarked and watermarked images using the prompts sampled from DiffusionDB (which has 1.8 million unique prompts) and a specific self-trained diffusion model different from the publicly available diffusion model checkpoints. The image dataset will be divided into many small shards and hosted on *Dropbox*.

The data loading tool will randomly download 10k watermarked images (2k images per watermark method). The submission tool will upload the 10k attacked images and trigger automatic evaluation. Each participating team will only be allowed to submit a small number of times per day.

After evaluation, the results, in the form of a 2D plot of combined detection performance vs. quality degradation, together with the current ranking, will be released through a *Hugging Face Space* application on a rolling, daily basis.

1.7 Website, tutorial, and documentation

The competition instructions and guidelines will be fully detailed on our dedicated website, `https://wavesbench.github.io/`, serving as an entry point and a central hub for all participants. The suite of baseline attacks and evaluation standards generally follow from the WAVES (An et al., 2024) benchmark.

The site will include a comprehensive timeline, registration details, submission guidelines, and comprehensive documentation and tutorials to aid in using our data loading, submission, and result visualization tools. Emails of all organizers will be available on the website. The website will be prepared as soon as possible after the acceptance notification.

The baseline attack strategies (see section 1.6) will be open-sourced and hosted on *Hugging Face* to serve as a code reference and starting point for participants to design and implement their attacks.

2 Organizational aspects

2.1 Protocol

The competition begins on July 18 and ends on October 31. Entries submitted past the deadline will result in automatic disqualification. Below is a description of the process for individual participants or team representatives to participate in the competition.

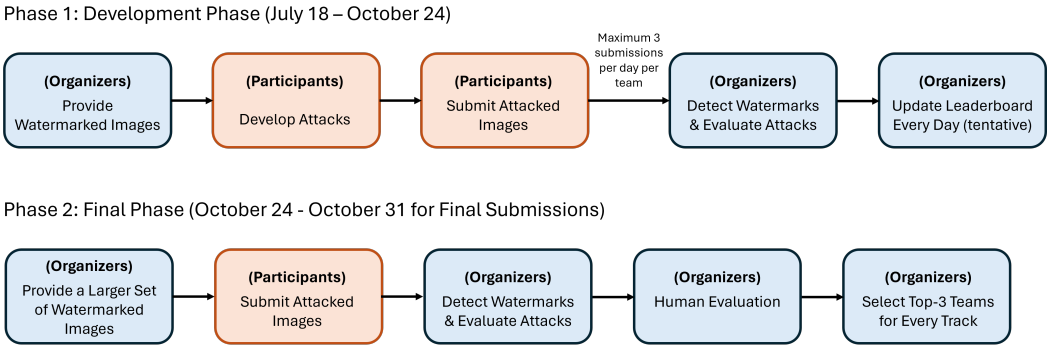

Figure 1: The general flow of this competition.

Protocol and dataset details.

- Participants must register on the official competition website one week before the end date of the competition (by October 24). They choose from two competition tracks, black-box and beige-box attacks, or both.
- The required documentation and a starter kit (as described in section 1.6) will be hosted on the competition website and accessible upon registration.
- The participants will be provided with API details from *Hugging Face* to download a dataset of 10k watermarked images (1k per five different types of watermarking methods) to perform the black-box or beige-box attacks.
- For the black-box attack track, the type of watermark method applied to each image is confidential, whereas for the beige-box attack track, this information will be provided along with images.
- The competition has two phases. During the development phase, the participants can upload their attacked images (each time on a random 10k images) a small number of times (temporarily set to three) per day. During the final phase, registration is closed, and participants need to upload their attacked images to a new and larger set of watermarked images (temporarily set to 50k images), as well as the attack descriptions through the competition interface.

Evaluation and validation.

- Participants are required to submit their attacked images for evaluation using our submission tool.
- Evaluation will be done on randomly selected 10k watermarked images (2k per watermark method).
- The evaluation results will be promptly released after each evaluation cycle (temporarily set to a day), contributing to a dynamic and continuously updating leaderboard.
- The outcome will be released as a 2D plot, with the X-axis representing aggregated image quality degradation metric and the Y-axis showing aggregated detection accuracy. As well as a leaderboard of attacks. These metrics and ranking criteria are detailed in section 1.5.
- The final evaluation will provide a comprehensive measure of the participants' methods against a vast and varied set of 50k watermarked and non-watermarked images.

Strategy and generalization of results.

- The strategy for the competition is to maintain a rolling leaderboard that reflects the participants' performance over time, with periodic updates after each evaluation cycle.
- The random sampling of the images ensures the generalization of results, preventing overfitting to a static set of images and reducing the risk of cheating across the black-box and beige-box attack tracks by cross-referencing the provided images.
- The final evaluation will assess the efficacy of the attacks over the entire dataset to ensure the robustness and generalization of the methods.

2.2 Rules and Engagement

Following is the draft of the contest rules.

- Open Format: This competition is open to all, with no age or area limitations. We encourage participants to share their approaches after the competition. The outstanding methods will be featured in a collaborative publication.
- Registration: Participants must be registered with their email address before submitting an entry. One account per participant. Double registration is not allowed. Every participant can only join one team. Teams may participate in multiple tracks. Organizers are not allowed to participate in the competition. There is no limit on team size.
- Submission Requirements: 1) Participants only need to submit images after their attacks. 2) The submitted images must follow the guidelines of naming and formatting. 3) During the competition period, we encourage general discussions, while teams are not allowed to share their codes with other teams.
- Prize Eligibility: We will provide monetary prizes to the top 3 teams for each track, subject to the reproducibility of the results. For prize eligibility and recognition, leading teams must at least provide their methods, code, and models to the organizers for verification. Leading teams will also be invited to give a talk in the competition workshop.

The competition's rules are designed to cultivate a fair, innovative, and collaborative environment. We will use a Google group to make announcements. We will create a Discord server for participants to ask questions, find teammates, and discuss.

2.3 Schedule and readiness

The following is the tentative schedule.

- Organizers prepare the event: April 17 - July 17
- Competition opens: July 18
- Development phase: July 18 - October 24
- Registration closes: October 24
- Final submission phase: October 24 - October 31
- Competition closes: October 31
- Organizers evaluate final submissions: November 1 - November 8
- Winning team announcement: November 9
- Organizers prepare the competition workshop: November 9 - December 9

At the time of writing this proposal, the code of some baseline methods and the code for evaluation metrics are ready. The watermarked images need to be generated, and the automatic evaluation pipeline remains to be prepared.

2.4 Competition promotion and incentives

To promote participation, we plan to distribute the cal through various channels.

- **Media.** We will distribute the call via Twitter, University mailing lists, and AI News providers.
- **Research Communities.** We will distribute the call to open research communities, including but not limited to https://mlcollective.org/, https://cohere.com/research, https://www.eleuther.ai/community, discord.gg/nousresearch. We already joined the Discords of the communities mentioned.
- **AI-image Communities.** We will distribute the call to communities that care about image watermarks, like https://www.midjourney.com/home and https://stability.ai/. We already joined the Discords of the communities mentioned.
- **Attracting Participants from Underrepresented Groups.** We plan to reach out to affinity groups that are underrepresented, including Black in AI (BAI), LatinX in AI (LXAI), Queer in AI, and Women in Machine Learning (WiML).
- **Presentation at Different Venues.** Organizers will include information about this competition in their upcoming presentations at different venues.

To incentivize participation, we plan to highlight the winning submissions and interesting submissions in many ways.

- **Prize.** We set prizes for the top three teams.
- **Publication** The outstanding methods will be featured in a collaborative publication.
- **Invited Talk.** The winning team will be invited to submit a short report of their methods. We will feature these reports on the competition website. The winning team will also be invited to give a talk in the competition workshop.
- **A Lightning Talk Session or a Poster Session.** We encourage all participants to share their approaches in the competition workshop either as a 3-minute lightning talk or a poster.

3 Resources

3.1 Organizing team

Our organizing team is composed of a multifaceted group of experts from various academic stages and institutions. It includes tenured faculty members at different points in their careers, as well as graduate and undergraduate students contributing a range of perspectives. Our team's graduate student organizers vary in experience from newcomers to those concluding their doctoral journeys. We are particularly honored to have Furong Huang and Bang An, two distinguished female scholars, playing central roles as competition coordinators and designers. The organizing institutions encompass the University of Maryland, Carnegie Mellon University, University of California - Santa Barbara, and Johns Hopkins University. Additionally, we have incorporated undergraduate students into the team to provide them with invaluable experience in prestigious academic environments like NeurIPS and to oversee our submission system.

Furong Huang (Coordinator, Platform Administrator) is an Assistant Professor in Computer Science at the University of Maryland. Specializing in trustworthy machine learning, AI for sequential decision-making, and high-dimensional statistics, Furong focuses on applying principles to solve practical challenges in contemporary computing to develop efficient, robust, scalable, sustainable, ethical, and responsible machine learning algorithms. Furong is recognized for her contributions with awards including best paper awards, the MIT Technology Review Innovators Under 35 Asia Pacific, the MLconf Industry Impact Research Award, the NSF CRII Award, the Microsoft Accelerate Foundation Models Research award, the Adobe Faculty Research Award, three JP Morgan Faculty Research Awards and Finalist of AI in Research - AI researcher of the year for Women in AI Awards North America. She serves as the IEEE Signal Processing Society Washington Chapter Chair. Furong has extensive experience organizing and chairing academic events, including NSF-Amazon "Fairness in AI" PI meeting January 2024, NSF-IEEE workshop: "Toward Explainable, Reliable, and Sustainable Machine Learning in Signal & Data Science" March 2023, "Dagstuhl Seminar on Tensor Computations: Applications and Optimization" 2020-2021 and 2021-2022, and "Matrix Factorization" Workshop at 5th Heidelberg Laureate Forum 2017.

Tom Goldstein (Coordinator, Data/Model Provider) is the Volpi-Cupal Endowed Professor of Computer Science at the University of Maryland, Director of the Maryland Center for Machine Learning, and co-PI of the NSF Institute for Trustworthy AI in Law and Society. His research focuses on the safety, security, and reliability of generative AI systems. Professor Goldstein has been the recipient of several awards, including SIAM's DiPrima Prize, a DARPA Young Faculty Award, and a Sloan Fellowship. Goldstein has previously organized eight workshops at major conferences, most recently "Trustworthy and Reliable Large-Scale Machine Learning Models" at ICLR 2023, "Adversarial Machine Learning in Computer Vision" at CVPR 2021, "ICML Workshop on Representation Learning for Finance and E-Commerce Applications" in 2021, and "Security and Safety in Machine Learning Systems" at ICLR 2021.

Soheil Feizi (Coordinator, Baseline Method Provider) is a faculty and the director of Reliable AI Lab in the Computer Science department at the University of Maryland, College Park (UMD). He has published over 100 peer-reviewed papers and given more than 50 invited talks. He has received multiple awards for his work including the ONR's Young Investigator Award, the NSF CAREER Award, the ARO's Early Career Program Award, two best paper awards, the Ernst Guillemin Thesis Award, a Teaching Award, and more than fifteen research awards from national agencies such as NSF, DARPA, ARL, ONR, DOE, NIST as well as industry such as Meta, IBM, Amazon, Qualcomm and Capital One. His recent work on the fundamental limits of practical attacks on watermarks was featured in Wired magazine.

Mucong Ding (Coordinator, Platform Administrator) is a fifth-year PhD student in Computer Science at the University of Maryland, College Park, advised by Dr. Furong Huang. His work broadly encompasses data efficiency, learning efficiency, graph and geometric machine learning, and generative modeling. His recent research focuses on designing a more unified and efficient framework for AI alignment and improving their generalizability to solve human-level challenging problems. He has published in top-tier conferences (NeurIPS, ICLR, CVPR, AISTATS, etc.), and some of his work has been recognized for Oral presentations and Spotlight papers (AISTATS 2021, Distribution Shifts Workshop at NeurIPS 2021, etc.).

Tahseen Rabbani (Coordinator, Platform Administrator) is a fifth-year Ph.D. candidate in Computer Science at the University of Maryland, College Park, advised by Dr. Furong Huang. His work broadly encompasses privacy, distributed learning, and learning efficiency, with a recent focus on watermarking for generative imagery. His work on fast, private, and memory-efficient learning strategies for resource-constrained clients has been published at ICLR, NeurIPS, and MSML. He has served on the program committees of the Privacy Regulation and Protection in Machine Learning Workshop at ICLR'24, ICDCS'24, and FL-NeurIPS'22. His research has earned him designations as an RSAC 2024 Security Scholar, COMBINE NSF Fellow (2021-2024), and 2022 Apple Scholar in AI/ML Nominee.

Bang An (Coordinator, Platform Administrator) is a fourth-year PhD student in Computer Science at the University of Maryland, advised by Prof. Furong Huang. She is a member of UMIACS. Before coming to UMD, she was a research staff member at IBM Research China. She received her bachelor's degree from Northeastern University (China) and master's degree from Tsinghua University. Her research focuses on Reliable Machine Learning with a particular interest in understanding and improving the robustness, fairness, generalization, and interpretability of deep learning models. She believes that reliable machine learning is a critical element in enabling AI to better serve and support humanity. Her recent works focus on the robustness of generative models, including AI-generated content detection, jailbreaking LLMs, and evaluating the robustness of image watermarks.

Vishal M. Patel (Coordinator) is an associate professor of electrical and computer engineering and a member of the Vision and Image Understanding Lab at Johns Hopkins University. His research interests are computer vision, machine learning, image processing, medical image analysis, and biometrics. Patel is an associate editor of the IEEE Transactions on Pattern Analysis and Machine Intelligence journal and chairs the conference subcommittee of IAPR Technical Committee on Biometrics (TC4). He has received a number of awards, including the 2021 IEEE Signal Processing Society (SPS) Pierre-Simon Laplace Early Career Technical Achievement Award, the 2021 NSF CAREER Award, the 2021 IAPR Young Biometrics Investigator Award (YBIA), the 2016 ONR Young Investigator Award, and the 2016 Jimmy Lin Award for Invention.

Yu-Xiang Wang (Evaluator) is an Associate Professor at UC San Diego in HDSI and CSE. Prior to UCSD, he was with CS@UCSB and co-founded the UCSB Center for Responsible Machine Learning. Yu-Xiang received his PhD in 2017 from Carnegie Mellon University (CMU). Yu-Xiang's research interests include statistical theory and methodology, differential privacy, reinforcement learning, online learning, and deep learning. His work has been supported by an NSF CAREER Award, Amazon ML Research Award, Google Research Scholar Award, and Adobe Data Science Research Award, and had received paper awards from KDD'15, WSDM'16, AISTATS'19, and COLT'21.

Lei Li (Evaluator) is an assistant professor at the Language Technology Institute, School of Computer Science at Carnegie Mellon University. His research interest lies in natural language processing, machine learning, and drug discovery. He has served as Associate Editor of TPAMI and organizer and area chair/senior PC for multiple conferences, including ACL, EMNLP, ICML, ICLR, NeurIPS, KDD, AAAI, IJCAI, WSDM, and CIKM. He has launched ByteDance's machine translation system (VolcTrans) and Xiaomingbot automatic writing system, and many of his algorithms have been deployed in production (Toutiao, Douyin, Tiktok, Xigua, Feishu/Lark), serving over a billion users. He has delivered five tutorials at ACL 2021, EMNLP 2019, NLPCC 2019, NLPCC 2016, and KDD 2010. He was a lecturer for the 2014 Probabilistic Programming for Advancing Machine Learning summer school in Portland, USA.

Mary-Anne Hartley (Evaluator) is an Assistant Professor in Biomedical Informatics and Data Science at Yale. Her research is focused on developing and validating novel data-driven tools designed to improve healthcare in low-resource settings, with a special interest in Africa. She is particularly interested in distributed and private methods for AI-based healthcare. She completed her undergraduate degrees at the Universities of Pretoria and Cape Town before moving to Switzerland, where she completed a PhD and MD at the University of Lausanne, with an MPH at the London School of Hygiene and Tropical Medicine. In 2019, she started the research group, "Intelligent Global Health" in the School of Computer Science at the Swiss Institute of Technology (EPFL) and continues this work in LiGHT (Laboratory for intelligent Global Health Technologies). Through these groups, she maintains a strong presence and partnership between EPFL and Yale through student exchange, research collaboration, and a visiting professorship.

Yuxin Wen (Baseline Method Provider, Evaluator) is currently pursuing a Ph.D. in Computer Science at the University of Maryland, College Park, under the guidance of Prof. Tom Goldstein. His research interests lie at the intersection of security, privacy, and machine learning, with a particular focus on generative models. Yuxin is dedicated to uncovering security and privacy vulnerabilities within machine learning systems, aiming to highlight their potential risks. Simultaneously, he is involved in developing innovative solutions to ensure user security and privacy in these systems.

Xuandong Zhao (Evaluator) is a fifth-year Ph.D. candidate in Computer Science at UC Santa Barbara, advised by Prof. Yu-Xiang Wang and Prof. Lei Li. He earned his B.S. in Computer Science from Zhejiang University in 2019. His research focuses on the intersection of machine learning and natural language processing, aiming to build solid theoretical foundations and practical algorithms for responsible generative AI. Specific projects include detecting AI-generated content via watermarking, evaluating and aligning large language models (LLMs) for safety, protecting the intellectual property of LLMs, and developing privacy-preserving language models. Xuandong has interned at Alibaba, Microsoft, and Google and is a recipient of the Chancellor's Fellowship from UC Santa Barbara.

Mehrdad Saberi (Evaluator) is a CS Ph.D. student at the University of Maryland since Spring 2023, advised by Prof. Soheil Feizi. His research mainly revolves around Foundation and Multimodal Models, covering topics such as image generation, vision-language models, and interpretability and robustness of machine learning methods. His work on the limits of practical attacks on watermarks was featured in Wired magazine and published at ICLR'24.

Souradip Chakraborty (Evaluator) is a 3rd year CS Ph.D. student at the University of Maryland, working on the Foundations of Trustworthy Machine Learning, with a focus on developing safe, reliable, deployable and provable RL methods for real-world applications such as robotics, healthcare, finance, etc. His research focuses on designing a reliable and unified framework for AI Alignment, focusing on statistical insights of biases in preference data along with their long-term ethical and societal implications. His recent research has focused on exploring the possibilities of AI-generated text detection using statistical and provable methods. He has co-authored top-tier publications and US patents in the field of Artificial Intelligence and Machine Learning. Recently received the Outstanding Paper Award (TSRML at Neurips 2022) and Outstanding Reviewer Awards, Neurips 2022, Neurips 2023 and AISTATS 2023.

Mo Zhou (Evaluator) is currently a Ph.D. candidate in the ECE department at Johns Hopkins University. He is advised by Prof. Vishal M. Patel. His research interest lies in computer vision and machine learning, especially securing deep neural network performance under distribution shifts. In particular, his research topics include adversarial attack, adversarial defense, adversarial attack detection, and watermarking image generative models. He organized or co-organized the AROW workshop at ICCV2023, and the AdvML workshop at CVPR2024.

Chenghao Deng (Evaluator) is currently a Ph.D. student in Electrical and Computer Engineering at the University of Maryland, College Park, under the guidance of Prof. Furong Huang. His research interests lie at the intersection of security, fairness, reinforcement learning, and foundation models. He is committed to identifying potential vulnerabilities within prevalent machine learning systems and proposing strategies to mitigate these risks.

Anirudh Satheesh (Beta Tester) is a first-year undergraduate student at the University of Maryland studying Computer Science (Machine Learning) and Applied Mathematics. He is currently researching cooperative multi-agent reinforcement learning and easy-to-hard generalization. Previously, he researched efficient data augmentations, with one paper accepted for publication at ICLR 2024, and denoising speech signals via GANs, with one paper published in the Journal of Student Research 2023. His interest in watermarking and security is leveraging adversarial attacks on input data to break watermarks and developing defenses to mitigate these attacks.

3.2 Resources provided by organizers

- **Prizes**:
  - First place: $1000;
  - Second place: $800;
  - Third place: $500.
- **Staff Support**:

  - Dedicated IT team from Center for Machine Learning, UMIACS at the University of Maryland;
  - Dedicated communication team from TRAILS AI Institute at the University of Maryland.
- **Fund Raising**: Concrete fundraising plans through established industrial points of contact at Microsoft, Google, OpenAI, Amazon, Capital One, JP Morgan, etc.
- **Hosting Server**: Center for Machine Learning's contemporary GPU node clusters and hosting on Hugging Face Spaces with funding provided.

3.3 Support requested

- Designated space at the conference for presentations and winner demonstrations.
- Video-graphy crew.
- Display panels for posters.
- On-site conference liaison.

References

Al-Haj, A. Combined dwt-dct digital image watermarking. Journal of computer science, 3(9): 740–746, 2007.

An, B., Ding, M., Rabbani, T., Agrawal, A., Xu, Y., Deng, C., Zhu, S., Mohamed, A., Wen, Y., Goldstein, T., et al. Benchmarking the robustness of image watermarks. arXiv preprint arXiv:2401.08573, 2024.

Fernandez, P., Couairon, G., Jégou, H., Douze, M., and Furon, T. The stable signature: Rooting watermarks in latent diffusion models. arXiv preprint arXiv:2303.15435, 2023.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems, 30, 2017.

Hore, A. and Ziou, D. Image quality metrics: Psnr vs. ssim. In 2010 20th international conference on pattern recognition, pp. 2366–2369. IEEE, 2010.

Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., and Aila, T. Improved precision and recall metric for assessing generative models. Advances in Neural Information Processing Systems, 32, 2019.

Saberi, M., Sadasivan, V. S., Rezaei, K., Kumar, A., Chegini, A., Wang, W., and Feizi, S. Robustness of ai-image detectors: Fundamental limits and practical attacks, 2024.

Tancik, M., Mildenhall, B., and Ng, R. Stegastamp: Invisible hyperlinks in physical photographs. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2117–2126, 2020.

Tu, S., Sun, Y., Bai, Y., Yu, J., Hou, L., and Li, J. Waterbench: Towards holistic evaluation of watermarks for large language models. arXiv preprint arXiv:2311.07138, 2023.

Wang, Z. J., Montoya, E., Munechika, D., Yang, H., Hoover, B., and Chau, D. H. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. arXiv preprint arXiv:2210.14896, 2022.

Wen, Y., Kirchenbauer, J., Geiping, J., and Goldstein, T. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. arXiv preprint arXiv:2305.20030, 2023.

Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., and Dong, Y. Imagereward: Learning and evaluating human preferences for text-to-image generation. arXiv preprint arXiv:2304.05977, 2023.

Zhang, K. A., Xu, L., Cuesta-Infante, A., and Veeramachaneni, K. Robust invisible video watermarking with attention. arXiv preprint arXiv:1909.01285, 2019.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In CVPR, 2018.