2025 NIST GenAI (Pilot) Evaluation Plan for Image Generators

Released: 2025-03-13

Updated: 2025-03-13

GenAI Eval Team

National Institute of Standards and Technology
100 Bureau Dr, Gaithersburg, MD 20899

Contact: genai-poc@nist.gov

Disclaimers

Certain commercial equipment, instruments, software, or materials are identified in this document to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor necessarily the best available for the purpose. The descriptions and views contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NIST or the U.S. Government.

Updates

2025-03-13	Initial Release

Table of Contents

1 Introduction

In recent years, digital content from generative Artificial Intelligence (AI), including deepfakes, has had unprecedented growth and proliferation across various modalities, including image, video, audio, and text. This surge in generative AI presents both opportunities and challenges. The technologies have facilitated creative expression enabling artists, designers, and writers to quickly generate visually stunning content as well as professional written content. On the other hand, it has raised concerns and issues regarding the authenticity and integrity of media in digital content. With the advancements in generative AI technology, it is becoming increasingly difficult to distinguish AI-generated content from human-generated content.

In this NIST Generative AI (GenAI) program, we invite and encourage participating teams from academia, industry, and other research labs to support research in Generative AI. GenAI is an evaluation series that provides a platform for testing and evaluation to measure the performance of AI content generators and AI content discriminators (e.g., detectors). The platform is intended to support multiple modalities and technologies enabled by both sides of the generative spectrum, "generators" and "discriminators."

Generator (G) teams will be tested on their system's ability to generate content that is indistinguishable from human-generated content. For the pilot study, the evaluation will measure the strengths and weaknesses in their approaches and may provide insights about how and when humans and/or AI can detect AI-generated content.

Discriminator (D) teams will be tested on their system's ability to differentiate between AI-generated content and human-generated content.

Lessons learned from both sides of teams should benefit future research directions and approaches to understand cutting-edge technologies as well as for providing recommendations and guidance for responsible and safe use of digital content.

The 2025 GenAI evaluation pilot study, discussed in this document, will focus on the image modality. In the pilot GenAI generator task, the objective of Image Generators (Image-G) is to automatically generate realistic images given a textual description. The textual descriptions will span a set of varied attributes across many categories representing realistic real-world scenarios. On the other side, the pilot Image Discriminators (Image-D) task is to detect if a target image was generated using a Generative AI system or not. The context of this evaluation assumes completely AI-generated content (cases where humans use AI tools to enhance content with no semantic changes such as resize, contrast, sharpness, etc. are not included). Please see the discriminator evaluation plan for the Image-D task for more details.

Participants are required to indicate if they are participating as a generator team, a discriminator team, or both. This document describes the task specification for "generator teams". Textual descriptions of scenarios for Generator participants to create their images will be provided by the NIST GenAI team, across different evaluation rounds. Generator teams are allowed to automatically supplement (if they desire) the given textual description prompt with sample images collected automatically to build their final prompt.

Any questions or comments concerning the GenAI evaluation series should be sent to genai-poc@nist.gov.

2 Tasks

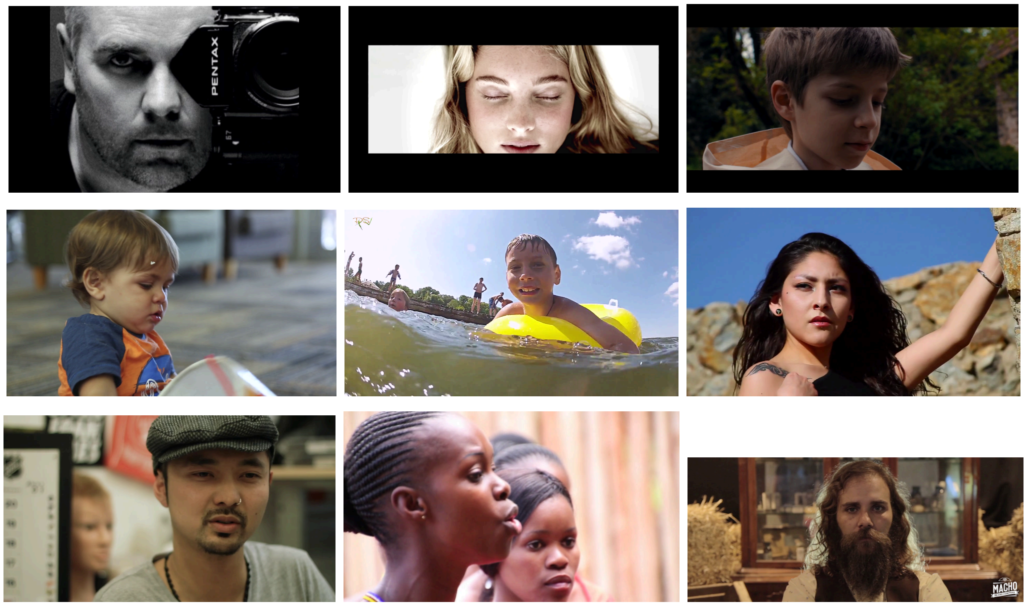
The primary goal of the pilot GenAI evaluations is to measure system behavior for creating AI-generated content. The Image-G participants will be given a list of topics in the form of a textual description for a set of images to be generated. Each text description is expected to be no more than 3 sentences per image.

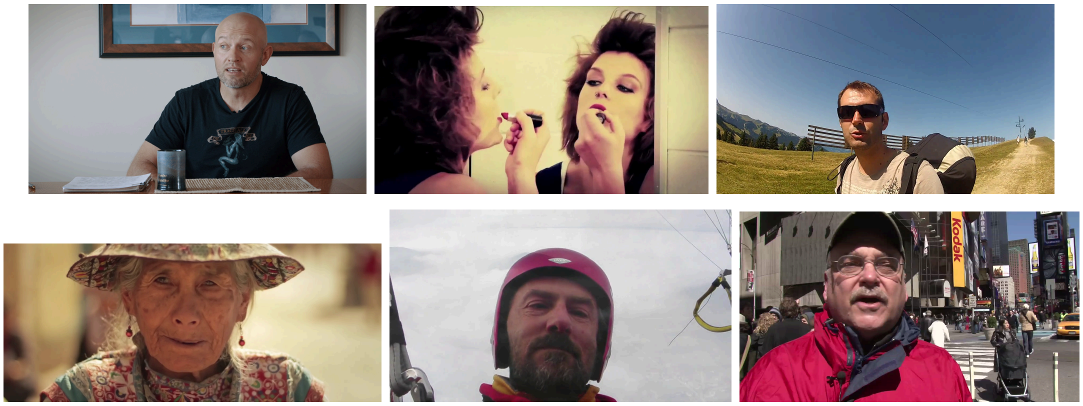
The task is to use Generative AI to automatically generate up to 10 realistic images per topic that could be viewed to satisfy the textual description of the topic. Those images may include a varied set of categories and attributes as illustrated in the table below and demonstrated by sample images under the table. Please note that the categories and attributes are not meant to be exhaustive but to illustrate a broad set of possible values.

Table 1: Examples of Category and Attributes

Category	**Attributes**
Sex	Male, Female
Age	Child (0-12 years) Adolescent (13-17 years) Young Adult (18-30 years) Middle-aged Adult (31-60 years) Senior (60+ years)
Lighting Conditions	Daylight Morning Noon Afternoon Night Artificial lighting Streetlights Poor lighting Evenly lit Backlit Low-light (indoors)
Environment	Indoor/Outdoor Day/Night Weather (e.g., sunny, rainy, snowy, cloudy, windy, foggy, ..) Home Office Factory/Warehouse Retail Urban Rural Beach Mountain Water (lake, river, ocean) Snow/ski slope Nature (forest, park, open field)
Background	Office/Workspace Factory/Warehouse Natural (mountain, forest)

	Urban cityscape Water (ocean, river) Beach Home interior Crowd/People Vehicles (cars, buses)
Facial Expression	Neutral Smiling Frowning Angry Surprised Sad Laughing
Accessories/Appearance	With glasses Without glasses Headgear (hat, helmet) Jewelry (earrings, necklace) Facial hair (beard, mustache) Makeup (heavy, light, no makeup) Hairstyle (short, long, tied, loose) Clothing type (formal, casual, work uniform)





This pilot evaluation does not address image manipulation contributing to deepfakes, however, this remains a potential research topic in future challenge problems.

For more information and details about the task specifics for generator teams, please refer to Appendix A.

2.1 Protocol and Rules

The participants are NOT allowed to use the test dataset for purposes of training, modeling, or tuning their algorithms. The participants may use any publicly available data that complies with applicable laws and regulations to train their models. All machine learning or statistical analysis algorithms must complete training, model selection, and tuning prior to running their system on the GenAI test data; learning/adaptation during processing is not permissible.

Each topic consists of a textual description. All topics must be processed **independently** of each other within a given round of evaluation and across all evaluations, meaning content extracted from the data must not affect the processing of another task's data.

Each submission (run) must contain results (images) for ALL topics (i.e. a submission should not miss or ignore a topic from the list of testing topics). A maximum of 10 images generated per topic per submission is allowed, where one of the 10 images MUST have applied the NIST GenAI topic textual description as is and without any modifications (this will be used as a baseline prompt, without an accompanying image(s)). The other up to 9 images can be generated using an automatically modified prompt (no manual intervention to modify prompts is allowed) to the provided official textual description.

Modified prompts by Generator teams may include the addition of 2 fixed image samples collected automatically by the system. This means that Generator systems may submit final images based on either a text only prompt, or text + image(s) prompt.

For all these 9 images, all applied prompts to generate the submitted images **MUST** be fixed across the set of images submitted per topic. Only the random number seed (used by the algorithm) can change. All applied prompts should be reported in the submission runs. Thus, for a topic in a submission, a team will be required to report the NIST prompt (textual description) used to generate an image X, and also the set of fixed prompts (text and any applied images) used to generate a set of up to 9 more images.

Each submission must only contain 1 run. A team will be allowed to submit multiple submissions as per the leaderboard submission guidelines. All images submitted must be in WebP lossless file format and have resolution (WxH) value only from the following set {1080x1080, 1138x720, 1280x534, 1280x560, 1280x570, 1280x720, 1280x960, 1920x1080, 1920x804, 1920x816, 3840x2160, 600x480, 640x360, 640x480, 720x540, 852x480, 960x720}. Resolution of images must vary within a topic. Meaning, the 10 images generated for a topic should have 10 different resolutions.

In addition, images submitted must preserve their EXIF metadata information. NIST will remove all EXIF information from generated images before releasing them to Image D-participants.

While participants may discuss their own results, they may not make advertising claims about their standing in the evaluation including results of other teams, regardless of rank, winning the evaluation, or claiming NIST endorsement of their system(s). The following language in the U.S. Code of Federal Regulations (15 C.F.R. § 200.113 (d)) shall be respected: NIST does not approve, recommend, or endorse any proprietary product or proprietary material. No reference shall be made to NIST or to reports or results furnished by NIST in any advertising or sales promotion which would indicate or imply that NIST approves, recommends, or endorses any proprietary product or proprietary material, or which has as its purpose an intent to cause directly or indirectly the advertised product to be used or purchased because of NIST test reports or results.

After all Image-G and Image-D evaluation rounds are concluded, NIST will generate a report summarizing the system results with anonymized team names. Participants may publish or otherwise disseminate these charts unaltered and with appropriate reference to the original NIST report results.

3 Data Resources

NIST will make all necessary data resources available to generator participants. Each team will receive access to data resources upon completion of the required data agreement forms and based on the published release dates for each task. Please refer to the published schedule for data release dates.

4 Submission Instructions

4.1 Data Generation Instructions

NIST GenAI team created a set of topics using a set of images covering a diverse set of attributes and categories (see Table-1 above and sample images). The goal of the descriptions and topics is to represent a real-world dataset of persons captured images as uploaded by general internet users.

Topics will be distributed by NIST. Only GenAI generator participants who have completed and submitted all required data agreement forms will be allowed access. As the example below shows, each topic includes an id (num), title, and the required topic prompt (textual description). Please check the published schedule (Section 4) for testing data release dates. All topics will be aggregated in one master XML file and released to Generator teams.

Example of topic:

```
<topic>
<num> topic_5445 </num>
<title> Young Man  </title>
<prompt>
Generate an image of young man with glasses looking upward to the sky
</prompt>
</topic>
```

4.2 Data Submission Guidelines

- Each team may submit up to 10 runs per evaluation round. Each run should include a maximum of 10 images per topic. A submission run should be in the format of a single compressed zip file to include the run results xml file (see below for specifications), any sample images used as part of the prompt, and all images generated for all topics. **No multiple zip files are allowed per run**.

- Please use the following folder naming conventions for your run submission:

  - images_prompts : folder name (as applicable) to contain any images used as part of any topic. Please use the following descriptive file names format to easily identify topics that images have been part of (topic.1.prompt.1.webp, topic.1.prompt.2.webp, etc).

  - images_generated : folder name to contain all images for all topics (topic.2.image.1.webp, topic.2.image.2.webp, … etc)

- Each run should contain images for all topics; a run can not skip a topic or submit images for a subset of the topics.

- All images must be submitted as webP lossless files, and follow the guidelines for aspect ratios and/or resolution allowed (see guidelines above for allowed values). EXIF metadata must be preserved in the images.

- Please remember to dedicate 1 image out of the 10 to be generated using the NIST-provided textual description. The other 9 images should all be generated using a fixed prompt built automatically (not manually).

- Modified system prompts may include sample images (i.e. final prompt is composed of text + image(s)). If a team decides to add sample images to the prompt, then a maximum number of 2 images should be used per prompt. These images are optional and a team may decide to only use text prompts only. To be specific, these image samples should be ***fixed*** for all prompts within a topic (a team must use the ***same*** 1 or 2 image samples as part of the prompt for all generated images within a topic). For example, given 150 topics, the maximum number of image prompts that a team may use will be 300 images.

- Image content should be free from offensive visuals or inappropriate remarks. Prohibited images include, but are not limited to, any image with explicit nudity, sexual content, hate symbols, child exploitation, or non-consensual images. NIST has the right to exclude any image or whole runs if it determines the content to be inappropriate for the general public. The GenAI team will have a process to automatically detect image content and any flagged images will be reviewed by a team of NIST personnel for final recommendation.

- Each run should include high-level metadata to characterize the generator system as requested by the below run format and DTD file. As explained in the DTD file, teams need to provide some required information/parameters, such as:

  - teamName: The name of the team as registered on the NIST GenAI website

  - trainingData: Name of training dataset or collection of different datasets or source data

- priority: The priority of the submitted run (the lower number, the higher the priority). For any required manual review of submissions, NIST may need to limit effort to only the highest priority runs. Priority should be an integer with a range between 1 to 10

- trained: A boolean (T or F) to indicate if the run was the output of a trained system by the team specifically for this task (T) or the output of an already existing system that the team used to generate the outputs (F)

- desc: Name of the base model used by the system.

- link: A link to the model used to generate the run (e.g. GitHub, etc)

- topic: The topic id (the “num” field in the topic xml file)

- elapsedTime: The processing time of the model per topic to generate the image after the topic was presented to it.

- usedImagePrompts: A boolean (T or F) to indicate if a topic has used (T) sample images (up to 2 images) as part of the prompt in addition to text, or did not use any images (F).

- filename: The name of the generated image file name. Please use the image file naming convention as demonstrated in the example below (eg. topic.1.image.1.webp, topic.1.image.2.webp, etc) to save and transmit your image files.

- prompt: the text description (prompt) used by the system to generate the corresponding image.

- NIST-prompt: A boolean to indicate if the generated image was based on the provided based on the official prompt (T), or based on a modified prompt by the team (F).

Example of a sample run:

```
<!DOCTYPE GeneratorResults SYSTEM "GeneratorResult.dtd">
<GeneratorResults teamName="participant_1">
  <GeneratorRunResult trainingData="OpenAI" priority="1" trained="T" desc="This run uses the
top secret x-component" link="TBD">
    <GeneratorTopicResult topic="topic_1" elapsedTime="5" usedImagePrompts=”F”>
      <Image filename="topic.1.image.1.webp" prompt="prompt_NIST_text_description"
NIST-prompt=”T”/>
      <Image filename="topic.1.image.2.webp" prompt="fixed_prompt_1" NIST-prompt=”F”/>
      <Image filename="topic.1.image.3.webp" prompt="fixed_prompt_1" NIST-prompt=”F”/>
       ….
       ….
    </GeneratorTopicResult>
    <GeneratorTopicResult topic="topic_2" elapsedTime="5" usedImagePrompts=”T”>
      <Image filename="topic.2.image.1.webp" prompt="prompt_NIST_text_description"
NIST-prompt=”T”/>
      <Image filename="topic.2.image.2.webp" prompt="fixed_prompt_2" NIST-prompt=”F”/>
      <Image filename="topic.2.image.3.webp" prompt="fixed_prompt_2" NIST-prompt=”F”/>

       ….
       ….
    </GeneratorTopicResult>
    <!-- ... -->
    <GeneratorTopicResult topic="topic_40" elapsedTime="5" usedImagePrompts=”F” >
      <Image filename="topic.40.image.1.webp" prompt="prompt_NIST_text_description"
NIST-prompt=”T”/>
      <Image filename="topic.40.image.2.webp" prompt="fixed_prompt_40" NIST-prompt=”F”/>
      <Image filename="topic.40.image.3.webp" prompt="fixed_prompt_40" NIST-prompt=”F”/>
       ….
    </GeneratorTopicResult>
  </GeneratorRunResult>
</GeneratorResults>
```

4.3 Generator Data Submission Validation

- NIST will provide, prior to submission dates, a validator script to participants to validate their output XML file format as well as content specific to the task guidelines (e.g. topic ids, empty required attributes, etc). All generator teams should validate their runs before submitting them to NIST. Example of available DTD validators (via a shell script): xmllint --valid simple_sample.xml

- **Submission instructions:** according to the published schedule, the submission form will be open and available (via the GenAI website) for teams to submit their data outputs based on the specified format. Please make sure to follow the schedule and submit on time, as extending the submission dates may not be possible.

- Upon submission, NIST will validate the data outputs uploaded and report any errors to the submitter.

- Please take into consideration that submitting your data outputs indicates and assumes your agreement to the data transfer agreement and rules of behavior.

4.4 Data Agreement

All generator teams submitting data generation outputs will be required to complete and sign a Data Transfer Agreement and a Data Usage Agreement before uploading their data outputs.

4.5 System Descriptions

The NIST GenAI team will request a system description for the top-performing systems in their submissions. Documenting a system is vital to interpreting evaluation results. Please make sure you document this information while developing your system and submitting your results. A system description should include, but is not limited to, the following information:

Section 1. Submission Identifier(s)

List the submission IDs for which system outputs were submitted; the GenAI team can help identify Submission IDs as needed.

Section 2. System Description

A brief technical description of your system and the system model used.

Section 3. System Hardware Description and Runtime Computation

Describe the computing hardware setup(s) and report the number of CPU and GPU cores. A hardware setup is the aggregate of all computational components used.

Section 4. Training Data and Knowledge Sources

List the resources used for system development and runtime knowledge sources, if any.

Section 5. References

List pertinent references, if any.

5 Performance Metrics For Generator Data Output

The detailed information about the metrics described below can be found in the image discriminators evaluation plan Section 6.

A Generator system A is considered successful against a discriminator system B if B has difficulty distinguishing between AI-generated content from A and human-created content (e.g., a photo of a real object/person, real artwork, or real-life actions). Generator A's performance against discriminator B will be summarized using the AUC and the BrierT score for B when challenged with a mix of human-created content and AI-generated content using A.

AUC values close to 1 (and BrierT and BrierN scores close to 0) indicate good performance of discriminator B against generator A but poor performance for generator A against discriminator B. On the other hand, AUC scores equal to 0.5 or lower indicate poor performance for discriminators but good performance for generators. An AUC score close to 0.5 suggests that the discriminator is unable to differentiate between the distribution of system scores for AI-generated content and the distribution of system scores for authentic (human-created) content.

Sometimes a generator system A may be able to induce a discriminator system B to assign confidence scores to the AI generator content that turn out to be lower than the confidence scores assigned to human-generated authentic content. In such cases, the AUC score can be less than 0.5 and, theoretically, can even be 0. This is the reason for calculating and displaying two separate Brier scores – (1) BrierT score for AI-generated content, and (2) BrierN score for human-created authentic content. The goal for generator systems would be to drive down the discriminator AUC score (less than or equal to 0.5 is better) and drive up the discriminator BrierT scores (closer to 1 is better) for AI-generated content against all participating discriminator systems. The goal for discriminator systems would be to get a high AUC score (closer to 1 is better) and low BrierT scores (closer to zero is better) against all participating generator systems and low BrierN scores for human-generated content.

Discriminator system detection scores will not be available until D-participants submit their results on G-participants' data submissions. Hence, immediately after each Round for G-participants, we will be able to report the performance based only on some baseline algorithms of interest. Please refer to the published schedule.

Appendix A Generator Validator Script

G-Validator Script Usage

```
# validate Image-G system output

$ python validate_generator_sysout.py -s /path/to/submission_directory -t /path/to/topic.xml -d /path/to/dtd_file
```