SOGAI

Subjective and Objective Quality Assessment of GenAI Content



Output of DALL-E when asked to visualize “AI generated content”.

This group is investigating subjective methods to evaluate the content produced by generative AI methods.

Related documents:

- Project page (https://vqeg.org/projects/sogai/)
- Kickoff meeting
- **Presentation schedule**

Description

The rapid advancement of Generative AI (GenAI) has introduced novel ways of creating text, video, audio, and image content, opening up transformative opportunities across various applications and industries. GenAI-generated content differs significantly from traditional, camera-captured media, necessitating new approaches for evaluating its quality, fitness for purpose, and user acceptability.

To address these needs, this initiative seeks to standardize both **subjective testing methodologies** and **objective metrics** for assessing the quality of AI-generated content.

Subjective evaluation requires rethinking traditional methodologies, questionnaires, and statistical analyses to accommodate the unique attributes of GenAI content. For example, questions like “Rate the quality of this video” must be expanded to address specific aspects such as translation accuracy, lip-sync quality, style adherence, prompt consistency, and overall plausibility. Before such aspects are considered, it is important to define them. Plausibility for example, is very much context dependent. If the “generator” asks for something not traditional (hence unplausible, for example Spiderman dancing in La la Land), it should be considered otherwise.

Furthermore, defining reliable objective metrics as proxies for subjective evaluations is critical, given the high scale and evolving nature of GenAI content production.

The scope of this effort encompasses a range of GenAI paradigms. For example:

- **Prompt-Driven Content Creation**: Generating entirely new videos based on user prompts (e.g., “Create a video of a cat dancing on ice”) or transforming existing media to align with specified styles or themes.
- **Automatic dubbing**: Translating videos between languages while ensuring accurate lip-sync and voice matching.
- **Style Transfer**: Modifying existing videos to adopt the style of a given image or video.

This initiative also acknowledges the challenges posed by subjective testing for GenAI content, such as diverse user expectations, the inherent subjectivity of evaluating “prompt adherence,” and the varying nature of questions across use cases. Engaging the broader GenAI community, including researchers and practitioners contributing to conferences like CVPR and ICIP, will be essential for ensuring alignment and adoption of standardized methodologies.

By establishing robust frameworks for subjective and objective evaluation, this group aims to support the reliable assessment of GenAI content, fostering its broader adoption and enhancing user trust in AI-generated media.

November 10, 2025

- Summary of October events (slides)
    - QoMEX
    - VQA Competition at ICCV
    - Wallenberg Forum on “Evaluation as a Driver of Generative AI”

September 22, 2025

- BMVA symposium on Image Quality in February 2026:
    - https://www.bmva.org/meetings/26-02-04-ImageQuality.html
- ICIP
    - Objective metrics as “holy grail”

August 25, 2025

- Transcript:
  - https://otter.ai/u/GxWGb8fRkEqG9zEmIL4KMQ9SoFY?utm_source=va_chat_link_1
- Presentation by Zhengzhong Tu on using VLMs for IQA
  - Slides
  - Other related papers:
    - Q-CLIP

July 28, 2025

- Transcript:
  - https://otter.ai/u/Ve-6pFoyHAcMaKLdiJxiRPky99w?utm_source=va_chat_link_1
- Luisa Verdoliva's "plenary talk at ICME 2025: "**Synthetic Media Verification in the Era of Generative AI**"
- "What makes an image realistic?"

June 30, 2025

- Transcript:
  - https://otter.ai/u/Y09NLx3WV2QrHCuoK02MOC86178?utm_source=va_chat_link_1
- CDVL
  - Support for larger files and files unsuitable for wider distribution
- Do a deep dive into the VLMs for IQA/VQA next time? (zhengzhong, abhijay)
- IQA for machine vision tasks?
  - Perceptual Quality Assessment for Embodied AI https://arxiv.org/abs/2505.16815 (maybe related to security tasks (margaret))

June 3, 2025

- Transcript:
  - https://otter.ai/u/MffgWQ2KY4cAnaVauKMVZPdxgWs?utm_source=va_chat_link_1
- Discord
  - https://discord.gg/Cv3F3rNE
- Move to seminar style meetings with presentations?
  - **Presentation schedule**
- Recordings on YouTube:
  - https://www.youtube.com/@vqegofficial3248/videos
- ICASP

April 1, 2025

- **Existing subjective test methodologies and objective metrics, and their shortcomings:**
    - Please contribute examples to this document:
      Objective metrics and subjective tests.pptx
    - VQEG presentations by Nabajeet/Abhijay/Saman et al.:
      VQEG-AIGC

- Do a small study with SOGAI experts?
- Interesting questions:
    - Define dataset and prompts
        - Sony
        - Lukas Krasula
            - Design standard dataset that's used by many groups to evaluate image models
            - Come up with challenging prompts that don't favor any specific model
    - Number of subjects required
- **Planned subjective studies:**
    - Prompt-driven content creation:
        -
    - Automatic dubbing:
        -
    - Image/video enhancement:
        - Evaluating submissions to Challenge on Learned I mage Compression [Lucas Theis]
            - Images: 2AFC with / without reference
            - Video: 2AFC vs ACR
            - 2AFC is also commonly used for text-to-image/video generations
            - Different rater pools:
                - Biased experts (participating teams)
                - Unbiased experts (CLIC organizers)
                - Crowd-sourced raters
- **Planned work on objective metrics:**
    -

February 24, 2025

- Next meeting: **1.4.2025**
    - Goal: Draft text highlighting problems of existing subjective and/or objective evaluations.
- Timeline and objectives?
    - Review work vs recommendation (white paper, standard)

        - Review helps us identify shortcomings of previous work
            - Field is moving quickly, can't spend too much time reviewing
            - Rely on collective wisdom already present in the group to create quick draft highlighting problems
        - Existing reviews (from VQEG members)
            - Quality Prediction of AI-Generated Images and Videos: Emerging Trends and Opportunities

          Some additional papers from team at TUIL:
            - Analysis of Appeal for Realistic AI-Generated Photos | IEEE Journals & Magazine
            - https://www.researchgate.net/publication/386880768_Appeal_prediction_for_AI_up-scaled_Images

- 1 "generation" track, 1 "enhancement" track
- **Example applications of GenAI**
    - **Prompt-driven content creation**
        - "Create a video of a cat dancing on ice"
        - Text input or image based prompting
            - Which dimensions are most relevant to evaluate?
    - Code generation and autocompletion
        - Examples: GitHub Copilot, JetBrains AI Assistant (integrated in PyCharm)
        - General feeling of the group: Let's consider text as input but not as output.
    - **Automatic dubbing** [Ioannis]
        - Meta's AutoDub
    - Style transfer
        - Modify an existing video to adopt the style of a given image or video
    - Character-based AI photo and video generation
        - https://photoai.com/
        - Consistency of multiple views
    - Real-time interactive avatars [Lucas]
        - http://anam.ai
    - Customer service calls [Lucas]
        - https://bland.ai
    - Computational photography
        - Inpainting
        - Super-resolution
    - Image correction and inpainting in mobile 3D scanning
        - Example: Poly.cam's AI-based auto-correction and inpainting *(Poly.cam is a mobile app that uses AI to enhance 3D scans by automatically fixing imperfections, completing missing parts, and improving textures.)*
    - "Extreme" compression
        - Text + sketch

    - **Image/Video Enhancements** [Abhijay]
        - Selective enhancement (denoising/deraining/deblurring) based on text
        - Selective compression for parts of the image/video
        - Deepfakes
    - Avatarisation of image and videos (e.g. sensitive data such as medical images) [Meriem_QAH]
    - AI-assisted anonymization and augmentation of medical images *(Use cases: modifying images to protect patient data while maintaining usability for AI training.)*
    - *Synthetic training video/image generation using text-prompts to train AI models such as for object detection, tracking and other models that require large video or image datasets. [Martin]*
- **Literature review**
    - **Subjective tests**
        - https://faculty.washington.edu/weicaics/paper/papers/KeWLAC2025.pdf
    - **Benchmarks**
        - VBench
        - Q-Bench
        - https://imgsys.org/
        - Holistic Evaluation of Text-To-Image Models
        - ImagenHub
        - EvalCrafter
    - **Metrics**
        - FVMD
        - AIGV-Assessor
        - https://arxiv.org/pdf/2501.18314

January 20, 2025

Description

- Generative AI (GenAI) is a new way of creating video content that is gaining popularity and offers multiple new opportunities for adoption in various products
- These videos have different characteristics from traditional, camera-captured videos and as such there is need to assess their quality, fitness for the purpose they have been created and acceptability of the final users of the service or product that has generated them
- The subjective assessment of these GenAI videos requires different methodologies, questionnaires, and statistical evaluation methods
- Defining objective metrics that can be used as proxies for subjective assessment is the ultimate goal of the video quality community that is even more important in this new

GenAI paradigm, given the very high scale of such videos that makes it very difficult to perform continuous subjective evaluation of evolving models

- This group attempts to address both methodologies and metrics that can be used to perform such subjective and objective assessment of GenAI content
- Paradigms of such GenAI include:
    - “AutoDub” (recent announcement from Meta) - translating existing videos from one language to another, offering visual content that matches the translated language
    - Modifying existing videos to match another image or video style (style transfer)
    - Completely new content generated by a user text, audio, image or video prompt (“Create a video of a cat that is dancing on ice”), or “Make a video that animates this image” (+an image input), or “Make this video look as if it was shot in 1923” (+a video input)
- Challenges:
    - The questions asked to subjects about GenAI content can be vastly different from the standard quality assessment of traditional videos (“Rate the quality of this video on a scale of 1-5”).
    - For the AutoDub application, the questions varied from “Is this translation accurate ?”, “Is the translated voice similar to the original language voice ?”, “Is there good lip-sync in the autodubbed video ?”, “Are you satisfied by the generated video ?”, “How good is the audio quality ?”
    - For newly generated content, the main question is typically “How plausible is the generated video ?”, or “How consistent is the generated video with the supplied prompt ?”, on top of “How good is the audio and video quality of the content ?”
    - The definition of “adherence to the prompt” is highly subjective and may create issues when aligning multiple subjective scores on this question.
    - The generative AI community does not necessarily follow the ITU standardization process. How do we reach this community? For example, challenges or papers at CVPR, ICIP.