REPRODUCIBLE SUBJECTIVE EVALUATION

Max Morrison,* Brian Tang, Gefei Tan & Bryan Pardo
Northwestern University
morrimax@u.northwestern.edu

ABSTRACT

Human perceptual studies are the gold standard for the evaluation of many research tasks in machine learning, linguistics, and psychology. However, these studies require significant time and cost to perform. As a result, many researchers use objective measures that can correlate poorly with human evaluation. When subjective evaluations are performed, they are often not reported with sufficient detail to ensure reproducibility. We propose Reproducible Subjective Evaluation (ReSEval), an open-source framework for quickly deploying crowdsourced subjective evaluations directly from Python. ReSEval lets researchers launch A/B, ABX, Mean Opinion Score (MOS) and MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) tests on audio, image, text, or video data from a command-line interface or using one line of Python, making it as easy to run as objective evaluation. With ReSEval, researchers can reproduce each other's subjective evaluations by sharing a configuration file and the audio, image, text, or video files.

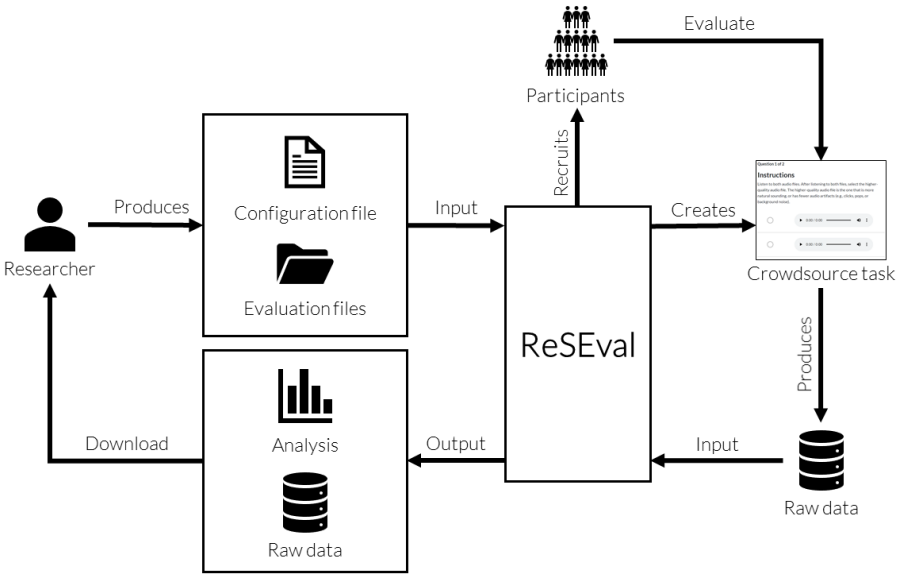

Figure 1: ReSEval system flow. A researcher creates a subjective evaluation by providing a configuration file and a directory of evaluation files as input. ReSEval creates a crowdsource task and recruits participants via, e.g., Amazon Mechanical Turk (MTurk). Participants complete the task, producing evaluation data for analysis. ReSEval analyzes the evaluation data and presents the researcher with a statistical analysis. With ReSEval, the researcher does not have to perform any web development. As well, aside from a one-time acquisition of API keys, the researcher does not have to interact with any third-party services (e.g., MTurk, Heroku, or Amazon Web services). Instead, ReSEval performs all of the necessary interactions with third-party services to configure and manage databases, servers, file storage, and crowdsourcing on behalf of the researcher.

*This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1842165.

1 INTRODUCTION

Subjective human evaluations of audio, image, text, or video data are a standard evaluation methodology in machine learning (Sheng & Zhang, 2019), linguistics (Cole et al., 2017), and psychology (Kazdin, 2021). For example, a machine learning researcher may want to know whether a conditionally generated image corresponds to its conditioning (Ramesh et al., 2021; Shoshan et al., 2021), and a linguist may be interested in measuring how humans perceive the pitch of speech (Cole et al., 2017).

Recruiting a large number of participants for in-person evaluations is a time-consuming process. For some types of tasks, comparable results can be achieved by performing evaluation on a crowdsourcing platform, such as Amazon Mechanical Turk (MTurk) or Prolific. As long as the evaluation is designed properly (Cartwright et al., 2016; Sai et al., 2020), such evaluations can be effective substitutes for in-person evaluations. However, researchers are often faced with strict page limits for conference submissions, and details of the evaluation important for replicating the study are routinely omitted from published papers. Frequently omitted details include the specifics of the methods used to prescreen participants, the wording of the instructions, and the type of sampling used to assign evaluation files to participants.

Even in the case where something like a prescreeing test is described in the publication, it can still be difficult to reproduce. A prescreening test is often performed to establish whether participants have the perceptual acuity to perform an evaluation task (e.g., passing a hearing test prior to evaluating a text-to-speech system). While crowdsourcing platforms such as MTurk offer some ability to filter participants based on qualifications, perceptual screening tests (Cartwright et al., 2016; Woods et al., 2017) are typically not available in the default platforms, requiring researchers to develop and deploy their own prescreening code, which others would need access to if they are to truly reproduce the work.

While MTurk contains many easy-to-use evaluation templates for some tasks, as of March 2022 they do not support many common research tasks in machine learning (e.g., A/B tests of images for image super-resolution, ABX tests of text for continuation tasks, or MUSHRA tests of audio quality). Finally, for machine learning researchers, platforms such as MTurk do not integrate easily into the typical Python-based software development workflow. Our contribution is more similar to psiTurk (Gureckis et al., 2016), a Python MTurk API wrapper and Flask web server used for research tasks in behavioral sciences. However, using psiTurk necessitates significant web development skills (at least HTML, CSS, and Javascript) and requires manual server configuration. As well, reproducing an experiment with psiTurk is significantly more cumbersome, as the steps necessary to reproduce an experiment vary substantially across experiments.

Given the issues associated with performing high-quality, reproducible subjective evaluation, many researchers have proposed proxy metrics that can be used to estimate human perception. However, whenever such an objective measure is proposed, subsequent research demonstrates its failures. For example, Frechét Inception Distance (FID) (Heusel et al., 2017), a common objective metric for image generation, is biased towards its training data and is not aligned with human perceptions (Jung & Keuper, 2021). Peak Signal-to-Noise Ratio (PSNR) is frequently used for image and vision denoising (Zhang et al., 2020), but weights each pixel as equally important to visual perception and is not able to measure the perceptual impact of various types of distortions (Wang & Bovik, 2009). Metrics for natural language understanding such as BLEU (Papineni et al., 2002) have been unable to produce even a moderate correlation with human evaluations of nautral language generation systems (Novikova et al., 2017; Sai et al., 2020). In speech synthesis, the DeepSpeech Distances (Bińkowski et al., 2019) of two state-of-the-art systems do not correlate with human judgments obtained via A/B testing (Morrison et al., 2022). Some works propose to learn an objective metric from human subjective evaluation data. For example, the Learned Perceptual Image Patch Similarity (LPIPS) is trained on A/B tests of image patches (Zhang et al., 2018), and the Contrastive Deep Perceptual Audio Metric (CDPAM) is trained on an A/B test of speech quality (Manocha et al., 2021), but the correlation of such measures with human perception is only guaranteed on the range of data on which the systems are trained.

What is needed is an approach to subjective evaluation that is truly reproducible and integrates well with machine learning development workflows. To this end, we propose Reproducible Subjective Evaluation (ReSEval), an open-source framework for performing crowdsourced subjective evalua-

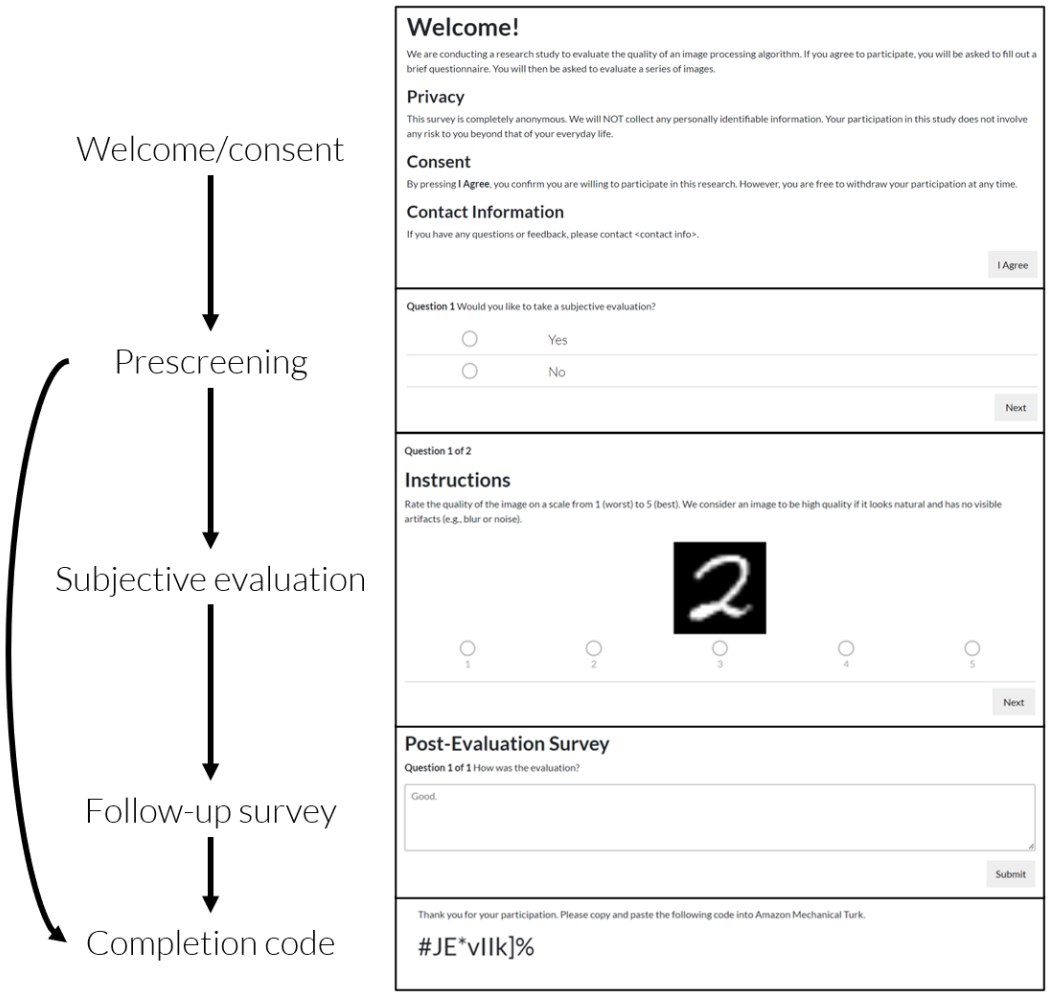

Figure 2: An overview of the stages of a subjective evaluation on ReSEval. After obtaining the consent of the participant, we perform a prescreening step. Participants who pass the prescreening step perform a subjective evaluation and take a follow-up survey. Participants who complete the follow-up survey or do not pass the prescreening are taken to the final page, which provides a completion code for participants to enter on the crowdsource platform to receive payment for their work. All of the text is configurable via Markdown. The type of test, type of question (free response or multiple choice), and number of questions are also configurable.

tions. ReSEval allows researchers to launch a subjective evaluation either from the command line or using one line of Python, making it as easy to run as objective evaluation. This enables complex machine learning evaluation pipelines, such as automatically evaluating whether a new generative model should replace a production model, or automatically determining when to stop training a model based on human preference scores. Further, ReSEval offers researchers a convenient way to make their subjective evaluations reproducible by their research community by simply releasing the configuration file and evaluation files they used when performing evaluation with ReSEval.

Papers-with-code and open GitHub repositories have provided a much-needed boost to machine learning research reproducibility. Deployment of trained models in online repositories, such as HuggingFace, has been another important step. ReSEval is an essential component to take the next step forward in reproducible research. Imagine online challenges or Papers-with-Code leader boards where all submissions are automatically evaluated using the exact same subjective study evaluation. Imagine being able to verify all the details of a study design because its exact implementation was

posted on GitHub. ReSEval makes this possible. ReSEval is available under an open-source license at `github.com/reseval/reseval`.

2 BEST PRACTICES FOR CROWDSOURCED SUBJECTIVE EVALUATIONS

We next discuss best practices for performing crowdsourced subjective evaluation, focusing on best practices that are included by default or simple to implement with ReSEval. For a more thorough treatment of best practices, see (Mason & Suri, 2012).

Use the correct test By default, ReSEval includes many of the most widely-used subjective evaluation test paradigms, including A/B, ABX, MOS, and MUSHRA-style tests.

- **A/B** - The participant is asked which of two audio, image, text, or video examples ranks higher along a perceptual attribute (e.g., audio quality or text sentiment).
- **ABX** - The participant is asked which of two stimuli (e.g. two images) is, e.g., more similar to a reference stimulus (a canonical image).
- **Mean opinion score (MOS)** - The participant rates a perceptual quality of an audio, image, text, or video example from 1 (worst) to 5 (best).
- **MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA)** - The participant is presented multiple audio, image, text, or video examples and uses sliders to rate each of the examples.We refer to these as MUSHRA-style, rather than MUSHRA tests, because MUSHRA denotes a specific standard audio evaluation methodology, which is detailed in the International Telecommunications Document BS.1534 (Liebetrau et al., 2014)

It is important that the hypothesis being tested is consistent with the test format. For example, while MOS and MUSHRA-style tests are conveniently capable of evaluating more than one pair of conditions at once, they are less able to demonstrate a statistically significant difference between two conditions when the perceptual differences between those conditions is subtle relative to other conditions. An A/B test can expose these subtleties as statistically significant, even with many fewer participants.

This has implications for the design of ablation studies in machine learning: if the ablation conditions are significantly worse than the proposed model, an MOS or MUSHRA-style test that includes the ablation conditions, the proposed model, and a ground-truth reference will skew ratings of the proposed model towards ground-truth. For an example of this in the field of speech synthesis, see (Kong et al., 2020) and (Morrison et al., 2022). However, MOS tests can be a cost-effective way of comparing many conditions at once. Therefore, one cost-effective way to show statistical significance is to first conduct an MOS (or MUSHRA-style) test, and then perform A/B tests on pairs of conditions where the test shows an inconclusive preference. Usually, this process requires researchers to manually review results and deploy a second user study. ReSEval makes it possible for the researcher to specify in code when to automate the deployment of, e.g., an A/B test, conditional on the results of a prior test (e.g., an MOS test).

MOS tests are typically cheaper than MUSHRA-style tests as they require less time for participants to complete. However, they also have limited ability to rank-order samples relative to MUSHRA-style tests, which can be useful data for downstream tasks, such as training a machine learning model to approximate human subjective preference. For both MOS and MUSHRA-style tests, using high- and low-anchor conditions can minimize changes in the range and variance of the results between evaluations with different conditions. These anchors are representative conditions of very good and bad examples for the specific task being evaluated. For example, text-to-speech uses ground-truth recorded speech as the high-anchor, and might use speech with added noise as a low-anchor.

Prescreen It is important that participants be well-qualified to perform evaluation. Depending on the specific evaluation, this can include fluency in a particular language, having no vision impairment, or having access to headphones and a quiet listening environment. Crowdsource platforms permit filtering participants by some of these criteria. Other filtering criteria require prescreening to be included as part of the survey design. Reproducing this prescreening step is critical for reproducing a subjective evaluation. Deploying a study through ReSEval makes perfect reproduction of the prescreening simple, as the full specification is included. For audio tasks, ReSEval also includes a listening test that ensures the participant's listening conditions are suitable.

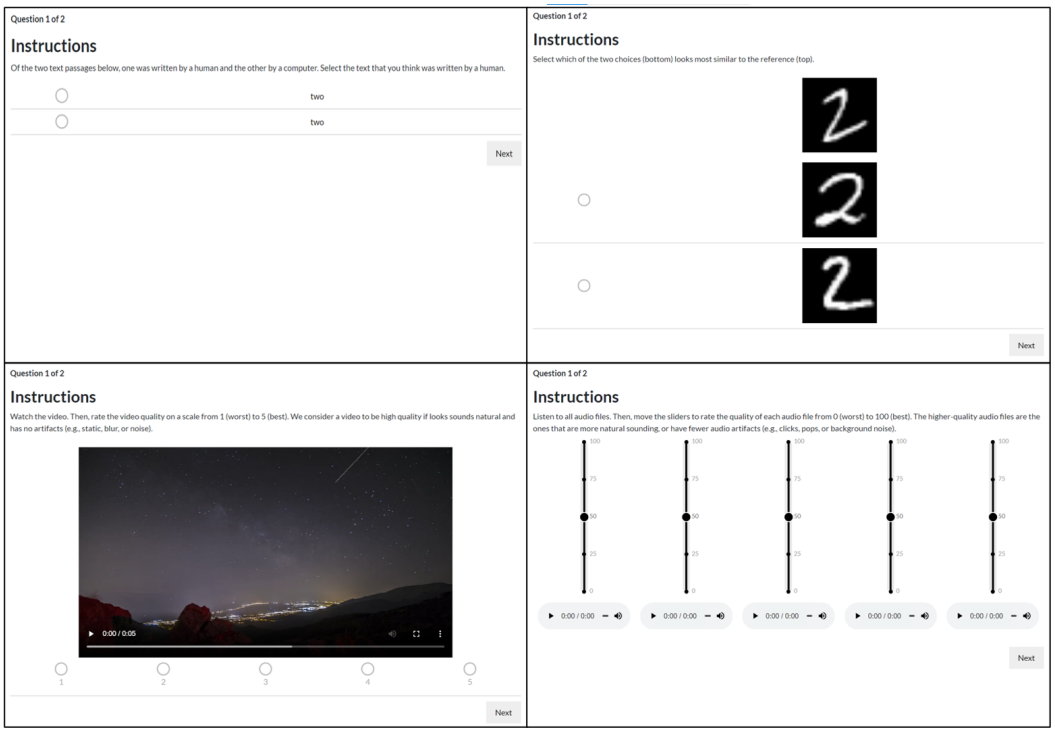

Figure 3: Four examples of subjective evaluation tests available with ReSEval. **(Top left)** An A/B test with text data. **(Top right)** An ABX test with images. **(Bottom left)** An MOS test on videos. **(Bottom right)** An audio MUSHRA test.

Account for learning curves and fatigue effects The order in which samples are presented to new participants as well as the number of samples presented can impact ratings (Schwarz et al., 2016). For example, while high-anchors and low-anchors can minimize changes in the range and variance of the results for MOS tests, a new participant performing their first evaluation question has no basis for comparison. Showing participants high- and low-anchor examples prior to evaluation can reduce this learning curve by pre-establishing a perceptual range. As well, it is important to limit the total length of the evaluation and not allow participants to repeat the evaluation in order to prevent fatigue effects from affecting evaluation results.

ReSEval randomly orders examples for participants to prevent learning and fatigue effects from biasing specific examples. Given a random seed, these assignments are deterministic, meaning that any learning and fatigue effects are replicable if the study is repeated with the same seed.

3 RESEVAL

We provide a brief overview of ReSEval[1]. ReSEval is a Python package that permits A/B, ABX, MOS, and MUSHRA-style tests on audio, image, text, and video data. Crowdsourced participants in a subjective evaluation are first presented with an introduction screen that describes the test, followed by an optional prescreening step, the evaluation, and an optional followup survey (Figure 2). Users of ReSEval can configure all of the text of the introduction, prescreening, followup survey, and evaluation instructions in a single configuration file via Markdown (see Appendix A for an example configuration file). This configuration also includes the type of test administered (e.g., A/B or MOS); the participant filtering criteria; the participant pay; and the cloud services used for file storage, database management, and server hosting. We designed this configuration to include all of the parameters necessary to fully replicate a crowdsourced subjective evaluation, provided access to the

[1]This paper describes ReSEval version `0.0.2` at commit `322e28e`

same audio, image, text, or video files being evaluated and the same version of ReSEval. After setup, a subjective evaluation can be launched in one line on the command-line given a configuration file `<config>` and a directory of evaluation files `<directory>`.

```
python -m reseval <config> <directory> --production
```

This command sets up storage, database, and server resources for a subjective evaluation, launches the subjective evaluation to crowdsourced participants, monitors progress, performs statistical analysis of results, pays participants, and shuts down compute resources once evaluation has finished. These steps can be performed individually for more control, and can also be called directly via our Python API. ReSEval can be run locally or in remote development mode (e.g., using the MTurk Sandbox) in order to debug evaluations before deployment. ReSEval runs on Linux, MacOS, and Windows and can be installed via pip[2].

4 Conclusion

We present Reproducible Subjective Evaluation (ReSEval), a framework for performing reproducible crowdsourced subjective human evaluations as easily as objective evaluation. ReSEval lowers the barrier-to-entry for performing high-quality crowdsourced subjective evaluations, a necessary step for many research tasks in machine learning, linguistics, and psychology. We are continuing to add features to ReSEval, such as a pre-test page that shows participants good and bad examples to reduce the initial learning curve, integrated support for performing attention checks, and support for additional cloud compute platforms (e.g., Amazon Web Services and Firebase) and crowdsourcing platforms (e.g., Prolific).

References

Mikołaj Bińkowski, Jeff Donahue, Sander Dieleman, Aidan Clark, Erich Elsen, Norman Casagrande, Luis C Cobo, and Karen Simonyan. High fidelity speech synthesis with adversarial networks. *arXiv preprint arXiv:1909.11646*, 2019.

Mark Cartwright, Bryan Pardo, Gautham J Mysore, and Matt Hoffman. Fast and easy crowdsourced perceptual audio evaluation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 619–623. IEEE, 2016.

Jennifer Cole, Timothy Mahrt, and Joseph Roy. Crowd-sourcing prosodic annotation. *Computer Speech & Language*, 45:300–325, 2017.

Todd M Gureckis, Jay Martin, John McDonnell, Alexander S Rich, Doug Markant, Anna Coenen, David Halpern, Jessica B Hamrick, and Patricia Chan. psiturk: An open-source framework for conducting replicable behavioral experiments online. *Behavior research methods*, 48(3):829–842, 2016.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf.

Steffen Jung and Margret Keuper. Internalized biases in fréchet inception distance. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021.

Alan E Kazdin. *Research design in clinical psychology*. Cambridge University Press, 2021.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033, 2020.

[2] `pip install reseval`

Judith Liebetrau, Frederik Nagel, Nick Zacharov, Kaoru Watanabe, Catherine Colomes, Poppy Crum, Thomas Sporer, and Andrew Mason. Revision of rec. itu-r bs. 1534. In *Audio Engineering Society Convention 137*. Audio Engineering Society, 2014.

Pranay Manocha, Zeyu Jin, Richard Zhang, and Adam Finkelstein. Cdpam: Contrastive learning for perceptual audio similarity. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 196–200. IEEE, 2021.

Winter Mason and Siddharth Suri. Conducting behavioral research on amazon's mechanical turk. *Behavior research methods*, 44(1):1–23, 2012.

Max Morrison, Rithesh Kumar, Kundan Kumar, Prem Seetharaman, Aaron Courville, and Yoshua Bengio. Chunked autoregressive gan for conditional waveform synthesis. In *International Conference on Learning Representations (ICLR)*, 2022.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. Why we need new evaluation metrics for nlg. *arXiv preprint arXiv:1707.06875*, 2017.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. A survey of evaluation metrics used for nlg systems. *arXiv preprint arXiv:2008.12009*, 2020.

Diemo Schwarz, Guillaume Lemaitre, Mitsuko Aramaki, and Richard Kronland-Martinet. Effects of test duration in subjective listening tests. In *International Computer Music Conference (ICMC)*, pp. 515–519. HKU University of the Arts Utrecht, HKU Music and Technology, 2016.

Victor S Sheng and Jing Zhang. Machine learning with crowdsourcing: A brief summary of the past research and future directions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 9837–9843, 2019.

Alon Shoshan, Nadav Bhonker, Igor Kviatkovsky, and Gerard Medioni. Gan-control: Explicitly controllable gans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14083–14093, 2021.

Zhou Wang and Alan C Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE signal processing magazine*, 26(1):98–117, 2009.

Kevin JP Woods, Max H Siegel, James Traer, and Josh H McDermott. Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, 79(7):2064–2072, 2017.

Fan Zhang, Angeliki V Katsenou, Mariana Afonso, Goce Dimitrov, and David R Bull. Comparing vvc, hevc and av1 using objective and subjective assessments. *arXiv preprint arXiv:2003.10282*, 2020.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

A EXAMPLE CONFIGURATION

Here, we provide an example configuration file for an A/B test on images.

```yaml
# A name to give to this evaluation configuration
name: ab-image-example

# The type of test to run. One of [ab, abx, mos, mushra].
test: ab

# The type of data to use. One of [audio, image, text, video].
datatype: image

# The location to store files used for evaluation. One of [aws].
storage: aws

# The third-party platform hosting the MySQL database. One of
# [heroku].
database: heroku

# The third-party platform hosting the server. One of [heroku].
server: heroku

# Crowdsourcing configuration
crowdsource:

  # The crowdsourcing platform used for evaluation. One of
  # [mturk].
  platform: mturk

  # The survey title shown to potential participants
  title: Title

  # The survey description shown to potential participants
  description: Description

  # Keywords that participants can use to find your survey
  keywords: Keywords

  # Filter participants
  filter:

    # Only allow participants from a certain countries
    countries: ['US']

    # Only allow participants who have previously completed at
    # least this number of tasks
    approved_tasks: 0

    # Only allow participants who have a sufficiently high
    # acceptance rating
    approval_rating: 0

  # How much you pay participants (in US dollars)
  # E.g., 2.00 is two dollars; 0.50 is fifty cents
  payment:

    # The amount that you pay even if they don't pass
    # prescreening
    base: 0.05

    # The additional amount that you pay participants who
    # complete evaluation
```

```
    completion: 0.45

  # How long to wait for things (in seconds)
  duration:

    # Total lifespan of the evaluation, after which the
    # evaluation is no longer available for participants to
    # take
    total: 86400

    # The maximum time you will allow a participant to spend on
    # your task
    assignment: 1800

    # Duration after which payment is automatically made
    autoapprove: 172800

# The number of participants
participants: 2

# The number of evaluations each participant performs
samples_per_participant: 2

# A seed to use for deterministic random sampling
random_seed: 0

# Introduction text to display on the first page participants
# visit
welcome_text: "
  # **Welcome!**\n
  We are conducting a research study to evaluate the
  quality of an image processing algorithm. If you agree to
  participate, you will be asked to fill out a brief
  questionnaire. You will then be asked to evaluate a series
  of images.\n
  ### **Privacy**\nThis survey is completely anonymous. We will
  NOT collect any personally identifiable information. Your
  participation in this study does not involve any risk to you
  beyond that of your everyday life.\n
  ### **Consent**\nBy pressing **I Agree**, you confirm you are
  willing to participate in this research. However, you are free
  to withdraw your participation at any time.\n
  ### **Contact Information**\nIf you have any questions or
  eedback, please contact <contact info>."

# Questions that participants must answer before they are
# permitted to perform evaluation. If a multiple choice question
# has correct_answer defined, the participant must select that
# answer to be able to continue to the evaluation.
prescreen_questions:

  # Test question
  - name: DummyQuestion

    # The type of question. One of
    # [free-response, multiple-choice].
    type: multiple-choice

    # Question text
```

```
    text: Would you like to take a subjective evaluation?

    # Possible answers
    answers: ['Yes', 'No']

    # Indicate a correct answer
    correct_answer: 'Yes'

# Instructions presented to the participant during evaluation
survey_instructions: "
  ## **Instructions** \nSelect the higher-quality image. The
  higher-quality image is the one that looks the most natural,
  or has fewer artifacts (e.g., blur or noise)."

# Questions presented to the participant after evaluation
followup_questions:

  # Follow-up question
  - name: DummyFollowup

    # The type of question. One of
    # [free-response, multiple-choice].
    type: free-response

    # Question text
    text: "How was the evaluation?"

    # Placeholder text
    placeholder: Good.
```