VitaEval: Open-source Human Evaluation Tool for Video-to-Text and Video-to-Audio Systems

Goran Topić[1] **Yuki Saito**[2] **Katsuhito Sudoh**[3] **Shinnosuke Takamichi**[4]
Hiroya Takamura[1] **Graham Neubig**[5] **Tatsuya Ishigaki**[1]
[1]AIST [2] The University of Tokyo [3] Nara Women's University
[4] Keio University [5] Carnegie Mellon University
{goran.topic, ishigaki.tatsuya}@aist.go.jp

Abstract

We present VitaEval, an open-source tool that streamlines the preparation process for human evaluation of video-to-text and video-to-audio systems. Evaluating such systems typically requires segmenting long videos, aligning subtitles and synthesized audio, and building custom interfaces for annotators—tasks that are time-consuming and often technically demanding. VitaEval addresses these challenges by automating video segmentation, synchronizing audio and subtitles, and generating web-based interfaces. Researchers can deploy evaluation setups without needing expertise in FFmpeg or HTML5, significantly lowering the barrier to conducting multimodal human evaluations. In a case study, we demonstrate that annotation setups using VitaEval can be created within one hour. The demo video is available at https://www.youtube.com/watch?v=TL4w1vWWaNY.

1 Introduction

Video-to-text generation, including dense captioning and real-time commentary systems (Krishna et al., 2017; Ishigaki et al., 2021, 2023), is gaining attention as large language models become increasingly capable of processing and describing dynamic visual content. Recent advances even extend these systems with text-to-speech technologies to produce natural-sounding commentary, enabling immersive experiences for users (Ishigaki et al., 2023). However, conducting human evaluation for such multimodal outputs remains a bottleneck. Evaluators must assess not just text quality, but also the alignment between video, text, and audio. This requires researchers to perform video segmentation, subtitle formatting, audio overlay, and interface design—steps that may seem lightweight individually but accumulate into a significant amount of manual effort and technical expertise in video and web technologies.

Existing annotation tools such as CVAT (Corporation, 2024), VIA (Dutta and Zisserman, 2019), and Label Studio (Tkachenko et al., 2020-2022) are built primarily for vision tasks and lack support for multimodal synchronization or interface automation. Commercial platforms offer flexibility but require extensive customization and cost. Thus, there is a strong need for a lightweight, customizable, and open-source solution tailored to multimodal evaluation.

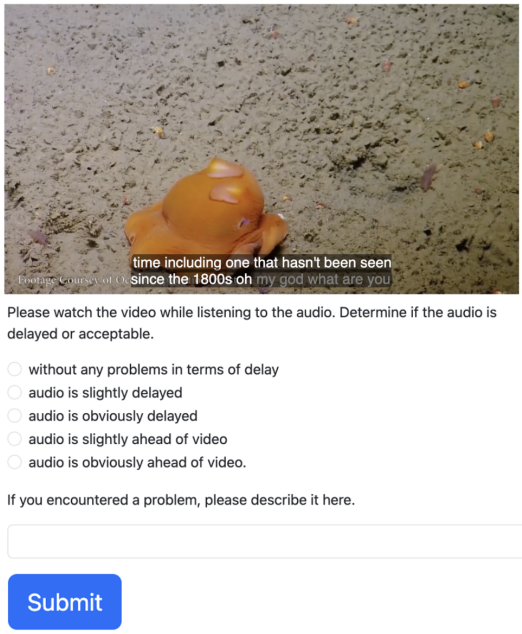

Figure 1: The evaluation interface, automatically generated by VitaEval.

Despite the importance of evaluation in generation research, many existing benchmarks rely on simplified setups where only the text output is evaluated in isolation. However, in realistic applications such as sports commentary, instructional videos, or accessibility services, the interplay between video, text, and audio is crucial for understanding user experience. Manual preparation of evaluation setups for such tasks not only slows down research cycles but also hinders reproducibility and scalability. There is a pressing need for

Proceedings of the 18th International Natural Language Generation Conference: System Demonstrations, pages 8–9
October 29–November 2, 2025.

tools that can lower this barrier and enable rapid, reproducible, and scalable evaluation for multimodal generation systems. Thus, we propose VitaEval.

2 VitaEval: System Overview

VitaEval is a Django-based web application that provides an end-to-end pipeline for setting up human evaluation tasks involving video, text, and audio. It supports two user roles: administrators who configure evaluation projects, and evaluators who annotate the video segments.

In the administrator workflow, to create an evaluation project, administrators upload a video file along with an optional JSON file specifying cut points, subtitle files (SRT/WebVTT), and commentary audio (MP3/WAV). VitaEval uses FFmpeg to cut the video and synchronize overlays. It then generates an evaluation interface where subtitles are rendered via HTML5 captioning and audio commentary is played back in sync with video.

Figure 1 shows an example interface automatically generated by VitaEval. Each evaluation screen presents a video segment with overlaid captions and audio, followed by one or more questions (e.g., "Is the audio delayed?") and optional free-text comments. Question types include radio buttons, checkboxes, and text fields. Results are saved in JSON and downloadable through the admin panel.

For scalability, VitaEval supports integration with Amazon Mechanical Turk (MTurk) for crowdsourcing, AWS S3 for media hosting, and scriptable endpoints for bulk data upload. Multiple administrators can manage tasks collaboratively. The system runs on Python 3.10 and can be hosted locally or on a cloud server.

3 Case Study

We applied VitaEval to annotate a racing game commentary dataset (Ishigaki et al., 2023), aiming to classify utterances as either subjective or objective. We spent 10 minutes to write a python script to convert SRT files to cut intervals, we deployed the evaluation interface within an hour by using VitaEval. In total, our tool reduces manual coding effort by at least 141 lines. The interface allowed annotators to make judgments using synchronized audio and subtitles, showcasing the tool's utility in real-world research workflows.

4 Conclusion

By automating video segmentation, audio/text synchronization, and interface creation, VitaEval eliminates the need for manual scripting or web development. Our system is easy to deploy and supports large-scale evaluation via crowdsourcing platforms and cloud storage. The source code and documentation are publicly available at: `https://github.com/aistairc/commentary_evaluator`. Further technical details can be found in the supplementary material (Topić et al., 2025) [1].

References

CVAT.ai Corporation. 2024. Computer vision annotation tool (cvat).

Abhishek Dutta and Andrew Zisserman. 2019. The via annotation software for images, audio and video. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 2276–2279, New York, NY, USA. Association for Computing Machinery.

Tatsuya Ishigaki, Goran Topić, Yumi Hamazono, Ichiro Kobayashi, Yusuke Miyao, and Hiroya Takamura. 2023. Audio commentary system for real-time racing game play. In *Proceedings of the 16th International Natural Language Generation Conference: System Demonstrations*, pages 9–10, Prague, Czechia. Association for Computational Linguistics.

Tatsuya Ishigaki, Goran Topić, Yumi Hamazono, Hiroshi Noji, Ichiro Kobayashi, Yusuke Miyao, and Hiroya Takamura. 2021. Generating racing game commentary from vision, language, and structured data. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 103–113, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Goran Topić, Yuki Saito, Katsuhito Sudoh, Shinnosuke Takamichi, Hiroya Takamura, Graham Neubig, and Tatsuya Ishigaki. 2025. Detailed technical descriptions of VitaEval: A open-source human evaluation tool for video-to-text and video-to-audio systems. In *arXiv*.

[1]This study is based on results obtained from a project in BRIDGE implemented by the Japanese government.