Making Accessible Movies Easily: An Intelligent Tool for Authoring and Integrating Audio Descriptions to Movies

Ming Shen
School of Software Technology,Zhejiang University
China
shenming2023@zju.edu.cn

Gang Huang
College of Computer Science and Technology,Zhejiang University
China
david.huang@zju.edu.cn

Yuxuan Wu
School of Software Technology,Zhejiang University
China
wux521@zju.edu.cn

Shuyi Song
DBAPPSecurity Ltd.,College of Computer Science and Technology,Zhejiang University
China
brendasoung@zju.edu.cn

Sheng Zhou
School of Software Technology,Zhejiang University
China
zhousheng_zju@zju.edu.cn

Liangcheng Li*
College of Computer Science and Technology,Zhejiang University
China
liangcheng_li@zju.edu.cn

Zhi Yu
School of Software Technology,Zhejiang University
China
yuzhirenzhe@zju.edu.cn

Wei Wang
College of Computer Science and Technology,Zhejiang University
China
wangwei_eagle@zju.edu.cn

Jiajun Bu
College of Computer Science and Technology,Zhejiang University
China
bjj@zju.edu.cn

Abstract

Blind and visually impaired (BVI) individuals encounter significant challenges in perceiving the visual content of movies. Audio descriptions (AD) are inserted into speech gaps to describe visual content and storyline for BVI individuals. However, the processes of authoring and integrating AD are laborious, involving tasks such as identifying speech gaps, authoring AD scripts, dubbing, and integrating them into the movie. To streamline these processes, we introduce EasyAD, an intelligent tool to automate these processes. EasyAD utilizes character recognition technology to identify speech gaps and utilizes speech synthesis technology for AD dubbing. EasyAD addresses the misidentification of the background music of existing methods, and for the first time applies a multimodal large language model in the tool to generate AD. EasyAD is currently operational at the China Braille Library and we invite 6 AD authors for a user study. The results demonstrate that with the use of EasyAD, the processing time for a medium-difficulty movie is reduced by nearly 50%, reducing the workload of AD authors and accelerating accessible movie production in China. EasyAD leverages the advantages of AI technologies, especially multimodal large language models, for accessible movie production and benefits BVI individuals.

CCS Concepts

• **Human-centered computing → Accessibility systems and tools**.

Keywords

accessible movie, accessibility, blind and visually impaired, audio descriptions, multimodal large language model

ACM Reference Format:
Ming Shen, Gang Huang, Yuxuan Wu, Shuyi Song, Sheng Zhou, Liangcheng Li, Zhi Yu, Wei Wang, and Jiajun Bu. 2024. Making Accessible Movies Easily: An Intelligent Tool for Authoring and Integrating Audio Descriptions to Movies. In *The 21st International Web for All Conference (W4A '24), May 13–14, 2024, Singapore, Singapore*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3677846.3677855

1 Introduction

"That of all the arts, the most important for us is the cinema."

–Vladimir Lenin

Movies serve as a significant form of art and play a crucial role in our lives. However, blind and visually impaired (BVI) individuals encounter significant challenges in perceiving the visual content of movies. To address this issue, audio descriptions (AD) [18], which describe visual content and storyline, are integrated into movies at speech gaps. The Web Content Accessibility Guidelines [9] recommends that AD should be provided for all online videos.

Nevertheless, the process of authoring and integrating AD is laborious. Authors have to invest considerable effort in identifying speech gaps, authoring AD scripts, dubbing and editing them into the movie [25]. Besides, condensing the crucial content for comprehension within the limited speech gap also requires advanced skills from authors.

[*] Corresponding author: liangcheng_li@zju.edu.cn

W4A '24, May 13–14, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-1030-8/24/05
https://doi.org/10.1145/3677846.3677855

To alleviate the workload for authors, various methods and tools have been proposed by researchers. For instance, tools like CinAD [5] automate the identification of speech gaps and the generation of AD by analyzing movie scripts and subtitle files. However, the challenge arises in obtaining movie scripts and subtitle files on many occasions. Alternatively, tools like 3PlayMedia [1] and descript [7] employ speech-to-text recognition, enabling authors to edit the text for AD creation, which are automatically dubbed and integrated. Nevertheless, there is a risk of misinterpreting background music as text. Other tools like CrossA11y [14], which identifies inaccessible segments through multimodal model alignment. However, these segments might not align perfectly with speech gaps. Methods such as AutoAD [10, 11] initially utilize multimodal large language models to generate AD, achieving commendable results. However, these methods have not yet been practically implemented in tools.

China has nearly 20 million BVI individuals [16], but current tools have technical limitations, especially in Chinese language support. Without specialized AD tools, the China Braille Library [13] uses general video editing software like PremierePro [3] and CapCut [6] for AD production, resulting in only 220 accessible movies on its website [13] due to the labor-intensive manual process.

To simplify the AD production, this paper introduces EasyAD, which automates speech detection, dubbing, editing, AD generation. To overcome the influence of background music, EasyAD employs character recognition technology to recognize subtitles and further identify speech gaps. As multimodal large language models have strong capabilities in video understanding and text generation, EasyAD applies them in a tool for the first time to generate high-quality AD. EasyAD utilizes speech synthesis technology to dub AD, employs FFmpeg to integrate them into the movie, and offers a user-friendly interface.

EasyAD is currently operational at the China Braille Library and we evaluated EasyAD in a user study with 6 AD authors. The result showcases that using EasyAD takes significantly less time than the previous tool, reduces the workload of AD authors, and accelerates accessible movie production in China. EasyAD leverages the advantages of AI technologies, especially multimodal large language models, for accessible movie production and benefits BVI individuals.

In summary, we contribute:

- An intelligent tool for authoring and integrating AD into Movies.
- The first to incorporate multimodal large language model-based AD generation into an AD auxiliary tool.
- A user study with AD authors from China Braille Library.

2 Related Work

Our work builds upon prior work in audio descriptions auxiliary tools and methods.

2.1 Audio Description

The accessibility of movies is a long-standing concern in the film industry [25]. AD are the "art of translating what is seen into what is heard; conveying visual information through the human voice and descriptive language" [18], and AD can only be placed during speech gaps in movie dialogues to avoid covering the original conversations [15]. AD have played a crucial role in facilitating movie access for BVI individuals, and the Web Content Accessibility Guidelines [9] recommend the provision of AD for all online videos. The general processes of creating AD involves reviewing the entire movie, identifying speech gaps, authoring AD scripts, dubbing, and editing them into the movie [25].

2.2 Audio Description Generation

Authoring AD is challenging because it involves describing crucial visual content within a limited length. To address this, researchers have proposed a series of methods to automatically generate AD [2]. Some methods such as CinAD [5] generate AD by analyzing movie scripts, but movie scripts are difficult to obtain. Most techniques rely on a cross-attention framework to align video clips with captions, demanding substantial labeled data [2, 27]. Leveraging multimodal large language models, recent approaches such as AutoAD [11] and AutoADII [10] utilize visual models (e.g., CLIP [22]) for feature extraction and large language models (e.g., ChatGPT [23]) for AD generation, addressing training data limitations and enhancing AD quality. Despite these advancements, multimodal large language model-based AD generation still has not been integrated into AD auxiliary tools. In this paper, we are the first to incorporate it into an AD auxiliary tool, ensuring state-of-the-art technology to benefit BVI individuals.

2.3 Audio Description Auxiliary Tools

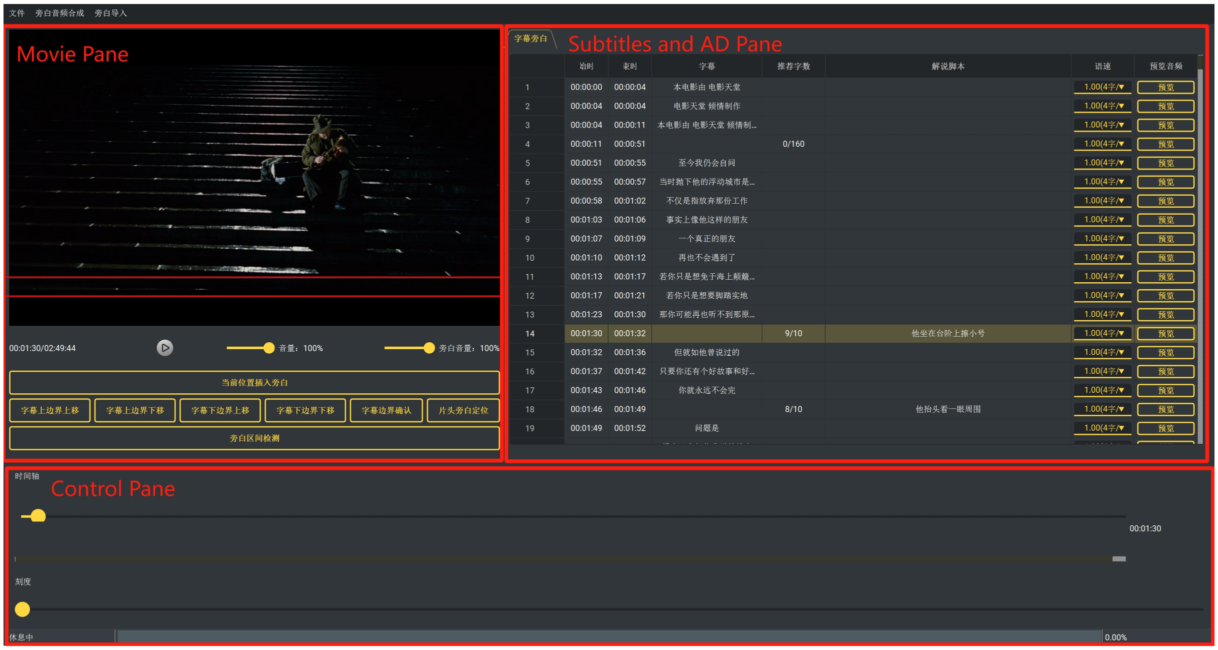
To simplify AD creation, tools like 3PlayMedia [1] and Descript [7] use speech-to-text for AD dubbing and integration, but still require manual speech gap detection and often misinterpret background music. EasyAD addresses these issues by recognizing subtitles to avoid the interference of background music. While tools like CrossA11y [14] detect segments with accessibility issues, they extend audio tracks unnecessarily and disrupt movie coherence. CinAD [5] identifies speech gaps via subtitle scripts, but these are hard to access without production rights. EasyAD solves this by recognizing on-screen subtitles, identifying speech gaps, and using multimodal models to generate and synthesize AD, streamlining the entire process.

3 Audio Description Auxiliary Tool

This section provides an overview of EasyAD, including interface, usage processes, pipeline, and implementation details.

3.1 Interactive Interface

Figure **??** shows the tool's interactive interface, featuring the Movie Pane, Subtitles and AD Pane, and Control Pane. The toolbar allows authors to create projects, import movies, and integrate dubbed AD. The Movie Pane displays the movie, enabling subtitle position adjustments, recognition, and speech gap identification. The Control Pane shows detection progress and the movie timeline for easy navigation. The Subtitles and AD Pane displays recognized subtitles, speech gap detection results, recommended word counts, and AD generated by the multimodal model, simplifying the authoring process.



	冶时	束时	字幕	推荐字数	解说脚本	语速	预览音频
1	00:00:00	00:00:04	本电影由 电影天堂			1.00(4字/▼	预览
2	00:00:04	00:00:04	电影天堂 倾情制作			1.00(4字/▼	预览
3	00:00:04	00:00:11	本电影由 电影天堂 倾情制...			1.00(4字/▼	预览
4	00:00:11	00:00:51		0/160		1.00(4字/▼	预览
5	00:00:51	00:00:55	至今我仍会自问			1.00(4字/▼	预览
6	00:00:55	00:00:57	当时抛下他的浮动城市是...			1.00(4字/▼	预览
7	00:00:58	00:01:02	不仅是指放弃那份工作			1.00(4字/▼	预览
8	00:01:03	00:01:06	事实上像他这样的朋友			1.00(4字/▼	预览
9	00:01:07	00:01:09	一个真正的朋友			1.00(4字/▼	预览
10	00:01:10	00:01:12	再也不会遇到了			1.00(4字/▼	预览
11	00:01:13	00:01:17	若你只是想免于海上颠簸...			1.00(4字/▼	预览
12	00:01:17	00:01:21	若你只是想要脚踏实地			1.00(4字/▼	预览
13	00:01:23	00:01:30	那你可能再也听不到那原...			1.00(4字/▼	预览
14	00:01:30	00:01:32		9/10	他坐在台阶上擦小号	1.00(4字/▼	预览
15	00:01:32	00:01:36	但就如他曾说过的			1.00(4字/▼	预览
16	00:01:37	00:01:42	只要你还有个好故事和好...			1.00(4字/▼	预览
17	00:01:43	00:01:46	你就永远不会完			1.00(4字/▼	预览
18	00:01:46	00:01:49		8/10	他抬头看一眼周围	1.00(4字/▼	预览
19	00:01:49	00:01:52	问题是			1.00(4字/▼	预览

Figure 1: EasyAD’s interface.

3.2 Usage Processes

The author can use the toolbar to create a project and import a movie file. As illustrated in Figure 1, EasyAD automatically detects subtitle positions, subject to author confirmation and modification. Detection results in Subtitles and AD Pane present start and end times, along with text content of subtitles or generated AD. The author can click on the start or end time of a speech gap to view the corresponding moment in the movie and edit the AD according to their understanding. After editing, the author can click the preview button to preview the dubbed audio of the AD. Upon completion, the author can integrate the AD into the original movie by clicking the audio synthesis button in the top-left corner.

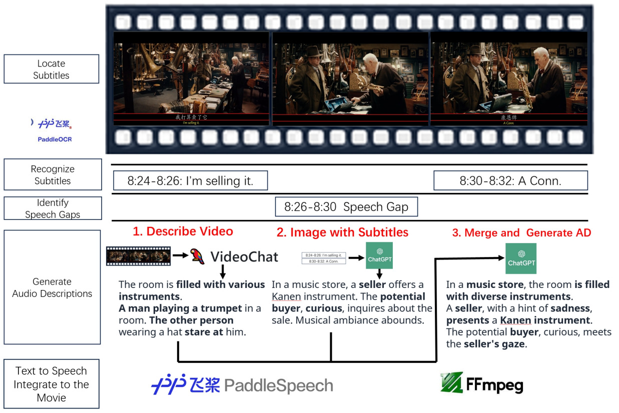
3.3 Pipeline of EasyAD

As shown in Figure 2, EasyAD’s pipeline comprises the following key processes. 1) Select frames to pinpoint the position of subtitles. 2) Recognize subtitles present in the movie. 3) Detect speech gaps by analyzing the subtitle results. 4) Use a multimodal large language model to comprehend movie segments and generate AD scripts. 5) Convert the generated AD scripts to speech and seamlessly integrate them into the movie.

3.4 Implementation Details

The tool uses the PyQt5 [21] framework for the interactive interface.

3.4.1 Subtitle Recognition. In China, almost all movies include subtitles of speech. To avoid misidentifying background music when transcribing speech, EasyAD obtains the speech content by recognizing subtitles. Since the subtitles appear at a fixed position, to improve the detection effect, at the beginning of recognition, EasyAD extracts several frames from the movie, detects the text, clusters the recognition results, and analyses the subtitle position. Authors can modify the subtitle position through the interface and EasyAD will pay more attention to the location area. EasyAD uses Paddle OCR [19] for text recognition, which is a framework of Optical Character Recognition (OCR) based on Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN).



The room is **filled with various instruments.** **A man playing a trumpet** in a room. **The other person** wearing a hat **stare at** him.

In a music store, a **seller** offers a Kanen instrument. The **potential buyer**, **curious**, inquires about the sale. Musical ambiance abounds.

In a **music store**, the room **is filled with diverse instruments.** A **seller**, with a hint of **sadness**, **presents** a **Kanen instrument.** The potential **buyer**, curious, meets the **seller's gaze**.

Figure 2: The overview of EasyAD. EasyAD utilizes a multimodal large language model to extract visual features, utilizes ChatGPT to imagine the picture based on contextual subtitles, and merges both to generate high-quality AD.

3.4.2 Speech Gap Detection. In the subtitle recognition process, EasyAD extracts data every 5 frames for subtitle recognition and records the current timestamp. If the subtitle recognition result is empty, it is identified as the start time of a speech gap. Otherwise, it is considered the start of a subtitle. When the subsequent subtitle recognition result matches the previous one, the end time of the preceding subtitle or speech gap is extended to the current timestamp. Conversely, if the recognition results differ, a new subtitle or speech gap is initiated.

3.4.3 Audio Description Generation. Multimodal large language models are trained on vast amounts of data and exhibit excellent generalization capabilities. When it comes to specific video understanding tasks, they can effectively utilize information from multiple modalities simultaneously, generating descriptions that more accurately capture the video content than previous models. EasyAD utilizes Video-Chat[12], a multimodal large language model, to generate a summary description of the speech gap and adjacent movie segments. Video-Chat is proficient in recognizing shallow information such as visual elements in videos but encounters difficulties in understanding deep semantics such as character actions and storylines, especially when the input video clips are short in length. As Figure 2 shows, Video-Chat exactly describes the environment (a room with instruments) and accurately identifies the characters (two men) and the main object (trumpet), but misunderstands the action of selling trumpet as playing trumpet. To address this issue, EasyAD utilizes ChatGPT[23] to guess the plot of the movie by the context dialogues (subtitles) and generate descriptions with rich semantic information (music store, seller and buyer, inquires about the sale). Finally, EasyAD merges the two descriptions together, using deep semantic information to correct the descriptions generated by Video-Chat and generate AD with both shallow visual information and deep semantic information.

3.4.4 Text to Speech and Integration. EasyAD uses Paddle Speech [20], which is a framework based on the transformer, for text-to-speech synthesis. Then this tool uses the video editing framework FFmpeg to integrate AD into the original movie.

4 User Study

EasyAD has been operational at the China Braille Library for three months. To evaluate EasyAD in simplifying authoring and integrating AD, we invite six AD authors from the China Braille Library to conduct a user study. They all have worked for accessible movie production for more than 3 years. They comprise 3 men and 3 women and their ages range from 28 to 45 years old. The user study involves both questionnaire investigation and face-to-face interviews. In the questionnaire investigation, participants are asked to write down the total and each process time-cost of a medium-difficulty movie using EasyAD and previous tools. In face-to-face interviews, we ask them if it is easy to learn and use EasyAD. Then, we inquire about their feelings regarding automated processes, including their acceptance of the time consumption, the accuracy of detecting speech gaps, the quality of automatically generated AD, and so on. Finally, we question them about the influence of EasyAD on their daily work and solicit their opinions and suggestions regarding EasyAD.

The processing time for a medium-difficulty movie is illustrated in Figure 3. Following the utilization of EasyAD, the total completion time has been reduced by almost 50%, with an 85% decrease in speech gaps identification, a 90% reduction in dubbing and editing, and a 25% decrease in authoring AD. EasyAD increases the speed of accessible movie production. In face-to-face interviews, participants fully affirmed the tool's contribution to accessible movie production. They expressed satisfaction with EasyAD's accuracy and convenience in identifying speech gaps and synthesizing speech. They praised that they no longer need to expend substantial effort on the above processes, and with just some clicks, everything is completed automatically by EasyAD. They said that the AD generated by EasyAD are impressive, which describe the important visual content and provide a significant reference for them. They mentioned that although the automatically generated AD may lack details such as facial expressions, it adequately fulfills the basic requirements of BVI individuals. Consequently, if time is limited, they have the option to publish directly without modifications, to swiftly release an accessible movie within 3 hours.

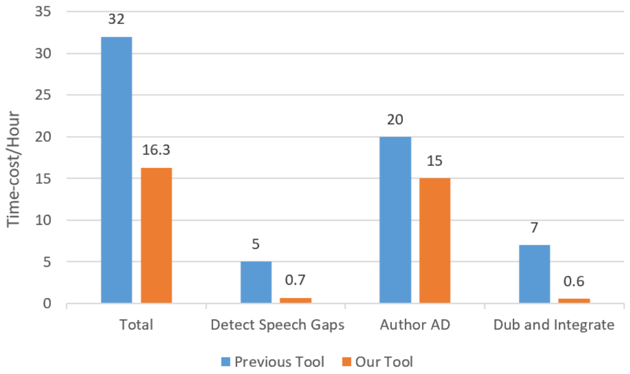


Figure 3: The average completion time (in hours) of authoring and integrating AD for an accessible movie with medium processing difficulty.

EasyAD has received positive user feedback and relieves AD authors from arduous tasks. In conclusion, EasyAD significantly speeds up accessible movie production, benefiting BVI individuals.

5 Conclusion and Future Work

This paper presents EasyAD, an automated tool for identifying speech gaps, generating, dubbing, and integrating AD. It addresses background music misidentification and, for the first time, applies a multimodal language model to produce higher-quality AD. Currently used at the China Braille Library, EasyAD reduces processing time by nearly 50%, receiving positive feedback. Its goal is to simplify AD production, fill the gap in China, and benefit BVI individuals.

Future improvements include further reducing time costs, enhancing visual detail in AD, adding speech transcription for movies without subtitles, and offering multilingual versions to serve BVI individuals globally.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (Grant No.62372408)

References

[1] 3PlayerMedia. 2023. 3PlayerMedia. https://www.3playmedia.com/.

[2] Jesus Perez-Martin adn Benjamin Bustos, Silvio Jamil F. Guimarães, Ivan Sipiran, Jorge Pérez, and Grethel Coello Said. 2022. A comprehensive review of the video-to-text problem. *Artificial Intelligence Review* (2022).

[3] Adobe. 2023. Premiere Pro. https://www.adobe.com/sg/products/premiere.html.

[4] Gabriel Reyes Amy Pavel and Jefrey P Bigham. 2020. Rescribe: Authoring and Automatically Editing Audio Descriptions. *Proceedings of the 33th Annual ACM Symposium on User Interface Software and Technology (UIST)* (2020).

[5] Virginia P Campos, Tiago MU de Araújo, Guido L de Souza Filho, and Luiz MG Gonçalves. 2020. CineAD: a system for automated audio description script generation for the visually impaired. *Universal Access in the Information Society* 19 (2020), 99–111.

[6] CapCut. 2023. CapCut. https://www.capcut.cn/.

[7] Descript. 2023. Descript. https://www.descript.com/.

[8] FFmpeg. 2023. FFmpeg. https://ffmpeg.org/.

[9] Accessibility Guidelines Working Group. 2023. Web Content Accessibility Guidelines (WCAG) 2.1. https://www.w3.org/TR/WCAG21/.

[10] Tengda Han, Max Bain, Arsha Nagrani, Gul Varol, Weidi Xie, and Andrew Zisserman. 2023. AutoAD II: The Sequel – Who, When, and What in Movie Audio Description. *International Conference on Computer Vision (ICCV)* (2023).

[11] Tengda Han, Max Bain, Arsha Nagrani, Gul Varol, Weidi Xie, and Andrew Zisserman. 2023. AutoAD: Movie Description in Context. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2023).

[12] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355* (2023).

[13] China Braille Library. 2020. China Braille Library. http://www.blc.org.cn/Index.aspx.

[14] Xingyu Bruce Liu, Ruolin Wang, Dingzeyu Li, Xiang Anthony Chen, and Amy Pavel. 2022. CrossA11y: Identifying Video Accessibility Issues via Cross-modal Grounding. *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (UIST)* (2022).

[15] China Society of Motion Picture and Television Engineers. 2022. Technical specification for productions of accessible film and television program. https://r.csmpte.com/photoAlbum/csmpte/page/20220111/2022011115274273362.pdf.

[16] China Assosiation of Persons with Visual Disabilities. 2023. China Assosiation of Persons with Visual Disabilities. https://www.cdpf.org.cn/.

[17] American Council of the Blind. 2023. The American Audio Description Project. https://adp.acb.org/.

[18] American Council of the Blind. 2023. The Audio Description Project. https://acb.org/adp/.

[19] PaddlePaddle. 2023. Paddle OCR. https://github.com/PaddlePaddle/PaddleOCR.

[20] PaddlePaddle. 2023. Paddle Speech. https://github.com/PaddlePaddle/PaddleSpeech.

[21] Python. 2023. PyQt5. https://www.pythonguis.com/pyqt5-tutorial/.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning (ICML)* (2021).

[23] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog* (2019).

[24] Pablo Romero Fresco. 2013. Accessible filmmaking:: Joining the dots between audiovisual translation, accessibility and filmmaking. *The Journal of Specialised Translation* 20 (2013), 201–223.

[25] Pablo Romero-Fresco. 2019. *Accessible Filmmaking: Integrating translation and accessibility into the filmmaking process*.

[26] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[27] Yujia Wang, Wei Liang, Haikun Huang, Yongqi Zhang, Dingzeyu Li, and Lap-Fai Yu. 2021. Toward automatic audio description generation for accessible videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–12.