SwissADT: An Audio Description Translation System for Swiss Languages

Lukas Fischer[UZH], **Yingqiang Gao**[UZH], **Alexa Lintner**[zh aw], **Annette Rios**[UZH], **Sarah Ebling**[UZH]

[UZH]Department of Computational Linguistics, University of Zurich, Switzerland
{fischerl, yingqiang.gao, rios, ebling}@cl.uzh.ch

[zh aw]School of Applied Linguistics, Zurich University of Applied Sciences, Switzerland
alexa.lintner@zhaw.ch

Abstract

Audio description (AD) is a crucial accessibility service provided to blind persons and persons with visual impairment, designed to convey visual information in acoustic form. Despite recent advancements in multilingual machine translation research, the lack of well-crafted and time-synchronized AD data impedes the development of audio description translation (ADT) systems that address the needs of multilingual countries such as Switzerland. Furthermore, most ADT systems are based only on text and it is unclear whether incorporating visual information from video clips improves the quality of ADT output. In this work, we introduce SwissADT, an **emerging** ADT system for three main Swiss languages and English, designed for future use by our industry partners SWISS TXT and the Swiss Broadcasting Corporation (SRG). By collecting well-crafted AD data augmented with video clips in German, French, Italian, and English, and leveraging the power of Large Language Models (LLMs), we aim to enhance information accessibility for diverse language populations in Switzerland by automatically translating AD scripts to the desired Swiss language. Our extensive experimental results, consisting of automatic and human evaluations of the quality of ADT, demonstrate the promising capability of SwissADT for the ADT task. We believe that combining human expertise with the generation power of LLMs can further enhance the performance of ADT systems, ultimately benefiting a larger multilingual target population. [1] [2]

1 Introduction

AD denotes the process of acoustically describing relevant visual information that renders streaming media content in television or movies and other art forms partly accessible to blind persons and persons with visual impairment (Bardini, 2020; Wang et al., 2021; Ye et al., 2024). This service involves the creation of textual descriptions, so-called "AD scripts", of key visual elements of a scene, such as actions, environments, facial expressions, and other important details that are not conveyed through dialogue, sound effects, or music (Snyder, 2005; Mazur, 2020). They are typically inserted into natural pauses that do not interfere with the ongoing narration. AD scripts are voiced by a professional human speaker or synthesized by a computer and mixed with the original audio.

Despite recent advancements in multilingual machine translation (Liu et al., 2020; Xue et al., 2021) and Large Language Models (LLMs) research (Brown et al., 2020; Achiam et al., 2023), two major challenges remain unsolved in developing well-performing ADT systems. Firstly, many ADT systems are built on pre-trained machine translation models that need texts in both the source and target languages as inputs. Training these ADT systems requires large amounts of manually crafted data, leading to high operational costs (Ye et al., 2024). Secondly, existing ADT systems are predominantly text-only machine translation models, neglecting the visual modality which is paramount for the ADT task and has proven to be useful as part of multimodal machine translation (Li et al., 2021).

In Switzerland, the primary target group of AD users comprises approximately 55,000 blind persons and 327,000 persons with visual impairment (Spring, 2020). Meeting the accessibility demands of Switzerland's multilingual population requires high-quality translation solutions.

In this work, we address the aforementioned challenges by developing an ADT system specifically for the three main languages of Switzerland, i.e., German, French, and Italian. To create train-

[1]This work was previously presented as a preprint (arXiv:2411.14967).

[2]A demo version of our system is hosted on GitHub. AD data will be made available via the GitHub link once data sharing agreements are finalized.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Industry Track), pages 370–379

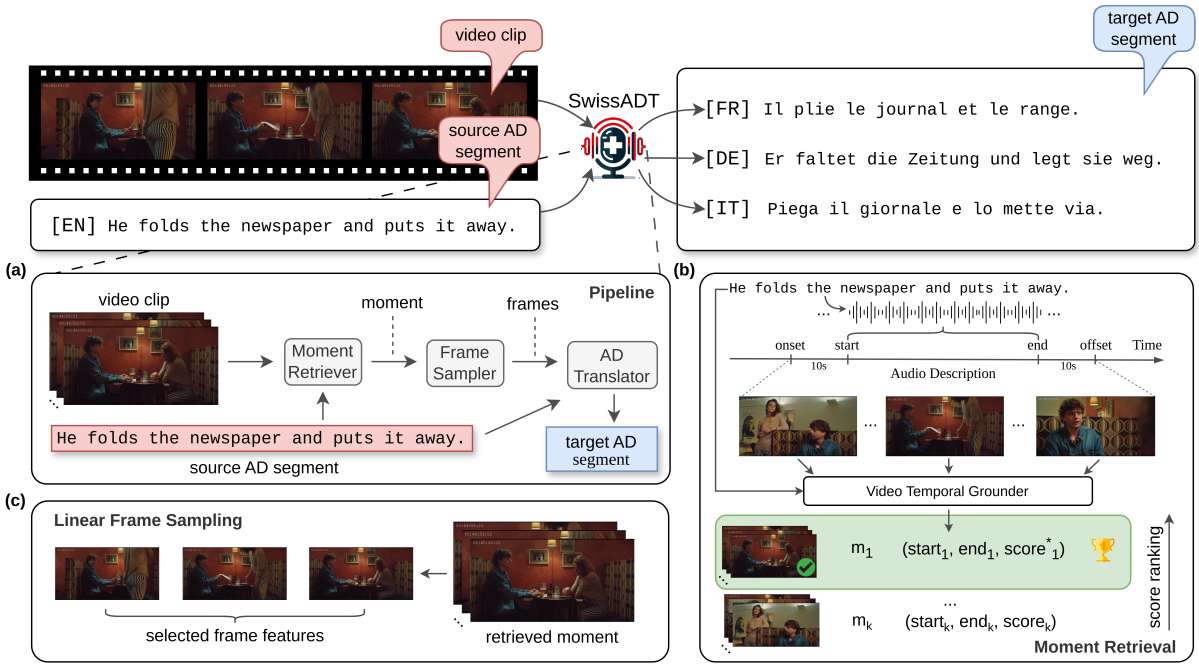


Figure 1: **(a) Overview of SwissADT**: An end-to-end pipeline that translates a given AD segment from English to the three main languages of Switzerland with the most salient video frames; **(b) Detail of the moment retriever**: it selects a moment, i.e., the most salient sequence of consecutive frames, to augment the translation inputs; (c) **Detail of the frame sampler**: it linearly interpolates the retrieved moment to obtain a cascade of frames used as inputs to the **AD translator**. In our implementation, we choose LLMs (GPT-4 models) as the AD translator due to their superior capabilities for performing multilingual machine translation tasks.

ing data for LLM-based ADT models with minimal human effort, we utilize DeepL[3] with English as an auxiliary language to generate AD scripts in the three Swiss languages. To verify if LLMs are a potential solution to ADT task, we conduct automatic and human evaluations of LLM-generated AD scripts. Additionally, to further improve the translation quality, we incorporate video clips as part of the inputs to the LLM-based ADT models.

Our contributions are: 1) We propose SwissADT, the first multilingual and multimodal ADT system for Swiss languages; 2) We conduct extensive evaluations of our ADT systems using both automatic and human quality assessments; 3) We highlight the system's emerging potential for real-world multilingual ADT applications; and 4) We provide the source code for SwissADT, which is easily installable for reproducibility.

2 Related Work

The automatic generation of ADs from video clips has been explored by both the natural language processing (NLP) and computer vision (CV) communities. This research is often conducted as part of tasks such as video captioning (generating descriptive text for a video) or video grounding (temporally aligning a text query with video segments).

In recent years, several datasets and models for ADs have been published, where many of them are movie subtitles or video descriptions (Chen and Dolan, 2011; Lison and Tiedemann, 2016; Xu et al., 2016; Lison et al., 2018). Oncescu et al. (2021) proposed QuerYD, an open-source dataset created for the text-video retrieval and event localization tasks, where ADs and video segments are annotated by human volunteers. Soldan et al. (2022) presented MAD, a large-scale benchmark dataset for video-language grounding, aggregated by aligning ADs with their temporal counterparts in videos. Zhang et al. (2022) introduced MovieUN, a large benchmark specifically designed for the movie understanding and narrating task in Chinese movies. Han et al. (2023b) released AutoAD, a model that leverages both text-only LLMs and multimodal vision-language models (VLMs) to generate context-conditioned ADs from movies. In another work of theirs (Han et al., 2023a), the authors further developed an extended model to address three crucial perspectives of AD generation, i.e., *actor identity (who)*, *time interval (when)*, and *AD*

[3] https://www.deepl.com/de/translator

Language	# Files	# Characters	Video Hours	AD Hours	Ratio
German	144	1,197,254	144:24:52	20:07:25	13.93%
French	30	569, 535	28:53:24	8:44:00	30.23%
Italian	23	486, 135	26:57:59	9:18:47	34.54%
Swiss German	95	945, 865	71:31:32	15:27:21	21.61%
total	292	3,168,789	271:47:48	53:37:33	19.73%

Table 1: Overview of our aggregated AD data.

content (what). Despite benefiting from existing large-scale corpora and state-of-the-art research in NLP and CV, these works are limited to monolingual applications. Consequently, they fail to meet the needs of Switzerland's multilingual population.

A second line of research explores the feasibility and suitability of applying machine translation models for ADT which was originally conceived as a human task. In the study conducted by Fernández-Torné and Matamala (2016), the *creation*, *translation*, and *post-editing* of English-Catalan AD script pairs were extensively investigated to assess whether machine-translated AD scripts achieved satisfactory quality. The authors found that machine translation models can serve as a feasible solution. Vercauteren et al. (2021) studied English-Dutch AD script pairs and found that errors were prevalent in the machine-translated AD scripts, indicating that post-editing by human experts was necessary.

In contrast to some of the above studies, we show that introducing visual inputs to ADT systems can lead to improved results, as verified by our AD professionals during the human evaluation.

3 SwissADT: An ADT System for Swiss Languages

SwissADT is a multilingual and multimodal LLM-based ADT system that translates AD scripts between English and the three main languages of Switzerland with visual and textual input. It contains three basic components:

Moment Retriever To identify the most relevant moment (that is, a sequence of consecutive frames) in a video clip for a given AD segment, we initially select a video segment that spans from ten seconds before the AD's start runtime (onset) to ten seconds after its end runtime (offset).[4] We then apply the video temporal grounder CG-DETR (Moon et al., 2023), which takes in both the AD script and the selected video segment and outputs the most relevant moment of variable length by providing the start and end times, along with a grounding score. The final moment is retrieved by selecting the highest-ranked moment with the highest grounding score from the pool of candidate moments.

Frame Sampler We linearly sample multiple video frames from the retrieved moment.[5] These frames are then utilized as visual inputs of the AD translator. We empirically report results on using four frames and every 50th frame.[6]

AD Translator We deploy multilingual and multimodal LLMs as the backbone AD translator of SwissADT. We conduct experiments with the fundamental GPT-4 models `gpt-4o` and `gpt-4-turbo`. We decide to apply zero-shot learning as part of a cost-effective solution.

Our modularized implementation of SwissADT streamlines the integration of state-of-the-art LLM research outcomes. This design allows for the seamless incorporation of cutting-edge moment retrievers and AD translators with minimal effort.

4 Data Collection

4.1 AD Scripts and Video Clips

We aggregate AD scripts from movies and TV shows that were aired on Swiss national TV stations, namely *Schweizer Radio und Fernsehen* (SRF), *Radio Télévision Suisse* (RTS), and *Radiotelevisione Svizzera* (RSI). Table 1 gives an

[4]Adding ten-second buffers ensures that the described moment is fully included in the video segment. Although ADs are usually synchronized with the described content, they may be shifted in dialogue-heavy scenes to fit no-speech segments. This buffer, recommended by our AD experts, sufficiently captures the described content even with such shifts.

[5]Linear sampling reliably includes frames that are representative of the entire segment. We leave other sampling methods for future research.

[6]In our system, the number of video frames can be manually set by the user.

overview of the aggregated AD scripts.

It is noteworthy that AD scripts in French and Italian occupy significantly more runtime in videos compared to those in German. This discrepancy arises from the data source: German ADs are predominantly derived from episodes of the TV game show *1 gegen 100*, which features relatively static scenes (same studio setting and moderator throughout, with only the game candidates varying), thereby reducing the necessity for extensive ADs. Conversely, French and Italian ADs are primarily sourced from movies and documentaries, which typically require more descriptive narration.

To facilitate the data storage, we use the SRT format (commonly used for subtitles) for ADs and mp4 format for videos. Figure 2 (Appendix A) demonstrates an AD passage from our dataset.

4.2 Synthetic ADs with DeepL

Due to a lack of parallel data, we use DeepL to generate synthetic AD scripts for each language pair of our system.

We translate all German, French, and Italian AD scripts into the other two Swiss languages, respectively, as well as into English. We include English as a mediating language in our ADT models to allow potential synergies with an AD script generation system developed by a research partner in our project. In addition, the moment retriever CG-DETR was trained on an English dataset, therefore, English is required as an intermediary language in our pipeline. For each source language, we shuffle the parallel ADs and randomly split them into train, dev, and test sets (see Table 2 for more detail). We limit the number of ADs in both the dev and test sets to 200 samples each to preserve training data for further experiments, given the 7,500-sample size for French and Italian. AD data is scarce, so we carefully balanced its usage between training and testing. Additionally, we maintained consistent sizes across all languages to ensure uniform evaluation.

We exclude Swiss German AD scripts due to the inadequate translation quality when using DeepL.

Source	Split	# ADs	# Characters
German	train	21,272	1,175,412
	dev	200	10,648
	test	200	11,194
French	train	7,099	538,063
	dev	200	15,533
	test	200	15,939
Italian	train	7,108	460,235
	dev	200	13,332
	test	200	12,568

Table 2: Dataset split for AD scripts of each source language. We use test sets for automatic ADT evaluation.

5 Evaluation Method

5.1 DeepL Translation Quality Estimation

We assess the quality of silver-standard AD scripts translated by DeepL using GEMBA-MQM (Kocmi and Federmann, 2023), an LLM-based metric that employs three-shot prompting with GPT-4 to identify and annotate error spans. This evaluation is conducted on test sets comprising 200 ADs for each source-target language pair, with weights assigned to *No Error*, *Minor Error*, *Major Error*, and *Critical Error* being 0, 1, 5, and 10, respectively. Table 3 presents the overall error weights of the DeepL-translated AD scripts.

	EN-trg	DE-trg	FR-trg	IT-trg
DE-src	**1.775**	-	2.465	2.925
FR-src	**1.585**	3.295	-	3.075
IT-src	**2.375**	3.525	3.815	-

Table 3: Quality estimation of the synthetic ADs generated by DeepL. Source languages are placed row-wise and target languages column-wise. All weights are below 4, indicating that translation errors do not exceed the major level requiring extensive modifications.

These results indicate that the errors in DeepL-translated AD scripts range from minor to major; therefore, they generally maintain a level of translation utility suitable for practical use in real-world scenarios, such as serving as the source language in our experiments.

5.2 Automatic ADT Evaluation

We use BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and CHRF (Popović, 2015) as automatic evaluation metrics for AD scripts translated by SwissADT, where the scores are calculated by comparing the generated AD scripts to the ground truths. Appendix C shows the prompts used for translation.

AD Translator	Input Modality	EN → DE			EN → FR			EN → IT		
		BLEU	METEOR	CHRF	BLEU	METEOR	CHRF	BLEU	METEOR	CHRF
gpt-4o	text-only	56.95	80.44	77.20	65.75	83.58	80.74	**63.30**	79.03	**78.66**
gpt-4-turbo	text-only	54.27	78.08	76.10	64.42	82.95	80.36	58.64	77.94	76.29
gpt-4o	text + 4 frames	**58.20**	**81.23**	**78.20**	**66.10**	83.37	**81.12**	63.15	79.24	78.31
gpt-4o	text + n frames	57.88	80.15	77.20	65.59	83.40	80.75	62.67	**79.75**	78.51
gpt-4-turbo	text + 4 frames	54.61	77.47	75.80	64.40	**83.70**	80.60	57.99	77.40	76.20
gpt-4-turbo	text + n frames	54.06	78.21	76.00	65.85	83.41	80.90	58.58	77.99	76.21

Table 4: Results of ADTs, where we highlight the best scores per system in bold. In the table, n represents the number of frames sampled at intervals of every 50 frames. Consequently, n varies depending on the duration of the retrieved moment (the average values of n are: EN→DE: 2.40, EN→FR: 3.48, EN→IT: 2.87).

5.3 Human Evaluation with AD Professionals

We conduct human evaluations with our AD experts[7] to assess the quality of AD scripts translated by SwissADT. Our objective is to verify the hypotheses that automatic evaluation scores reflect the human judgments well, and that multimodal inputs improve translation quality.

We utilize Microsoft Forms[8] to conduct our study. Following the Scalar Quality Metric (SQM, Freitag et al. (2021)) evaluations, we assess each AD pair (both source and target languages) along three dimensions: *fluency*, *adequacy*, and *usefulness* for audio description (i.e., how well the German target text suits the AD genre). AD experts rate these dimensions on a seven-point scale (0 to 6). The assessment is conducted online, and we compensate the AD experts at a rate of 85 CHF per working hour. We compare the translations of our best AD translator, `gpt-4o`, for two input modalities: text-only, and text with four frames as inputs for this assessment.

Due to challenges in hiring AD experts with sufficient English proficiency for French and Italian, we focus on evaluating German AD scripts. We recruit three AD experts (A, B, and C), all with translation degrees as well as professional experience ranging from three to over ten years. Furthermore, AD experts B and C are also professionally trained post-editors.

For the human evaluation, we randomly sample 30 consecutive blocks of 10 AD segments from our German dataset. We choose consecutive AD segments, so AD experts have more context to judge the translations. To minimize bias, each AD expert evaluates the same 30 blocks, in randomized order.

We use `gpt-4o` to translate the English silver AD segments back to German. We randomly select one of two strategies for each segment: text-only and text + four frames. The AD experts are presented with the English source segment and the German translation of `gpt-4o`, without knowing which input modality was used for the translations.

We report weighted Cohen's kappa (Cohen, 1968) for inter-evaluator agreement.

6 Results and Discussions

6.1 AD Translations

Table 4 presents the automatic evaluations of various AD translators. We observe that

- `gpt-4o` outperforms `gpt-4-turbo`;
- GPT-4-based results demonstrate promising performance in the ADT task, as indicated by high evaluation scores. This finding supports the effectiveness of applying machine translation models to address the ADT task, which is aligned with previous literature;
- Augmenting source ADs with corresponding video frames generally enhances translation quality, with the inclusion of more input frames leading to improved results. This suggests that it is beneficial to incorporate the visual modality into the ADT pipeline to utilize the power of fundamental LLMs.

The slightly better performance of `gpt-4o` with text-only on EN→IT may be due to language-specific factors, the small dataset size or varying multilingual zero-shot capabilities, as the differences are minimal. This result does not undermine the hypothesis that multimodal input improves translation quality overall, as other language pairs show the expected benefits. For examples where visual input is beneficial, refer to Appendix D.

[7] We plan to gather feedback from visually impaired users in the future, once SwissADT reaches a sufficient quality level.

[8] https://forms.office.com

text-only	A&B	B&C	A&C	
fluency	0.30	0.22	0.21	
adequacy	0.38	0.25	0.33	
usefulness	0.21	0.18	0.35	
text-only	A	B	C	**avg.**
avg. fluency	5.28	4.95	5.50	**5.24**
avg. adequacy	5.53	5.74	5.77	**5.68**
avg. usefulness	5.18	5.38	5.76	**5.44**

(a) AD translator with only texts as inputs.

text + 4 frames	A&B	B&C	A&C	
fluency	0.29	0.25	0.20	
adequacy	0.35	0.40	0.39	
usefulness	0.14	0.38	0.18	
text + 4 frames	A	B	C	**avg.**
avg. fluency	5.37	5.16	5.61	**5.38**
avg. adequacy	5.62	5.77	5.70	**5.70**
avg. usefulness	5.12	5.27	5.78	**5.39**

(b) AD translator with 4 video frames as inputs.

Table 5: Pairwise inter-evaluator agreement scores on AD fluency, adequacy, and AD usefulness, measured with Cohen's weighted Kappa (Cohen, 1968). We also report both the average evaluation scores for individual AD experts and the overall average scores across all AD experts.

Given that training human AD experts requires completing a curriculum that encompasses numerous essential competences and skills (Matamala and Orero, 2007; Jankowska, 2017; Colmenero et al., 2019), there is a persistent shortage of AD experts available to AD producers. Consequently, implementing automatic ADT systems based on multilingual and multimodal LLMs followed by human post-editing could leverage AD production.

6.2 Human Evaluation

Table 5 presents the inter-evaluator agreement results conducted with our AD experts as well as the average evaluation scores given by each AD expert, respectively. First, we see that our AD experts demonstrate a fair level of agreement overall, highlighting the inherent difficulty in evaluating AD translations even among professionally trained individuals. Given this subjective variability among human evaluators, we contend that automatic evaluation metrics remain essential, as they offer an additional objective assessment independent of the evaluators' training.

We also observe that AD scripts translated with four frames as input are rated higher in fluency (i.e., 5.38), and adequacy (i.e., 5.70) as compared to the text-only input translations (fluency: 5.24, adequacy: 5.68). These results verify our hypothesis that multimodal input improves translation quality. The dimension AD usefulness, however, is rated slightly higher for the AD scripts translated with the text-only input (i.e., 5.44) as compared to the four-frames translations (i.e., 5.39).

In future research, we aim to refine the definition of "usefulness" and develop more explicit guidelines to improve the consistency and accuracy of assessments.

Additionally, we plan to involve the target group in the next round of evaluations to obtain even more relevant and meaningful feedback. We will also incorporate the videos into the evaluation process to create a more realistic viewing experience, ensuring that the assessments better reflect the real-word use case.

7 Conclusions and Future Work

In this work, we present SwissADT, a multilingual and multimodal ADT system designed to support three Swiss languages and English. Our findings demonstrate that leveraging LLMs to address the ADT task represents a significant initial step towards achieving information accessibility, as validated by our experienced AD experts. This system provides a viable solution for enhancing accessibility for blind and visually impaired individuals in multilingual settings.

Future research will focus on fine-tuning LLMs for Swiss languages, improving system robustness to real-world data variability, and deploying the system with our industry partners. Additionally, we plan to conduct post-editing studies to further validate SwissADT's potential for real-world applications, ensuring high-quality outputs that minimize human effort while supporting professional workflows. Post-editing data will also be used to refine and improve the models over time.

We believe that integrating human expertise into the LLM pipeline for the ADT task will more effectively meet end users' expectations and satisfaction. As with any accessibility technology, it is paramount that it serves the needs of the end users.

Limitations

The limitations of our work are the following: 1) Due to the lack of high-quality data, we do not include Romansh as a target AD language, despite it being an official language of Switzerland that has nearly 35,000 native speakers;[9] 2) Given the difficulty in sourcing AD experts for French and Italian, we are unable to conduct human evaluations for these two languages. However, we expect the results to be comparable to German ADs, as indicated by the comparable translation results of our best AD translator `gpt-4o`; 3) The multimodal nature of ADs has not been taken into account in the human evaluation, which would require AD experts to have access to the visual inputs; 4) We do not utilize the Swiss German part of our dataset, as the absence of standardized spelling rules in Swiss German still poses a challenge for machine translation systems. This is primarily due to the fact that each word in Swiss German can have multiple spelling variations, resulting in an expanded vocabulary size.

Ethics Statement

To ensure privacy protection and data anonymization, we formally obtained informed consent for data collection of human ratings as per the guidelines of the Zurich University of Applied Sciences.

Acknowledgments

This work was funded by the Swiss Innovation Agency (Innosuisse) Flagship Inclusive Information and Communication Technologies (IICT) under grant agreement PFFS-21-47. We thank our industry partners SWISS TXT and SRG, particularly, Daniel McMinn and Veronica Leoni, for providing us with the use case and making data available.

We thank the anonymous reviewers for their constructive comments.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Floriane Bardini. 2020. Audio Description and the Translation of Film Language into Words. *Ilha do Desterro A Journal of English Language, Literatures in English and Cultural Studies*, 73:273–296.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

David Chen and William B Dolan. 2011. Collecting Highly Parallel Data for Paraphrase Evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Luque Colmenero, M Olalla, and Silvia Soler Gallego. 2019. Training Audio Describers for Art Museums. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 18:166–181.

Anna Fernández-Torné and Anna Matamala. 2016. Machine Translation in Audio Description? Comparing Creation, Translation and Post-Editing Efforts. *SKASE Journal of Translation and Interpretation*, 9(1):64–87.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Tengda Han, Max Bain, Arsha Nagrani, Gul Varol, Weidi Xie, and Andrew Zisserman. 2023a. Autoad II: The Sequel-Who, When, and What in Movie Audio Description. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13645–13655.

Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. 2023b. AutoAD: Movie Description in Context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18930–18940.

Anna Jankowska. 2017. Blended Learning in Audio Description Training. *Między Oryginałem a Przekładem*, (38):101–124.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting Translation Quality Error Spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775.

[9] Source: Swiss Federal Statistical Office

Jiaoda Li, Duygu Ataman, and Rico Sennrich. 2021. Vision Matters When It Should: Sanity Checking Multimodal Machine Translation Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8556–8562.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical Rescoring of Sentence Alignments in Large, Noisy Parallel Corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Anna Matamala and Pilar Orero. 2007. Designing A Course on Audio Description and Defining the Main Competences of the Future Professional. *Linguistica Antverpiensia, New Series–Themes in Translation Studies*, 6.

Iwona Mazur. 2020. Audio description: Concepts, theories and research approaches. *The Palgrave handbook of audiovisual translation and media accessibility*, pages 227–247.

WonJun Moon, Sangeek Hyun, SuBeen Lee, and Jae-Pil Heo. 2023. Correlation-guided Query-Dependency Calibration in Video Representation Learning for Temporal Grounding. *arXiv preprint arXiv:2311.08835*.

Andreea-Maria Oncescu, Joao F Henriques, Yang Liu, Andrew Zisserman, and Samuel Albanie. 2021. Queryd: A Video Dataset with High-Quality Text and Audio Narrations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2265–2269. IEEE.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Maja Popović. 2015. chrF: Character N-gram F-score for Automatic MT Evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Joel Snyder. 2005. Audio Description: The Visual Made Verbal. In *International congress series*, volume 1282, pages 935–939. Elsevier.

Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba, Chen Zhao, Silvio Giancola, and Bernard Ghanem. 2022. MAD: A scalable dataset for language grounding in videos from movie audio descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5026–5035.

Stefan Spring. 2020. Sehbehinderung, Blindheit und Hörsehbehinderung: Entwicklung in der Schweiz. Eine Publikation zur Frage: Wie viele sehbehinderte, blinde und hörsehbehinderte Menschen gibt es in der Schweiz? – Berechnungen 2019. Technical report, Schweizerischer Zentralverein für das Blindenwesen SZBLIND.

Gert Vercauteren, Nina Reviers, and Kim Steyaert. 2021. Evaluating the Effectiveness of Machine Translation of Audio Description: the Results of Two Pilot Studies in the English-Dutch Language Pair. *Revista Tradumàtica: Traduccio i Tecnologies de la Informacio i la Comunicacio*, (19):226–252.

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*, pages 1–12.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Xiaojun Ye, Junhao Chen, Xiang Li, Haidong Xin, Chao Li, Sheng Zhou, and Jiajun Bu. 2024. MMAD: Multimodal Movie Audio Description. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11415–11428.

Qi Zhang, Zihao Yue, Anwen Hu, Ziheng Wang, and Qin Jin. 2022. MovieUN: A Dataset for Movie Understanding and Narrating. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1873–1885.

A Audio Description Scripts

We make use of a common format for subtitles, namely SRT, where we treat ADs as subtitles. See Figure 2 for detailed data schema.

```
7
00:01:13,240 -> 00:01:16,720
$ Eine wuchtige Rolls Roice
Luxus-Limousine.  * Ein Händler
kommt:

8
00:01:42,240 -> 00:01:45,360
Chris nickt lächelnd.
$$ Der Händler öffnet die
Autotüren.

9
00:01:46,200 -> 00:01:51,360
UT: Toll.  Es gibt nicht viele
Autos für so grosse Menschen wie
mich.  So viel Beinfreiheit.
```

Figure 2: An example of a German AD script with spoken subtitles and special characters used in our data schema. The presence of a dollar sign ($) signifies a constrained timeframe of faster pace of speech. An asterisk sign (*) indicates a scene change within the script. Spoken subtitles are marked by UT as an abbreviation for "Untertitel" in German.

B Pricing

To estimate the cost of translating large datasets of ADs, we provide the calculations in Table 6 based on our dataset. Notice that OpenAI's pricing policy is subject to change, and that other factors, such as resolution and size of the input frames, as well as frequency and length of AD segments have great influence on the total price.

Model	Pricing	Cost for 190 ADs	
		text-only	text + 4 frames
gpt-4o	5.00 $ / 1M input tokens	$0.06	$4.28
	15.00 $ / 1M output tokens	$0.06	$0.06
	total	**$0.11**	**$4.33**
gpt-4-turbo	10.00 $ / 1M input tokens	$0.11	$8.55
	30.00 $ / 1M output tokens	$0.11	$0.11
	total	**$0.23**	**$8.66**

Table 6: Expected translation costs for an average AD script (assuming a video duration of 56 minutes, 190 AD segments). We resize the input frames to 960x540 pixels, which results in roughly 4,500 total input tokens (including text prompt) for a single ADT with 4 frames. The average lenght of text-only prompts is 60 tokens, and the average output length is 20 tokens. Pricings of `gpt-4o` and `gpt-4-turbo` are as of 12 July 2024.

text-only
Translate the following audio description from {source_language} to {target_language}. Respond with the translation only. This is the audio description to translate: {audio_description}
text + frames
Translate the following audio description for the frames of this video from {source_language} to {target_language}. Respond with the translation only. If the audio description does not match the image, please ignore the image. Respond with a translation only. This is the audio description to translate: {audio_description}

Table 7: Prompts used for translation with `gpt-4o` and `gpt-4-turbo`. The placeholders {source_language} and {target_language} denote the respective Swiss languages, while {audio_description} refers to the AD script to be translated. Prompts used for **text + frames** target both text + 4 frames and text + n frames configurations. The instruction to ignore irrelevant images addresses potential noise from linear sampling.

C Prompts

Table 7 demonstrates the empirical prompts that we used in our experiments for `gpt-4o` and `gpt-4-turbo` AD translators.

D Examples

The following examples demonstrate how multimodal input enhances translation quality by offering extra context. The relevant frames are shown in Figure 3.

Grammatical Ambiguity The Italian audio description *Volta la testa verso un treno che avanza sui binari* presents multiple translation possibilities. The verb *volta* can be interpreted in two ways:

- **3rd person singular indicative**: *He/she turns his/her head towards a train moving on the tracks.*
- **2nd person singular imperative**: *Turn your head!*

This ambiguity is resolved through the visual context of a man sitting on a train platform, as shown in Figure 3a.

Lexical Ambiguity The French audio description *Le phare éclaire deux chevreuils* presents two

(a) Visual context for the AD: ***Volta*** *la testa verso un treno che avanza sui binari.* (EN: ***He*** *turns* ***his*** *head towards a train moving on the tracks.*)

(b) Visual context for the AD: *Le* ***phare*** *éclaire deux chevreuils.* (EN: *The* ***spotlight*** *illuminates two deer.*)

Figure 3: Two examples of ambiguity that require additional context for resolution. The words that are correctly disambiguated by the visual input are highlighted in bold. Examples taken from the TV shows *Neumatt* (3a) and *Passe-moi les jumelles* (3b).

possible translations:

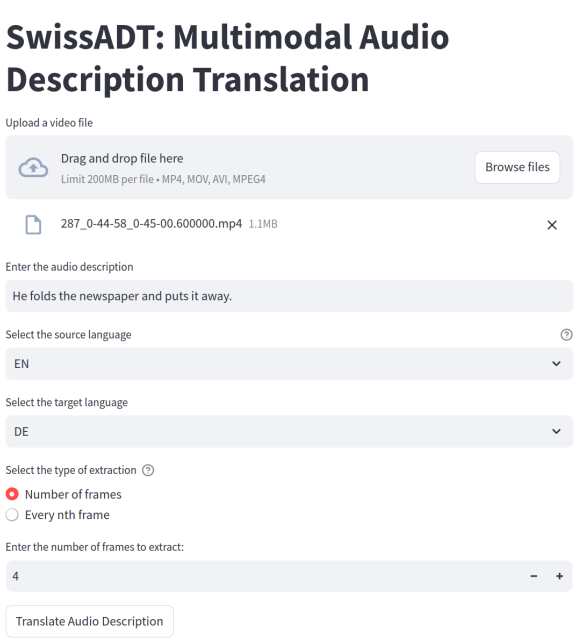
- *The lighthouse illuminates two deer.*
- *The spotlight illuminates two deer.*

The second frame in Figure 3b clearly shows that, in this context, *phare* should be translated as *spotlight*.

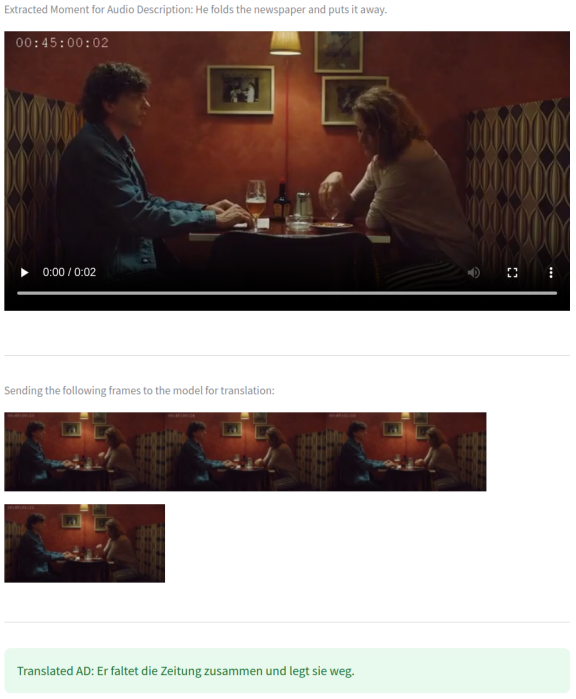
E System Demonstration

Our system demonstration for SwissADT (see Figure 4 for the system appearance) is hosted at `https://github.com/fischer192/swissADT`. Please follow our detailed instructions on our project page to set up the demo.

In addition, our demo also runs on our department server at `https://pub.cl.uzh.ch/demo/swiss-adt` which can be visited without configurations. We have also recorded a YouTube video explaining how to use the demo, which can be accessed at `https://youtu.be/5PQs8DscubU`.



(a) **Demonstration of SwissADT**. To generate the translated AD script from English to German, the user would upload the video clip and provide the AD script in the source language. Additionally, the user would input the number of frames to be sampled from the retrieved moment.



(b) **Generated AD in German**. We display the retrieved moment that best aligned with the source AD script in English, as well as the frames that are linearly sampled from the retrieved moment used by our best AD translator `gpt-4o`.

Figure 4: User interaction interface for SwissADT. We use `Streamlit` and `Docker` to implement the user interaction platform.