RNIB

See differently

CENTRE FOR TRANSLATION STUDIES

UNIVERSITY OF SURREY

Evaluation of AI-generated audio description for factual TV/media genres

Authors:
Sabine Braun, Shenbin Qian, Yuan Zou, Constantin Orasan
Centre for Translation Studies and Institute for People-Centred Artificial Intelligence,
University of Surrey

Commissioned by the Royal National Institute of Blind People

Table of contents

Introduction

1.1 Context

Access to digital media is vital for fostering an inclusive society and reducing inequalities. However, as demand for accessible media content continues to grow—driven by the expansion of audiovisual content and evolving accessibility legislation—traditional workflows for creating accessible content are increasingly strained. Current practices, relying exclusively on human expertise, are both costly and time-intensive. This includes workflows for creating audio-described content, which is the focus of this report.

At present, the legal requirement for qualifying broadcasters in the UK is to provide **audio description (AD)** for at least 10% of their programming. In 2012, major broadcasters such as the BBC, Channel 4, ITV, and Sky voluntarily committed to audio describing at least 20% of their content—a commitment recognised by Ofcom. While many broadcasters exceed the minimum threshold of 10% or 20% respectively, channels with high repeat rates still feature a significant amount of undescribed content, often due to limited resources, as noted in RNIB's 2023 broadcast content project 2023.[1]

While human-generated AD remains the gold standard for quality, there is growing interest in using computer vision, machine learning, and generative AI techniques to assist or automate AD production and/or delivery. Advances in AI-based automated video scene understanding, video captioning, and speech technologies, coupled with the instruction-following capabilities of large language models and multimodal models, have enabled progress in automating character, object, and action recognition, generating AD scripts, detecting gaps in the original audio track for AD insertion, and voicing AD scripts.

However, much of this research and development remains at an experimental stage. Engagement with users is currently limited, and the risks associated with generative AI, such as the variability in the quality of its output and inherent bias, raise concerns about its appropriateness for AD. Further scrutiny is needed to determine

[1] https://www.rnib.org.uk/professionals/research-and-data/reports-and-insight/rnib-broadcast-content-project-2023/

the role that AI can play in producing and delivering AD reliably and at scale.

1.2 The present study

Overall aim

Against the background outlined above, the overall aim of this study was to assess the feasibility of using current AI models to automatically generate AD for a selected media/broadcast genre; to evaluate the AI outputs in terms of **accuracy**, **intelligibility**, **coherence, relevance** and **robustness (safety)**; and subsequently to develop **initial recommendations** for further AI development in this area.

Specific objectives

1. Identify a broadcast/media genre and a data set for a feasibility study; the genre and data set should be low-risk and low-complexity content.
2. Identify 3 to 4 recent AI models developed to generate verbal descriptions from video scenes; apply them to the selected data to assess their capability in producing AD.
3. Evaluate the AI outputs through linguistic analysis; develop a set of evaluation criteria for this purpose, informed by AD guidelines and previous research on both human and AI-generated AD.
4. Consult with professional audio describers to gain insight into their strategies for describing the selected genre, providing an additional benchmark for evaluating the AI descriptions.
5. Gather feedback from blind and partially sighted users on the AI-generated output to evaluate how well it meets their needs and expectations.
6. Triangulate the results, develop conclusions and initial recommendations, with specific reference to the analysed genre and implications for other genres, and with reference to ethical implications of AD automation.

Sections 2, 3 and 4 of this report will focus on the evolution of AI models for AD, before sections 5 to 8 present the methodology and findings of our evaluation. Section 9 presents the conclusions, and section 10 outlines the recommended next steps. While this study focuses on AI, it should not be viewed as the sole solution for improving access. Other efforts remain essential, including

continued research and practical initiatives aimed at enhancing the quality of AD, as well as exploring alternative approaches—such as integrating enhanced commentary directly into TV programmes.

2 AI – Challenge and opportunity for expanding audio description

Introducing their recent *AI Opportunities Action Plan*, the Government noted that AI "[i]t is not a technology that is coming; a future revolution on the horizon. It is already here, materially changing lives" and that "in the coming years, there is barely an aspect of our society that will remain untouched" by AI.[2]

In the context of accessible content creation, the use of AI currently poses a number of challenges that need to be addressed in order to progress its use in a safe, meaningful and useful manner.

1. **Understanding audiovisual content can be highly complex**: Engaging meaningfully with audiovisual material and the narratives shared through it often entails complex language-based and multimodal interactions to derive meaning. This becomes more difficult when a modality employed by the creator is not accessible to the audience. To bridge this gap, various forms of multimodal translation (such as in AD: converting visual elements into verbal descriptions) are employed to provide a comparable experience for the audience. Successful translation/audio description requires a profound comprehension of the source material as well as critical decisions regarding what is most salient and relevant in the specific context.
2. **AI systems currently lack contextual awareness/understanding**: The current limitations of AI in terms of contextual awareness and understanding mean that AI cannot replicate the human ability to grasp linguistic and cultural nuances, resulting in misleading or inappropriate outputs, potentially undermining meaningful accessibility and user trust. In addition, AI text generation typically lacks the narrative depth and emotional resonance of human communication.

[2] Department for Science, Innovation and Technology, AI Opportunities Action Plan: government response, CP1242 (January 2025): https://www.gov.uk/government/publications/ai-opportunities-action-plan-government-response/ai-opportunities-action-plan-government-response

3. **AI systems are black boxes and can be biased**: Further challenges associated with the use of AI-generated AD include potential bias in training data. AI systems, including Large Language Models, are black boxes. A lack of transparency in text generation makes it difficult to predict how AI descriptions are generated, potentially resulting in critical errors.
4. **Limited user involvement in AI research & development**: User engagement with the development of AI tools for AD is currently limited. Where users are involved, there is little co-creation of solutions, further undermining user trust.

As a result of these challenges, AI-generated descriptions of visual content have, up until recently, often exhibited factual errors in object recognition and classification and have primarily focused on individual images rather than visual sequences. These limitations have hampered action recognition and classification, as AI has struggled to identify connections between images/frames, leading to absence of temporal sequencing, detection of narrative patterns in visual sequences and cohesive storytelling.

At the same time, the future potential of AI should not be evaluated only in terms of today's limitations. Generative AI has shown remarkable potential for innovation. The emergence of language and multimodal AI models, alongside advancements in Natural Language Processing/Generation, presents significant opportunities to enhance the fluency and cohesiveness of automatically generated AD. Thus, further scrutiny of the role AI can play in AD is both timely and necessary.

3 From video captioning to automating AD: Learnings from previous research

3.1 Video captioning with deep learning models

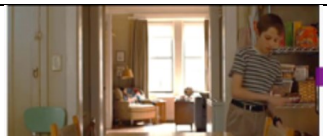
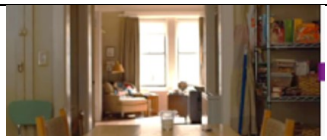
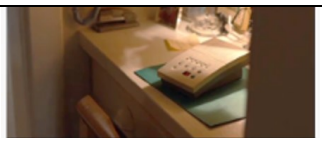
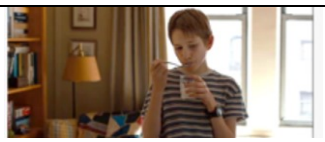
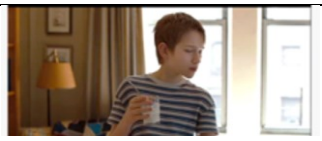
Automatic captioning of video content has been a long-term aim of the computer vision research community. Around 10-12 years ago, advancements in machine learning—particularly **deep learning** and **neural networks**—led to a surge in research activity in this field. Convolutional Neural Networks (CNNs) were used for object detection and classification in images through feature extraction, while Recurrent Neural Networks (RNNs) were employed for text generation. Machine learning methods relied on **supervised training** using **annotated image datasets** to establish

connections between images and text descriptions. The focus was on single-frame captioning, lacking continuity features such as character/object tracking, action recognition, and temporal sequencing.

Image data sets, such as Microsoft's MS COCO (Li et al., 2015), which provided the basis for training AI models, included non-iconic still images and limited motion images. MS COCO, for example, has 330,000 images containing 1.5 million objects. Identification and classification of objects in MS COCO was done via crowdsourcing, where annotators provided one-sentence captions for the images. However, these data sets and the process of development had severe limitations, rendering them unsuitable for supporting the automation of AD. For example, annotators received only sparse instructions and minimal training, while quality control was limited, leading to variable description quality and descriptions that were too clumsy or generic for the purposes of AD, which needs to be succinct and precise (Braun & Starr, 2019).

The European Horizon 2020 project MeMAD (Methods for Managing Audiovisual Data, 2018-21; grant no. 780069) developed a framework for video captioning using deep learning (Sjöberg et al., 2018) and evaluated its suitability for AD in the context of media accessibility. An example of the captions generated in MeMAD is given in Table 1 below. The table presents seven consecutive frames from a key scene in the film Extremely Loud and Incredibly Close (2011). Set in New York on 11 September 2001, the scene follows young Oskar Schell after he has been sent home early from school due to the unfolding events of the 9/11 terrorist attacks. Sensing that something terrible has happened, Oskar enters the stillness of his family's flat and attempts to ground himself in the familiar routine of getting a snack. He moves restlessly in and out of the kitchen, and listens to answerphone messages from his father, who is trapped in the World Trade Center. He then switches on the television to watch the news. The scene creates a powerful juxtaposition between the ordinary actions of an innocent child and the immense tragedy in the city.

Table 1: AI-generated video captions for an extract of the film Extremely Loud and Incredibly Close (2011), created with the MeMAD project's DeepCaption tool.

Frame number	Visual presentation of the frame	Description of the frame content (our own)	AI-generated caption
1		A young boy, Oskar, stands in the kitchen of his family's apartment, holding a yogurt in one hand and turning towards the table.	A woman is sitting on a bed and smiling.
2		The kitchen is now empty of Oscar, with only the yogurt remaining on the table. In the background, a wide doorway opens into a brightly lit living room with a large window.	A kitchen with a window and a window.
3		Oskar is back in the kitchen. He stands at the table, sniffing the yogurt.	A man is dancing in a room.
4		The kitchen is empty of Oscar again. The view towards the living room reveals a large armchair and an overhead floor lamp.	A woman is sitting in a chair and looking out the window.
5		On a light-coloured desk near the living room, a telephone with an answering machine shows a blinking red light.	A bookshelf with a teddy bear sitting on it.
6		Now in the living room, Oskar is eating his yogurt with a spoon.	A woman is holding a baby and looking at it.
7		Oskar holds his yogurt in one hand and a television remote control in the other.	A man is dancing in a room.

The often inaccurate and, at best, bland descriptions were typical of the video captions generated for a series of film extracts used in the MeMAD project (the MeMAD500 corpus; Braun et al., 2020), partly due to a mismatch between the film content and the available training data, but also owing to the limitations of the deep learning models available at the time. The findings of the MeMAD project evaluation are summarised in Table 2 below.

Table 2: MeMAD evaluation summary

Problems in AI-generated video captions	Recommendations for improvement (Braun & Starr, 2019, 2022)
Inaccurate object, action, character recognition	Improved algorithms & feature extraction, use of scripts, subtitles, dialogue
Errors in object scale & perspective	Context-aware object recognition & topic detection
Gender bias in character identification	Facial recognition, voice diarisation, film databases
Simplistic action labelling	Training on moving images & diverse datasets
Limited vocabulary & expression	High-quality datasets with expressive language
Repetition in captions	Reduce redundancy based on similarity check
No narrative continuity or temporal cohesion	Moving image datasets & cross-frame tracking
Lack of storytelling coherence	Thematic & character-driven training
No relevance filtering (describes everything)	Focus on key narrative elements, omission of basic actions (talking)

3.2 Refinements in video captioning

Subsequent developments and improvements in video captioning included the following:

- **Character tracking** (Rohrbach et al., 2015, 2017), i.e., the development of algorithms that could track and re-identify characters across different frames to provide more cohesive and contextually relevant descriptions.
- A shift from single-frame to **cross-frame captioning** (Venugopalan et al., 2015), which allowed for a more holistic understanding of the video content, generating descriptions that reflect the progression of events (temporal sequencing).

- **Multi-sentence descriptions** (Yu et al., 2016) allowing for richer, more detailed and more comprehensive descriptions that capture the richness of the content.
- **Dense video captioning** (Aafqq et al., 2019; Krishna et al., 2017) focussing on identifying and describing multiple events/actions within a video segment.
- **Video data sets for training (as opposed to image data sets)**: domain-specific (e.g. cooking shows), then generic (e.g. movies, activities) and audio descriptions (e.g., LSMDC dataset) which further improved the quality of descriptions.
- **Visual storytelling** (Huang et al., 2016), a study aimed at identifying a storyline in sequenced visual content (photo series) by identifying key visual elements and their contribution to the storyline.
- **Audio description generation** (Wang et al., 2021), the first system for automating AD, based on dense video captioning and using similarity detection to reduce redundancy in auto-generated descriptions. The system included automatic gap detection in the audio track. Their study also included user testing.

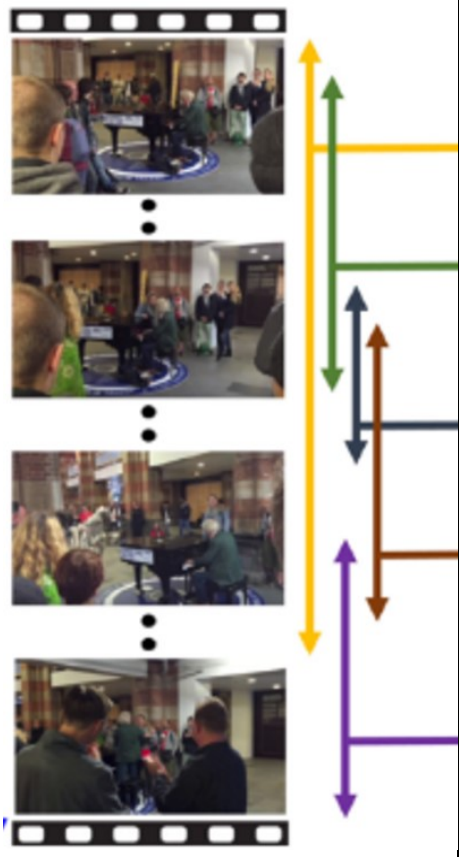
Table 3 below, reproduced from (Aafqq et al. (2019), illustrates three different methods for automatically generating video descriptions that emerged from the research outlined above: "Video captioning," "Single-frame captioning," and "Dense video captioning." At the centre of Figure 1 is a vertical film strip with four sequential frames from a video. The frames show an elderly man playing a piano in a public lobby as a crowd gathers and interacts with him. The three captioning methods are shown in the columns to the left and right of the film strip and are linked to the film strip as follows:

- The left column, under the heading "video captioning," contains a single sentence: "An old man is playing piano in a hall room in front of many people." It is meant to be a concise summary of the whole film strip.
- The next column, labelled "single-frame captioning," provides several simple, disconnected descriptions, such as "A person is playing piano" and "People clapping." Each one describes only a single frame, i.e. a moment in isolation.
- The third column, labelled “film strip,” shows the film strip with the four frames.

- The column on the right, labelled "Dense video captioning," presents a detailed narrative sequences of the events in the film strip. The set of coloured, overlapping arrows brackets different segments of the film strip and points to corresponding captions that describe specific actions. For example, one arrow links multiple frames to the caption, "Another man starts dancing to the music," showing how this method captures the duration and context of individual events to tell a complete story.

Figure 1: Video captioning methods (based on Aafqq et al., 2019)

Video captioning	Single-frame captioning	Film strip with four frames	Dense video captioning

An old man is playing piano in a hall room in front of many people.	Frame 1: A person is playing piano. Frame 2: A group of people standing in a room. Frame 3: A man playing a piano. Frame 4: People clapping.		Frame group 1: An elderly man is playing the piano in front of a crowd Frame group 2: A woman walks to the piano and briefly talks to the elderly man. Frame group 3: The woman starts singing along with the pianist. Frame group 4: Another man starts dancing to the music, gathering attention from the crowd. Frame group 5: Eventuall

			y the elderly man finishes playing and hugs the woman, And crowd applause.

Figure 2 below, taken from Huang et al. (2016), demonstrates the authors' approach to visual storytelling, which was based on still images rather than video scenes. This is illustrated at the top of Figure 3, which presents a series of five photographs, showing the following:

Photo 1: A family gathered around a dinner table, smiling.
Photo 2: A close-up shot of a white plate holding several cooked oysters.
Photo 3: A black dog standing on a sandy beach and a person's hand reaching down towards it.
Photo 4: A person and a small dog playing together in the ocean surf.
Photo 5: A yellow and white hang glider soaring high above the ocean on a clear day.

The table below the photo sequence shows the output quality of four different AI models that were tasked with creating a coherent story based on this photo sequence. The models are labelled "+Viterbi," "+Greedy," "-Dups," and "+Grounded."

Figure 2: Visual storytelling (Huang et al., 2016)

AI model	Descriptions generated by the model
+*Viterbi*	1. This is a picture of a family. 2. This is a picture of a cake. 3. This is a picture of a dog.

	4. This is a picture of a beach. 5. This is a picture of a beach.
+ *Greedy*	1. The family gathered together for a meal. 2. The food was delicious. 3. The dog was excited to be there. 4. The dog was enjoying the water. 5. The dog was happy to be in the water.
- *Dups*	1. The family gathered together for a meal. 2. The food was delicious. 3. The dog was excited to be there. 4. The kids were playing in the water. 5. The boat was a little too much to drink.
+*Grounded*	1. The family got together a cookout. 2. They had a lot of delicious food. 3. The dog was happy to be there. 4. They had a great time on the beach. 5. They even had a swim in the water.

Finally, the automated system for audio description (AD) generation developed by Wang et al. (2021) is worth highlighting. Unlike the other approaches outlined above, it includes features that are specific to AD. For example, the system first analyses the video's audio track to identify silent gaps where AD can be inserted. It then generates multiple candidate descriptions for each video segment corresponding to a gap in the audio. From this list, the system selects the most suitable description using evaluation criteria such as irrelevance cost, diversity cost, and perplexity cost. The final output consists of the selected AD, accompanied by a timestamp indicating where it should be inserted in the video.

One of Wang et al.'s examples involves a video segment with a woman in an art museum looking at paintings, for which the system generated the following candidate descriptions:

1. A woman is standing next to a large painting
2. A woman is seen speaking to the camera
3. A woman is seen standing before a man
4. A woman is standing behind a bar
5. A woman is standing in a room.

The description selected in the system's optimisation process is the first one, suggesting that this process is capable of selecting the most relevant description for the given context. While the ability to produce timed descriptions and the optimisation process are strengths of this system, its focus on video segments that correspond to gaps in the audio track is a limitation, as it means that segments with audio are not taken into account analysed nor described. This contrasts with the human AD creation process, in which the audio describer takes all visual content into account, not just the content coinciding with silent intervals, and applies strategies to fit essential visual information into the available audio gaps.

3.3 Summary

Early AI-based video captioning, prior to the Transformer era, primarily relied on computer vision and machine learning techniques, using neural networks to recognise objects and actions within video frames. While this enabled cross-frame captioning, challenges persisted, such as a lack of narrative depth, poor narrative coherence and a lack of contextual awareness. Additionally, the format—i.e. continuous written captions—was not suitable for AD, with the notable exception of Wang et al. (2021), although their study also exhibited several limitations.

4 Recent AI models and experimental AD automation

4.1 Transformer models and advanced language processing

Recent advancements in generative AI, particularly through sophisticated Transformer models, have significantly improved many tasks, including video scene understanding and description. These models are typically pre-trained using self-supervised learning techniques. While this reduces the need for extensive manual data annotation, it also introduces risks, such as the amplification of biases in uncurated training data. Notable innovations in this space include Vision-Language Models and Large Language Models, each contributing in distinct ways to the automatic description of audiovisual content:

- **Vision-Language Models (VLMs)** map visual input to text representations, modelling temporal relationships between

frames, and enabling cross-frame visual analysis for narrative continuity.
- **Large Language Models (LLMs)** generate fluent and relatively cohesive text descriptions.

Among the main impacts are improved narrative flow, potentially overcoming previous issues related to cohesion and coherence, and enhanced accuracy through comprehensive foundation models that cover a wide range of domains.

4.2 Zero-shot AD generation

In research focussed on automating AD, the developments outlined above have led to **zero-shot AD generation** using off-the-shelf pre-trained VLMs and LLMs. This approach requires no additional training or fine-tuning, relying instead on prompting strategies that leverage the models' inherent capabilities. A notable methodology is the two-stage approach proposed by Chu et al., 2024 and Xie et al. 2024:

- **Stage 1** involves the deployment of a VLM, which is tasked with generating a rich and detailed description of the video content, including characters, actions, interactions, locations, and objects present within the video.
- **Stage 2** employs an LLM to distil this information into a concise, single-sentence summary, tailoring the style and tone of the summary to align with specific prompts or requirements (e.g., different AD styles).

This approach appears to be effective for very short video clips, ranging from a few seconds to one or two minutes in length. Figure 3, adapted from Xie et al. (2004), illustrates it. At the top of Figure 3 is a horizontal film strip containing several frames from the TV show The Big Bang Theory. Three characters are visible: a man in a grey shirt (Sheldon) is standing behind a sofa, while a man in a green shirt (Leonard) and a woman in a floral top (Penny) are seated on it. The film strip also contains circles around the characters' faces—Sheldon's is circled in green, Penny's in red, and Leonard's in blue—based on a separate character identification module within the system.

Underneath the film strip, a table presents the two stages of the approach: "Stage 1: VLM-based video description" in the left and centre columns, and "Stage 2: LLM-based AD summary" in the right-hand column. For Stage 1, both the input (a multi-question

prompt) and the output (a dense video description) are shown. Under Stage 2, the prompt provided to the LLM is displayed. It instructs the model to generate a one-sentence description from the dense video captions produced by the VLM, to prioritise visually meaningful actions, and to ensure the description fits within the available gap in the audio track.

Figure 3: Zero-shot AD generation (Xie et al., 2004)

Stage 1: VLM-based video description – input (multi-question prompt)	**Stage 1: VLM-based video description – output (dense video description)**	**Stage 2: LLM-based AD summary**
Please describe the video in following steps: 1. Identify main characters. Possible characters: Sheldon Cooper (green), Penny (red), Leonard Hofstadter (blue). 2. Describe actions of characters. 3. Describe interactions between characters. 4. Describe facial expressions of characters.	1. Main characters: Sheldon Cooper, Penny, Leonard Hofstadter 2. Actions: Sheldon Cooper is standing up. Penny and Leonard Hofstadter are sitting on the couch. 3. Interactions: Leonard Hofstadter and Penny look at each other. 4. Facial expressions: no obvious expressions.	Please summarise the following description into one sentence. Description: {output from Video-VLM}" • **Character description:** Use characters first name if available, otherwise use pronouns such as 'He'. • **Action summary:** Focus on most attractive actions including {verb list}, ignore actions such as “talking” and “sitting”. • **Length adjustment:** Adjust the length

		of the output according to the video duration. Duration: {duration of the video}" • **Examples:** For example, an AD of duration 1.6s is: 'Paul looks at his wife lovingly.' **3. Predicted AD:** Leonard and Penny exchange stares.

4.3 Further recent developments

The most recent research has focused on enhancing automatically generated descriptions through the integration of multimodal and contextual information. Several models have experimented with incorporating context from various sources—including the programme's narrative, subtitles, prior AD segments, and pre-trained components—as seen in AutoAD (Han et al., 2023). Attention mechanisms have been used to refine the overall quality of descriptions by prioritising relevant visual and textual information, as demonstrated by DistinctAD (Fang et al., 2024). More advanced architectures such as MMAD (Ye et al., 2024) have introduced bespoke modules to address specific challenges in AD generation: an audio-aware module to detect and respond to audio cues, a subtitle module for analysing dialogue and aligning it with visual descriptions, and an actor-tracking module to enhance character continuity. These innovations aim to improve the accuracy, coherence, and narrative consistency of AI-generated AD. However, it remains to be seen whether the improvements are incremental or whether they mark a shift in the quality of automated descriptions. These models are currently not available for testing.

4.4 Summary

Transformer-based zero-shot AD generation has been explored using short clips from genres such as TV soaps, TV dramas and

feature films. The models are limited to description of short video segments (ranging from a few seconds to approximately 2 minutes); they cannot describe entire programmes or episodes.

The developers of the AI models have demonstrated improvements over previous models, using automated evaluation metrics commonly employed in computer vision/computer science to measure and quantify progress. A qualitative analysis of the limited number of examples provided in the developers' papers also suggests tangible improvements in areas such as character identification, object and action labelling, and cohesiveness, although challenges persist, particularly in the accuracy of object and action recognition, the relevance of descriptions, and over-description relative to the available time.

Contemporary TV soaps, TV dramas and feature films may be particularly challenging for the automatic generation of AD. While these genres differ from one another in some respects—for example, the simpler camera work, frequent repetition, and greater focus on domestic settings and personal relationships in TV soaps generally require less detailed AD than TV dramas and films—the fast-paced nature of modern productions often results in dense audio tracks with near-continuous dialogue, leaving limited space for AD. This entails a highly distilled presentation of information, based on a nuanced understanding of the audiovisual content to determine what is most relevant for inclusion in the AD.

The aim of our evaluation was twofold: to test extracts of potentially simpler TV genres and to use qualitative evaluation methods for assessing the **accuracy**, **intelligibility**, **coherence, relevance,** and **robustness (safety)** of audio descriptions created with recent AI models. Testing long-form audiovisual content is beyond the scope of this study.

5 Evaluation methodology

5.1 Overall approach

The evaluation comprised the following parts:

- Selection of suitable TV materials to test AI-generated descriptions
- Survey and selection of AI models
- Creation of the AI descriptions, using three AI models

- Linguistic analysis (qualitative analysis) of the AI-generated descriptions
- A focus group with audio description users
- Informal feedback sessions with audio describers.

Given the relatively small data sample, no quantitative analysis was carried out.

5.2 Materials

The materials for this study were provided by the BBC, matching the genres considered for this study, namely **factual television** with a focus on information and instruction (*Gardener's World*), **factual reality TV** (*Bargain Hunt, DIY SOS*) and a TV drama, for comparative purposes (*Casualty*). The selection was based on the following criteria:

- They are popular genres among blind and partially sighted audiences.
- Compared to other popular genres, they are of lower complexity (despite differences; with the reality shows being more 'dramatic' than Gardener's World).
- The risk of misunderstanding the visual content and therefore the risk of critical errors in AI-generated descriptions is low (again, despite differences/caveats).
- They are low-to-moderate-budget productions.

We selected one episode from each of the four BBC programmes. The episodes of *Bargain Hunt* and *Casualty* run for roughly 40–50 minutes, whereas *DIY SOS* and *Gardeners' World* episodes last about 60 minutes, giving a total of circa 210 minutes of video. The source files were supplied in MP4 format, and the BBC provided professionally written audio descriptions in XML, which served as 'ground-truth' references.

5.3 Data preparation

The automatic generation of AD with AI typically unfolds in three steps:

1. Segment detection – identifying the start/end timecodes of portions that need AD;
2. Text generation – producing a concise description for each detected segment;
3. Speech synthesis and reintegration – converting the text to audio and melding it back into the video.

As indicated earlier, no single system in academia or industry currently handles this pipeline flawlessly and seamlessly. Most research focuses on step 2—generating accurate descriptions—as it is the most technically demanding. In machine learning terms, it is a video understanding problem. Here, the model ingests video frames (created by sampling from a video clip to fit a model's input requirements) and outputs a textual summary.

Following this paradigm, we pre-processed our dataset to extract the relevant clips based on the professionally written audio descriptions and their timecodes:

1. **Parse ground-truth XML** – Extracted 163 AD segments with their exact timecodes.
2. **Create aligned clips** – Cut each episode on those timecodes, yielding 163 short clips that match the human ADs.

These clips were paired with prompts instructing the models to generate audio descriptions. Model selection and prompt design are described next.

5.4 AI model survey and selection

We reviewed relevant papers published in 2024 and shortlisted open source contenders:

Candidate	Model	Publicly deploy-able	Training required	Reason for exclusion (if any)
MMAD (Ye et al., 2024)	ESResNeXt (Donahue et al., 2015) for audio encoder, CLIP (Radford et al., 2021) for visual encoder, and Llama2 (Touvron et al., 2023) for text generation	Yes	yes	Requires paid API for Llama-2 from a third party
AutoAD Zero (Xie et al., 2024)	VideoLLaMA2 (Cheng et al., 2024) for video understanding and Llama3 (Grattafiori et al., 2024) for text summarisation	yes	no	—
AutoAD III (Han et al., 2024)	BLIP2 (Li et al., 2023) and MiniGPT-4 (Zhu et al., 2023) for video understanding and OPT (Zhang et al., 2022) and Llama2 for text generation	no	yes	Code or model unreleased
LLM-AD (Chu et al., 2024)	GPT-4V (Yang et al., 2023)	no	no	Code or model unreleased

Only **AutoAD Zero** could be fully implemented. To benchmark proprietary systems, we also tried Microsoft's *AI Audio Descriptions* project (GPT-4o-backed) as well as Google **Gemini-2.0-Flash** and

Gemini-2.5-Pro. The Microsoft project generated no descriptions after installation and was dropped from the evaluation.

5.5 Model setup

The *AutoAD Zero* paper introduces a prompting-based approach to guide a foundation vision-language model, such as VideoLLaMA2 to understand and describe the input videos. In capturing character interactions within videos, their method incorporates a facial recognition module to identify main characters, using this information to enrich the input prompts. Once the vision-language model generates textual descriptions of the scenes, a large language model (LLM)—LLaMA-3—is employed to summarise these descriptions into concise audio descriptions that align with the video's timing. In our experiments, we omitted the facial recognition module and made slight modifications to their prompt (see below) to suit our specific objectives. All other components remained consistent with the original methodology described in their paper.

> "Please describe the TV series clip in the following four steps: "
> "1. Identify whether there are persons in the scene, and you call them he or she; "
> "2. Describe the actions of persons in one sentence, i.e., who is doing what, focusing on the movements; "
> "3. Describe the interactions between persons in one sentence, such as looking; "
> "4. Describe the facial expressions of the persons in one sentence. "
> "Make sure you do not hallucinate information. "
> "###ANSWER TEMPLATE###: 1. Main persons: "; 2. Actions: "; 3. Person interactions: "; 4. Facial expressions: "."

For the two Google Gemini models—Gemini-2.0-Flash and Gemini-2.5-Pro, both general-purpose LLMs with vision capabilities—we used the Google Generative AI library to call their APIs, running them with all hyperparameters set to default. We prompted the models to perform the AD task using our video clips, following the prompt template proposed by Chu et al. (2024) with minor modifications. The final prompt template we used is as follows:

#You are an expert movie audio description writer. Your task is to correctly describe and write audio description for the scene in the provided movie clip. The task is to make audio description (AD) for the whole clip step by step. Adhere to AD guidelines.
Step 1 Procedure: select and identify 2 frames for each second.
Step 2 Describe: Describe what is happening in every frame in detail. You are not only required to describe what is happening in every particular frame but also need to pay attention to the scenes or events that are happening continuously across multiple frames.
Step 3 Summarize: Use the frame description in the previous step and the visual clues in the identified video frames to summarize the movie clip and generate a one-sentence audio description. The number of words in the sentence cannot exceed 25 words. Only include the first name of the characters if necessary.

5.6 Evaluation criteria

The following set of criteria was developed to evaluate the quality of the AI-generated AD. The criteria refer to the scripting/script only, not to the voicing/ delivery.

1. **Accuracy**

- **Correct Identification of key visuals** – Does the AD correctly describe people, objects, actions, and locations? (e.g., plant species in *Gardeners' World*, antiques in *Bargain Hunt*, tools in *DIY SOS*)
- **Description of embodied behaviour** – Does the AD effectively convey important embodied/non-verbal details such as expressions, gestures?
- **Precision in descriptions** – Are descriptions specific rather than vague? (e.g., "A mahogany 19th-century writing desk" is more effective than "a brown old desk.")
- **Effective conveyance of atmosphere** – Does the AD help the viewer understand the setting, such as describing a beautifully landscaped garden or a busy auction house?
- **Avoidance of misinterpretation** – Does the AD avoid making assumptions or subjective interpretations that could mislead viewers?

2. **Clarity/intelligibility**

- Clear grammar and syntax
- Avoidance of ambiguity
- Concise delivery – Is the information delivered in a way that is informative but not overly detailed or wordy?
- Use of simple, accessible language – Does the AD avoid overly technical or complex language when unnecessary?
- Appropriate use of terminology – When specialist terms are necessary, are they explained clearly?
- Avoidance of redundancy – Does the AD avoid repeating information that is already conveyed through dialogue or narration?

3. **Cohesion**

- **Continuity within and across AD segments** – Do the AD segments feel connected?
- **Seamless integration** – Is the AD complementary to, or does it fit smoothly with, the dialogue/narration without feeling intrusive?

4. **Relevance**

- **Prioritisation of essential / salient information** – Does the AD focus on details that contribute to understanding the show's content/storytelling? (e.g., mentioning the appearance of a rare flower in *Gardeners' World* rather than background elements)

5. **Engagement & enjoyability**

- **Capturing the tone of the programme** – Does the AD reflect the mood of the show? (e.g., warm, informative for *Gardeners' World*; enthusiastic, engaging for *Bargain Hunt*)
- **Maintaining viewer interest, enhancing, not distracting** – Does the AD support the viewer's experience without overwhelming or detracting from the main content?

6. **Handling of written/graphical info**

- **Explanation of on-screen text and graphics** – Are crucial graphics, captions or other written text described where necessary?

7. **Timing**

- **Length** – Would the AD fit naturally into pauses without overlapping important dialogue or sound cues, while being delivered at a pace matching the programme's rhythm?

The criteria were applied to the **linguistic analysis** of the sample material (see section 6). While they also guided the **focus group with AD users** and the **informal discussions with audio describers**, the qualitative data collected from these sessions was analysed thematically (see sections 7 and 8, respectively.)

6 Linguistic analysis

A linguistic analysis of the AI-generated descriptions was conducted using sample material from three of the BBC programmes outlined in Section 5.1: *Gardeners' World, Bargain Hunt,* and *DIY SOS*. These programmes were selected to maintain a focus on the factual and factual reality genres. The analysis followed a qualitative, functional approach, guided by the principles of AD and the evaluation criteria outlined in Section 5.5. This section presents a non-exhaustive summary of the findings.

6.1 Accuracy

The AI-generated descriptions generally featured grammatically well-formed sentences and a level of detail that compared favourably with descriptions produced before the advent of pre-trained Transformer models. However, closer analysis revealed that, despite their surface fluency, the AI outputs were often inaccurate. While not all errors were serious, examining them is essential, as they can disrupt the viewing experience, cause confusion for audiences, and—critically in this context—sometimes go undetected.

6.1.1 Character identification and description

Accurate character identification and description has long been recognised as a key challenge in automating AD, and early research began addressing this issue. The AutoAD Zero model tested in this study is a recent example of models that tackle the problem by integrating a face recognition module into its pipeline. This allows it to identify a programme's main characters by matching their faces with relevant film databases. In tests conducted by the model's developers (Xie et al., 2024), character

identification and naming worked well, provided a pre-identified character's face was clearly visible on screen.

As noted earlier, we did not implement a face recognition module in our evaluation, as our focus was on broader issues of character identification—specifically, the extent to which current AI models are able to accurately describe and re-identify characters within a scene. For instance, we examined whether the models used a character's appearance or clothing to support consistent re-identification. In this regard, all three models demonstrated significant limitations, as illustrated by **Example 1** below, which includes two consecutive description segments.

Example 1a: Character identification – stage-2 descriptions (summary descriptions [Gardeners' World]

	Segment 1	Segment 2
BBC AD	**Monty** pushes a wheelbarrow along a grassy path, a strip of thin meadow that runs between tall, well-trimmed hedges.	**He** arrives at some wooden raised beds.
AutoAD Zero	**He** watches **the dog** run ahead of him.	**He** walks through the garden, with **her** following behind him.
Gemini 2.0	In a green garden, **a man** pushes a wheel-barrow down a path, followed by **a dog**.	**A man** wanders through a lush green garden maze, passing by an apple tree while **another man** pushes a wheelbarrow in the background.
Gemini 2.5	**A person** pushes a wheelbarrow along a woodland path followed by **a dog**, then walks with **the dog** down a formal path lined with hedges.	**A man** walks down a hedged garden path; later, **someone** pushes a wheelbarrow behind a close-up view of an apple tree branch.

The BBC AD demonstrates a classic technique for reintroducing a character (here: Monty) who appeared earlier but has not been mentioned for some time: it refers to Monty by name in the first segment and then with a pronoun in the subsequent one, consistently presenting him as a known character and successfully re-identifying him.

The AutoAD Zero model uses a pronoun in both segments, suggesting that it, too, is attempting to re-identify Monty based on his earlier appearance. However, in the first segment, the description also refers to Monty's dog as "the dog"—that is, as already known—even though the dog had not been introduced at that point. In the second segment, the model refers to the dog as "she." An analysis of the model's stage-1 output (i.e. the intermediary dense video description) helps explain these inconsistencies. In the stage-1 description for the first segment,

shown below, both Monty and the dog are initially introduced as new (“a man” and “a dog”) and are then repeatedly re-identified as “the man” and “the dog.” This suggests that the video-language model deployed in stage 1 was capable of maintaining character continuity within a scene, but that this information was not reliably carried over into the stage-2 description. The use of “she” in the second segment stems from a more serious character identification error: the model appears to identify a woman, even though no woman is present in the scene. This misidentification propagates through to the final (stage-2) description, further highlighting the model’s limitations in maintaining accurate character tracking.

The two Gemini models introduce both Monty and the dog as new in both segments, making the lack of re-identification across segments more apparent than in the AutoAD Zero model. The stage-1 descriptions reveal that, like AutoAD Zero, the two Gemini models struggle with character identification—for example, by misidentifying Monty repeatedly as “another person”. A notable observation regarding the stage-1 outputs of the Gemini models is, however, the level of detail in the character descriptions. While these details are often accurate, they seem insufficient for reliable character re-identification across segments.

Example 1b: Character identification – Stage-1 descriptions (dense video descriptions) [Gardeners’ World]

	Segment 1	**Segment 2**
Auto AD Zero	1. Main persons: **a man and a dog**; 2. Actions: **the man** is walking through a garden, and **the dog** is running ahead of him; 3. Person interactions: **the man** is looking at **the dog**; 4. Facial expressions: **the man** is smiling.	1. Main persons: **a man** and **a woman**. 2. Actions: **The man** is walking through a garden while **the woman** is walking behind him. 3. Person interactions: None. 4. Facial expressions: None.
Gemini 2.0	*Frame 1:* The scene begins with a lush forest view, thick with trees and foliage. The sunlight filters through the leaves. A narrow path leads through the woods, lined with dense greenery. *Frame 2:* The camera stays	*Frame 1:* The back of **a man** wearing a dark jacket walks forward between two tall green hedge walls. **The man** is in a garden, and green potted plants stand between the hedges. *Frame 2:* The back of **the man** walking, with potted plants, is

	focused on the path ahead, winding slightly through the trees. A large tree trunk stands to the right of the path, partially obscuring the view. *Frame 3:* The camera pans slightly to the left, revealing more of the thick undergrowth on either side of the path. The sunlight becomes more diffused, giving a soft glow to the scene. *Frame 4:* The camera's focal point shifts to a twig in the foreground, with the forest path slightly out of focus in the background. The view is framed by the twig, with leaves and trees surrounding it. *Frame 5:* **A man** pushes a wheelbarrow along a green path lined with tall, manicured hedges. **A dog** trots alongside him. *Frame 6:* **The man** continues pushing the wheelbarrow down the path, his back to the camera. **The dog** stays close by his side as they move forward.	shown. *Frame 3:* A close-up shot shows an apple tree with ripe apples. *Frame 4:* The camera focuses on an apple tree. **A person** behind the tree walks forward pushing a wheelbarrow. *Frame 5:* A close-up shows the apple tree, and a person passes behind it with a wheelbarrow.
Gemini 2.5	*Frame 1:* View through dense, sunlit green leaves and thin branches of a woodland path. A thick tree trunk is prominent on the right, its bark textured and dappled with shadow. The path is covered in fallen leaves and leads away into the brighter background. *Frame 2:* The scene remains the same static shot of the woodland path viewed through the foliage. Sunlight filters through the canopy, creating bright spots on the leaves and	*Frame 1:* Seen from behind, **a man** wearing a dark jacket walks away from the camera down a grassy path. The path is bordered on both sides by tall, dense, neatly clipped green hedges. Large terracotta pots containing cone-shaped topiary bushes line the path intermittently. Ahead, a taller, pointed tree stands against a bright blue sky. *Frame 2:* The scene cuts to a close-up of a branch on an apple tree, laden with green leaves. A

	the ground. *Frame 3:* The view is unchanged, but white text appears overlaid at the bottom center: "#GardenersWorld". *Frame 4:* Still the same view of the path through the leaves, with the "#GardenersWorld" text present. *Frame 5:* The front wheel and handles of a dark-coloured wheelbarrow enter the frame from the right, moving left along the path. The "#GardenersWorld" text remains. *Frame 6:* The wheelbarrow continues moving left, showing more of its dark basin. It navigates the leaf-strewn path. *Frame 7:* **A person** wearing dark trousers, a dark jacket or top, and tall boots walks into view behind the wheelbarrow, pushing it leftwards along the path. They are partially obscured by the foreground foliage. *Frame 8:* **The person** continues pushing the wheelbarrow. A medium-sized, brown dog appears behind the person, following them along the path. *Frame 9:* **The person** pushing the wheelbarrow is now mostly out of frame to the left. The brown dog continues to follow, trotting along the path. *Frame 10:* **The person** and wheelbarrow have exited the frame left. The dog continues walking left along the path. [abridged]	single apple, mostly green with a patch of red blush, hangs prominently on the right side of the frame. Other smaller, less ripe apples are visible deeper within the foliage. The background consists of more blurred green garden plants. *Frame 3:* The close-up on the apple branch continues. In the blurred background, the top edge of a grey wheelbarrow with yellow handles appears, moving from left to right behind the leaves. *Frame 4:* The wheelbarrow continues its movement in the background. More of it is visible, along with the arms and torso of **a person** wearing a dark blue shirt pushing it. The apple remains in focus in the foreground. *Frame 5:* **The person** pushing the wheelbarrow moves further across the background. Their dark blue shirt and hands gripping the yellow handles are visible behind the apple tree foliage. The apple in the foreground remains the focal point. *Frame 6:* **The person** pushing the wheelbarrow continues moving towards the right side of the frame in the background, becoming slightly more obscured by leaves again. The close-up on the apple branch persists. [abridged]

In conclusion, all three models display notable shortcomings in character identification and re-identification. These limitations impede the generation of coherent and contextually accurate descriptions, particularly in content where character continuity is essential. The two-stage models offer useful insights into the models' capabilities and highlight that not every seemingly accurate description stems from correct pre-processing in stage one—raising important concerns around reliability and trust.

In addition to the challenges outlined above, the audio describers with whom we had informal discussions about AI also highlighted the sensitive issues that can arise from misrepresenting and misgendering characters in AD and felt that this is an area where AI can currently not be trusted. One AD user contended that "the problem is when you know someone's being sort of misgendered, then you probably don't necessarily trust the rest of it so much" (AD user 3).

6.1.2 Object and action identification and description

The AI models also struggled with accurate object and action identification and/or description. While improvements were evident compared to older descriptions, the final (stage-2) descriptions often still default to vague terms such as "a bird," "an object," or "some items". This can reduce the usefulness and informativeness of the descriptions, particularly in genres like nature or lifestyle programming where detail is important.

There were notable differences in the level of detail and accuracy of object descriptions across the three AI models, as illustrated by **Example 2** below. Object recognition errors—such as the AutoAD Zero model misidentifying a large wooden spoon as a frisbee—risk disrupting narrative coherence in a programme like Bargain Hunt, making it more difficult for viewers to follow who or what is being referred to. Gemini 2.0 incorrectly identifies the object as two separate items, a spoon and a bowl, while Gemini 2.5 describes it as a ladle—which, although not entirely accurate, may be considered acceptable.

As with Example 1, the stage-1 descriptions provide insight into how the final (stage-two) outputs were produced. These reveal that both AutoAD Zero and Gemini 2.0 initially identified two separate objects (frisbee/spoon and bowl/spoon, respectively), while Gemini 2.5 described the object as a ladle with a bowl. In the stage-2

descriptions, AutoAD Zero omits the spoon entirely, Gemini 2.0 retains it, and Gemini 2.5 focuses solely on the ladle, omitting the bowl. While analysing the stage-1 outputs helps identify the source of error in the AutoAD Zero and Gemini 2.0 outputs, the rationale behind what is retained and omitted in the stage-2 descriptions remains unclear.

Example 2: Object identification [Bargain Hunt]

BBC AD	A giant wooden spoon.
AutoAD Zero	**Stage-1 description:** 1. Main persons: Two men in red shirts. 2. Actions: One man is holding **a frisbee**, and the other man is holding **a wooden spoon**. 3. Person interactions: They are talking to each other. 4. Facial expressions: They are smiling.
	Stage-2 description: Tom holds **a frisbee** and talks to Alex.
Gemini 2.0	**Stage-1 description:** Two young men in red polo shirts stand under an orange tent at an outdoor market, examining **a wooden bowl** and **wooden spoon**. Numerous other people and stalls are visible in the background.
	Stage-2 description: Under an orange tent at an outdoor market, two young men in red polo shirts are inspecting **a wooden bowl and spoon**.
Gemini 2.5	**Stage-1 description:** Outdoors at a bustling market under a bright orange canopy tent. Two young men in matching red polo shirts stand side-by-side looking at items on a table. The man on the left has styled light brown hair. The man on the right has curly brown hair and glasses, and he holds **a large, light-brown wooden ladle with a long handle**, examining its **bowl**. Shelves next to them display various items. Other shoppers browse in the background.
	Stage-2 description: Two young men in red shirts stand under an orange market tent examining **a large, antique wooden ladle** among other wares.

Errors are also evident in the description of complete actions, which sometimes include inaccurate object references as well. In the video scene described in **Example 3**, the character (Monty) is shown pushing a wheelbarrow along a dirt path. To his right, he passes several other wheelbarrows lined up against a cluttered background of garden supplies. His dog follows behind him. However, the description generated by AutoAD Zero misrepresents both the objects Monty walks past and the dog's movement, incorrectly stating that the dog is walking in the opposite direction, when in fact it is following him.

Example 3: Action identification [Gardeners' World]

AutoAD Zero	He walks past **a pile of tires** and **a greenhouse**, while **the dog walks in the opposite direction**.

6.2 Clarity and intelligibility

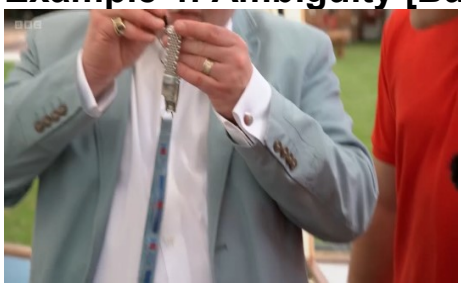
In addition to inaccuracies, the AI-generated descriptions frequently lacked clarity, largely due to structural ambiguities and unconventional phrasing. The Gemini outputs also featured a notable amount of description relating to cinematic techniques. This is not typical of AD and offers opportunities for condensation—particularly valuable given that some of the AI-generated descriptions were overly long for the time available.

6.2.1 Ambiguous descriptions

Some descriptions were ambiguous. For instance, the Gemini 2.0 description in **Example 4** below is structurally ambiguous due to the placement of the two prepositional phrases "with a chain" and "in front of an outdoor market scene." It is unclear whether the chain is part of the object or related to the act of inspection. Similarly, the phrase "in front of an outdoor market scene" could refer either to the man's position or to the location of the object. Although this ambiguity can be resolved with general world knowledge, hearing the description delivered rapidly as part of an AD track could cause confusion for users. Notably, the stage-1 description in this case is clearer than the final output. The Gemini 2.5 description is unambiguous but omits details such as the chain. Interestingly, however, its stage-1 output is more detailed than that of the other AI models, and therefore more likely to avoid ambiguity in the final version.

Beyond the structural ambiguity—which is the focus here—this example also highlights other issues of clarity and issues of accuracy: AutoAD Zero provides an incorrect description, while Gemini 2.0 uses vague language (“small metallic object”), and Gemini 2.5 includes a similarly imprecise reference (“a metal object he holds outdoors”).

Example 4: Ambiguity [Bargain Hunt]



BBC AD	Thomas looks closely at the watch through a **loupe**.
AutoAD Zero	**Stage-1 description:** 1. Main persons: A man wearing a suit and tie. 2. Actions: The man is holding **a pair of glasses** up to his face and adjusting them. 3. Person interactions: None. 4. Facial expressions: The man's facial expressions are not visible in the video.
	Stage-2 description: He adjusts **his glasses**.
Gemini 2.0	**Stage-1 description [abridged]:** A man in a light blue blazer and colorful eyeglasses closely examines **a small, metal object with a chain**, holding it up to his eye. The background shows an outdoor market with tents and people.
	Stage-2 description: A man in a light blue blazer inspects **a small metallic object with a chain in front of an outdoor market scene.**
Gemini 2.5	**Stage-1 description [abridged]:** The man consistently uses a magnifier attached to a metal object (like a watch band on a lanyard) to inspect it closely, holding it up to his right eye throughout the clip.
	Stage-2 description: A man in a blazer uses an attached magnifier to closely inspect **a metal object** he holds outdoors.

6.2.2 Unusual wording

Besides cases of ambiguity, some of the AI-generated descriptions had unusual phrasings, as illustrated in **Example 5**. The description in Example 5a, for instance, is slightly awkward due to the sequence of events and mixed temporal references. The verb

"hang" is a stative verb; it describes a state or condition, not an action with a defined beginning or end. On the other hand, using "before" implies a temporal boundary or sequence — it cues the listener to expect a change or eventual transition. That contrast can make the sentence feel semantically mismatched: the berries are in a continuous state (hanging), but “before” suggests that this state concludes just prior to the next event — which is logically and visually odd, since the berries are not shown to stop hanging. The last part of the sentence is slightly ambiguous — it is not immediately clear whether it is the garden, the person, or the entire scene that is being viewed. “Foreground flowers” is a confusing phrase.

Example 5a: Unusual/opaque wording [Gardeners’ World]

Gemini 2.5	Small red berries hang on a branch before a wide shot reveals a person tending a lush vegetable garden, viewed later through foreground flowers.

The description also includes a reference to cinematic technique ('wide shot'), which appeared frequently in the Gemini outputs. This suggests that general-purpose models like Gemini either require more effective prompting or fine-tuning to avoid using cinematic language where it is inappropriate or unnecessary. In the focus group, one AD user found such language “quite technical” (Participant 1). However, others noted that references to cinematic technique can be useful when the technique is particularly relevant to the action (Participant 2), and that in the context of a gardening programme, viewers are likely to want more information about what is visible and what the garden looks like (Participant 3).

The description in Example 5b illustrates further issues with clarity, particularly due to the use of abstract, figurative language such as "from a wheelbarrow's perspective". In an audio format, this may be confusing, as users must rapidly interpret the phrase as metaphorical rather than literal. In addition, the subsequent phrase "following a dog down a path" introduces ambiguity about who is following the dog, which may disrupt the user’s understanding of the scene.

Example 5b: Unusual/opaque wording [Gardeners’ World]

Gemini 2.5	From a wheelbarrow's perspective following a dog down a path, the view changes to show a man pushing it through a garden.

6.3 Cohesion

Cohesion refers to how different parts of a text are connected through language, e.g. through markers of temporal sequence, enabling a smooth flow and continuity of meaning. In the context of AD, cohesion is important both within and across individual descriptions, and in relation to the original audiovisual content. Producing cohesive descriptions is challenging for AI models for several reasons: first, the issues outlined above—such as inaccuracies in identifying characters, objects, and actions, a lack of specificity in the descriptions as well as problems with clarity—undermine cohesion; second, although dense video captioning has enabled some cross-frame description, the connections AI models can establish remain limited; and third, without contextual awareness, AI models are unable to reliably assess the cohesiveness of the descriptions they generate.

Example 6 below illustrates several cohesion-related issues in the AI-generated descriptions. The BBC AD links the two concurrent actions in the scene—the robin perching nearby and Monty transferring the dahlia's roots—maintaining cohesion and clarity. In contrast, the stage-1 description produced by AutoAD Zero is both generic and partially inaccurate, depicting two actions as seemingly unrelated and referring vaguely to "a plant". This lack of specificity and connection contributes to the vagueness of the resulting stage-2 description. For Gemini 2.0, the more detailed stage-1 description appears to support the generation of a more cohesive final description. However, unlike AutoAD Zero, the temporal sequencing of actions is incorrect in the final Gemini 2.0 output. The reason for this is not transparent. The Gemini 2.5 stage-1 description is significantly more elaborate than those of the other models and even includes brief intermediary summaries ("continuous action" descriptions). This appears to contribute to a highly accurate and cohesive final output. However, references to the robin perching nearby—prominent in the early part of the stage-1 description—are omitted from the stage-2 description, likely because they are not reinforced later in the stage-1 output. Nevertheless, the final (stage-2) description, while lacking mention of the robin, remains cohesive in itself.

Example 6: Cohesion [Gardeners' World]

BBC AD	There's a robin perching nearby as Monty transfers the dahlia's tuberous roots to a rubber bucket.
AutoAD Zero	**Stage-1 description:** 1. Main persons: a man and a bird; 2. Actions: the man is digging in the garden, planting a plant, and picking up a plant; 3. Person interactions: the bird is perched on a fence; 4. Facial expressions: the man is smiling.
	Stage-2 description: [Monty] plants a plant and picks up another, while a bird perches on the fence nearby.
Gemini 2.0	**Stage-1 description:** *Frame 1:* A robin sits on a wooden fence post in a lush green garden. *Frame 2:* In a flower bed, a person kneels, digging up plants. *Frame 3:* Close-up of hands carefully extracting a plant from the soil, revealing its roots. *Frame 4:* Hands place a plant with cut stems into a large black bucket already containing similar plants and name tags.
	Stage-2 description: A robin perched on a garden fence then Monty digs up and replants dahlia plants into a bucket.
Gemini 2.5	**Stage-1 description:** *Frame 1:* A European robin with a distinct orange-red breast perches on a horizontal bamboo cane tied to a vertical stake in a lush garden. Large green leaves of plants fill the background and middle ground. A blurred black hose lies coiled in the foreground left. The BBC logo is in the top left corner. *Frame 2:* The robin remains perched, facing slightly right. The garden scene is unchanged. *Frame 3:* No change in the robin or the garden setting. *Frame 4:* Still no significant change. The robin is alert on its perch. *Frame 5:* The robin stays in the same position amidst the garden greenery. ***Continuous action: * A robin sits peacefully on a bamboo stake in a garden.** *Frame 6:* Cut to a wider shot. A man with grey hair, wearing a navy blue sweater, dark trousers, and green wellington boots, bends over a raised garden bed filled with dark soil and wilting dahlia plants. A tall green hedge forms the background. A black bucket rests on the paved path to the left. *Frame 7:* The man leans further over, using his bare hands to dig into the soil at the base of a dahlia plant. A few yellow and reddish flowers are visible on the wilting plants.

	Frame 8: He continues digging, concentrating on the plant's base. Part of a golden retriever dog can be seen in the background near the hedge. ***Continuous action: * A man begins digging up a dahlia plant in a raised garden bed next to a path.** *Frame 9:* Close-up on the man's hands in the soil. He works his fingers around the thick, fleshy tuberous roots of the dahlia plant. He wears a watch on his left wrist. *Frame 10:* His hands carefully loosen the soil around the dahlia tubers, gripping the base of the stems. *Frame 11:* He continues to gently excavate the root ball, separating it from the surrounding soil. *Frame:* The man lifts the dahlia tubers slightly, preparing to remove the plant. ***Continuous action:* Close-up on the man's hands carefully digging out the dahlia tubers from the soil.** *Frame 13:* Cut to a close-up inside the black bucket. The dug-up dahlia plant, complete with soil-covered tubers and cut-back green stems, sits inside. Two white plastic labels stick out from the soil. *Frame 14:* A hand reaches into the bucket and adjusts one of the white plant labels next to the dahlia stems. *Frame 15:* The hand withdraws, leaving the dahlia plant with its labels sitting securely in the bucket. ***Continuous action:* The excavated dahlia plant is placed in a bucket, and its label is adjusted.**
	Stage-2 description: In a garden, a man carefully digs up a dahlia plant, exposing its tubers, and places it into a black bucket.

6.4 Saliency and relevance

The preceding examples have illustrated how the AI models generate descriptions through a multi-stage process. This process typically begins with highly detailed visual analysis, followed by a summarisation phase to produce the actual description. However, this summarisation does not mirror the selective information prioritisation performed by professional audio describers. Rather than focusing on the most narratively or contextually relevant elements of a scene, the AI tends to condense content without clear consideration of narrative coherence or audience needs.

Of course, the question of what to prioritise for description is complex, even for human describers, and the answer is not always

clear-cut. This can be illustrated by **Example 7**. In the first example, presenter Monty Don is working in a walled garden and then walks towards a specific section of the wall where a rose is climbing. As he stops in front of the rose, he says, “I’m going to prune this rose, which is a job that I’m looking forward to. This is a climbing rose…”.

Example 7a: Relevance [Gardener’s World]

BBC AD	Monty drops a pair of garden loppers into his wheelbarrow alongside some gloves and a roll of string. Ned falls in behind him as he pushes the barrow to an overgrown wall.
AutoAD Zero	He pushes a wheelbarrow, and she smiles back at him. He walks past a pile of tires and a greenhouse, while the dog walks in the opposite direction.
Gemini 2.0	Monty fills a wheelbarrow with gardening tools and transports it to another part of the garden, parking it next to other wheelbarrows. A person pushes a wheelbarrow down a garden path, with a focus shifting to leafy foliage.
Gemini 2.5	In a walled garden, Monty retrieves shears from a wheelbarrow, then pushes it across the brick patio towards other parked wheelbarrows. A person in wellies pushes a wheelbarrow past others; the focus shifts to a plant branch as the person walks away into the blurred background.

In relation to this example, the participants in our focus group agreed that the reference to the wall was helpful in the description, because, as one of the AD users put it, “the roses could be claiming a fence … with that [description] you know it's a wall it's climbing up as opposed to something else” (AD user 1). Interestingly, the BBC audio description places a reference to the wall at the end of the segment, following Monty’s journey through the garden and concluding at the point where he stops. This structure creates a strong link with his subsequent spoken line about pruning a climbing rose. By contrast, only one of the AI-

generated descriptions (Gemini 2.5) briefly mentions the walled garden—and only at the beginning of the description.

In the second example, taken from the opening scene of the same Gardeners' World episode, Monty is seen picking runner beans and verbally refers to the action, saying, "I think I've got all the runner beans that are left to pick," before the scene cuts to a close-up of a basket into which he is placing the beans.

Example 7b: Relevance [Gardener's World]

BBC AD	In the vegetable plot at Longmeadow, Monty's with his dog Ned picking runner beans and putting them in a basket.
AutoAD Zero	He tends to the garden.
Gemini 2.0	The clip opens with vibrant red berries before revealing a tranquil garden where an elderly man is carefully tending to his plants.
Gemini 2.5	Small red berries hang on a branch before a wide shot reveals a person tending a lush vegetable garden, viewed later through foreground flowers.

The BBC AD includes the action of picking runner beans. However, unlike the earlier "wall" example—where the reference in the description *complements* Monty's verbal mention of the climbing rose—this description creates a degree of redundancy with what Monty says. One of the describers commented that they would still choose to describe the runner beans, following the standard principle of describing what is seen on screen and taking into account the time available in the scene. One AD user noted, "in one of the descriptions it says he's picking runner beans, but he tells you that anyway" (AD user 1). The user contended that while there was enough time to include this visual detail, had time been more limited, it might have been preferable to prioritise other information.

In many cases, relevance is created through a level of nuance that in the description that the AI-generated versions typically do not capture. For instance, in the following scene from *Bargain Hunt*, presenter Christina sits in a parked car, mimicking driving by turning the steering wheel and imitating engine sounds. This playful act, which adds a humorous element to the show, is missed by the AI-generated descriptions. They focus on the appearance of the car (described correctly as orange with blue flame designs) and/or literal actions—such as turning the steering wheel—without conveying the fact that the car is stationary and the driving is being feigned. This omission results in a loss of the scene's intended humour.

Example 7c: Relevance [Bargain Hunt]

BBC AD	Christina's turning the steering wheel of a **parked** car.
AutoAD Zero	A woman is driving a car.
Gemini 2.0	A woman in sunglasses sits in the driver's seat of an orange car with blue flame designs.
Gemini 2.5	A blonde woman in sunglasses and driving gloves sits in an orange classic car, holding the steering wheel before turning with a slight smile.

The examples discussed in this section highlight the nuanced judgement required—something current AI summarisation processes do not reliably replicate. They also underscore the inherent subjectivity involved in all forms of translation or transfer, whether between languages or modalities. A core professional element of audio description (and translation), therefore, is intersubjectivity: the ability to base description on informed decisions about what the audience is likely to value or need.

6.5 Engagement and enjoyability

A question closely linked to achieving relevance in audio descriptions is how to make them engaging for the audience. This

emerged as another area in which the current AI-generated descriptions face significant challenges.

One example discussed in the focus group was the description of the robin appearing in a scene from Gardeners' World. The AI models sometimes described it only "a bird". However, the AD users unanimously noted that in a nature programme like Gardeners' World, they expect the description to include details about the animals featured. For example, rather than simply being told that a bird appears, participants expressed a preference for being told what kind of bird it is—or, if the species is rare and unlikely to be familiar to the audience, for a description of its appearance to be included. As one user explained, this level of specificity is "more relevant in this genre of programme" (AD user 3), and another noted that it contributes to a more authentic outdoor viewing experience (AD user 1). This feedback underscores the link between relevance, engagement, and enjoyment: making genre-appropriate choices—such as naming animals or describing their appearance—can enhance both the relevance and emotional resonance of the description.

Similarly, one of the describers also drew a connection between character identification and audience engagement, noting that the use of appropriate character identifiers—such as including the name of Monty's dog, Ned, who frequently appears in the show and is familiar to regular viewers—can foster emotional connection and enhance viewer engagement.

While current AI models can be prompted to adopt particular tones or styles, additional training of AI models will be needed to align their outputs with what users of specific programme types find engaging and meaningful.

6.6 Handling of written text and graphics

While this aspect was not examined in detail in the current study, our analysis indicates that the treatment of on-screen text varies across AI models and audiovisual content, with no discernible pattern. For instance, in the Gardeners' World episode, both Gemini-generated descriptions initially omitted the word "World" from the programme title but included it in the subsequent segment (**Example 8**). This inconsistency has the potential to confuse users.

The handling of graphical elements such as programme logos varied as well, as illustrated in the example below. The AutoAD Zero model did not include a description of the logo, appearing to prioritise the identification of characters and their actions or interactions. As there were no characters present in this scene, the model seemed unable to generate a meaningful description. AD users remarked that they would have preferred a description of the logo, particularly because such branding elements can change over time. While the BBC's AD also omitted the logo, users appreciated that it referenced the programme's hashtag, which was shown on screen.

Example 8: Programme logo and title [Gardeners' World]

	Segment 1	Segment 2
BBC AD	BBC. Gardeners' World.	Hashtag Gardeners World.
AutoAD Zero	The rake and trowel are arranged neatly beside the tree.	There is no action or character interaction in this 2.23-second video clip.
Gemini 2.0	A tree drawing grows on screen surrounded by gardening tools and is revealed to be the title of the show "Gardeners."	The "Gardeners' World" title card fades into a close-up of vibrant pink flowers with yellow centers in a soft, green garden setting.
Gemini 2.5	Framed by autumn leaves, a white tree silhouette grows and gardening icons appear, followed by the golden title "Gardeners".	The "Gardeners' World" title appears framed by autumn leaves, followed by a close-up of vibrant pink cosmos flowers blooming against a blurred green background.

6.7 Timing

Effective AD should be carefully timed to fit naturally into pauses in the programme, avoiding any overlap with dialogue or important sound cues. The length of each segment must align with the rhythm and pacing of the original content, ensuring the description can be delivered clearly and at a natural pace. The AI-generated descriptions in our sample material had a tendency of being too long for the available space, potentially resulting in rushed delivery or omitted content. In some cases, there was also a tendency toward over-description, where unnecessary detail may hinder

rather than enhance comprehension. This is illustrated in **Example 9** below, where the BBC AD offers a concise summary, stating simply that the players are playing Lacrosse, while the AI-generated descriptions provide much more detail on how the game is played. However, participants in our focus group expressed appreciation for the additional detail offered by some AI-generated descriptions when compared with the BBC AD. What needs to be tested further is whether this higher level of detail remains desirable over longer periods—such as during a full programme or episode—rather than in the short clips used in the focus group setting.

Example 9: Length of AD segments [Bargain Hunt]

BBC AD	Girls playing lacrosse.
AutoAD Zero	[They] run determinedly, trying to get the ball.
Gemini 2.0	On a green field, women play lacrosse and attempt to gain control of the ball.
Gemini 2.5	A player in white carries the ball but is tackled by a red defender, causing her to fall and lose possession on the lacrosse field.

6.8 Summary

Current transformer-based, zero-shot AD models and pipelines show significant advances over earlier models, particularly in generating descriptions for very short video clips. These models have improved in areas such as character identification, re-identification, and overall narrative flow. However, in our sample material, the accuracy of the descriptions remained inconsistent, with meaning shifts, object misidentification, and occasional misgendering or omission of character names still occurring in considerable measure. Descriptions also often focussed on non-salient details or perspectives, reducing their relevance and clarity.

While some AI outputs contain more detail than traditional AD, this can result in over-description that exceeds the available time,

particularly in broadcast contexts. There are also issues with cohesion and narrative continuity across segments. Descriptions sometimes include technical or unconventional phrasing, such as referring to camera movements. Although the generation process with AI appears to be faster, the linguistic analysis suggests that the reliability and scalability of current AI models for full-length programmes remain uncertain.

7 Feedback from AD users

A focus group with three experienced and active audio description users was conducted online on 24 April 2025. Participants were presented with AI-generated descriptions from the three selected models—AutoAD Zero, Gemini 2.0 Flash, and Gemini 2.5 Pro—alongside the corresponding original audio descriptions. The various versions were discussed and evaluated by the group. To maintain consistency across the samples and eliminate potential bias introduced by voice performance, which was not part of the evaluation, all samples—including the original audio descriptions—were voiced using the same synthetic voice, generated via ElevenLabs.

7.1 Accuracy

The users felt that the AI-generated descriptions were often better than expected but they also identified several problems such as a lack of accuracy and lack of useful detail in the AI descriptions:

> “I think yeah, it exceeded what I was expecting from the short AI clips I've heard before, they tend to be very, I don't know, I just see them as very binary sometimes, but that one was a lot better. It painted a much better picture of what was actually on the screen, much better than what I was expecting to hear.” (Participant 1; 33:04)

> “Think of it's a nature programme you'd like to know that it would be accurate about telling you what the animals, the birds were, it would identify them. Although I know it's about gardens, but you know it's the whole outdoors nature thing. You may think that it would get those accurate and tell you what kind of bird that is. Different seasons and all that kind of thing.” (P1; 49:46)

> “Yeah, I think I agree with [P1] on that one. Yeah, I think it's all nice to know. So ‘Robin’, rather than just sort of a generic bird sort of thing.” (P3; 50:05)

7.2 Attitudes towards AI-AD

The users’ overall assessment was that the AI descriptions are not ready for use without human review:

> “Probably at the moment I would say it probably still does need a human sort of maybe to check its accuracy, because obviously there was in sort of all of them at times a little bit of sort of misdirection and maybe not accurate audio description.” (P3; 1:08:01)

> “But yeah, I mean, it's definitely, definitely better than I thought. Like [P1] said earlier, and if it's a way of sort of adding more audio description to sort of different programmes that might not have audio description at all, I think it's definitely promising.” (P3; 1:08:17)

> “So yeah, I don't think it's quite there. Yeah, I think it still needs a bit of review by a human, but it's a great step forward, but audio description is a real art form and you know, describers who are really good at it are also good at putting emotion into it. You know they can make their voices work to the type of programme or film you're watching. They can put emotion, excitement into it.” (P1; 1:08:37)

> I don't think AI will get that quite yet. At the moment I think you know it takes a bit of emotion and thought to do something like that, but I was really impressed by the short clips we've had so far.” (P1; 01:08:37)

7.3 Social inclusion

Users reiterated the importance of AD beyond its basic function of describing a programme’s content, but also as a tool for promoting social inclusion. Accurate AD enables users to fully engage with and enjoy media, which in turn supports their meaningful participation in society:

> “Goes back to just being accurate about what's actually being seen on the screen. If it's going to be part of that social chat because some of the things that described you were saying completely inaccurate.” (P3; 1:10:57)

7.4 Trust

The overall conclusion was that, although AI-generated descriptions can aid basic understanding of media content, AI remains too unreliable to be used without human oversight:

> “You can't fully trust it at the moment.” (P3; 1:11:55)

8 Feedback from audio describers

Two experienced audio describers were invited to provide informal feedback on the same sample materials reviewed by the AD users. These sessions took place on 24 and 28 April 2025. Following the discussions, one of the describers independently assessed the level of effort required to post-edit the AI-generated scripts. This section presents a thematic summary of the audio describers’ feedback. The final section addresses the viability of post-editing the AI-generated scripts.

8.1 Human expertise and subjectivity in AD

- Both describers emphasised the complex editorial decisions involved in AD—what to describe, how, and when.
- Subjective elements like emotion, affection, or gender representation appeared particularly challenging for the reviewed AI-generated scripts; currently best handled by human describers.
- The use of specific names, places, and tone was cited as a way humans create emotional resonance—something AI struggles to replicate meaningfully.
- One describer highlighted that the reviewed AI descriptions were not non-sensical, comparing some elements in them to the descriptions produced by untrained describers/beginners.
- The other describer was more sceptical, emphasizing that even basic errors (e.g., misgendering) can be complex and problematic, and not easily “fixed” post hoc.

8.2 AI Capabilities and limitations

- AI lacks nuance and struggles with interpreting emotion or subjective content.
- AI-generated AD can produce excessive or irrelevant description (over-description) and does not always follow AD conventions.
- Both interviewees saw some promise in hybrid models, especially where AI is involved in non-creative tasks or high-volume, low-stakes content, factual programmes.
- Pointed to potential use of AI for legacy material. Would have to be clearly labelled “AI-generated”.
- One describer was more open to using AI-generated scripts as editable drafts, potentially reducing workload if properly integrated.
- The other describer remained sceptical, believing it would be quicker for professionals to work from scratch.

8.3 AI Integration in workflows

- Current AI tools (e.g., for gap detection or scripting) are not sufficiently accurate or workflow-compatible.
- Consensus that genre- or programme-specific AI models could be more effective than generic solutions.
- Review processes require nuanced human input, especially to maintain clarity, timing, and appropriateness.
- Enthusiasm for experimenting with AI-assisted drafting and editing varied between the interviewees.
- Potential benefit: coverage of more content including legacy content
- Risk: quick/light-touch review of AI content may become the norm, reducing quality standards

8.4 Ethical and professional considerations

- Both describers thought that describers’ professional judgment must remain central to the process.
- The idea of releasing unreviewed AI-generated AD was viewed as dangerous, ethically questionable, and legally risky.
- Both stressed the need to clearly label any AI-generated AD and involve professionals in AI training to ensure safe, effective use.
- One interviewee was interested in AI's potential if used responsibly and with human oversight.

- The second was more consistently sceptical, warning against assumptions that users can detect accuracy issues.

8.5 Future direction: Hybrid human-AI models

- Both describers felt that a hybrid, human-centric model is the most realistic and responsible way forward.
- This is not just to protect the describers' jobs, but to ensure accuracy, ethical use, and content quality.
- Adoption should be cautious and iterative, guided by describer input and practical testing—not driven solely by cost-cutting or automation trends.

Both feedback sessions underline that while AI may support certain technical or repetitive aspects of AD, its current limitations in nuance, emotion, and ethical reliability mean that human describers remain essential. There is space for innovation, particularly in legacy content and lower-stakes genres, but any AI integration must be transparent, carefully reviewed, and shaped by professional describers' expertise.

8.6 Viability of post-editing AI-generated AD scripts

One of the two audio describers also reflected on the potential for post-editing AI-generated scripts, based on two extracts from the *Gardeners' World* episode analysed in this study. Following the focus group and feedback sessions—where one of the AutoAD Zero descriptions was perceived as less effective than the Gemini versions—we presented both Gemini descriptions to the describer for further evaluation. Their feedback is reported below.

Overview: In the describer's view, it is currently more time-efficient to script this type of programme from scratch than to post-edit AI-generated scripts. This is largely due to the relatively straightforward nature of scripting such shows, which often allows for near real-time scripting during viewing. For example, editing and refining a 10-minute extract of the Gardeners' World episode took approximately 1 hour, whereas scripting AD from scratch for a 60-minute programme would typically take around three hours. This was despite the describer already being familiar with the material from an earlier informal feedback session.

Comparison of AI-generated versions: Selecting the superior AI-generated version proved challenging. Both versions exhibited

a consistent flaw: the failure to recognise and re-identify the main character, Monty Don, instead treating each appearance as a separate individual. Gemini 2.5 showed greater stylistic sophistication in places; however, the audio describer highlighted that this was less suited to lifestyle programming, which benefits from clear, concise AD that closely aligns with the visual content.

Timing and pacing issues: A key concern with both AI-generated versions was the pacing of the descriptions. The audio describer noted that much of the content was delivered too quickly for comfortable listening, with descriptions often occurring in rapid succession. Addressing these issues would require substantial editing—not just minor adjustments—thereby increasing overall production time.

Potential AI Applications: One promising area for AI integration could be the generation of character and location descriptions. These could be produced independently of the main programme and made available to viewers in advance, thereby enhancing accessibility. For example, in theatrical audio description, audiences are often provided with character descriptions and touch tours prior to the performance. Similarly, AI could assist in creating descriptions for key locations—such as hotel lobbies, gardens, or homes—that are often difficult to incorporate into the main AD track due to time constraints.

9 Conclusions and implications

9.1 Progress

Advances in video scene understanding and video captioning using large vision-language and language models—have enhanced image-text representations in audiovisual content. While early models primarily focused on single-frame captioning, newer approaches increasingly incorporate multi-frame captioning, improving character tracking and the recognition of actions and interactions (Chu et al., 2024; Han et al., 2023a, 2023b; Ye et al., 2024). These advancements have contributed to greater continuity in descriptions compared to earlier methods (Braun et al., 2020).

In experimental work on the automatic generation of AD, two-stage approaches that combine dense video captioning with subsequent summarisation have yielded richer and more detailed descriptions than previous techniques and have taken into account the specific

requirements for AD such as its need to fit into gaps in the audio track (Chu et al., Shen et al., 2023; Wang et al., 2021).

Meanwhile, complex multimodal pipelines are beginning to integrate additional sources of input—such as scripts, screenplays, subtitles, audio tracks, and contextual data—in the AD generation task for greater accuracy and relevance of the descriptions (Campos et al., 2020; Chu et al., 2024; Han et al., 2023a, 2023b; Wang et al., 2024; Ye et al., 2024).

9.2 Limitations and risks of AI in AD

AI-generated audio description (AD) remains experimental. Current state-of-the-art models continue to struggle with contextual awareness, accuracy, nuance, cohesion, and narrative depth. As such, it is unlikely that AI-generated descriptions—without human review and post-editing—can yet support a coherent understanding of different audiovisual genres or their narrative content. In fact, post-editing appears to take longer than writing AD from scratch, although further research and testing are needed to confirm this.

These models also perform poorly on longer video segments, which compromises continuity. It remains uncertain whether current AI AD models, developed for short clips, can be effectively scaled to support cohesive storytelling across extended sequences.

Additional limitations include limited user involvement during the design and development stages (rather than only in evaluation), and insufficient engagement with AD research from the humanities and social sciences. These gaps risk undermining both genuine accessibility and user trust. The feedback from AD users gathered in our study clearly highlights the need to better account for user priorities when determining how to shape in AD.

9.3 Implications

- Given the current state of development, AI-generated AD is deemed—by end users and describers—to be unsuitable for use without human supervision and review.
- The risks associated with AI-based video scene understanding and language generation mean that, for now, these techniques

cannot be reliably deployed in AD, particularly without human oversight.

- An AI-driven workflow would also not be feasible from a practical point of view, the current limitations of AI models regarding the length of video segments they can process.
- However, the rapid advancement of AI suggests significant future opportunities for its use as an assistive tool in the AD production workflow.
- When combined with human expertise, AI has the potential to broaden the availability of AD while ensuring that AD outputs remain of high quality and provide meaningful media access for blind and partially sighted audiences.
- At this stage, further feasibility testing and research across a wider range of AI models, pipelines, and genres is essential. This will help identify opportunities for the safe integration of AI tools in AD workflows.
- Such testing and research could pave the way for scalable and more reliable AI-enabled solutions.
- If AI-generated descriptions are used—whether with or without human review, for instance as a temporary solution or for legacy content—they must be clearly labelled as such. Additionally, the limitations of AI should be communicated to end users to ensure transparency. However, this approach has inherent risks: even with a warning, users have no reliable way of identifying where the AI may have made errors.
- Organisations producing and offering/delivering AD, organisational processes and policies should to be developed to mitigate and minimise the risks arising from the use of AI in AD workflows.

10 Recommended next steps

We recommend that further research, development and testing should be done under three strands.

10.1 Strand 1: Further exploration of the viability of hybrid workflows

1. Further testing of hybrid human-AI workflows
    1. Some professional describers already use specialised authoring tools such to create and edit AD, although opinions on their accuracy and usefulness vary. Integrating AI into these platforms, could - over time - enable the generation of

first-draft AD scripts, which professional describers can then refine, ensuring that tone, accuracy, and clarity are preserved.

Notes:
This hybrid approach supports scalability without compromising quality and the human touch that is central to high-quality AD.
To be viable in real-world settings, AI would need to process longer video segments effectively.
A key area for testing is identifying which types of content or production scenarios would benefit most from AI-assisted drafting combined with human post-editing, evaluated both in terms of time savings and maintaining quality.

2. AI could also be explored as a **quality assurance and review tool**, for example, checking consistency of descriptions within episodes and across episodes created by different describers, or reviewing descriptions against specific templates or editorial guidelines.

3. Another potential use of AI is for generating initial draft scripts for 'uncontroversial' **legacy material**, followed by light-touch human review; given that legacy material is not normally prioritised for AD. This would, however, require the development of clear labelling conventions (e.g., AI-generated, AI-generated with human review) to signal the nature and reliability of the description to end users. (See also previous section for limitations of this approach.)

Next step: Building on the pilot, explore collaboration with all stakeholders including describers, broadcasters, software providers to develop prototype workflows and hold discovery sessions to determine the viability and usefulness of different hybrid workflows.

2. Improving AI quality through prompt engineering
A further area to explore within the hybrid workflow strand would involve prompt engineering and refinement, experimenting with different phrasing, formats, and styles to get the best results from the model for different genres.

Next step: Develop a prompt-testing log to record which prompt styles work best for different programmes and genres.

Over time, build a prompt library and best-practice prompt guide for AD generation.

3. Expanding the range of genres

Feedback from describers confirms that current AD coverage remains low, at around 20%. AD users have consistently stated that most programmes without audio description or an equivalent such as an enhanced commentary integrated into a TV programme are effectively inaccessible to them.

It would be useful to explore which genres could benefit most from AI support by assessing their specific demands against the current capabilities of AI. For instance, genres such as sports and quiz shows—often rich in on-screen text and data—may be particularly well suited to AI-assisted description.

Next step: After initial workflows are refined, expand testing further to new genres to assess model performance under more complex conditions.

10.2 Strand 2: Further (technical) development

We recommend the following areas of further development, as a basis for improving the quality of AI-generated draft scripts that become useful for human post-editing.

4. Development of viable advanced and multimodal pipelines for AD generation

A further very promising avenue is the integration of additional input sources into AI pipelines. This could include a **face recognition** module to improve character identification, by drawing on external information about the characters, or modules capable of processing **existing scripts, programme guides** and/or the programme's original **narration/ dialogue** (e.g. via captions), and feed this information into the core description module, to make the descriptions more accurate and relevant in the context in which they occur.

Another useful addition could be a module that processes the **preceding description segment** to improve the cohesion and continuity of the subsequent description segments.
More advanced pipelines should also incorporate **object detection/classification tools** that identify objects particularly

relevant to the programme—such as, plants and animals in a nature programme. For example, Google's recently released open-source model SpeciesNet can identify animal species from images. If these models provide confidence scores, low-certainty predictions can be filtered out to preserve accuracy.

A more complex challenge for AI is deciding whether an object is familiar enough to be simply labelled (e.g. *a robin*), or whether it is too specialised for an audience (e.g. *a resplendent quetzal*) and should be described (e.g. *a brightly coloured bird with iridescent green feathers and a long tail*).

Next step: Explore multimodal pipelines and additional modules, and identify the improvements compared to the models/pipelines used in the current project.

5. **Development of AI models for describing longer video segments**

Currently, AI can generate descriptions for only short clips, not entire programmes or episodes. It needs to be tested whether existing models—primarily zero-shot approaches based on pretrained transformer architectures—can be extended to generate descriptions for longer clips, or whether new models and processing pipelines are required.

The potential of agentic AI—i.e., autonomous systems capable of planning, using tools and adapting over time—should be explored to evaluate whether they improve the accuracy and narrative continuity of descriptions across full-length programmes. Agentic AI may also be capable of integrating the preprocessing steps in current models into the AI pipeline.

Model context protocols (MCPs) may provide further opportunities. MCPs are open protocols that enable seamless integration between Agentic AI systems and external tools and sources. They could be employed to provide the AI model with relevant internal and external contextual information—such as information about a TV programme, prior episodes, prior scenes, character references, or background knowledge from external sources—to guide the model's behaviour, improve continuity across individual AD segments, and improve the relevance of the descriptions in their narrative context. Techniques such as retrieval-augmented generation (RAG) may play a key role by dynamically supplying

the model with pertinent background knowledge or previously generated content.

Next Step: Determine whether current models can be scaled up to generate AD scripts for longer clips and entire programmes, whether the current limitations are of a practical nature and can be overcome, or whether alternative models capable of generating AD for long-form content are needed.

6. **Fine-tuning existing AI models using AD-specific training data**

One of the most promising avenues for improvement of AI descriptions may be the post-training (e.g., fine-tuning) of generic AI models with AD-specific data (i.e. existing audio descriptions) and with AD guidelines and standards. This may allow AI to generate AD that is more accurate, context-aware, and stylistically aligned with industry norms.

Next step: Explore the feasibility of curating existing audio description datasets and their transparent use to train AI models. Develop protocols and ethical procedures for using audio descriptions generated by human professionals in order to train machines.

10.3 Strand 3: Evaluation methods

7. **Development of evaluation framework**

Development of a unified framework for benchmarking AD quality; a framework that facilitates evaluation of AD regardless of how it was generated. This will involve systematic and comprehensive working and testing with AD users to derive quality parameters and thresholds.

References

Braun, S., Starr, K., 2019. Finding the Right Words: Investigating Machine-Generated Video Description Quality Using a Corpus-Based Approach. Journal of Audiovisual Translation 2, 11–35. https://doi.org/10.47476/jat.v2i2.103

Braun, S., Starr, K., 2022. Automating Audio Description, in: Taylor, C., Perego, E. (Eds.), The Routledge Handbook of Audio Description. Routledge, London, pp. 391–406.

https://doi.org/10.4324/9781003003052-30
Braun, S., Starr, K., Delfani, J., Tiittula, L., Laaksonen, J., Braeckman, K., Van Rijsselbergen, D., Lagrillière, S., Saarikoski, L., 2021. When Worlds Collide: AI-Created, Human-Mediated Video Description Services and the User Experience, in: Lecture Notes in Computer Science. Springer, Washington D.C./Virtual, pp. 147–167. https://doi.org/10.1007/978-3-030-90328-2_10
Braun, S., Starr, K., Laaksonen, J., 2020. Comparing human and automated approaches to visual storytelling, in: Braun, S., Starr, K. (Eds.), Innovation in Audio Description Research. Routledge, London, pp. 159–196. https://doi.org/10.4324/9781003052968-9
Campos, V.P., de Araújo, T.M.U., de Souza Filho, G.L., Gonçalves, L.M.G., 2020. CineAD: a System for Automated Audio Description Script Generation for the Visually Impaired. Universal Access in the Information Society 19, 99–111. https://doi.org/10.1007/s10209-018-0634-4
Capel, T., Brereton, M., 2023. What is Human-Centered about Human-Centered AI? A Map of the Research Landscape, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. ACM, Hamburg Germany, pp. 1–23. https://doi.org/10.1145/3544548.3580959
Chang, R.-C.C., Sheshadri, S., Guo, A., Hara, K.-T.K., Natalie, R., Chang, R.-C.C., Sheshadri, S., Guo, A., Hara, K.-T.K., 2024. Audio Description Customization. ASSETS 2024 - Proceedings of the 26th International ACM SIGACCESS 1, 19. https://doi.org/10.1145/3663548.3675617
Cheng, Z. et al., 2024. VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs. *arXiv preprint*. https://arxiv.org/abs/2406.07476
Chu, P., Wang, J., Abrantes, A., 2024. LLM-AD: Large Language Model based Audio Description System. https://arxiv.org/abs/2405.00983v1
Donahue, J. et al., 2015. Long-term recurrent convolutional networks for visual recognition and description. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2625–2634.
Fryer, L., 2018. Staging the Audio Describer: An Exploration of Integrated Audio Description. Disability Studies Quarterly 38. https://doi.org/10.18061/dsq.v38i3.6490
Grattafiori, A. et al., 2024. The Llama 3 Herd of Models. *arXiv preprint*. https://arxiv.org/abs/2407.21783

Han, T. et al., 2024. AutoAD III: The Prequel -- Back to the Pixels. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pages 18164–18174, 2024.
Han, T., Bain, M., Nagrani, A., Varol, G., Xie, W., Zisserman, A., 2023a. AutoAD: Movie Description in Context. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2023-June, 18930–18940. https://doi.org/10.1109/CVPR52729.2023.01815
Han, T., Bain, M., Nagrani, A., Varol, G., Xie, W., Zisserman, A., 2023b. AutoAD II: The Sequel – Who, When, and What in Movie Audio Description. Proceedings of the IEEE International Conference on Computer Vision 13599–13609. https://doi.org/10.1109/ICCV51070.2023.01255
Johansson, R., Rastegar, T., Lyberg-Åhlander, V., Holsanova, J., 2024. Event Boundary Perception in Audio Described Films by People Without Sight. Applied Cognitive Psychology 38, e4228. https://doi.org/10.1002/acp.4228
Li, J. et al., 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv preprint*. https://arxiv.org/abs/2301.12597
Nadeem, A. et al., 2024. NarrativeBridge: Enhancing Video Captioning with Causal-Temporal Narrative. *arXiv preprint*. https://arxiv.org/abs/2406.06499
Natalie, R., Loh, J., Tan, H.S., Tseng, J., Chan, I.L.Y.-R., Jarjue, E.H., Kacorri, H., Hara, K., 2021. The Efficacy of Collaborative Authoring of Video Scene Descriptions, in: Proceedings of the 23rd International ACM SIGACCESS. ACM, New York, NY, USA, pp. 1–15. https://doi.org/10.1145/3441852.3471201
Pavel, A., Reyes, G., Bigham, J.P., 2020. Rescribe, in: Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology. ACM, New York, NY, USA, pp. 747–759. https://doi.org/10.1145/3379337.3415864
Radford, A. et al., 2021. Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint*. https://arxiv.org/abs/2103.00020
Shen, X., Li, D., Zhou, J., Qin, Z., He, B., Han, X., Li, A., Dai, Y., Kong, L., Wang, M., Qiao, Y., Zhong, Y., 2023. Fine-grained Audible Video Description, In Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10585–10596. 2023. https://doi.org/10.1109/CVPR52729.2023.01020
Starr, K.L., Braun, S., Delfani, J., 2020. Taking a Cue from the Human: Linguistic and Visual Prompts for the Automatic Sequencing of Multimodal Narrative. Journal of Audiovisual

Translation 3, 140–169. https://doi.org/10.47476/jat.v3i2.2020.138
Touvron, H. et al., 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint*. https://arxiv.org/abs/2307.09288
Vaswani, A. et al., 2017. Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6000–6010.
Wang, H., Tong, Z., Zheng, K., Shen, Y., Wang, L., 2024. Contextual AD Narration with Interleaved Multimodal Sequence. https://arxiv.org/abs/2403.12922v1
Wang, Y., Liang, W., Huang, H., Zhang, Y., Li, D., Yu, L.-F., 2021. Toward Automatic Audio Description Generation for Accessible Videos, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 1–12. https://doi.org/10.1145/3411764.3445347
Xie, J. et al., 2024. AutoAD-Zero: A Training-Free Framework for Zero-Shot Audio Description. In: *Proceedings of the 17th Asian Conference on Computer Vision*, pp. 2265–2281.
Yang, Z. et al., 2023. The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision). *arXiv preprint*. https://arxiv.org/abs/2309.17421
Ye, X., Chen, J., Li, X., Xin, H., Li, C., Zhou, S., Bu, J., 2024. MMAD: Multi-modal Movie Audio Description, in: Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (Eds.), Proceedings of the 2024 LREC-COLING 2024), Torino, Italia, pp. 11415–11428.
Zhang, S. et al., 2022. OPT: Open Pre-trained Transformer Language Models. *arXiv preprint*. https://arxiv.org/abs/2205.01068
Zhu, D. et al., 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *arXiv preprint*. https://arxiv.org/abs/2304.10592