Bridging the accessibility gap? The role of AI in the future of audio description generation

Sabine Braun
Centre for Translation Studies
University of Surrey

@drsabinebraun.bsky.social
@drsabinebraun

UNIVERSITY OF SURREY

Centre for Translation Studies

OVERVIEW

- AI in audio description: How we got here
- AI Video captions: Early models, evaluation, refinements
- From AI captions to AI audio description: Recent models, evaluation
- Further recent developments
- Limitations and risks of AI in AD
- Alternatives to (full) automation
- Where do we go from here?

AI IN AD: HOW WE GOT HERE

- **Rising demand for accessible media content** due to
  - Increase in audiovisual content, growing accessibility awareness, broader legislation
- Traditional AD workflows face challenges in meeting this demand
  - High quality but slow and costly
- Human-generated AD remains the gold standard, but difficult to achieve for:
  - Real-time/streaming content, user-generated content
- In parallel, computer vision/AI techniques for automatic description of visual content
  - **To what extent is AI a viable solution?**

AI VIDEO CAPTIONS – EARLY DAYS 1/2

10-12 years ago…
Deep Learning models for video captioning:

- **Neural networks:** Convolutional Neural Networks (CNNs) to detect & classify objects in images through **feature extraction**; Recurrent Neural Networks (RNNs) to generate text (Aafaq et al., 2019)
- **Machine learning:** Networks were trained using **supervised** methods relying on annotated image **data sets** to learn connections between objects in images and text descriptions (e.g., Rohrbach et al., 2013)
- **Single-frame captioning**; no continuity, i.e. no character/object tracking, no action recognition, temporal sequencing etc.

Saving Mr Banks (2013), Walt Disney Pictures

AI VIDEO CAPTIONS – EARLY DAYS 2/2

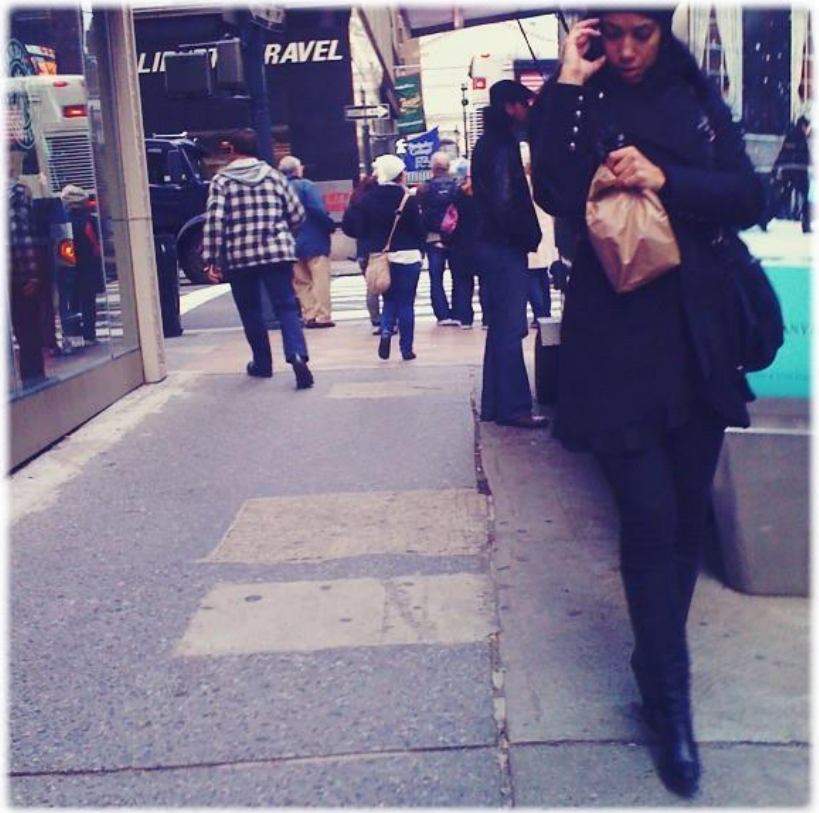
- **Image banks as training data sets:**
  - Non-iconic still images or limited motion images
  - Object segmentation: boundary boxes
  - Object labelling through crowdsourcing: **one-sentence captions**
  - e.g. **MS COCO** (Lin et al., 2015): 330k images with 1.5m objects; 5 captions per image
  - Sparse instructions, little training for annotators, limited quality control -> partly amateurish descriptions
- **Mismatch with requirements for AD**

MS COCO dataset

EXAMPLE: CROWDSOURCED IMAGE CAPTIONS 1/2

MS COCO: Flickr photos with five crowd-sourced captions

1. there is a lot of foot traffic on this street during the day.
2. people walking down a sidewalk near a road and a building.
3. a street with various people walking by a building.
4. there are people that are walking on the street
5. an image of a person walking down the street on her phone

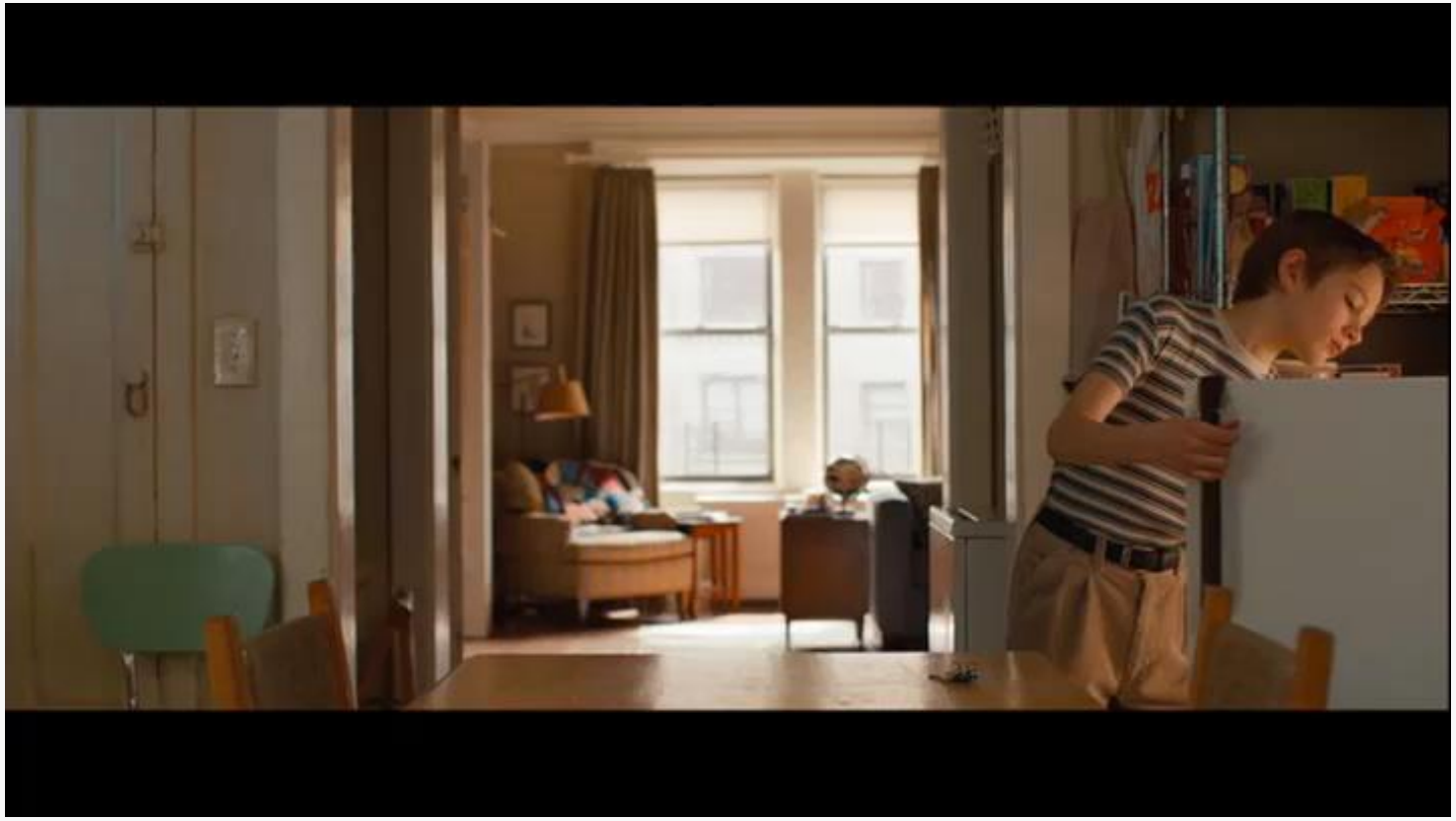


EXAMPLE: CROWDSOURCED IMAGE CAPTIONS 2/2

MS COCO: Flickr photos with five crowd-sourced captions

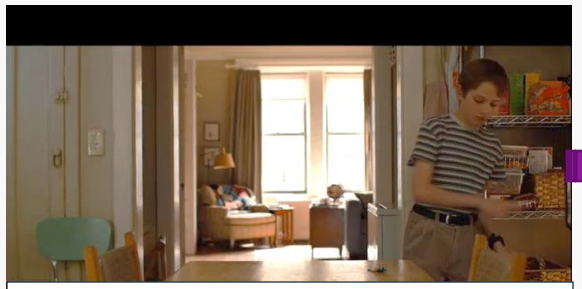
1. a grandmother standing next to a child in a kitchen.
2. baby trying to open wooden cabinets under the sink.
3. a woman and child stand in the kitchen.
4. an older woman is standing in the kitchen with a child.
5. the little girl is trying hard to open the cabinets.

EXAMPLE: HUMAN AD

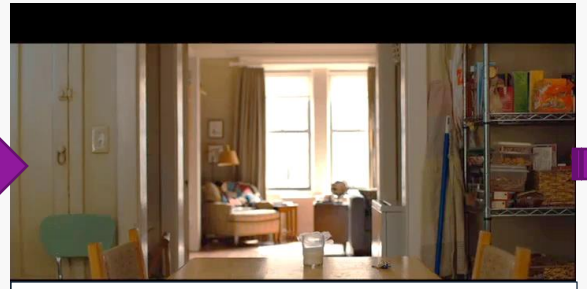


Extremely Loud and Incredibly Close, 2011

EXAMPLE: AI VIDEO CAPTIONS, 2018

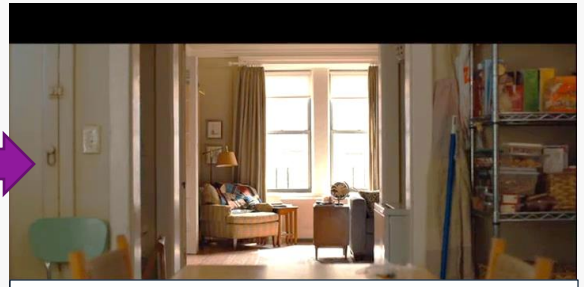


1 [100902:52]
A woman is sitting on a bed and smiling.



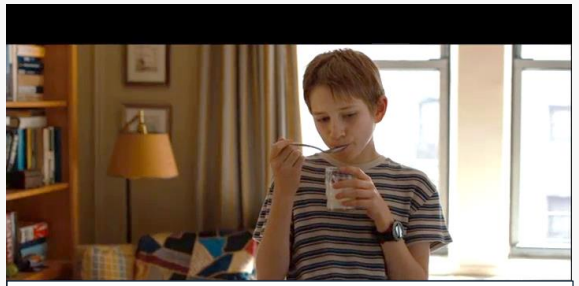
2 [100902:162]
A kitchen with a window and a window.

3 [100902:338]
A man is dancing in a room.

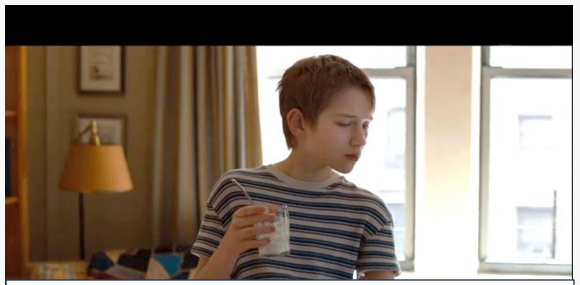


4 [100902:788]
A woman is sitting in a chair and looking out the window.

5 [100902:1191]
A book shelf with a teddy bear sitting on it.



6 [100902:1362]
A woman is holding a baby and looking at it.



7 [100902:1538]
A man is dancing in a room.

AI VIDEO CAPTIONS – EVALUATION (MEMAD)

Challenges for AI AD (2019/2020)	**Recommendations** (Braun & Starr, 2019, 2022)
Inaccurate object, action, character recognition	Improved algorithms & feature extraction, use scripts, subtitles, dialogue
Errors in object scale & perspective	Context-aware object recognition & topic detection
Gender bias in character identification	Facial recognition, voice diarization, 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, omit basic actions (talking)

MeMAD: Methods for Managing Audiovisual Data; EU Horizon 2020 grant No 780069
Surrey Research group: Kim Starr, Jaleh Delfani, Arianna Carloni, Sabine Braun

MeMAD
Methods for Managing Audiovisual Data

AI VIDEO CAPTIONS – REFINEMENTS

- **Character tracking / re-identification** (Rohrbach et al., 2015, 2017)
- From single-frame to **cross-frame captioning** (Venugopalan et al., 2015)
- **Multi-sentence descriptions** (Yu et al., 2016)
- **Dense video captioning** (Krishna et al., 2017)
- **Video data sets for training**: domain-specific (e.g. cooking shows), then generic (e.g. movies, activities) and audio descriptions (LSMDC)
- **Audio description generation** based on dense video captioning and similarity detection to reduce redundancy; including automatic gap detection; user testing (Wang et al., 2021)
- **Visual storytelling** (Huang et al., 2016)

EXAMPLE: VISUAL STORYTELLING

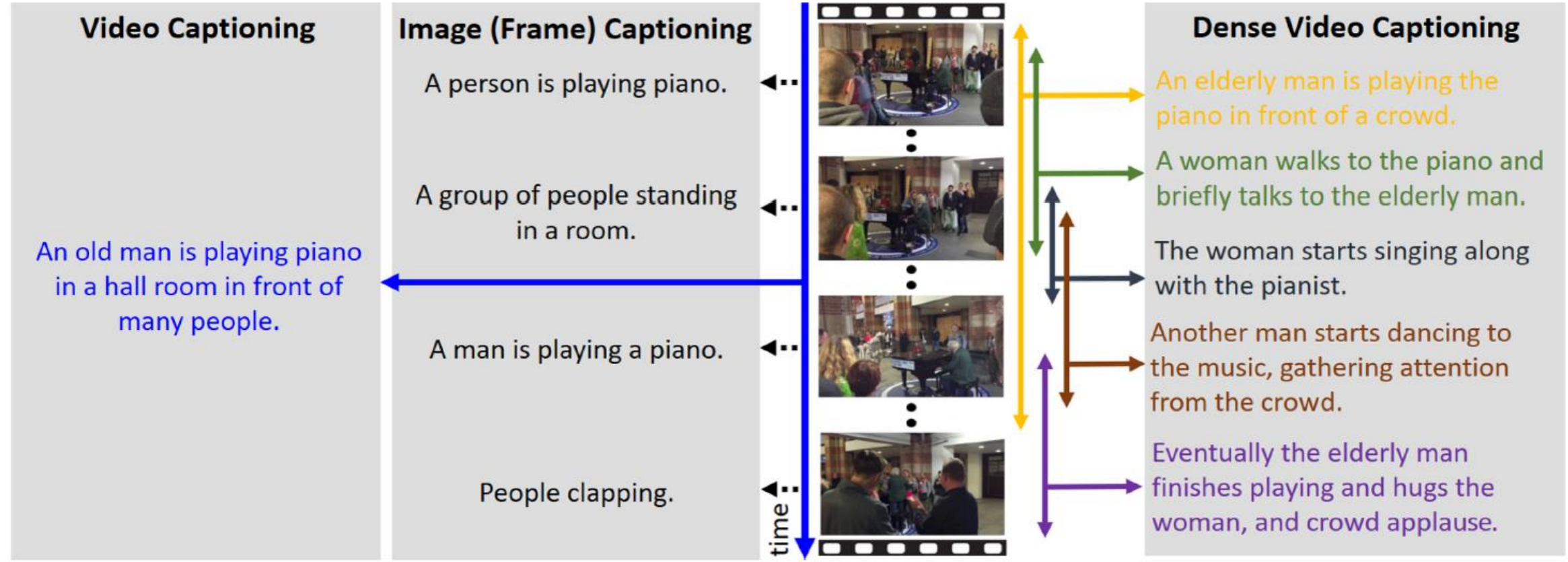




+Viterbi	This is a picture of a family. This is a picture of a cake. This is a picture of a dog. This is a picture of a beach. This is a picture of a beach.
+Greedy	The family gathered together for a meal. The food was delicious. The dog was excited to be there. The dog was enjoying the water. The dog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

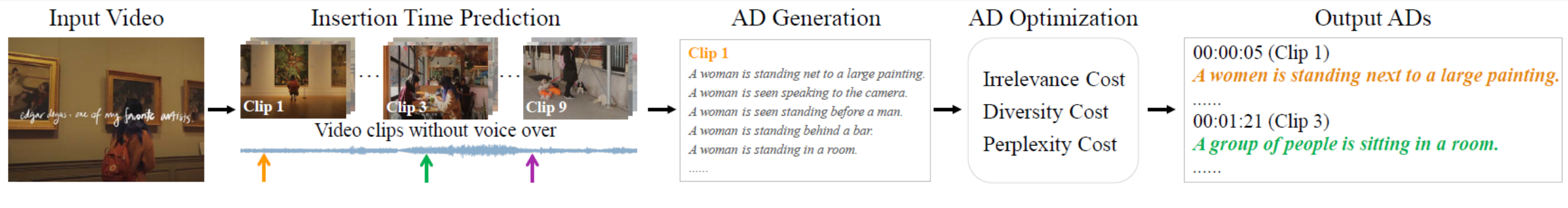
Huang et al., 2016

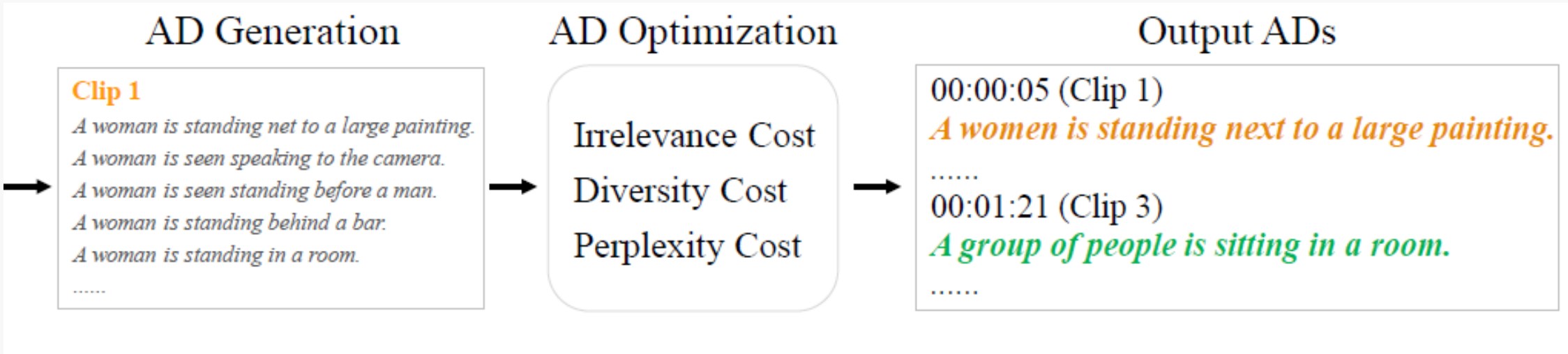
EXAMPLE: VIDEO CAPTIONING TYPES



Aafaq et al., 2019

EXAMPLE: AD GENERATION





Wang et al., 2021

R E C A P

Early AI Video Captioning incl. refinements (Pre-Transformer Era)

- **Focus on Computer Vision & Machine Learning:**
  - Methods relied on **neural networks** and **machine learning** for character/object recognition
  - **Action recognition** developed through video data sets to detect **interactions and movement** in video scenes; enabling cross-frame captioning
- **Challenges:**
  - Lacked **narrative coherence**—could recognize objects but not their **storytelling significance**
  - Limited **contextual awareness**—struggled with understanding events **across multiple frames**
  - Format **not ready for AD**—no gap detection in dialogue/audio track
    - **Exception: Wang et al. (2021), but other challenges remain**

FROM AI VIDEO CAPTIONS TO AI AUDIO DESCRIPTION

Rise of Transformer Models & Advanced Language Processing

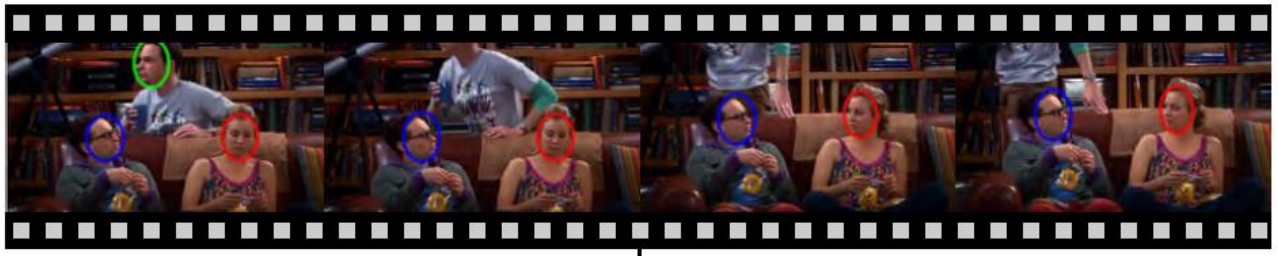
- **Transformers, pre-trained through self-supervised learning** (i.e. without manual data annotation), facilitate AD generation:
  - **Vision-Language models (VLMs)** map visual input to text representations; model **temporal relationships** between frames and allow **cross-frame visual analysis** for **narrative continuity**
  - **Large Language Models (LLMs)** generate **relatively fluent and cohesive text descriptions**
- **Impact:**
  - Improved **narrative flow**—potentially overcoming problems with **cohesion and coherence**
  - Enhanced **accuracy**—more comprehensive foundation models cover **wide range of domains**

RECENT MODELS FOR AI AUDIO DESCRIPTION

- **Zero-Shot AD generation:**
  - **Off-the-shelf pre-trained VLMs and LLMs** employed with **prompting strategies** seem to allow for AD generation without further training or fine-tuning
- **Two-stage approaches** (e.g., Chu et al., 2024; Xie et al., 2024)
  - Stage 1: a VLM produces comprehensive/dense video description (characters, actions, interactions, locations, objects, etc)
  - Stage 2: an LLM summarises the stage-1 description into a single sentence, adjusting content and style according to specific prompts
  - For very short video clips (a few seconds)

EXAMPLE: TWO-STAGE MODEL

Stage I: VLM-based video description



Visual-Language Model (VLM) for videos

Input
(Multi-question prompt)

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.*

Output
(Dense video description)

*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.*

Stage II: LLM-based AD summary

Please summarise the following description into one sentence.
Description: {output from Video-VLLM}

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.'.*

Large Language Model (LLM)

Predicted AD: *"Leonard and Penny exchange stares"*

Xie et al., 2024

AI MODEL EVALUATION 1/3

BBC Bargain Hunt

BBC AD:

Girls playing lacrosse.

Oxford AutoAD Zero with VideoLlama2 (Xie et al., 2024):

Step 1: Main persons: Two women playing lacrosse. 2. Actions: They are running, trying to get the ball. 3. Person interactions: They are playing against each other. 4. Facial expressions: They are focused and determined.

Step 2: Samantha and Rachel run determinedly, trying to get the ball.

Our prompt, adapted from LLM-AD (Chu et al., 2024), with Gemini:

Step 1: On a green field, women in white and red uniforms play lacrosse, using their sticks to pass the ball around. A woman in white falls.

Step 2: On a green field, women play lacrosse and attempt to gain control of the ball.

Evaluation of AI Models for AD: with financial support from the Royal National Institute of Blind People, supported by the BBC; **Surrey research group**: Shenbin Qian, Yuan Zou, Constantin Orasan, Sabine Braun

RNIB

See differently

AI MODEL EVALUATION 2/3

BBC Bargain Hunt

BBC AD:

A giant wooden spoon.

Oxford AutoAD Zero with VideoLlama2 (Xie et al., 2024):

Step 1: 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.

Step 2: **Tom holds a frisbee and talks to Alex.**

Our prompt, adapted from LLM-AD (Chu et al., 2024), with Gemini:

Step 1: 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.

Step 2: **Under an orange tent at an outdoor market, two young men in red polo shirts are inspecting a wooden bowl and spoon.**

Evaluation of AI Models for AD: with financial support from the Royal National Institute of Blind People, supported by the BBC; **Surrey research group**: Shenbin Qian, Yuan Zou, Constantin Orasan, Sabine Braun

RNIB

See differently

AI MODEL EVALUATION 3/3

BBC Bargain Hunt

BBC AD:

Thomas looks closely at the watch through a loupe.

Oxford AutoAD Zero with VideoLlama2 (Xie et al., 2024):

Step 1: 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.

Step 2: **He adjusts his glasses.**

Our prompt, adapted from LLM-AD (Chu et al., 2024), with Gemini:

Step 1: 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.

Step 2: **A man in a light blue blazer inspects a small metallic object with a chain in front of an outdoor market scene.**

Evaluation of AI Models for AD: with financial support from the Royal National Institute of Blind People, supported by the BBC; **Surrey research group**: Shenbin Qian, Yuan Zou, Constantin Orasan, Sabine Braun

RNIB
See differently

R E C A P

Transformer-based zero-shot AD generation

- **Currently designed for very short clips**
- **Genres:**
  - **TV drama/movies vs. factual reality shows:** different levels of narrative cohesion; amount of time available
- **Improvements:**
  - Character identification/tracking, re-identification; cohesiveness/narrative flow; accuracy
  - Less time consuming
- **Challenges:**
  - **Accuracy improved but still varied**—meaning shifts still occur; unreliable
  - **Language generation**—over-described for time available
  - **Long-form video**—not clear whether scalable

EXAMPLE: LONGER VIDEO 1/2

Comprehensive video description
(here: Google Gemini 2.0 Flash)

[00:00:00] A boy, wearing a striped t-shirt and tan pants, stands to the right of the frame, opening a refrigerator. The shot looks from a kitchen table and chairs through an open doorway to a sitting room with windows.
[00:00:02] He puts a glass on the table.
[00:00:04] He retrieves a small bag.
[00:00:05] He looks at the bag.
[00:00:07] He picks up a small pair of scissors.
[00:00:08] He cuts the bag open.
[00:00:10] He spoons the contents into the glass.
[00:00:11] He stirs the contents.
[00:00:12] He walks towards the sitting room.
[00:00:13] He continues forward.
[00:00:16] His hand reaches to open the sitting room door.
[00:00:17] He walks out of view.
[00:00:19] Empty room view.
[00:00:22] Empty room view.
[00:00:29] Boy walks to the sitting room to the right.
[00:00:31] Boy moves forward and towards the right.
[00:00:33] Closer shot of the table.
[00:00:36] The camera is focused on a telephone on a table.
[00:00:37] The telephone remains in view.
[00:00:40] Boy is eating from a glass while standing in the living room.
[00:00:43] He looks towards something out of view.
[00:00:48] He continues to look.
[00:00:50] TV with news headlines.

EXAMPLE: LONGER VIDEO 2/2

Summary (Gemini 2.0 Flash)

A boy prepares a drink in the kitchen, then walks to the living room and sits, while a phone message plays. The message hints at a crisis, and as the boy looks towards the television, a breaking news headline is revealed.

AD with time codes / identification of gaps in audio track (Gemini 2.0 Flash)

[00:00:00 to 00:00:01] A boy in a striped t-shirt stands in a kitchen, reaching into a refrigerator.
[00:00:01] He places a glass on the table, and grabs a small bag.
[00:00:07] He retrieves scissors and opens the bag.
[00:00:09] He adds the contents to the glass, then stirs it.
[00:00:12] He walks towards a living room through a doorway.
[00:00:16] The door is opened.
[00:00:17 to 00:00:18] Message one.
[00:00:19 to 00:00:23] The kitchen is visible from the doorway to the living room.
[00:00:23 to 00:30] Are you there? Hello. If you're there darling, pick up.
[00:00:35 to 00:37] Focus shifts to a telephone on a desk.
[00:00:40] The boy stands in the living room, stirring the glass.
[00:00:47] He drinks from the glass, then looks towards something out of frame.
[00:00:51] A television displays a breaking news headline.

FURTHER RECENT DEVELOPMENTS

- **Integration of Multimodal & Contextual Information**
  - **Context from movie, subtitles, previous AD, pre-trained components** (AutoAD; Han et al., 2023)
  - **Attention mechanisms** to refine overall AD quality (DistinctAD; Fang et al., 2024)
  - **Additional modules** (MMAD; Ye et al., 2024), e.g. audio-aware module (audio cues), subtitle module (analysis, alignment with AD), actor-tracking module (improved character continuity)
  - Aim: to improve accuracy, continuity, overall quality of the generated AD
  - **Incremental** improvements or **transformational** changes?
- **Automation of other components of the AD workflow**
  - **Automatic gap detection for AD insertion** (commercial software; Wang et al., 2021 and others)
  - **Synthetic voicing AD scripts** (recently: RNIB, 2024)

LIMITATIONS & RISKS OF AI IN AD

AI-generated AD remains in an experimental stage

Models (still) struggle with contextual awareness, accuracy, nuances, narrative depth, coherence

Bias in AD descriptions—models sometimes reinforce stereotypes

Limited user engagement in development (as opposed to only evaluation)

Limited engagement with AD scholarship from the humanities/social sciences

Impact: Can undermine genuine accessibility and user trust

ALTERNATIVES TO (FULL) AUTOMATION

Alternatives to full automation:

- **User-driven & interactive AD systems,** allowing users to request AD on demand—enabled greater flexibility (Describe Now; Cheema et al., 2024)
- **Post-editing of AI-generated AD scripts** (MeMAD Flow platform; Braun et al., 2021; Campos et al., 2020)

Alternatives to automation:

- **Integrative AD** (Fryer, 2018; Romero Fresco, 2019)
- **Collaborative AD production through volunteer upskilling** (Natalie et al., 2021)

Not always possible

AI IN AD: WHERE DO WE GO FROM HERE

- Overall aim: **broadening media access by increasing availability of AD, but without sacrificing quality**
  - Building strong AI models as part of a comprehensive approach
  - Human-centric principles for AI essential for ethical, high-quality AD
  - Continued commitment to humanities/social science-based AD research
- **Future Directions:**
  - Hybrid AI-Human Models—**integrating human storytelling principles** into AI (prompting, interacting/'reasoning' with AI, explainable AI models)
  - More diverse datasets—ensuring AI understands **different genres, emotions, and cultural contexts**
  - Real-time user feedback loops—allowing AI to **improve through direct user interaction**
  - Stronger evaluation frameworks—**unified approach** regardless of how AD was created

UNIVERSITY OF
SURREY