

# Watching videos without vision: Challenges, techniques, and the future of video accessibility

Candidacy Exam  
December 13, 2022

Presented by: Gaurav Jain  
Dept. of Computer Science  
Columbia University

# Outline

## 1 Challenge in video accessibility.

What is the core problem in making videos accessible to blind people?

## 2 Existing techniques.

How do systems facilitate video accessibility?

## 3 Future work.

What are opportunities for future work in video accessibility?

# Outline

## 1 Challenge in video accessibility.

What is the core problem in making videos accessible to blind people?

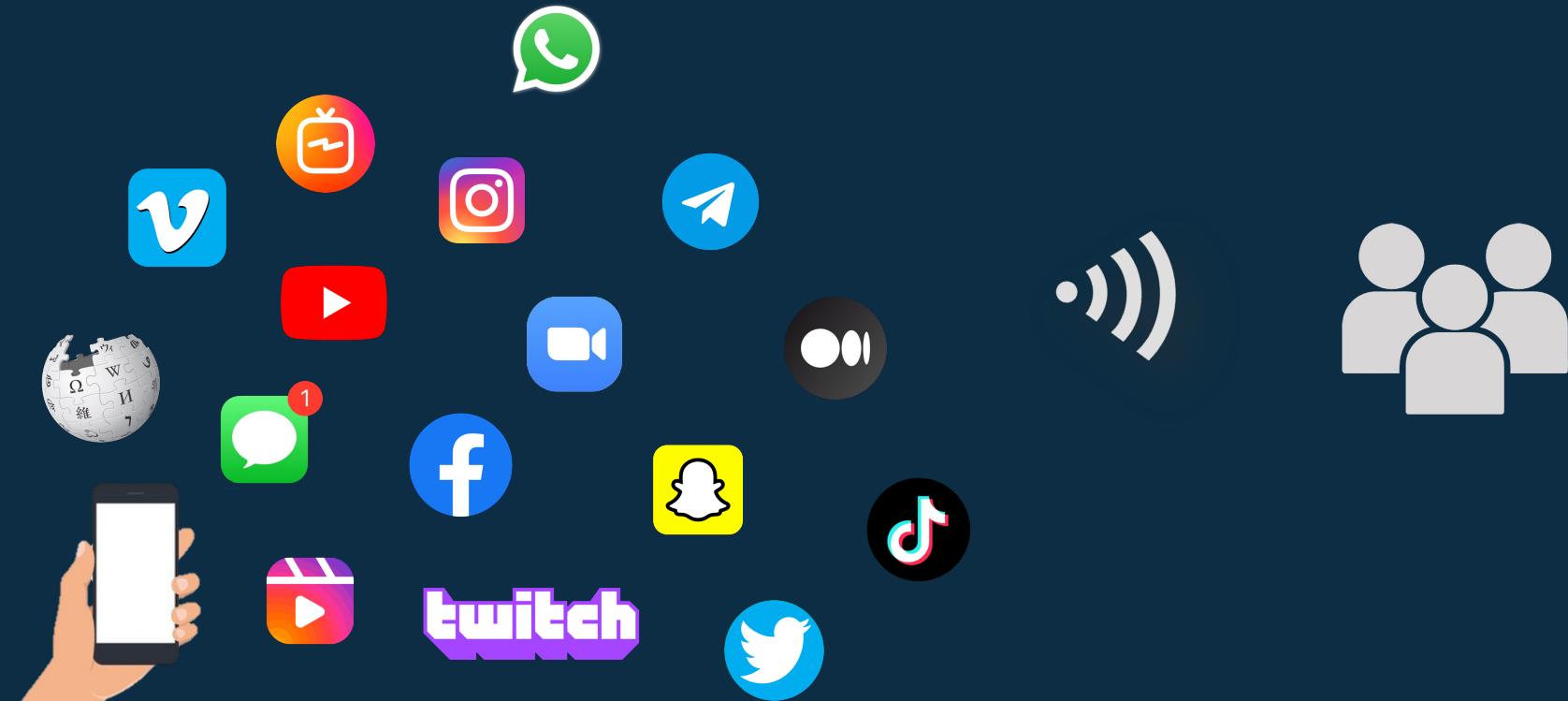
## 2 Existing techniques.

How do systems facilitate video accessibility?

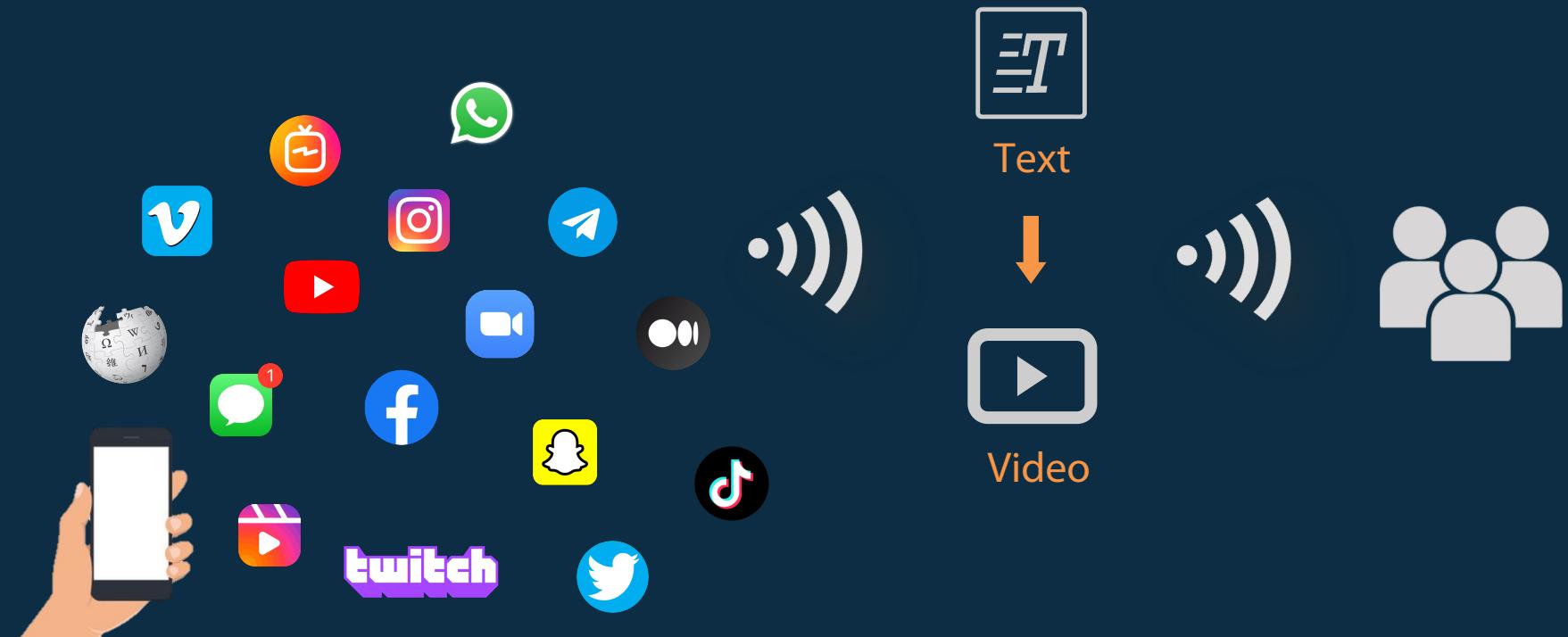
## 3 Future work.

What are opportunities for future work in video accessibility?

# Digital media



# Increasingly, information is shared as videos



# Increasingly, information is shared as videos

## Instructions

Text

Jollof Rice  
Recipe From Tunde Wey  
Adapted by The New York Times

Time 1 hour  
Rating 5 ★★★★★ (983)  
Notes Read 93 community notes ↗  
Featured in Nigerian Food, Crash-and-Burn Style  
Learn How to Make Rice

Save Give Print

Source: NY Times

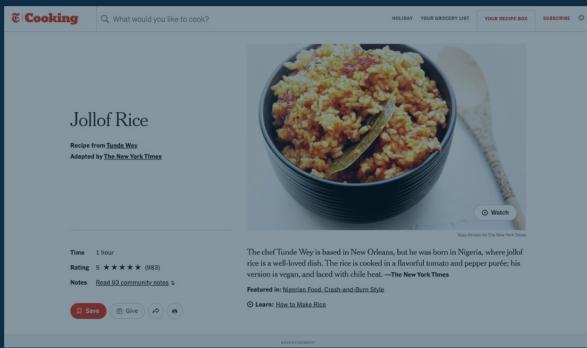
## Demonstrations



Credits: NYT Cooking | YouTube

# Increasingly, information is shared as videos

## Instructions



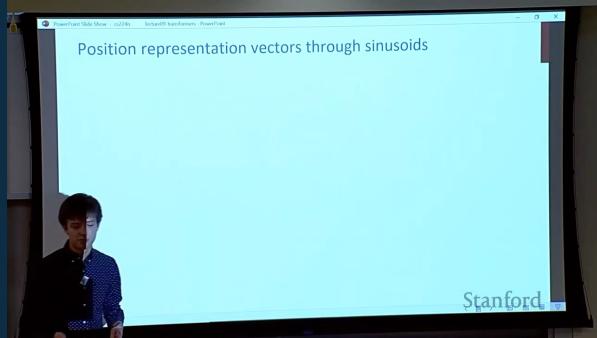
## Articles

A screenshot of an article titled "A Beginner's Guide to ChatGPT: Understanding What it Is, Why it Matters, and When/Where to Use It" by OpenAI. The article features a prominent OpenAI logo at the top. Below the logo, the text explains what ChatGPT is, how it uses GPT-3, and its potential applications. The source is cited as Medium.

## Demonstrations



## Tutorials



Text

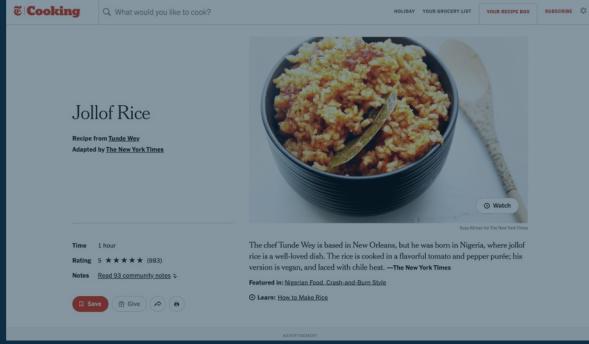
Video

Credits: NYT Cooking | YouTube

Credits: Stanford | YouTube

# Increasingly, information is shared as videos

## Instructions



## Demonstrations



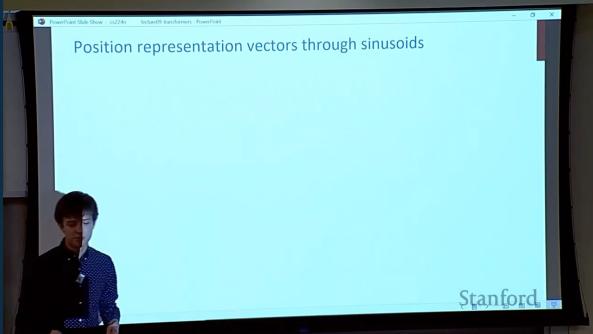
Credits: NYT Cooking | YouTube

## Articles

A screenshot of a Medium article titled "A Beginner's Guide to ChatGPT: Understanding What It Is, Why It Matters, and When/Where To Use It". The article features a purple header with the OpenAI logo. The main content discusses the basics of ChatGPT, its capabilities, and how it can be used. It includes several sections of text and a sidebar with user profiles and a comment section.



## Tutorials



Credits: Stanford | YouTube

## Blogs

A screenshot of a blog post titled "Top 5 Things to See & Do in Amsterdam". The post features a large image of a canal scene in Amsterdam. The text discusses five things to do in the city, including visiting the Van Gogh Museum and the Anne Frank House. It includes a sidebar with user profiles and a comment section.



Source: Nomadicmatt.com



Credits: Max Nomad | YouTube

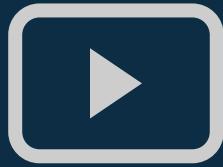
Text

Video



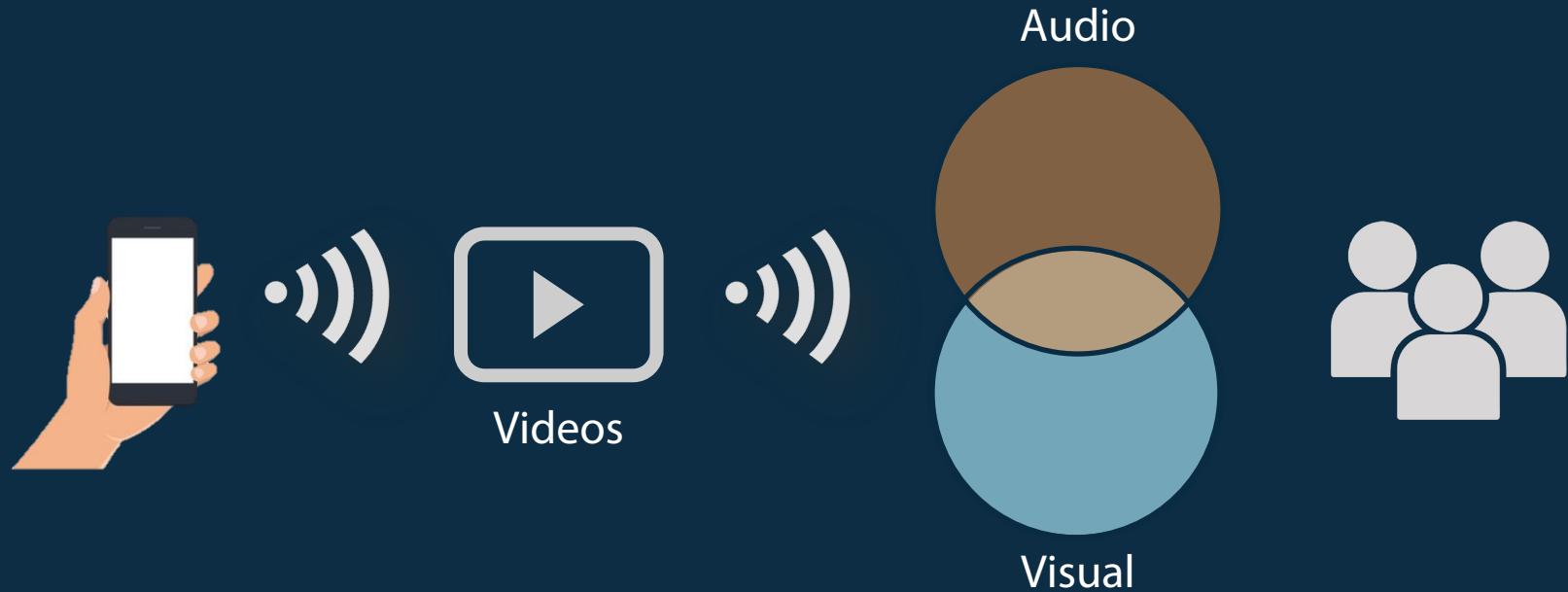
Text





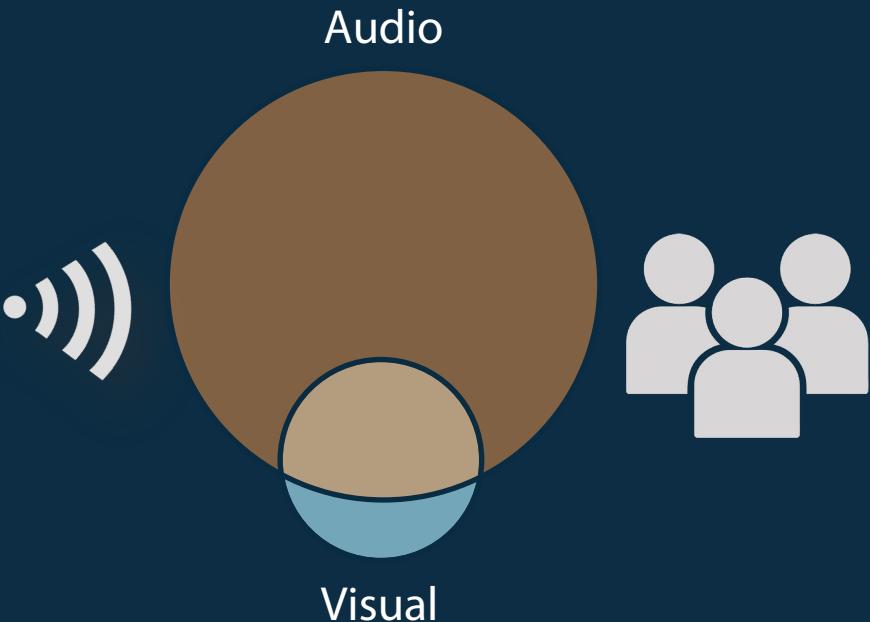
Videos







Credits: Ali Abdali | YouTube

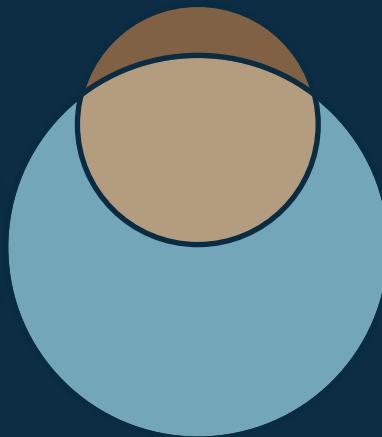




Credits: Yummy Treats | YouTube

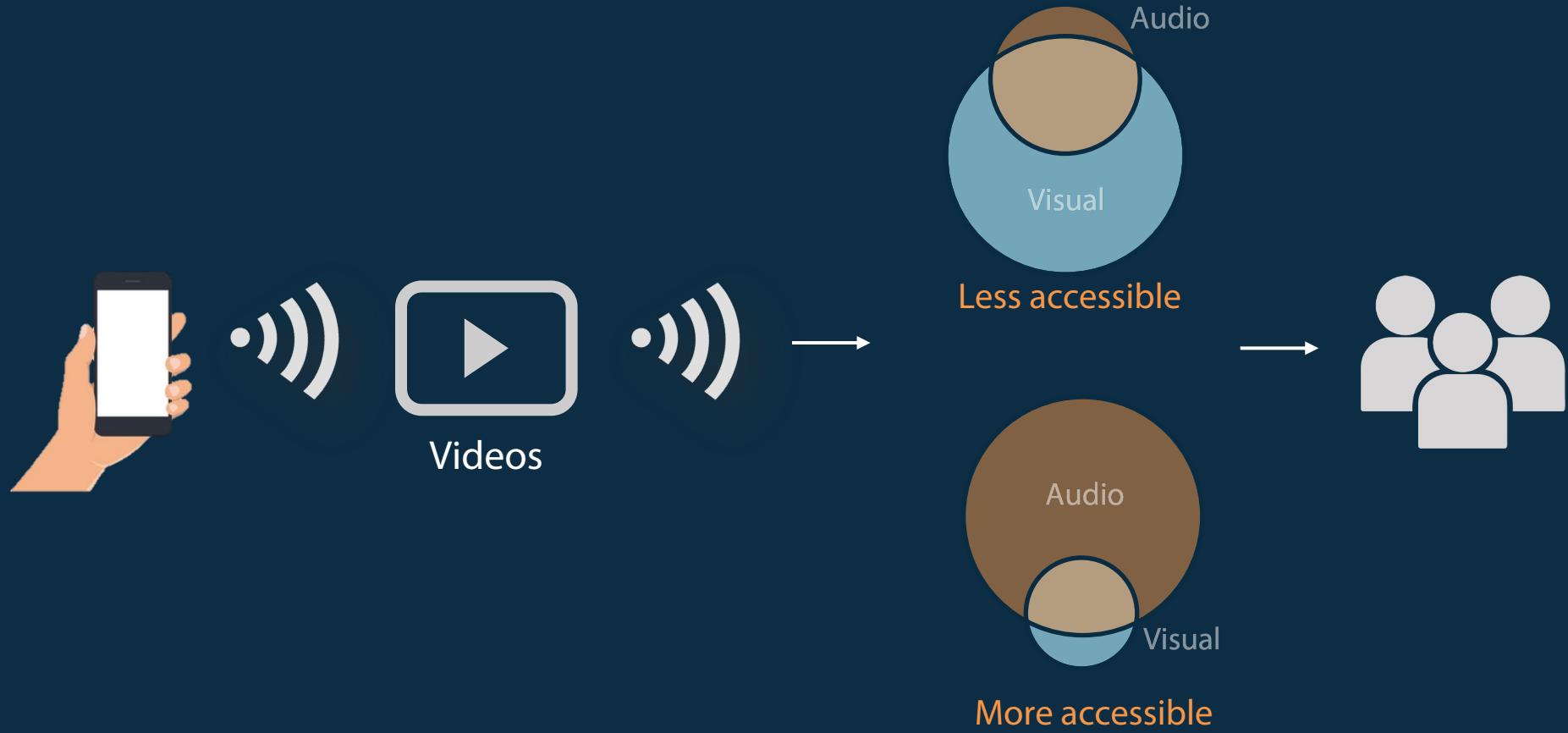


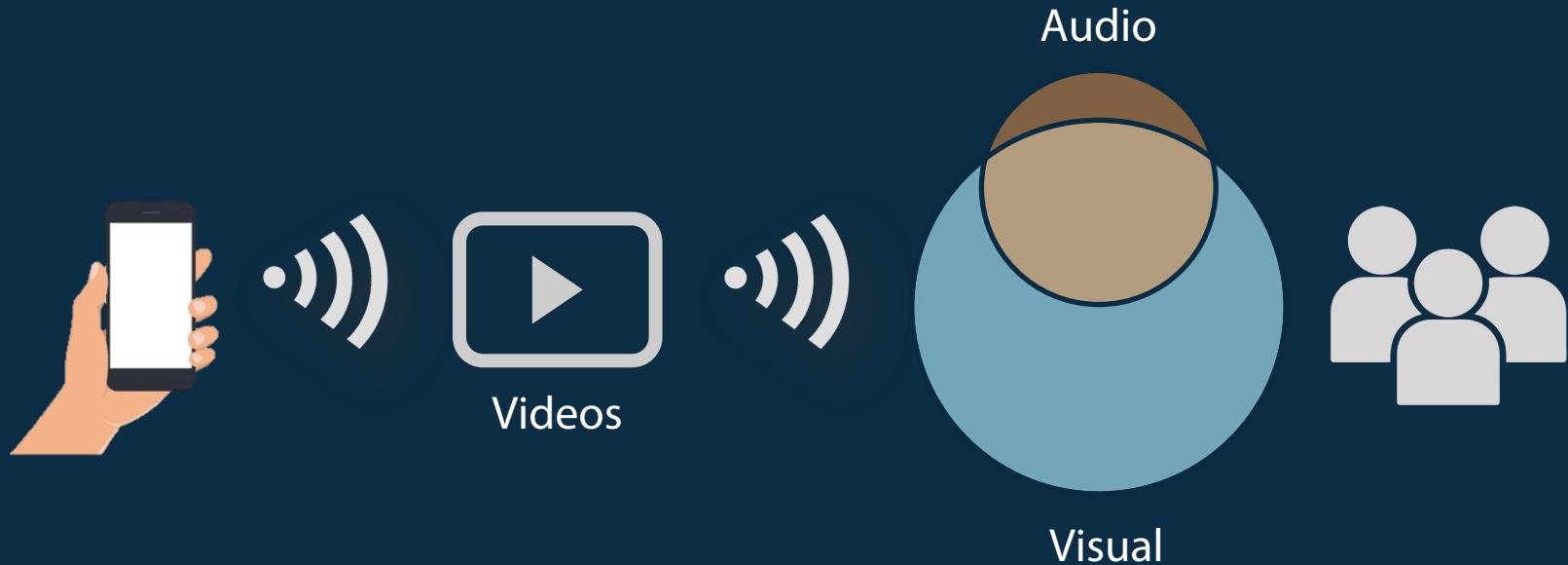
Audio



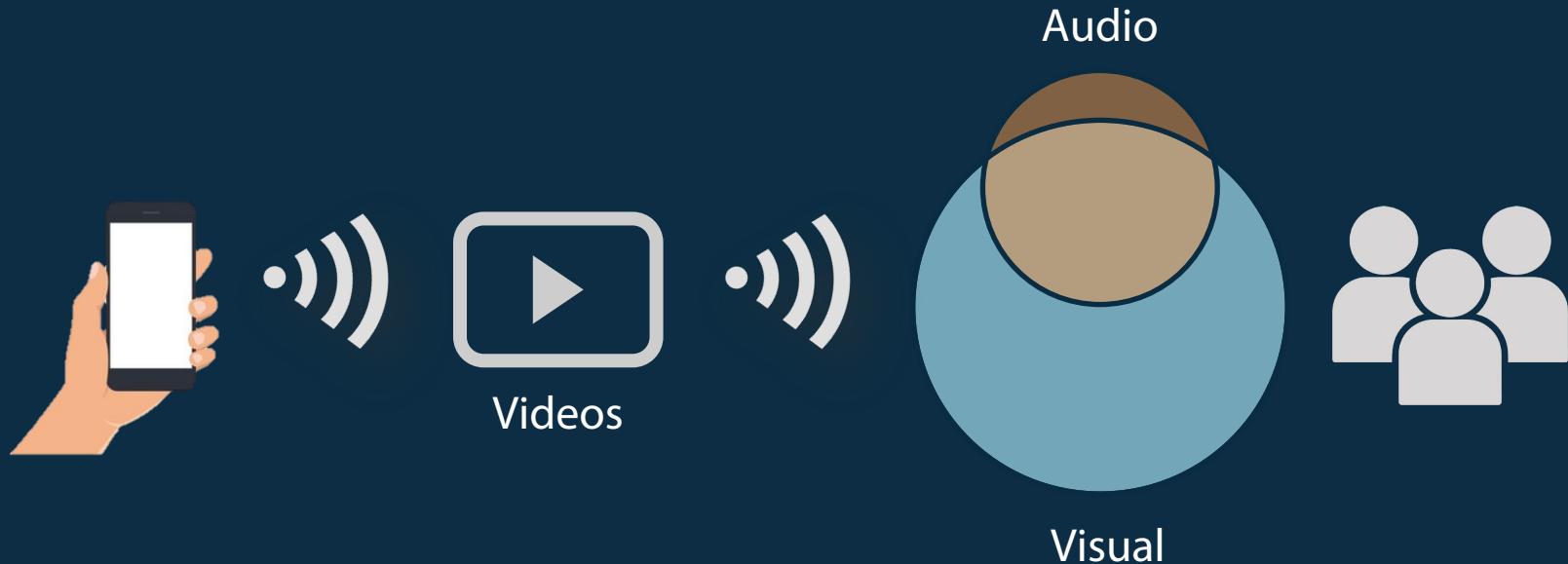
Visual



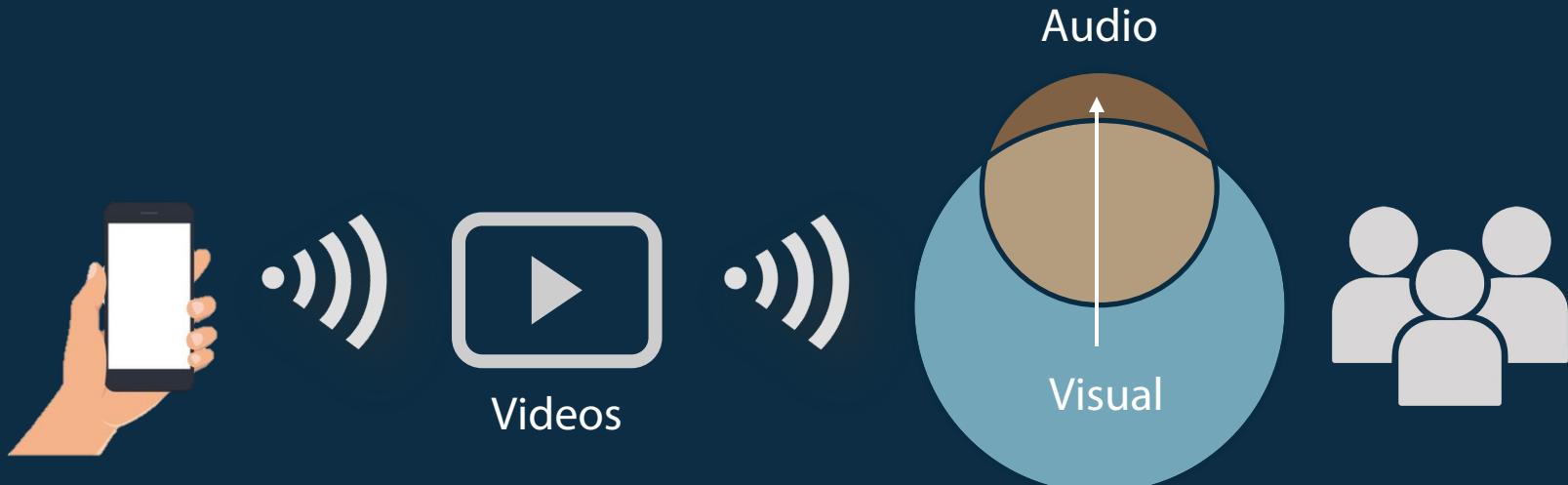




# Audio descriptions



# Audio descriptions



# Frozen

(audio described)



**Audio describing  
videos is complex**

**Videos are produced  
on a massive scale**

**60 hours**  
are needed to describe a  
two hour video.  
(Lakritz & Salway, 2006)



**720K hours**  
of video content is  
uploaded every day.  
(YouTube, 2021)

Audio describing  
videos is complex

60 hours  
are needed to describe a  
two hour video.  
(Lakritz & Salway, 2006)

Videos are produced  
on a massive scale



720K hours  
of video content is  
uploaded every day.  
(YouTube, 2021)

Core challenge:

Scaling audio descriptions to the  
massive video generation rates.

# Outline

## 1 Challenge in video accessibility.

Scaling audio descriptions to the massive video generation rates.

## 2 Existing techniques.

How do systems facilitate video accessibility?

## 3 Future work.

What are opportunities for future work in video accessibility?

# Outline

## 1 Challenge in video accessibility.

Scaling audio descriptions to the massive video generation rates.

## 2 Existing techniques.

How do systems facilitate video accessibility?

## 3 Future work.

What are opportunities for future work in video accessibility?

Existing techniques.

# The process of generating audio description.

#1.  
Identify  
a11y issues

Watch through the video to identify inaccessible video segments.

Existing techniques.

# The process of generating audio description.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions

Watch through the video to identify inaccessible video segments.

Describe the relevant visuals concisely.

Existing techniques.

# The process of generating audio description.



Watch through the video to identify inaccessible video segments.

Describe the relevant visuals concisely.

Record and fit the descriptions into the appropriate parts of the video.

## Existing techniques.

#1.  
**Identify  
a11y issues**

Watch through the video to identify inaccessible video segments.



#2.  
**Write  
descriptions**



#3.  
**Insert  
descriptions**

## Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

## Gaps in speech



Gagnon 2010

## Existing techniques.

#1.  
Identify  
a11y issues

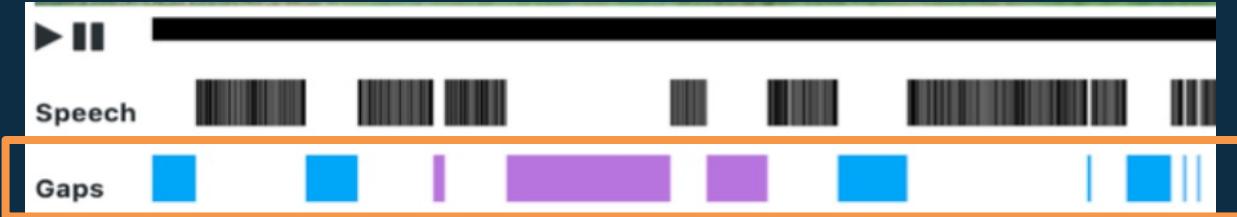


#2.  
Write  
descriptions



#3.  
Insert  
descriptions

### Gaps in speech



Gagnon 2010

- + simple
- + easy to fit descriptions

- inaccurate

## Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions

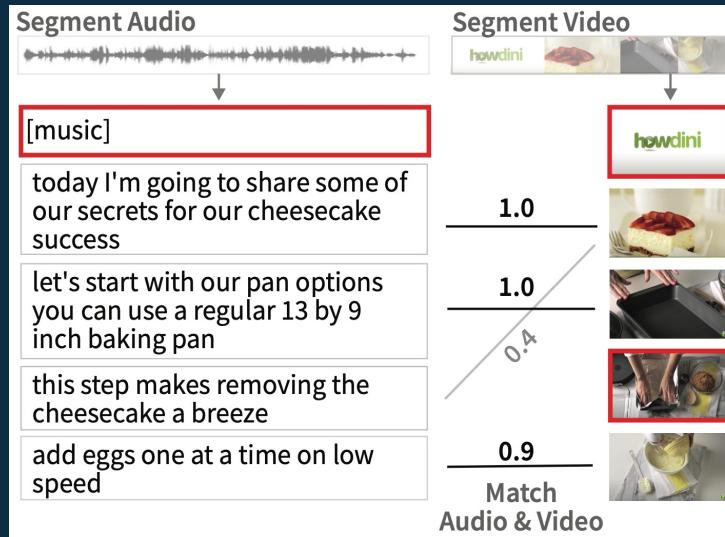


#3.  
Insert  
descriptions

### Gaps in speech



### CrossA11y



# Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions

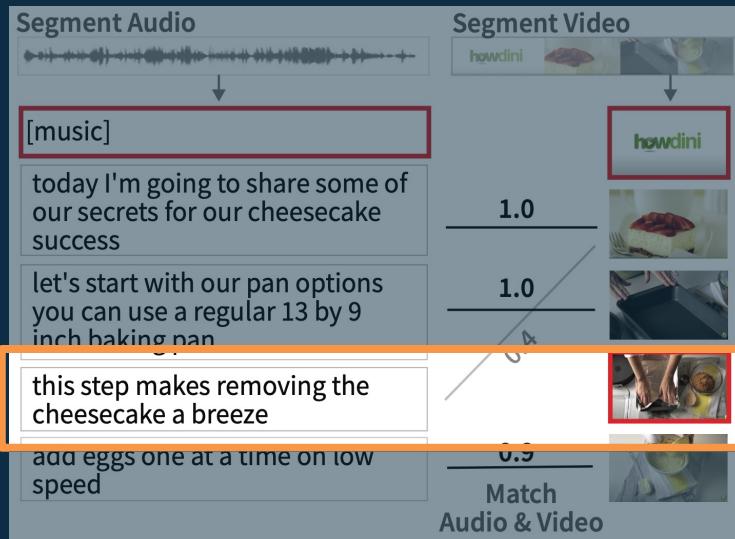


#3.  
Insert  
descriptions

## Gaps in speech



## CrossA11y



# Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



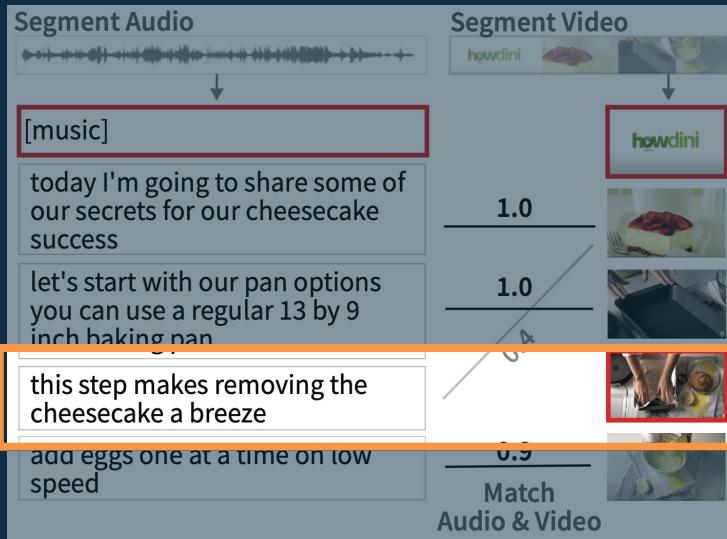
#3.  
Insert  
descriptions

## Gaps in speech

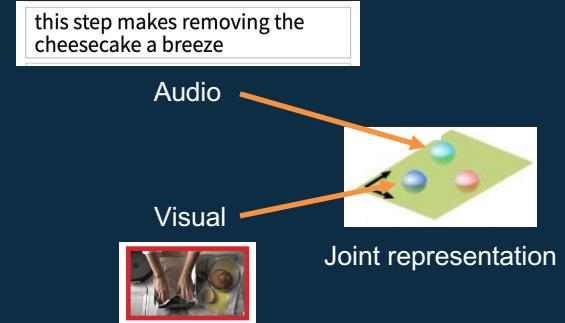


Gagnon 2010

## CrossA11y



Liu 2022



Alayrac 2020

# Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions

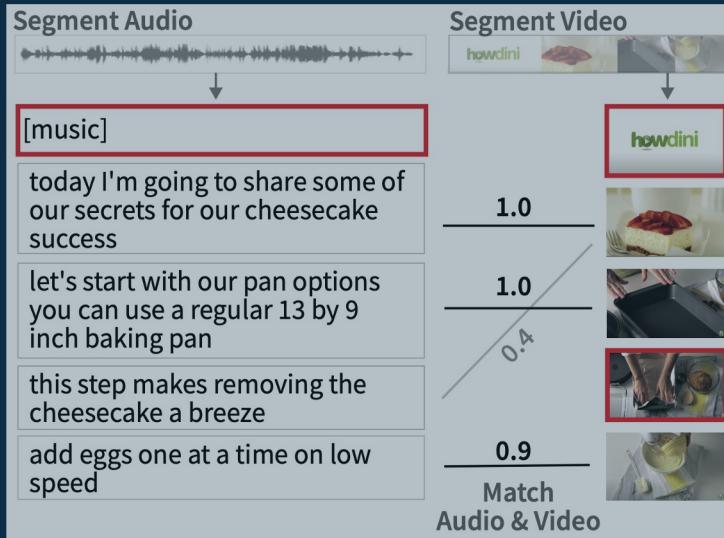


#3.  
Insert  
descriptions

## Gaps in speech



## CrossA11y



	Random	Gaps	CrossA11y
Precision	0.275	0.833	<b>0.694</b>
Recall	0.390	0.385	<b>0.984</b>
F1	0.323	0.526	<b>0.814</b>

+ accurate

- computationally expensive  
- Insertion is complex

## Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions

Describe the relevant visuals  
concisely.



#3.  
Insert  
descriptions

## Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions



Video

Automated tools.



Audio  
descriptions

Authoring support tools.



Author

## Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

## Automated tools.



Video



Audio  
descriptions

# Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions



Video



Visuals

+



[trumpet playing]

Sound

+

Sam: "I went for lunch today."

Spoken dialogues



Audio  
descriptions

+

Script, Title, etc.

Meta-data

## Existing techniques.

### Using movie script.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions



Visuals

+



[trumpet playing]

Sound

+

Sam: "I went for lunch today."

Spoken dialogues

+

Script, Title, etc.

Meta-data



## Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

## Using movie script.

SCRIPT	AD SCRIPT
<p><b>Recording, leaving the door.</b></p> <p>RECORDING (VS - cont)</p> <p>You always say this is foolish, that this is typical. But I want to be just like everyone else. Is it so hard to understand? Is it so hard to understand that I just want to have a life...</p>	<p>11 00:02:38,500 --&gt; 00:02:40,300 Recording, leaving the door.</p>
<p><b>Now we see the entire setting from the outside: it is an old trailer with worn and rusted walls. The door of the trailer is open. Above is a sign that reads: GREAT AMERICAN CIRCUS</b></p>	<p>12 00:02:54,000 --&gt; 00:02:57,498 The door of the trailer is open.</p>

## Existing techniques.

#1.  
Identify  
a11y issues

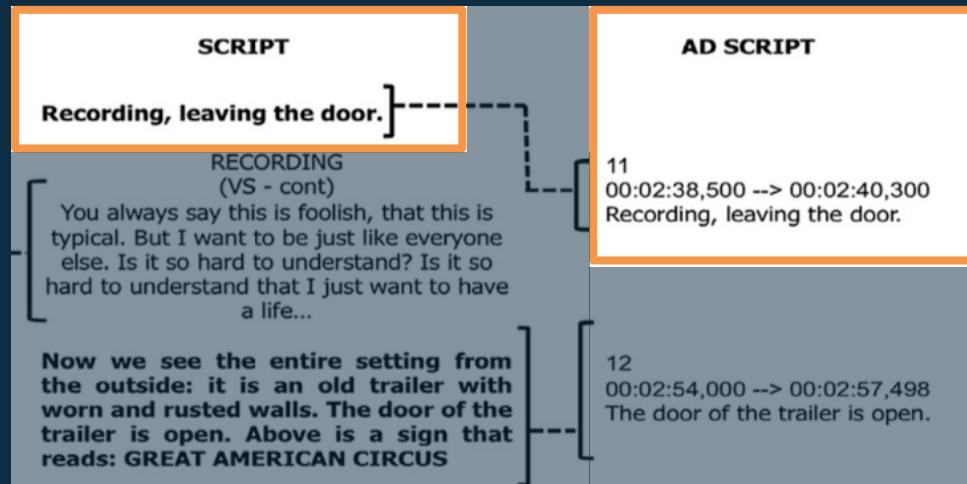


#2.  
Write  
descriptions



#3.  
Insert  
descriptions

## Using movie script.



- + better than no AD
- + improved comprehension

- poor quality & consistency
- limited applicability

## Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

## Using audio and visuals.



Visuals

+



[trumpet playing]

Sound

+

Sam: "I went for lunch today."

Dialogues

+

Script, Title, etc.  
Meta-data



## Existing techniques.

#1.  
Identify  
a11y issues

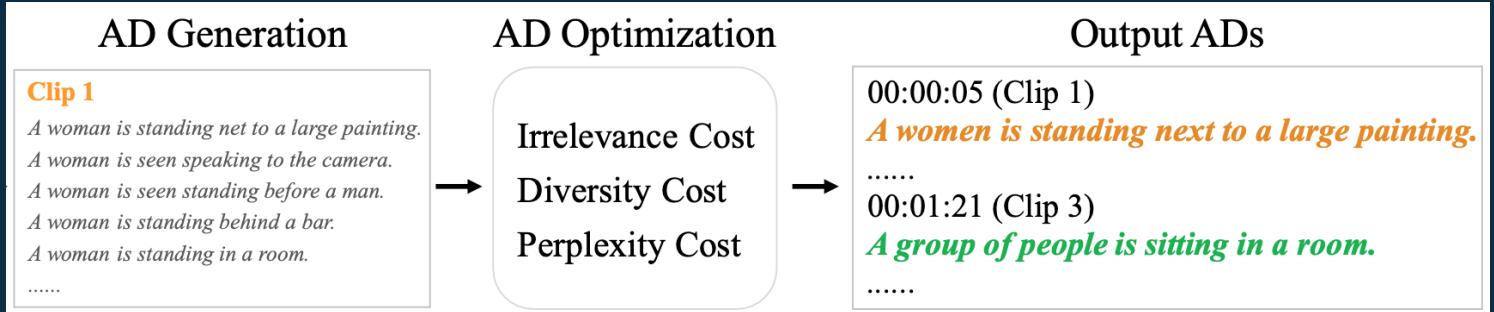


#2.  
Write  
descriptions



#3.  
Insert  
descriptions

## Using audio and visuals.



## Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

## Using audio and visuals.

### AD Generation

#### Clip 1

*A woman is standing next to a large painting.  
A woman is seen speaking to the camera.  
A woman is seen standing before a man.  
A woman is standing behind a bar.  
A woman is standing in a room.  
.....*

### AD Optimization

Irrelevance Cost  
Diversity Cost  
Perplexity Cost

### Output ADs

00:00:05 (Clip 1)

*A women is standing next to a large painting.*

.....

00:01:21 (Clip 3)

*A group of people is sitting in a room.*

.....

Event descriptions using  
dense video captioning

## Existing techniques.

### Using audio and visuals.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

#### AD Generation

##### Clip 1

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

#### AD Optimization

Irrelevance Cost  
Diversity Cost  
Perplexity Cost

#### Output ADs

00:00:05 (Clip 1)

*A women is standing next to a large painting.*

.....

00:01:21 (Clip 3)

*A group of people is sitting in a room.*

.....

Event descriptions using  
dense video captioning

Optimization using  
dynamic programming

# Existing techniques.

## Using audio and visuals.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

### AD Generation

#### Clip 1

*A woman is standing next to a large painting.  
A woman is seen speaking to the camera.  
A woman is seen standing before a man.  
A woman is standing behind a bar.  
A woman is standing in a room.  
.....*

### AD Optimization

Irrelevance Cost  
Diversity Cost  
Perplexity Cost



### Output ADS

00:00:05 (Clip 1)

*A women is standing next to a large painting.*

.....

00:01:21 (Clip 3)

*A group of people is sitting in a room.*

.....

Event descriptions using  
dense video captioning

Optimization using  
dynamic programming

## Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

## Using audio and visuals.

### AD Generation

#### Clip 1

*A woman is standing next to a large painting.  
A woman is seen speaking to the camera.  
A woman is seen standing before a man.  
A woman is standing behind a bar.  
A woman is standing in a room.  
.....*

### AD Optimization

Irrelevance Cost  
Diversity Cost  
Perplexity Cost



### Output ADS

00:00:05 (Clip 1)  
*A women is standing next to a large painting.*  
.....  
00:01:21 (Clip 3)  
*A group of people is sitting in a room.*  
.....

Event descriptions using  
dense video captioning

Optimization using  
dynamic programming

+ better than no AD

- inaccurate & confusing

## Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions



Video

Automated tools.



Audio  
descriptions

Authoring support tools.



Author

# Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

Fried Potato Balls

Scene 0 0:0.0 to 0:5.8 minutes [Replay Scene](#)

Description a close up of food

Text On-Screen Potato balls

Scene 1 0:5.8 to 0:21.4 minutes [Replay Scene](#)

Description a close up of a metal bowl

Text On-Screen put potatoes in water and boil them after about 30 minutes, when they are cooked (Check it with a knife).

Scene 2 0:21.4 to 0:27.6 minutes [Replay Scene](#)

Description a close up of a bowl

Text On-Screen then put them into a bowl peeled and chopped

Scene 3 0:27.6 to 0:37.4 minutes [Replay Scene](#)

Description and softened butter...

Text On-Screen ...a pinch of salt...  
...and black pepper.

# Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

AI-generated  
descriptions



Fried Potato Balls

Scene	Time Range	Description	Text On-Screen	Action
Scene 0	0:0.0 to 0:5.8 minutes	a close up of food	Potato balls	<a href="#">Replay Scene</a>
Scene 1	0:5.8 to 0:21.4 minutes	a close up of a metal bowl	put potatoes in water and boil them after about 30 minutes, when they are cooked (Check it with a knife)'	<a href="#">Replay Scene</a>
Scene 2	0:21.4 to 0:27.6 minutes	a close up of a bowl	then put them into a bowl peeled and chopped	<a href="#">Replay Scene</a>
Scene 3	0:27.6 to 0:37.4 minutes		and softened butter... ...a pinch of salt... ...and black pepper.	<a href="#">Replay Scene</a>

# Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

AI-generated  
descriptions



Optical  
character  
recognition

Fried Potato Balls

Scene 0 0:0.0 to 0:5.8 minutes Replay Scene

Description a close up of food

Text On-Screen Potato balls

Scene 1 0:5.8 to 0:21.4 minutes Replay Scene

Description a close up of a metal bowl

Text On-Screen put potatoes in water and boil them after about 30 minutes, when they are cooked (Check it with a knife)'.

Scene 2 0:21.4 to 0:27.6 minutes Replay Scene

Description a close up of a bowl

Text On-Screen then put them into a bowl peeled and chopped

Scene 3 0:27.6 to 0:37.4 minutes Replay Scene

Description

Text On-Screen and softened butter...  
...a pinch of salt...  
...and black pepper.

# Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

Scene	Time Range	Description	Text On-Screen	Action
Scene 0	0:0.0 to 0:5.8 minutes	a close up of food	Potato balls	Replay Scene
Scene 1	0:5.8 to 0:21.4 minutes	a close up of a metal bowl	put potatoes in water and boil them after about 30 minutes, when they are cooked (Check it with a knife)'.	Replay Scene
Scene 2	0:21.4 to 0:27.6 minutes	a close up of a bowl	then put them into a bowl peeled and chopped	Replay Scene
Scene 3	0:27.6 to 0:37.4 minutes		and softened butter... ...a pinch of salt... ...and black pepper.	Replay Scene

- + reduced time & effort
- + marginally improved description quality

- inaccurate automated suggestions

AI-generated  
descriptions

+

Optical  
character  
recognition

## Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

# Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

Welcome to ViScene  
A collaborative Audio Description Authoring Tool

The interface includes a video player showing a woman walking on a street, a timeline from 0:24 to 0:46, and a sidebar for selecting participants and entering feedback. Annotations with labels a through g point to various UI elements: a points to the participant selection dropdown; b points to the feedback text area; c points to the timeline; d points to the video controls; e.1 points to the closed caption text area; and e.2 points to a specific closed caption entry: "[SD] A woman with a purple scarf is looking at the road sign and taking her phone from her pocket". The feedback for this entry suggests: "Explain what kind of expression she made when she looked at the sign." Other feedback entries include: "[CC] Good design means sufficient contrast between foreground background and colors. That's not just text and images but links, icons, and buttons.", "[SD] The button with white background becomes clearer. The girl then continue her way to the destination.", and "[CC] If it's important enough to be seen, then it".

Time	Closed Caption	Scene Description	Feedback
0:24	[SD] A woman with a purple scarf is looking at the road sign and taking her phone from her pocket	interacting a real pain	Explain what kind of expression she made when she looked at the sign.
0:32	[CC] Good design means sufficient contrast between foreground background and colors. That's not just text and images but links, icons, and buttons.		
0:41	[SD] The button with white background becomes clearer. The girl then continue her way to the destination.		This is too long. Please shorten it.
0:46	[CC] If it's important enough to be seen, then it		

# Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

Time	Closed Caption	Scene Description	Feedback
0:24	interacting a real pain	[SD] A woman with a purple scarf is looking at the road sign and taking her phone from her pocket	Explain what kind of expression she made when she looked at the sign.
0:32	[CC] Good design means sufficient contrast between foreground background and colors. That's not just text and images but links, icons, and buttons.		
0:41		[SD] The button with white background becomes clearer. The girl then continue her way to the destination.	This is too long. Please shorten it.



Person 1:

Novice author

# Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

Time	Closed Caption	Scene Description	Feedback
0:24	interacting a real pain	[SD] A woman with a purple scarf is looking at the road sign and taking her phone from her pocket	e.2 Explain what kind of expression she made when she looked at the sign.
0:32	[CC] Good design means sufficient contrast between foreground background and colors. That's not just text and images but links, icons, and buttons.		
0:41		[SD] The button with white background becomes clearer. The girl then continue her way to the destination.	This is too long. Please shorten it.



Person 1:  
Novice author



Person 2:  
Reviewer

# Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

Time	Closed Caption	Scene Description	Feedback
0:24	interacting a real pain	[SD] A woman with a purple scarf is looking at the road sign and taking her phone from her pocket	e.2 Explain what kind of expression she made when she looked at the sign.
0:32	[CC] Good design means sufficient contrast between foreground background and colors. That's not just text and images but links, icons, and buttons.		
0:41		[SD] The button with white background becomes clearer. The girl then continue her way to the destination.	This is too long. Please shorten it.



**Person 1:**  
Novice author



**Person 2:**  
Reviewer

+ feedback improved descriptions  
+ cheaper than professional

- time consuming  
- requires two people

# Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

Time	Closed Caption	Scene Description	Feedback
0:24	interacting a real pain	[SD] A woman with a purple scarf is looking at the road sign and taking her phone from her pocket	Explain what kind of expression she made when she looked at the sign.
0:32	[CC] Good design means sufficient contrast between foreground background and colors. That's not just text and images but links, icons, and buttons.		
0:41		[SD] The button with white background becomes clearer. The girl then continue her way to the destination.	This is too long. Please shorten it.



Themes	Sub-themes
<b>Quality</b>	Descriptive Objective Succinct Learning Sufficient Interest Clarity Accurate Referable
<b>Speech Act</b>	Instructions Question Warning Compliment
<b>Required Action</b>	Revision Add information Fix grammar
<b>Guidance</b>	Suggestion Example Clarification



Person 1:  
Novice author



Person 2:  
Reviewer

# Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

Time	Closed Caption	Scene Description	Feedback
0:24	interacting a real pain	[SD] A woman with a purple scarf is looking at the road sign and taking her phone from her pocket	Explain what kind of expression she made when she looked at the sign.
0:32	[CC] Good design means sufficient contrast between foreground background and colors. That's not just text and images but links, icons, and buttons.		
0:41		[SD] The button with white background becomes clearer. The girl then continue her way to the destination.	This is too long. Please shorten it.



Themes	Sub-themes
<b>Quality</b>	Descriptive Objective Succinct Learning Sufficient Interest Clarity Accurate Referable
<b>Speech Act</b>	Instructions Question Warning Compliment
<b>Required Action</b>	Revision Add information <b>Fix grammar</b>
<b>Guidance</b>	Suggestion Example <b>Clarification</b>



Person 1:  
Novice author



Person 2:  
Reviewer

# Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

Time	Closed Caption	Scene Description	Feedback
0:24	interacting a real pain	[SD] A woman with a purple scarf is looking at the road sign and taking her phone from her pocket	e.2 Explain what kind of expression she made when she looked at the sign.
0:32	[CC] Good design means sufficient contrast between foreground background and colors. That's not just text and images but links, icons, and buttons.		
0:41		[SD] The button with white background becomes clearer. The girl then continue her way to the destination.	This is too long. Please shorten it.



Themes	Sub-themes
<b>Quality</b>	Descriptive Objective Succinct Learning Sufficient Interest Clarity Accurate Referable
<b>Speech Act</b>	Instructions Question Warning Compliment
<b>Required Action</b>	Revision Add information <b>Fix grammar</b>
<b>Guidance</b>	Suggestion Example <b>Clarification</b>



**Person 1:**  
Novice author

**Person 2:**

Automated  
feedback!

## Existing techniques.

### Including blind users

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

## Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

# Including blind users ... request descriptions

The screenshot shows the YouDescribe platform interface. At the top, there's a search bar with the word "tennis" and a "Search" button. To the right are links for "RECENT DESCRIPTIONS", "WISH LIST", and "SUPPORT", along with a user profile icon. Below the header, a blue bar says "DESCRIBED VIDEOS". Underneath are three video thumbnails:

- Learning to Play Tennis** by Buttered Side Down (4:42)
- Action Audio | Tennis Australia** by AKQA (2:13)
- Top Table Tennis Points of 2021 | Best...** by World Table Tennis (5:19)

Below this is another blue bar labeled "NON-DESCRIBED VIDEOS". Underneath are five video thumbnails:

- RAFA NADAL: 2022 ATP Highlight Reel** by Tennis TV (18:44)
- Tennis - One Night with the Valet (Official Mu...** by Tennisinc (1:54)
- Rafael Nadal Being ROASTED by Other...** by Tennistic Productions (9:42)
- ATP WINNING MOMENTS! 2022 SEASON** by Tennis TV (41:19)
- Tennis DOESN'T get much better! 🎾** by US Open Tennis Championships (0:22)

Each video thumbnail includes a blue "Describe" button at the bottom right.

## Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

## Including blind users ... write descriptions

Metrics		
Seconds:	11.097	
State:	2	
Log		
<pre>---- Inside Out 1 ---- Nav Play 0.000 AD *** Desc 59.981 57 // Another marble rolls down.  ---- Inside Out 2 ---- Nav Pause 3.147 Nav Play 3.129 AD *** Desc 13.934 AD *** Desc 27.703 AD *** Desc 45.432 46 // A marble rolls through a chute. AD *** Desc 58.913 57 // Another marble rolls down.  ---- Inside Out 3 ---- AD *** Desc 11.969 AD *** Desc 22.160 23 // Riley flips her plate. AD *** Qstn 39.201 Q: There's Riley, what are these other characters? // A: emotions, <u>human-esque</u> but not actually human Q: How do you spell the main character's name? // A: Riley</pre>		

Inside Out - Disgust and Anger  
Abundant Dialogue



### Navigation Controls

Play | Pause | Rewind | Forward | Timestamp | Replay

### Description Controls

Description | Question | Transcript

Q:

A:

Submit

## Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

## Including blind users ... write descriptions

Inside Out - Disgust and Anger  
Abundant Dialogue



Navigation Controls

Play Pause Rewind Forward Timestamp Replay

Description Controls

Description Question Transcript

Q:

A:

Submit

Accessible AD, a wizard-of-Oz prototype.

## Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

## Including blind users ... write descriptions



Navigation Controls

Play Pause Rewind Forward Timestamp Replay

Description Controls

Description Question Transcript

Q:

A:

Submit

Accessible AD, a wizard-of-Oz prototype.

## Existing techniques.

#1.  
Identify  
a11y issues

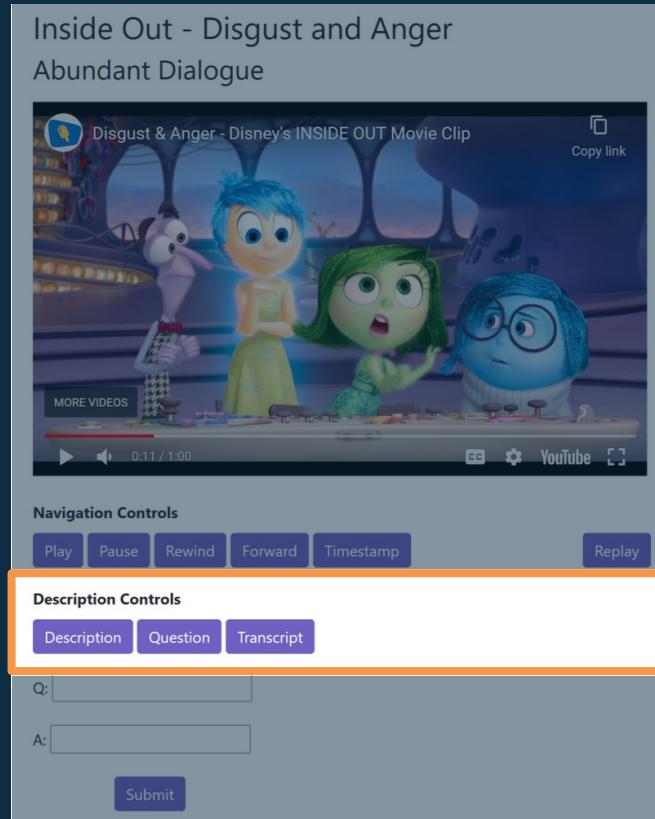


#2.  
Write  
descriptions



#3.  
Insert  
descriptions

## Including blind users ... write descriptions



Accessible AD, a wizard-of-Oz prototype.

Manually written baseline description, transcript, and visual question answering.

## Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

## Including blind users ... write descriptions

The screenshot shows a user interface for the Accessible AD prototype. On the left, there's a sidebar with buttons for 'Play', 'Description', and 'Caption'. The main area has two sections: 'Metrics' and 'Log'. The 'Metrics' section displays 'Seconds: 11.097' and 'State: 2'. The 'Log' section contains three entries labeled 'Inside Out 1', 'Inside Out 2', and 'Inside Out 3'. Each entry lists actions (Nav, Play, Desc) and their corresponding times. The 'Inside Out 3' entry includes a question and answer session:

- Inside Out 1 ----
  - Nav Play 0.000
  - AD \*\*\*\* Desc 59.981
  - 57 // Another marble rolls down.
- Inside Out 2 ----
  - Nav Pause 3.147
  - Nav Play 3.129
  - AD \*\*\*\* Desc 13.934
  - AD \*\*\*\* Desc 27.703
  - AD \*\*\*\* Desc 45.432
  - 46 // A marble rolls through a chute.
  - AD \*\*\*\* Desc 58.913
  - 57 // Another marble rolls down.
- Inside Out 3 ----
  - AD \*\*\*\* Desc 11.969
  - AD \*\*\*\* Desc 22.160
  - 23 // Riley flips her plate.
  - AD \*\*\*\* Qstn 39.201
  - Q: There's Riley, what are these other characters? // A: emotions, human-esque but not actually human
  - Q: How do you spell the main character's name? // A: Riley

Accessible AD, a wizard-of-Oz prototype.

Manually written baseline description, transcript, and visual question answering.

Navigation

Play

Description

Caption

Automation is challenging

## Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

## Including blind users ... write descriptions

The screenshot shows a user interface for a prototype called "Accessible AD". It has two main sections: "Metrics" and "Log".

**Metrics:**

- Seconds:** 11.097
- State:** 2

**Log:**

- Inside Out 1 ----
  - Nav Play 0.000
  - AD \*\*\*\* Desc 59.981
  - 57 // Another marble rolls down.
- Inside Out 2 ----
  - Nav Pause 3.147
  - Nav Play 3.129
  - AD \*\*\*\* Desc 13.934
  - AD \*\*\*\* Desc 27.703
  - AD \*\*\*\* Desc 45.432
  - 46 // A marble rolls through a chute.
  - AD \*\*\*\* Desc 58.913
  - 57 // Another marble rolls down.
- Inside Out 3 ----
  - AD \*\*\*\* Desc 11.969
  - AD \*\*\*\* Desc 22.160
  - 23 // Riley flips her plate.
  - AD \*\*\*\* Qstn 39.201
  - Q: There's Riley, what are these other characters? // A: emotions, human-esque but not actually human
  - Q: How do you spell the main character's name? // A: Riley

Accessible AD, a wizard-of-Oz prototype.

Manually written baseline description, transcript, and visual question answering.

Character identities (race, gender) & actions, setting or location

Automation is challenging

# Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

Metrics	
Seconds:	11.097
State:	2
Log	
---- Inside Out 1 ----	
Nav	Play 0.000
AD ****	Desc 59.981
57 // Another marble rolls down.	
---- Inside Out 2 ----	
Nav	Pause 3.147
Nav	Play 3.129
AD ****	Desc 13.934
AD ****	Desc 27.703
AD ****	Desc 45.432
46 // A marble rolls through a chute.	
AD ****	Desc 58.913
57 // Another marble rolls down.	
---- Inside Out 3 ----	
AD ****	Desc 11.969
AD ****	Desc 22.160
23 // Riley flips her plate.	
AD ****	Qstn 39.201
Q: There's Riley, what are these other characters? // A: emotions, <a href="#">human-esque</a> but not actually human	
Q: How do you spell the main character's name?	
// A: Riley	

+ gives agency to blind people

Inside Out - Disgust and Anger

Abundant Dialogue

Disgust & Anger - Disney's INSIDE OUT Movie Clip

Copy link

MORE VIDEOS

0:11 / 1:00

cc YouTube

Navigation Controls: Play, Pause, Rewind, Forward, Timestamp, Replay

Description Controls: Description, Question, Transcript

Q:  A:

Submit

Accessible AD, a wizard-of-Oz prototype.

Manually written baseline description, transcript, and visual question answering.

Character identities (race, gender) & actions, setting or location

- automation is challenging

## Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

Record and **fit** the descriptions  
into the appropriate parts of the  
video.

## Existing techniques.

### Record descriptions.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

Automate via  
text-to-speech  
in machine-voice

vs.

Manual  
recording in  
the human-voice

# Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

# Fit descriptions.

## A Navigate to video gaps



hey guys  
Garvey's guide to Ohio is coming your way so this is the first video guide for gobby's guys so we're gonna take you through where we stayed what we ate what we did yeah like our favorite things to eat at each of the restaurants you'll feel like you've got the lay of the land or you'll be most efficient with your time for sure many many a time keep I used to wear those inside no so oh I is kind of like a place where people go to like heal themselves and actually so it's very much like a hippie town it's holistic it's also just this really quiet sleepy town of mazing food great golf just a really chill place to get away

so really the first thing we did when we got to Ohio with went to the farmers market because what trip is a trip without the farmers market I am shots of strawberry if you want to skip farmers markets you pick the wrong girl I will

## B Script descriptions

	Title: Gabys Guide to Ojai
	A montage of bright footage from Ojai
	Shots of lavender in a farmers market
	Red flowers against a white house and blue sky.
	A courtyard and a pool.
	Gaby bikes along a path
	Close up of tater tots and french fries

## D Render composition

	Title: Gabys Guide to Ojai
	A montage from Ojai
	Shots of lavender
	Red flowers
	A courtyard and a pool.
	Gaby bikes
	Close up of french fries

## C Record descriptions

	Red flowers against a white house and blue sky.
<input checked="" type="radio"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	

## E Refine composition

	Red flowers against a white house and blue sky.
<input checked="" type="radio"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	

Rescribe

# Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

## Fit descriptions.

### A Navigate to video gaps



### B Script descriptions

	Title: Gabys Guide to Ojai
	A montage of bright footage from Ojai
	Shots of lavender in a farmers market
	Red flowers against a white house and blue sky.
	A courtyard and a pool.
	Gaby bikes along a path
	Close up of tater tots and french fries

### D Render composition

	Title: Gabys Guide to Ojai
	A montage from Ojai
	Shots of lavender
	Red flowers
	A courtyard and a pool.
	Gaby bikes
	Close up of french fries

### C Record descriptions

	Red flowers against a white house and blue sky.
	● ✓ ⏴ ○ ⓘ

### E Refine composition

	Red flowers against a white house and blue sky.
	● ✓ ⏴ ○ ⓘ

Rescribe

# Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

# Fit descriptions.

## A Navigate to video gaps



00:00:00.000  
Speech Gaps |  
  
hey guys  
Garvey's guide to Ohio is coming your way it so this is the first video guide for gobby's guys so we're gonna take you through where we stayed what we ate what we did yeah like our favorite things to eat at each of the restaurants you'll feel like you've got the lay of the land or you'll be most efficient with your time for sure many many a time keep I used to wear those inside no so oh I is kind of like a place where people go to like heal themselves and actually so it's very much like a hippie town it's holistic it's also just this really quiet sleepy town of mazing food great golf just a really chill place to get away  
  
so really the first thing we did when we got to Ohio with went to the farmers market because what trip is a trip without the farmers market I am shots of strawberry if you want to skip farmers markets you pick the wrong girl I will

## B Script descriptions

Title: Gabys Guide to Ojai
A montage of bright footage from Ojai
Shots of lavender in a farmers market
Red flowers against a white house and blue sky.
A courtyard and a pool.
Gaby bikes along a path
Close up of tater tots and french fries

## D Render composition

Title: Gabys Guide to Ojai
A montage from Ojai
Shots of lavender
Red flowers
A courtyard and a pool.
Gaby bikes
Close up of french fries

## C Record descriptions

Red flowers against a white house and blue sky.
● ✓ ⌂ ○ ⓘ

## E Refine composition

Red flowers against a white house and blue sky.
● ✓ ⌂ ○ ⓘ

Rescribe

# Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

# Fit descriptions.

## A Navigate to video gaps



hey guys  
Garvey's guide to Ohio is coming your way it so this is the first video guide for gobby's guys so we're gonna take you through where we stayed what we ate what we did yeah like our favorite things to eat at each of the restaurants you'll feel like you've got the lay of the land or you'll be most efficient with your time for sure many many a time keep I used to wear those inside no so oh I is kind of like a place where people go to like heal themselves and actually so it's very much like a hippie town it's holistic it's also just this really quiet sleepy town of mazing food great golf just a really chill place to get away

so really  
the first thing we did when we got to Ohio with went to the farmers market because what trip is a trip without the farmers market I am shots of strawberry if you want to skip farmers markets you pick the wrong girl I will

## B Script descriptions



Title: Gabys Guide to Ojai



A montage of bright footage from Ojai



Shots of lavender in a farmers market



Red flowers against a white house and blue sky.



A courtyard and a pool.



Gaby bikes along a path



Close up of tater tots and french fries

## D Render composition



Title: Gabys Guide to Ojai



A montage from Ojai



Shots of lavender



Red flowers



A courtyard and a pool.



Gaby bikes



Close up of french fries

## C Record descriptions



Red flowers against a white house and blue sky.



## E Refine composition



Red flowers against a white house and blue sky.



Rescribe

Pavel 2020

# Existing techniques.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

+ required minimal editing  
+ easy to use

# Fit descriptions.

**A Navigate to video gaps**

hey guys  
Garvey's guide to Ohio is coming your way it so this is the first video guide for gobby's guys so we're gonna take you through where we stayed what we ate what we did yeah like our favorite things to eat at each of the restaurants you'll feel like you've got the lay of the land or you'll be most efficient with your time for sure many many a time keep I used to wear those inside no so oh I is kind of like a place where people go to like heal themselves and actually so it's very much like a hippie town it's holistic also just this really quiet sleepy town of mazing food great golf just a really chill place to get away

so really  
the first thing we did when we got to Ohio with went to the farmers market because what trip is a trip without the farmers market I am shots of strawberry if you want to skip farmers markets you pick the wrong girl I will

**B Script descriptions**

Title: Gabys Guide to Ojai
A montage of bright footage from Ojai
Shots of lavender in a farmers market
Red flowers against a white house and blue sky.
A courtyard and a pool.
Gaby bikes along a path
Close up of tater tots and french fries

**C Record descriptions**

Red flowers against a white house and blue sky.
Red flowers against a white house and blue sky.

**D Render composition**

Title: Gabys Guide to Ojai
A montage from Ojai
Shots of lavender
Red flowers
A courtyard and a pool.
Gaby bikes
Close up of french fries

**E Refine composition**

Red flowers against a white house and blue sky.
Red flowers against a white house and blue sky.

Rescribe

- unnatural fit for long descriptions

## Existing techniques.

# Review.

#1.  
**Identify  
a11y issues**



#2.  
**Write  
descriptions**



#3.  
**Insert  
descriptions**

Watch through the video to identify inaccessible video segments.

Describe the relevant visuals concisely.

Record and fit the descriptions into the appropriate parts of the video.

## Existing techniques.

# Review.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

Gaps in speech (Gagnon 2010)  
CrossA11y (Liu 2022)

## Existing techniques.

# Review.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

Gaps in speech (Gagnon 2010)  
CrossA11y (Liu 2022)

Automated tools (Campos 2018; Wang 2021)  
Authoring support tools (Yuksel 2020; Natalie 2021a,b; Jiang & Ladner 2022; YouDescribe)

## Existing techniques.

# Review.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

Gaps in speech (Gagnon 2010)  
CrossA11y (Liu 2022)

Automated tools (Campos 2018; Wang 2021)  
Authoring support tools (Yuksel 2020; Natalie 2021a,b; Jiang & Ladner 2022; YouDescribe)

Automatic vs. Manual (Kobayashi 2010)  
Rescribe (Pavel 2020)

# Outline

## 1 Challenge in video accessibility.

Scaling audio descriptions to the massive video generation rates.

## 2 Existing techniques.

Support the process of audio description generation.

## 3 Future work.

What are opportunities for future work in video accessibility?

# Outline

## 1 Challenge in video accessibility.

Scaling audio descriptions to the massive video generation rates.

## 2 Existing techniques.

Support the process of audio description generation.

## 3 Future work.

What are opportunities for future work in video accessibility?

Existing techniques.

# Where do we stand?

Core challenge:

Scaling audio descriptions to the massive video generation rates.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

Gaps in speech (Gagnon 2010)  
CrossA11y (Liu 2022)

Automated tools (Campos 2018; Wang 2021)  
Authoring support tools (Yuksel 2020; Natalie 2021a,b; Jiang & Ladner 2022; YouDescribe)

Automatic vs. Manual (Kobayashi 2010)  
Rescribe (Pavel 2020)

Existing techniques.

# Where do we stand?

Core challenge:

Scaling audio descriptions to the massive video generation rates.

#1.  
Identify  
a11y issues



#2.  
Write  
descriptions



#3.  
Insert  
descriptions

Gaps in speech (Gagnon 2010)  
CrossA11y (Liu 2022)

Automated tools (Campos 2018; Wang 2021)  
Authoring support tools (Yuksel 2020; Natalie 2021a,b; Jiang & Ladner 2022; YouDescribe)

Automatic vs. Manual (Kobayashi 2010)  
Rescribe (Pavel 2020)

# Where do we stand?

#2.  
Write  
descriptions

Automated tools (Campos 2018; Wang 2021)  
Authoring support tools (Yuksel 2020; Natalie 2021a,b; Jiang & Ladner 2022; YouDescribe)

Scalability

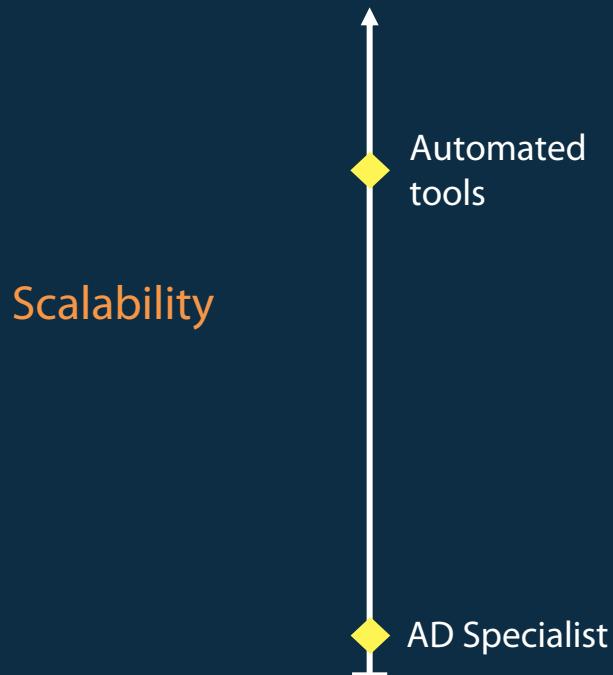


AD Specialist

# Where do we stand?

## #2. Write descriptions

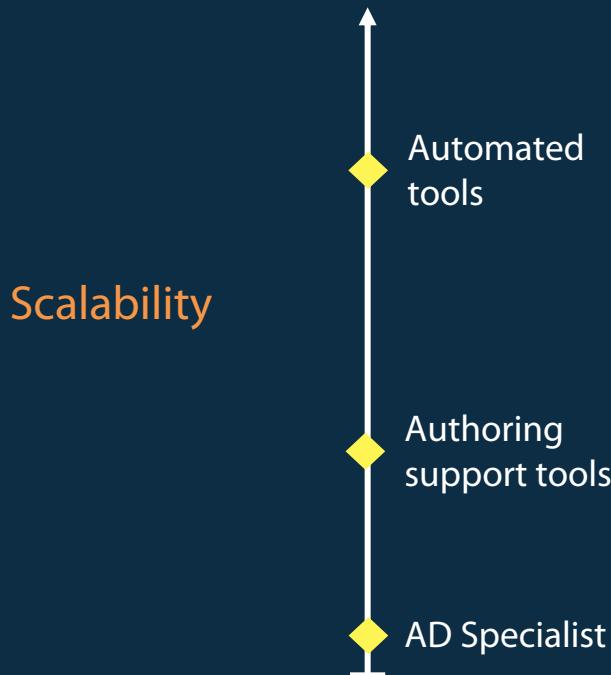
Automated tools (Campos 2018; Wang 2021)  
Authoring support tools (Yuksel 2020; Natalie 2021a,b; Jiang & Ladner 2022; YouDescribe)



# Where do we stand?

## #2. Write descriptions

Automated tools (Campos 2018; Wang 2021)  
Authoring support tools (Yuksel 2020; Natalie 2021a,b; Jiang & Ladner 2022; YouDescribe)



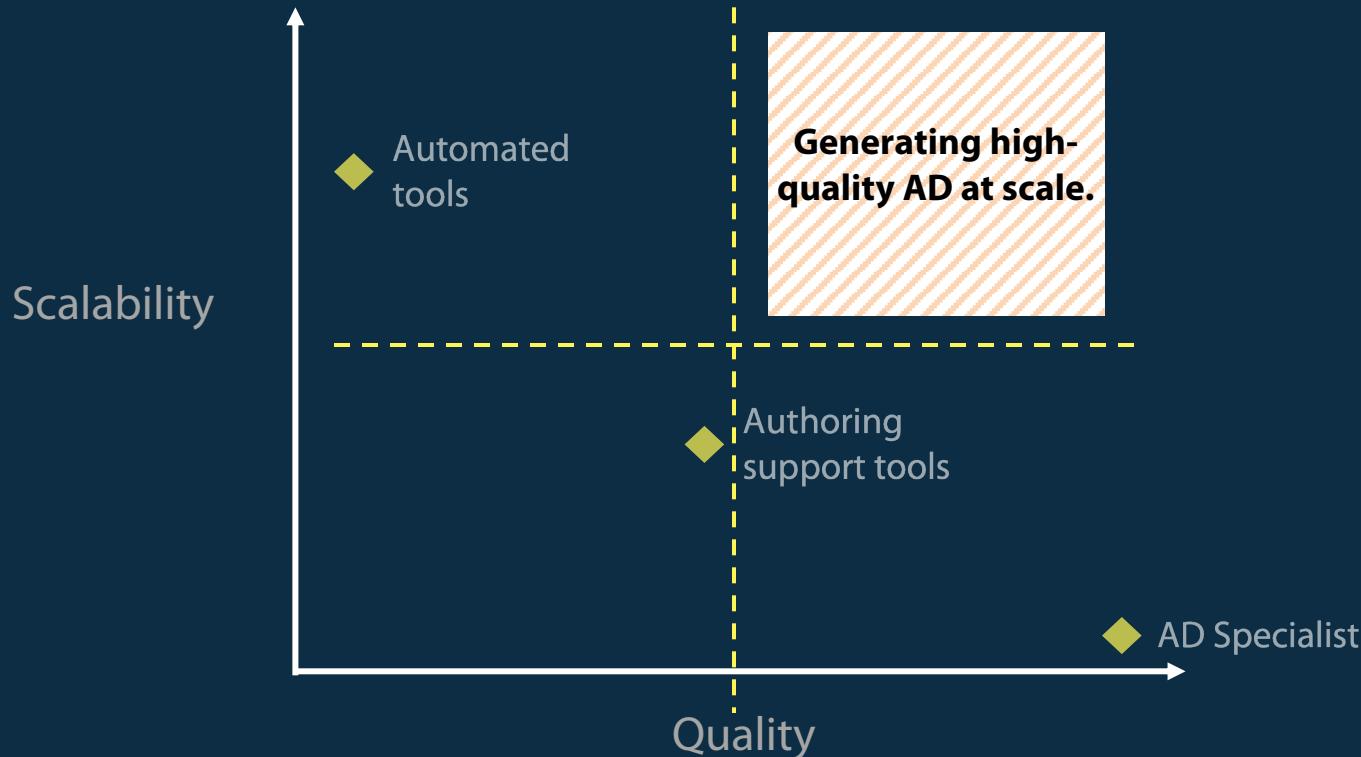
# Where do we stand?

## #2. Write descriptions

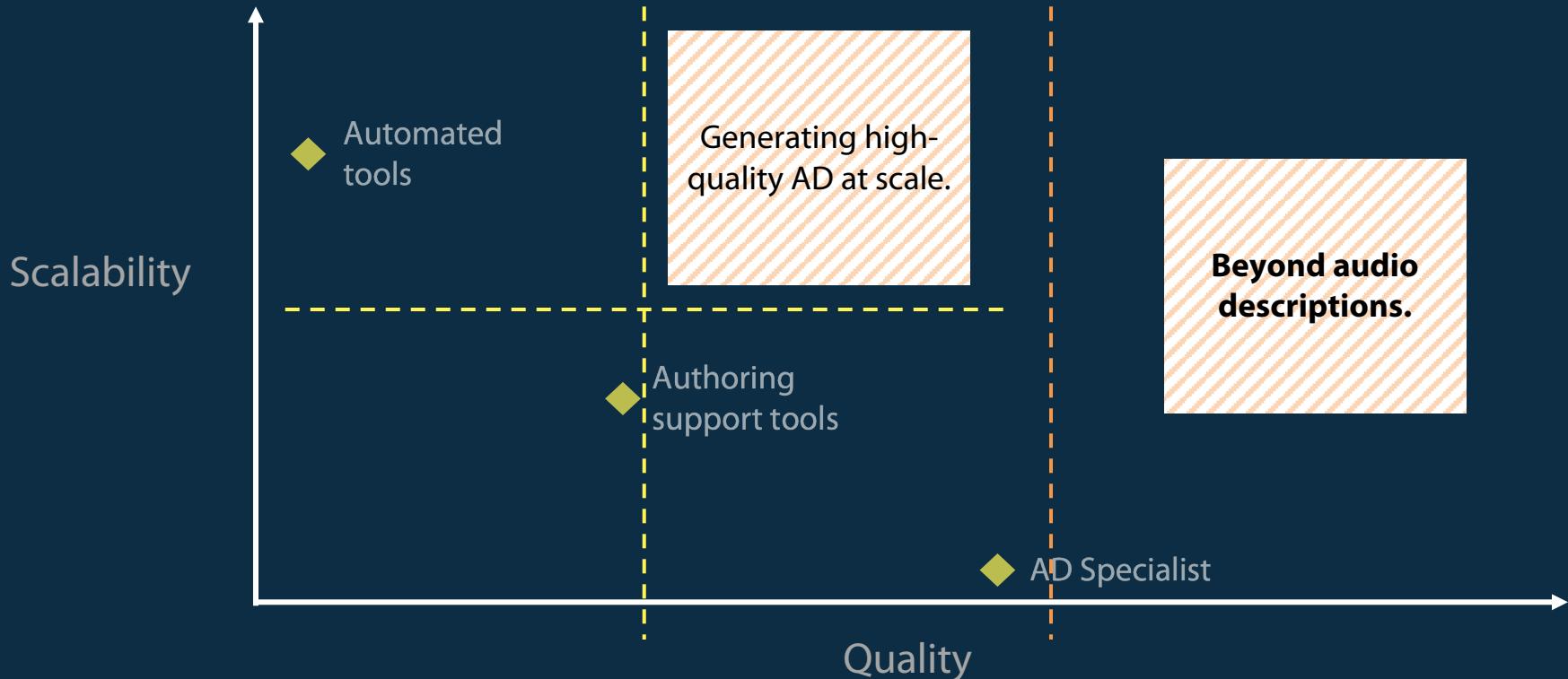
Automated tools (Campos 2018; Wang 2021)  
Authoring support tools (Yuksel 2020; Natalie 2021a,b; Jiang & Ladner 2022; YouDescribe)



# What next?



# What next?



# Generating high-quality AD at scale.

*What we learned...*

AI → cannot write descriptions (for now)

Individuals → individuals cannot achieve scalability

# Generating high-quality AD at scale.

*What we learned...*

AI



cannot write descriptions (for now)  
can process descriptions

Individuals



cannot be scalability  
can write descriptions

# Generating high-quality AD at scale.

*What we learned...*

AI → cannot write descriptions (for now)  
can process descriptions

Individuals → cannot be scalability  
can write descriptions

Crowdsourcing x AI?

# Generating high-quality AD at scale.

Crowdsourcing x AI?

Salisbury 2017

Proceedings of the Fifth Conference on Human Computation and Crowdsourcing (HCOMP 2017)

**Toward Scalable Social Alt Text: Conversational Crowdsourcing as a Tool for Refining Vision-to-Language Technology for the Blind**

**Elliot Salisbury,\* Ece Kamar,<sup>+</sup> Meredith Ringel Morris<sup>+</sup>**  
\*University of Southampton, <sup>+</sup>Microsoft Research  
[e.salisbury@ecs.soton.ac.uk](mailto:e.salisbury@ecs.soton.ac.uk), <sup>+</sup>[{eckamar, merrie}@microsoft.com](mailto:{eckamar, merrie}@microsoft.com)

**Abstract**

The access of visually impaired users to imagery in social media is constrained by the availability of suitable alt text. It is unknown how imperfections in emerging tools for automatic caption generation may help or hinder blind users' understanding of social media posts with embedded imagery. In this paper, we study how crowdsourcing can be used both for evaluating the value provided by existing automated approaches and for enabling workflows that provide scalable and useful alt text to blind users. Using real-time crowdsourcing, we designed experiences that varied the depth of interaction with the crowd in assisting visually impaired users at caption interpretation, and measured trade-offs in effectiveness, scalability, and reusability. We show that the shortcomings of existing AI image captioning systems frequently hinder a user's understanding of an image they cannot see to a degree that even clarifying conversations with sighted assistants cannot correct. Our detailed analysis of the set of clarifying conversations collected from our studies led to the design of experiences that can effectively assist users in a scalable way without the need for real-time interaction. They also provide lessons and guidelines that human captioners and the designers of future iterations of AI captioning systems can use to improve labeling of social media imagery for blind users.

**Introduction**

Social media is becoming pervasive in American culture; as of 2014, 74% of online adults in the U.S. use social networking sites (Duggan et al. 2015). The opportunity to engage with social media is an important part of social, professional, and political life, making it important that people who are blind or visually impaired (BVI) can access the entirety of content shared in social media. For example, Twitter

Twitter recently began to offer limited capabilities to augment images with alternative text (a.k.a. alt text or captions) that can be read aloud by the screen reader technology (e.g., JAWS, VoiceOver, Narrator, etc.) that provides computer access to people who are BVI (Kloots 2016); however, while no official numbers on alt text compliance on the web in general is low (Bigham et al. 2006; Goodwin et al. 2011; Shi 2006), and this trend is likely to be exacerbated by quickly-created, user-generated content such as tweets.

Recently, automated approaches that combine computer vision and natural language processing to describe image content have emerged as a potential solution for improving the accessibility of social media imagery for BVI users. Examples include the automatic alt text system deployed by Facebook (Wu, Pique, and Wieland 2016) and automated image captioning systems (Fang et al. 2015; Karpathy and Fei-Fei 2015). Although assisting blind users is a motivating application domain for these systems, the value these imperfect systems provide to BVI users is unclear. While existing systems are tested in the lab within constrained data sets, the performance of these systems in the context of social media (which incorporates a wide variety of professional and casual quality imagery and covers a range of subjects and styles) is not yet studied. The levels of detail, accuracy, or confidence expected from BVI users may not be attainable with current vision-to-language technologies. Unexpected imperfections in automated system output may degrade user trust, or may hurt users instead of helping them.

In this work, we explore ways for combining crowd input and existing automated approaches to assist BVI users in accessing social media with visual content. Our studies focus on the following research questions: (1) What ad-

# Beyond audio descriptions.



Audio  
descriptions

# Beyond audio descriptions.



Audio  
descriptions



Access to  
visual information

# Beyond audio descriptions.



Audio  
descriptions



Access to  
visual information



Experience of consuming  
visual information

# Beyond audio descriptions.



Audio  
descriptions



Access to  
visual information  
(Level of understanding)

Campos 2020



Experience of consuming  
visual information

# Beyond audio descriptions.



Audio  
descriptions



**Access to**  
**visual information**  
(Level of understanding)



**Experience of consuming**  
**visual information**  
(Immersion, enjoyment,  
engagement)

# Beyond audio descriptions.

Wilken & Kruger 2016



Audio  
descriptions



Access to  
visual information  
(Level of understanding)



Experience of consuming  
the information  
(Immersion, enjoyment,  
engagement)

*Across Languages and Cultures* 17 (2), pp. 251–270 (2016)  
DOI: 10.1556/084.2016.17.2.6

## PUTTING THE AUDIENCE IN THE PICTURE: **MISE-EN-SHOT AND PSYCHOLOGICAL IMMERSION IN AUDIO DESCRIBED FILM**

NICOLE WILKEN<sup>1</sup>, JAN-LOUIS KRUGER<sup>2</sup>

<sup>1</sup>School of Languages, North-West University  
Vaal Triangle Campus Vanderbijlpark, South Africa  
Phone: +27 826854701, fax +27 169103463  
E-mail: 20398026@nwu.ac.za

<sup>2</sup>Macquarie University,  
Balaclava Road, NSW 2109 Sydney, Australia  
Phone: +61 449630802  
E-mail: janlouis.kruger@mq.edu.au

**Abstract:** Audio description (AD) often emphasises the visual elements of a film rather than the way these elements are presented. However, what is seen and the way it is shown are equally important for creating meaning in film. The term *mise-en-shot* refers to the way in which visual aspects are shown to the audience. In order to determine whether the stylistic elements of film created by means of *mise-en-shot* could influence the reception of audio described film, the article investigates the effect of the presence or absence in the AD of these elements on the immersion of a sighted audience into the fictional world. Immersion is measured by means of sub-scales on character identification as well as transportation. In order to measure the effect of stylistic elements, the self-reported immersion of one group of sighted participants who sees a scene with the original soundtrack is compared to that of another sighted group who only hears the audio-described soundtrack of the scene. The findings suggest that although the absence of some *mise-en-shot* elements in the audio described version of the film does not influence transportation, it does influence the way in which a sighted audience identifies with characters in the film. It would therefore seem that these stylistic elements do have an important role in the immersion of audiences, which could have significant implications for AD.

**Keywords:** audio description, *mise-en-scène*, *mise-en-shot*, transportation, identification, immersion

## 1. INTRODUCTION

# Beyond audio descriptions.

Wilken & Kruger 2016



Audio  
descriptions



Access to  
visual information  
(Level of understanding)



Experience of consuming  
the information  
(Immersion, enjoyment,  
engagement)

*Across Languages and Cultures* 17 (2), pp. 251–270 (2016)  
DOI: 10.1556/084.2016.17.2.6

## PUTTING THE AUDIENCE IN THE PICTURE: **MISE-EN-SHOT AND PSYCHOLOGICAL IMMERSION IN AUDIO DESCRIBED FILM**

NICOLE WILKEN<sup>1</sup>, JAN-LOUIS KRUGER<sup>2</sup>

<sup>1</sup>School of Languages, North-West University  
Vaal Triangle Campus Vanderbijlpark, South Africa  
Phone: +27 826854701, fax +27 169103463  
E-mail: 20398026@nwu.ac.za

<sup>2</sup>Macquarie University,  
Balaclava Road, NSW 2109 Sydney, Australia  
Phone: +61 449630802  
E-mail: janlouis.kruger@mq.edu.au

**Abstract:** Audio description (AD) often emphasises the visual elements of a film rather than the way these elements are presented. However, what is seen and the way it is shown are equally important for creating meaning in film. The term *mise-en-shot* refers to the way in which visual aspects are shown to the audience. In order to determine whether the stylistic elements of film created by means of *mise-en-shot* could influence the reception of audio described film, the article investigates the effect of the presence or absence in the AD of these elements on the immersion of a sighted audience into the fictional world. Immersion is measured by means of sub-scales on character identification as well as transportation. In order to measure the effect of stylistic elements, the self-reported immersion of one group of sighted participants who sees a scene with the original soundtrack is compared to that of another sighted group who only hears the audio-described soundtrack of the scene. The findings suggest that although the absence of some *mise-en-shot* elements in the audio described version of the film does not influence transportation, it does influence the way in which a sighted audience identifies with characters in the film. It would therefore seem that these stylistic elements do have an important role in the immersion of audiences, which could have significant implications for AD.

**Keywords:** audio description, *mise-en-scène*, *mise-en-shot*, transportation, identification, immersion

## 1. INTRODUCTION

# Beyond audio descriptions.

Walczak & Fryer 2017



Audio  
descriptions



Access to  
visual information  
  
(Level of understanding)



Experience of consuming  
the information  
  
(Immersion, enjoyment,  
engagement)

Check for updates

Research Article

**Creative description: The impact of audio description style on presence in visually impaired audiences**

British Journal of Visual Impairment  
2017, Vol. 35(1) 6–17  
© The Author(s) 2016  
Reprints and permission:  
[sagepub.co.uk/journalsPermissions.nav](http://sagepub.co.uk/journalsPermissions.nav)  
DOI: 10.1177/0264619616661603  
[journals.sagepub.com/home/bvi](http://journals.sagepub.com/home/bvi)

**SAGE**

**Agnieszka Walczak**  
Universitat Autònoma de Barcelona, Spain

**Louise Fryer**  
University College London, UK

**Abstract**  
This article presents a study that tested the impact of audio description (AD) style on dimensions of presence (spatial presence, ecological validity, engagement, and negative effects) in blind and visually impaired audiences. The participants were shown two fragments of a naturalistic drama with two styles of description: 'standard' and 'creative'. While the former followed the principle of objectivity, the latter was an innovative type of AD that included elements of camera work and subjective descriptions of the characters, their actions, and scenes crucial to the plot. The findings show that the emotive AD prompted higher levels of presence for all participants. Overall, the new AD style seemed more natural, especially to participants with recent sight loss. The results suggest that creative scripts may stimulate presence and thus increase the chances of AD users having a more immersive viewing experience.

**Keywords**  
Accessibility, audio description, audiovisual translation, blind and visually impaired, creative description, presence

# Beyond audio descriptions.

How can we provide blind people an  
**equivalent experience** when watching  
videos?

# Beyond audio descriptions.

How can we provide blind people an equivalent experience when watching videos?

CHI 2018 Paper

CHI 2018, April 21–26, 2018, Montréal, QC, Canada

**Rich Representations of Visual Content for Screen Reader Users**

Meredith Ringel Morris<sup>1</sup>, Jazette Johnson<sup>1,2</sup>, Cynthia L. Bennett<sup>1,3</sup>, Edward Cutrell<sup>1</sup>

<sup>1</sup>Microsoft Research, Redmond, WA, USA  
<sup>2</sup>Vanderbilt University, <sup>3</sup>University of Washington  
 [{merrie, cutrell}@microsoft.com](mailto:{merrie, cutrell}@microsoft.com),  [jazette.johnson@vanderbilt.edu](mailto:jazette.johnson@vanderbilt.edu),  [bennec3@uw.edu](mailto:bennec3@uw.edu)

**ABSTRACT**

*Alt* text (short for “alternative text”) is descriptive text associated with an image in HTML and other document formats. Screen reader technologies speak the *alt* text aloud to people who are visually impaired. Introduced with HTML 2.0 in 1995, the *alt* attribute has not evolved despite significant changes in technology over the past two decades. In light of the expanding volume, purpose, and importance of digital imagery, we reflect on how *alt* text could be supplemented to offer a richer experience of visual content to screen reader users. Our contributions include articulating to operate their computers and mobile devices. Most major operating systems come with built-in screen readers that can be enabled in the accessibility settings (e.g., Apple’s VoiceOver, Google’s ChromeVox and TalkBack, Microsoft’s Narrator), and many people also choose to install third-party screen readers such as JAWS or NVDA. Screen readers render on-screen text as audio, and the user can navigate among different parts of the interface using shortcut keys (on a desktop or laptop computer) or gestures such as taps or swipes (on a tablet or smartphone).

Screen readers cannot render an image as audio unless the

# Beyond audio descriptions.

How can we provide blind people an equivalent experience when watching videos?

Interactions



Lee 2022

Representations



Ohshima 2018

Morris 2018

# Beyond audio descriptions.



Audio  
descriptions



Access to  
visual information  
(Level of understanding)



Lee 2022

Ohshima 2018



Experience of consuming  
visual information  
(Immersion, enjoyment,  
engagement)

Rich interactions and  
representations

# Outline

## 1 Challenge in video accessibility.

Scaling audio descriptions to the massive video generation rates.

## 2 Existing techniques.

Support the process of audio description generation.

## 3 Future work.

Generate high-quality AD at scale & thinking beyond audio descriptions.