JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

École des Ponts
ParisTech

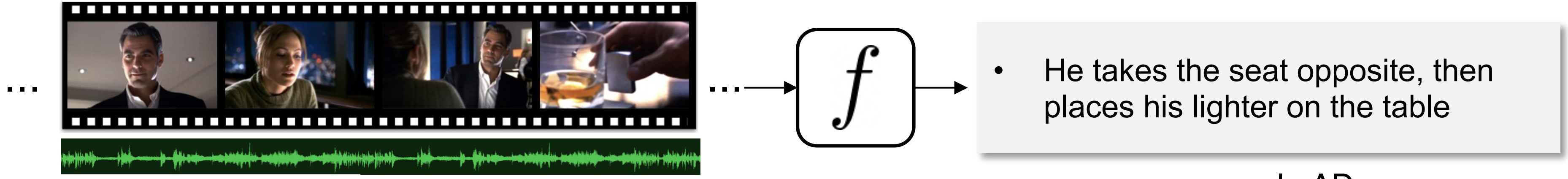
AutoAD:

Movie Description in Context

Tengda Han* Max Bain* Arsha Nagrani Gül Varol Weidi Xie Andrew Zisserman

Highlight@CVPR2023. Tag: THU-AM-234

What is AutoAD: Automatic Audio Description



- He takes the seat opposite, then places his lighter on the table

movie clips

example AD

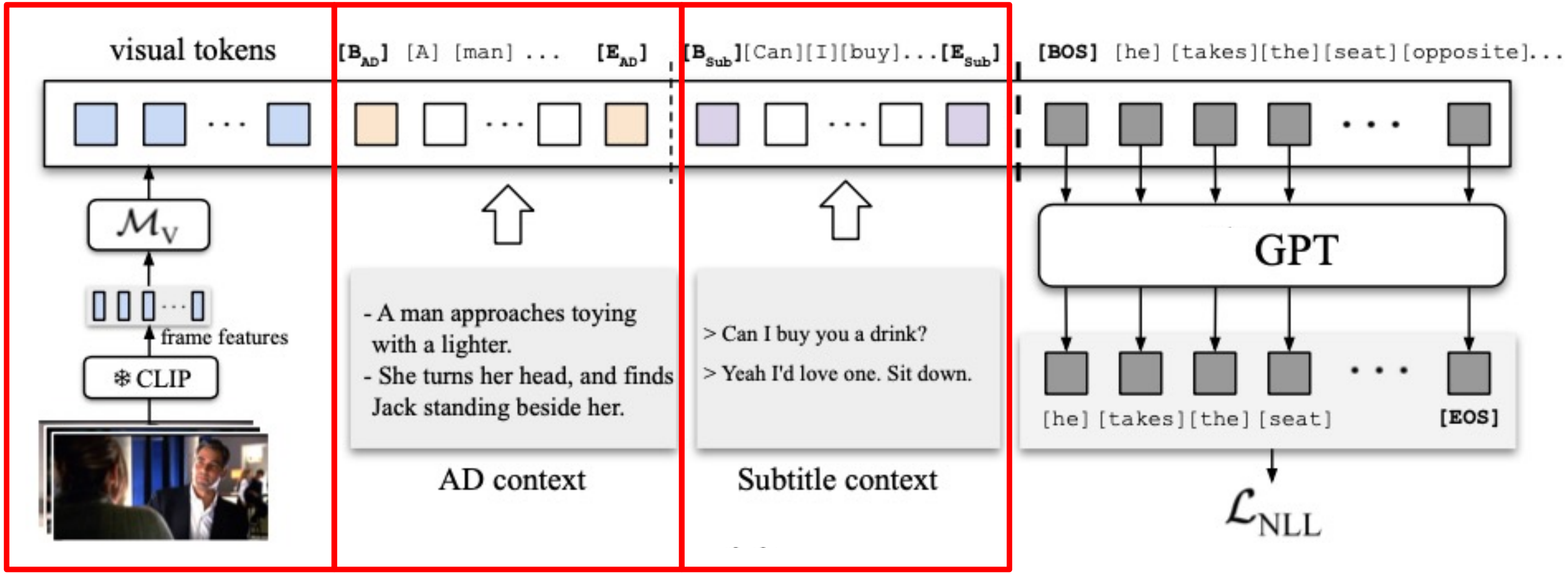
Audio Description (AD):

- Narration describing <u>visual</u> elements in movies, complementary to audio
- Developed to aid visually impaired audience

AutoAD – Automatic Audio Description:

- Aim to generate such descriptions with computer vision models automatically

Model Architecture


visual tokens
[B_AD] [A] [man] ... [E_AD]
[B_Sub][Can][I][buy]...[E_Sub]
[BOS] [he] [takes][the][seat][opposite]...
M_V
frame features
❄ CLIP
- A man approaches toying with a lighter.
- She turns her head, and finds Jack standing beside her.
AD context
> Can I buy you a drink?
> Yeah I'd love one. Sit down.
Subtitle context
GPT
[he] [takes][the] [seat]
[EOS]
L_NLL

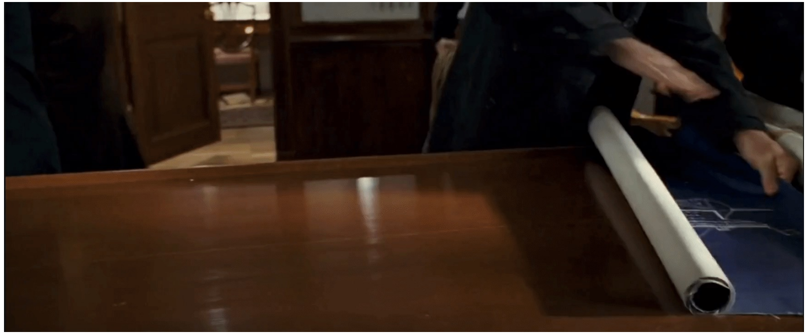
- Prompt-tuning GPT-2 for visual description
- We feed in visual, contextual AD, movie subtitles into the model

Pre-training with Partial Data

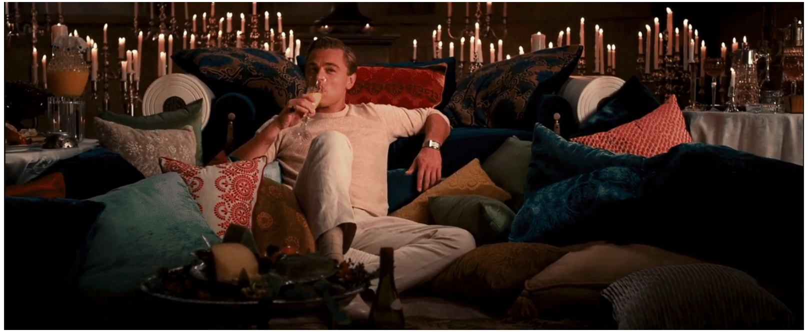
Dataset	Visual data	Text Descriptions	Subtitle	Size
MAD	✅	✅	✅	~500 movies
Conceptual Caption	✅	✅	❌	3M images
WebVid	✅	✅	❌	3M short videos
AudioVault-AD	❌	✅	✅	~8000 movies

- Complete movie data is very limited
- We pretrain our submodules on partial data

Qualitative Results



Context AD: ...The master-at-arms carts Jack away. In the chartroom, Andrews unrolls the ship's blueprint.
Ground-truth AD: Andrews Smith and others study the blueprint.
Prediction: They look at the map.



Context AD: Nick and Daisy smile and Gatsby gestures towards the ballroom. Klipspringer a wild-haired young man with glasses, plays the organ.
Ground-truth AD: Gatsby reclines on cushions as Nick and Daisy dance in the ballroom, which is lit by hundreds of candles.
Prediction: A man and a woman dance in a circle.

Samples from Titanic (1997) & The Great Gatsby (2013)

Overview of the Details

- What is AD data
- Method:
  - Prompt-tuning GPT-2
  - Partial-data Training
- Data Processing
- Results

What is movie Audio Description (AD)?

- Example of the original movie clip

Out of Sight (1998)

What is movie Audio Description (AD)?

- Example of the original movie clip

Out of Sight (1998)

What is movie Audio Description (AD)?

- Example of the AD

Out of Sight (1998)

What is movie Audio Description (AD)?

- Example of the AD

Out of Sight (1998)

What is movie Audio Description (AD)?

…

…

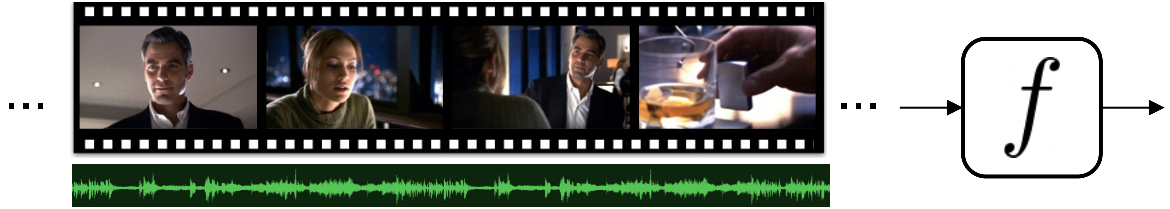
movie clips

- He takes the seat opposite, then places his lighter on the table

example AD

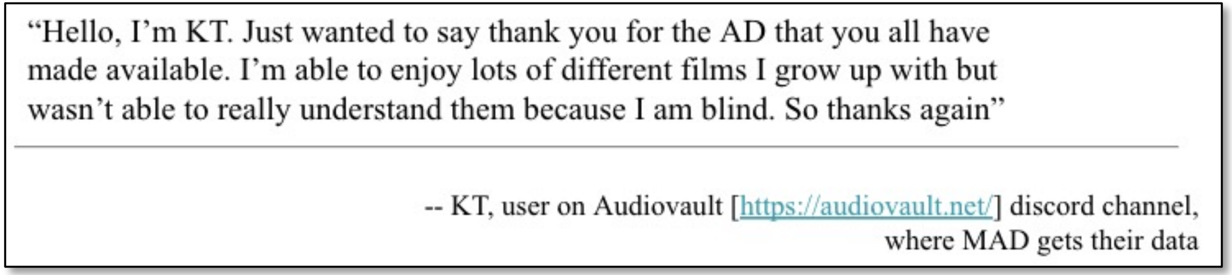
- **What**: narrations describing visual elements in movies
- **How**: typically generated by experienced annotators:
  - Dense descriptions over time
  - Complementary to the raw audio track
  - Aims at storytelling: includes characters’ name, emotion, action, etc.
- **Why**: developed to aid visually impaired audiences
  - AudioVault: https://audiovault.net/
  - the size of data is growing

Our objective: Automatic AD generation



- He takes the seat opposite, then places his lighter on the table

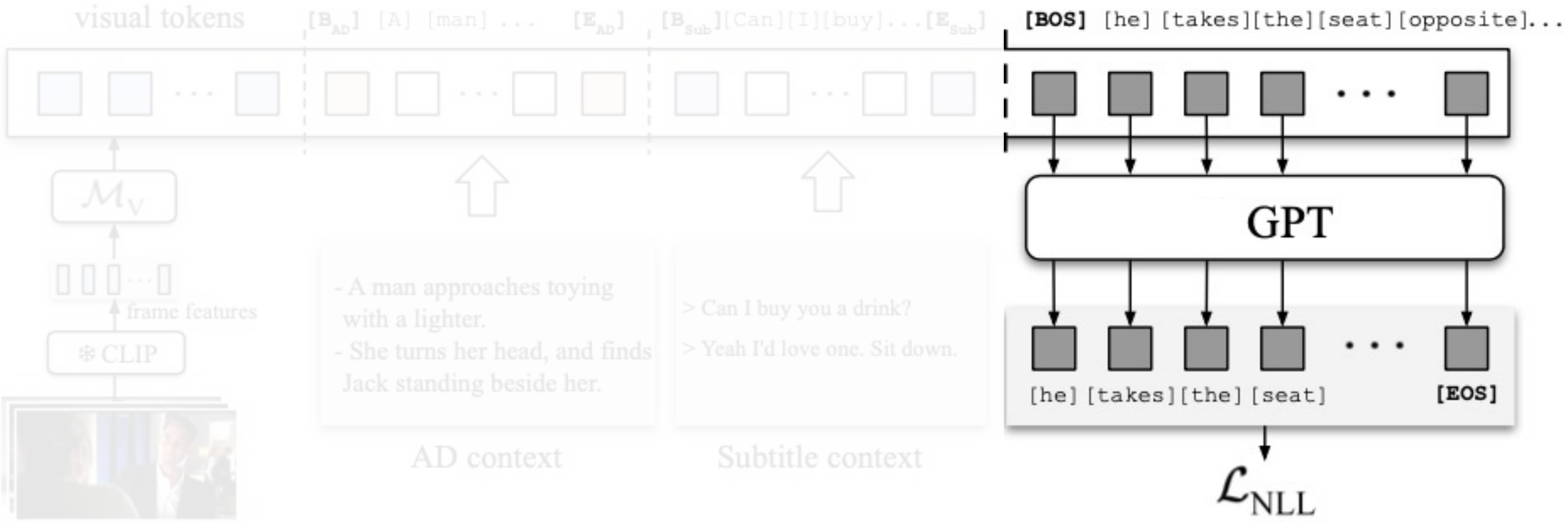
- A new way to evaluate movie understanding abilities
  - Long-term understanding, Multi-modal understanding, Fine-grained recognition
- Social impact:



"Hello, I'm KT. Just wanted to say thank you for the AD that you all have made available. I'm able to enjoy lots of different films I grow up with but wasn't able to really understand them because I am blind. So thanks again"

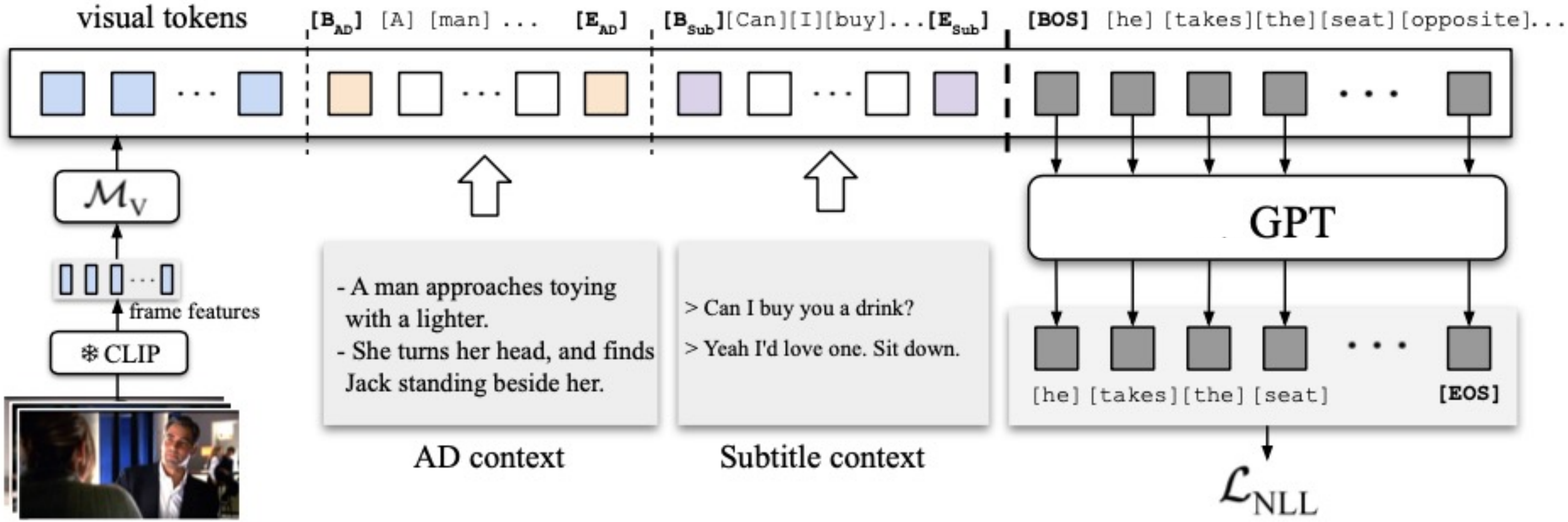
-- KT, user on Audiovault [https://audiovault.net/] discord channel,
where MAD gets their data

Method: Video Captioning with Long Multimodal Context


visual tokens
[B_AD] [A] [man] . . . [E_AD]
[B_Sub][Can][I][buy]...[E_Sub]
[BOS] [he] [takes][the][seat][opposite]...
$\mathcal{M}_V$
frame features
❄ CLIP
- A man approaches toying with a lighter.
- She turns her head, and finds Jack standing beside her.
> Can I buy you a drink?
> Yeah I'd love one. Sit down.
AD context
Subtitle context
GPT
[he] [takes] [the] [seat]
[EOS]
$\mathcal{L}_{NLL}$

- We use a pretrained GPT for text generation

Method: Video Captioning with Long Multimodal Context


visual tokens
$[B_{AD}]$ [A] [man] . . . $[E_{AD}]$
$[B_{Sub}]$[Can][I][buy]...$[E_{Sub}]$
[BOS] [he] [takes][the][seat][opposite]...
$\mathcal{M}_V$
frame features
❄ CLIP
GPT
[he] [takes][the] [seat]
[EOS]
$\mathcal{L}_{NLL}$

- A man approaches toying with a lighter.
- She turns her head, and finds Jack standing beside her.

AD context

> Can I buy you a drink?

> Yeah I'd love one. Sit down.

Subtitle context

- We use a pretrained GPT for text generation
- All the conditions are added as a prompting vectors
  - Visual features (CLIP), contextual AD, movie subtitles

Challenge: the lack of training data

web videos

- User-uploaded videos on platforms, e.g. YouTube, Shutterstock).
- About **<u>82 years</u>** of videos uploaded to YouTube every day [1].

movies

- About **<u>3.2 hours</u>** of movies produced every day [2].
- Most of them not accessible due to copyright restrictions.

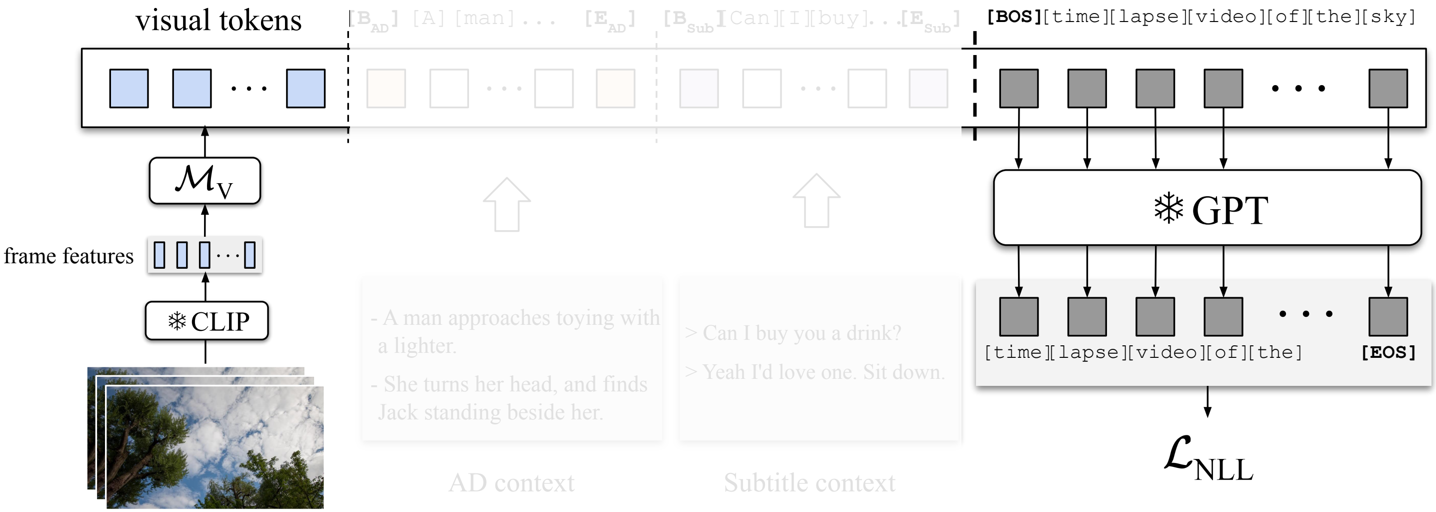
complete movie data

The movie data with corresponding visual, subtitles and description elements are very limited in size

[1] YouTube Official Blog. [2] https://www.imdb.com/search/title/?year=2022&title_type=feature&

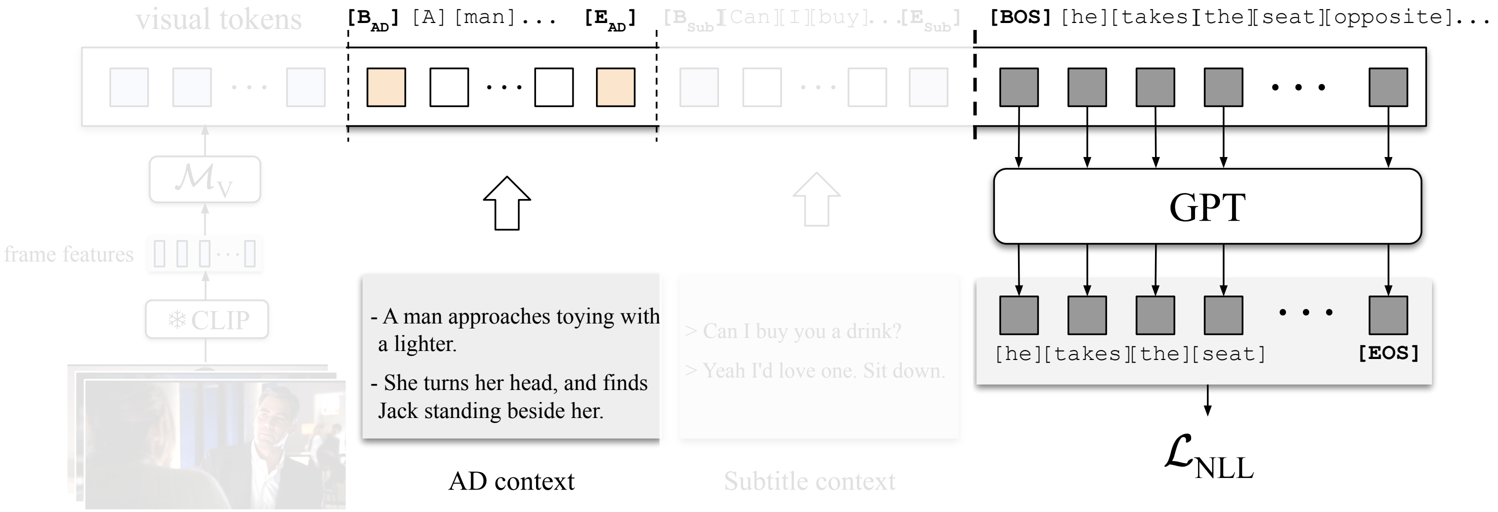
Pretrain with Partial Data

- The ‘complete’ movie dataset is limited in size, but we have:
  - Paired visual-textual data (without temporal context): CC3M, WebVid
  - Movie AD data (without visual information): downloaded from AudioVault
- We can use partial data to pretrain some of the modules:
  - visual only pretraining


visual tokens
[B_AD] [A] [man] . . . [E_AD]
[B_Sub][Can][I][buy]. . .[E_Sub]
[BOS][time][lapse][video][of][the][sky]
$\mathcal{M}_V$
frame features
CLIP
GPT
- A man approaches toying with a lighter.
- She turns her head, and finds Jack standing beside her.
> Can I buy you a drink?
> Yeah I'd love one. Sit down.
[time][lapse][video][of][the]
[EOS]
$\mathcal{L}_{NLL}$
AD context
Subtitle context

Pretrain with Partial Data

- The ‘complete’ movie dataset is limited in size, but we have:
  - Paired visual-textual data (without temporal context): CC3M, WebVid
  - Movie AD data (without visual information): downloaded from AudioVault
- We can use partial data to pretrain some of the modules:
  - text only pretraining



- A man approaches toying with a lighter.

- She turns her head, and finds Jack standing beside her.

> Can I buy you a drink?

> Yeah I'd love one. Sit down.

Examples of MAD-v1 dataset

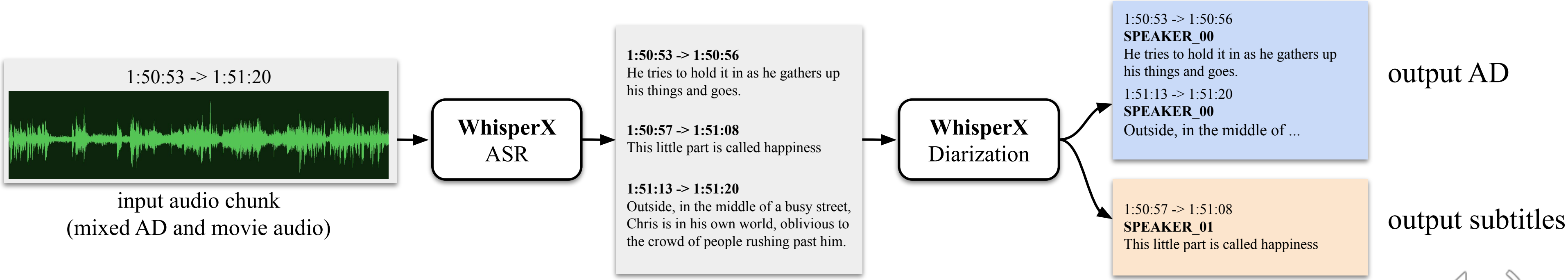
Manual Verification	She stands and the little warrior takes in her size, about twice his own.	Leia sits on a moss covered log.
MAD-v1	Angola, she stands in the Little Warrior, takes in her size about twice his own.	I'm not gon na. Leah sits on a Moss covered log.

Red color means erroneous AD

MAD Dataset: https://github.com/Soldelli/MAD
Samples from Star Wars VI: Return of the Jedi (1983)

Dataset preparation

- Denoise **MAD**
  - 488 movies with visual features, subtitles and AD
  - Original version has low-quality ASR and many dialogue leakages
- Collect & Denoise **AudioVault**
  - 7057 movies with subtitles and AD, but without visual features
  - Raw data downloaded is a single audio file with mixed movie soundtrack and AD



1:50:53 -> 1:51:20

input audio chunk
(mixed AD and movie audio)

WhisperX
ASR

1:50:53 -> 1:50:56
He tries to hold it in as he gathers up his things and goes.

1:50:57 -> 1:51:08
This little part is called happiness

1:51:13 -> 1:51:20
Outside, in the middle of a busy street, Chris is in his own world, oblivious to the crowd of people rushing past him.

WhisperX
Diarization

1:50:53 -> 1:50:56
SPEAKER_00
He tries to hold it in as he gathers up his things and goes.
1:51:13 -> 1:51:20
SPEAKER_00
Outside, in the middle of ...

output AD

1:50:57 -> 1:51:08
SPEAKER_01
This little part is called happiness

output subtitles

For both datasets, we use the same pipeline to collect the textual data from the raw audio

https://github.com/m-bain/whisperX

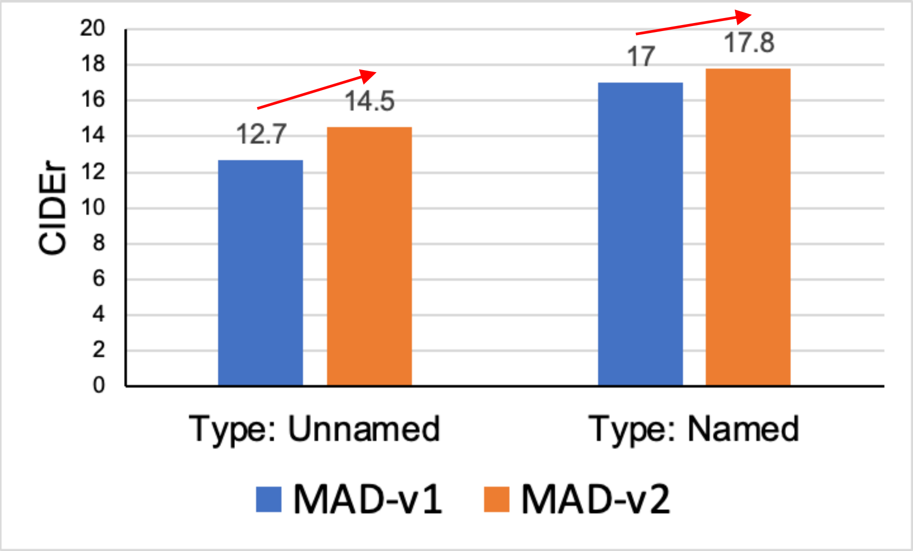
Qualitative comparison of MAD-v1 and v2

Manual Verification	She stands and the little warrior takes in her size, about twice his own.	Leia sits on a moss covered log.
MAD-v1	Angola, she stands in the Little Warrior, takes in her size about twice his own.	I'm not gon na. Leah sits on a Moss covered log.
MAD-v2 (ours)	She stands and the little warrior takes in her size about twice his own.	Leia sits on a moss-covered log.

Red color means errone AD

Samples from Star Wars VI: Return of the Jedi (1983)

Results: Denoising MAD dataset



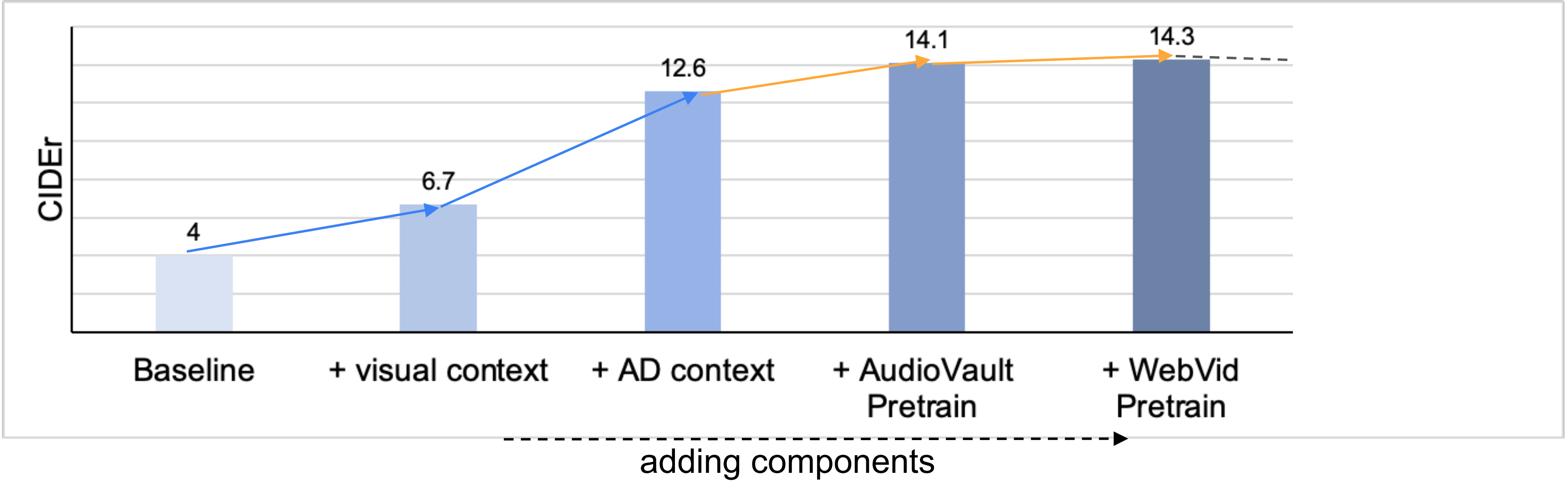
Named AD: Leia sits on a moss-covered log.

↕

Unnamed AD: SOMEONE sits on a moss-covered log.

↗ In general, training on the cleaner MAD-v2 performs better than MAD-v1

Results: context and pretraining

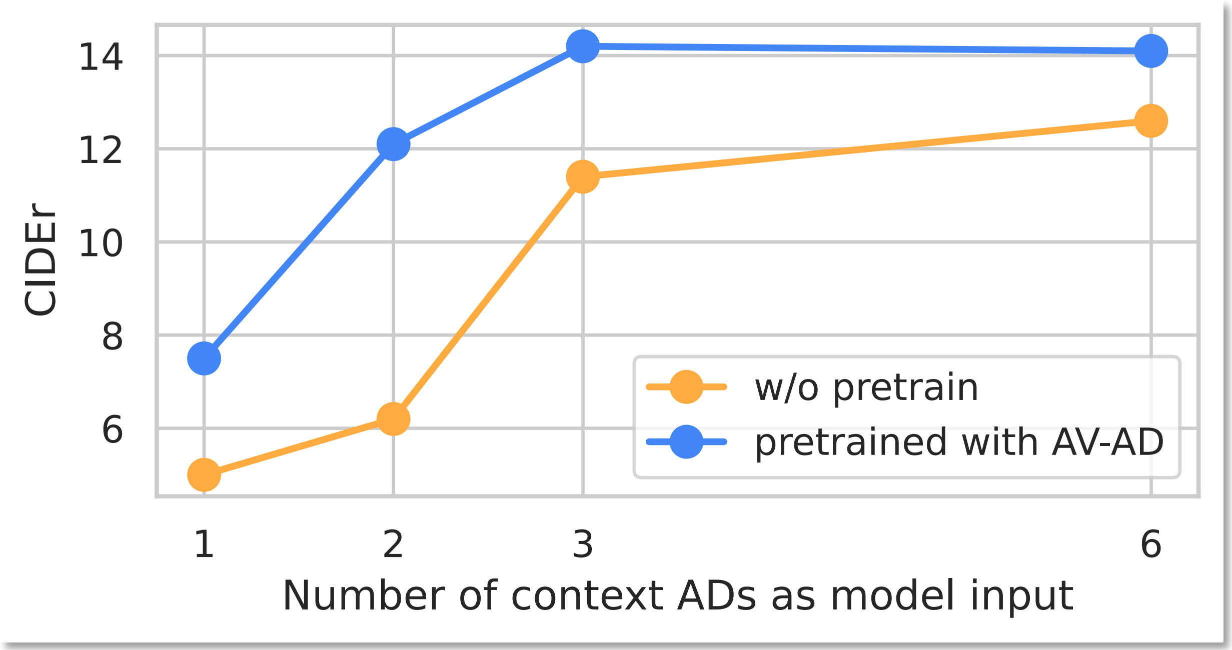


↗ Visual context, AD context is helpful

↗ Partial-data pretraining is helpful

↘ However, subtitle input does not help

Results: length of AD context



Qualitative Results

Context AD: Professor Snape approaches behind Harry. Snape takes Harry down to his storeroom. Snape raises his wand. Harry body goes rigid.

Ground-truth AD: His mind fills with terrifying memories.

Prediction: His eyes widen.

Context AD: Lovejoy walks alongside Jack and slips the heart of the ocean into Jack's coat pocket...The steward removes Jack's coat, while the master-at-arms frisks him.

Ground-truth AD: The steward pulls the necklace from the pocket.

Prediction: He takes the necklace and puts it in his pocket.

Context AD: Surrounded by gushing fountains and ornamental palms, they look up at the house. Gatsby looks at Daisy framed by the fountain. It's an orange-squeezing machine.

Ground-truth AD: Daisy Gatsby and Nick swim on his private beech.

Prediction: A man swims in the pool.

Samples from Harry Potter and the Order of the Phoenix (2007), Titanic (1997), The Great Gatsby (2013)

Achievements and Limitations

😀

- Define the AD generation task – meaningful to the visually impaired
- Collect and denoise datasets for AD
- Propose models and training methods

🤔

- Cannot reference character names
- Does not tackle the task of “when” to generate AD
- Fine-grained scene understanding and verb recognition needs improvement

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

École des Ponts
ParisTech

Thank you!

Project page: https://www.robots.ox.ac.uk/~vgg/research/autoad/