Using Video-Language models to Generate Audio Descriptions for Movies

Andrew Zisserman

January 2024

CVPR 2023, ICCV 2023: Research with Max Bain, Tengda Han, Arsha Nagrani, Gül Varol, Weidi Xie

Introduction: Video Understanding

Buster Keaton `Cops' (1922)

- What is happening in the video?
  - Who is in the video?
  - What are they doing?
  - What is the scene?
  - Where is it?
- What is the story?

What is movie Audio Description?

- Narration describing visual elements in the movie to aid the visually impaired

Movie clip from
`Out of Sight’ (1998)
with Audio Description

What is movie Audio Description (AD)?

Narration describing visual elements in movies to aid the visually impaired:

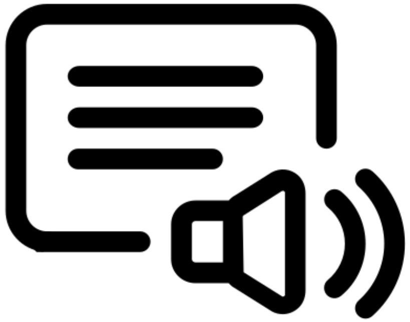
- Complementary to the raw audio track (no need to describe the audio)
- Aim is storytelling: includes character names, emotion, actions, …
- Dense descriptions over time (previous context very important)

- "A Dataset for Movie Description", A Rohrbach, M Rohrbach, N Tandon, B Schiele, CVPR 2015
- Large Scale Movie Description Challenge (LSMDC) 2015-2021

Objective: Automated Audio Description generation

…  … → f →

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



Text to speech

- A new way to evaluate movie understanding abilities
  - Long-form videos; multi-modal; fine-grained recognition

- Societal 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”

  - Available from AudioVault (https://audiovault.net/), provided by volunteers

Train a Visual-Language Model to generate the AD

Video frames

Visual-language model (VLM)

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

- 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

Outline

1. Background on visual language models

- Two types of network architecture using adapters

2. A basic AD model, data, and training

- Adapting pre-trained vision and language models to this task

3. Improving the `who’ in generated AD

- Supplying supplementary information on characters

4. Improving the `what’ in generated AD

- Adapting pre-trained video-language models to this task
- Evaluating performance

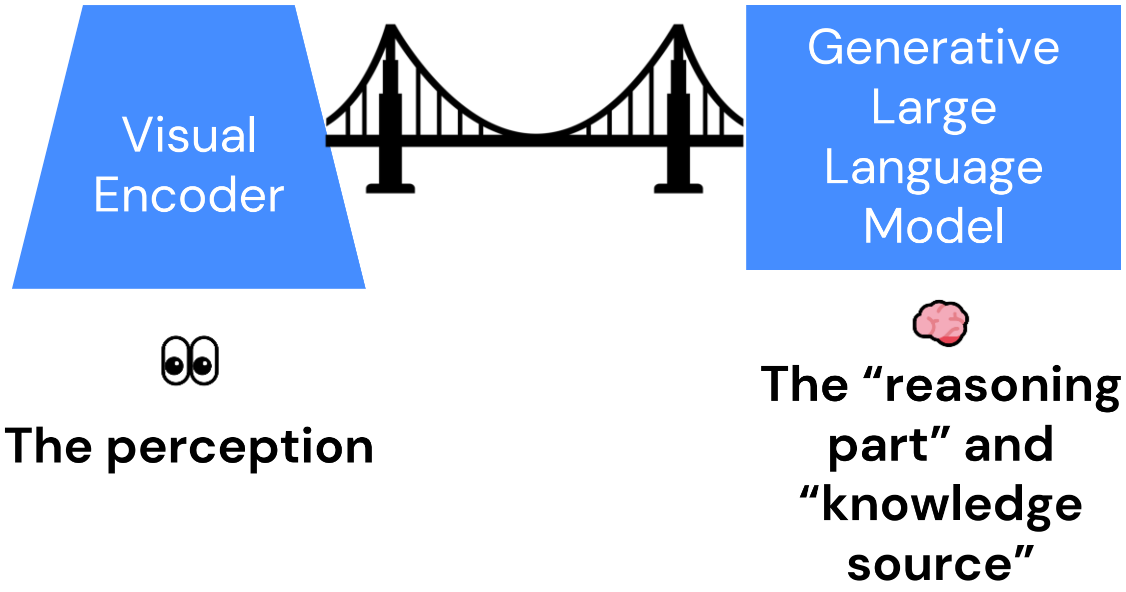
How to splice a visual encoder into a language model

Language model:

- pre-trained transformer decoder (e.g. GPT2 like)
- Pre-trained and Frozen

Visual encoder:

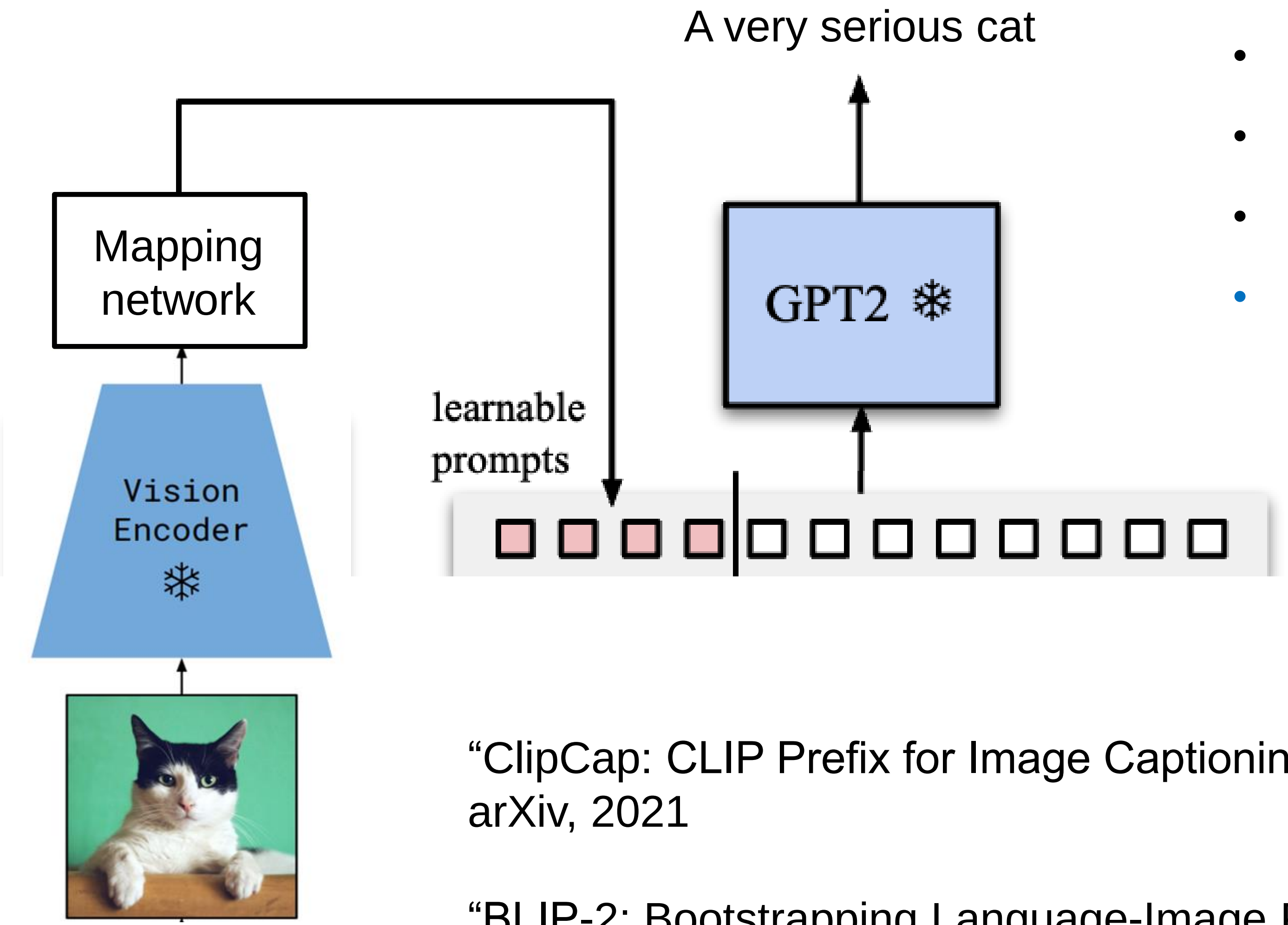
- Can be Convolutional or Visual Transformer (ViT)
- Pre-trained and Frozen



Visually-conditioned Language Models:

- Method 1: Prompt Tuning
- Method 2: Cross-attention

Overview of Architecture 1: Prompt-tuning GPT

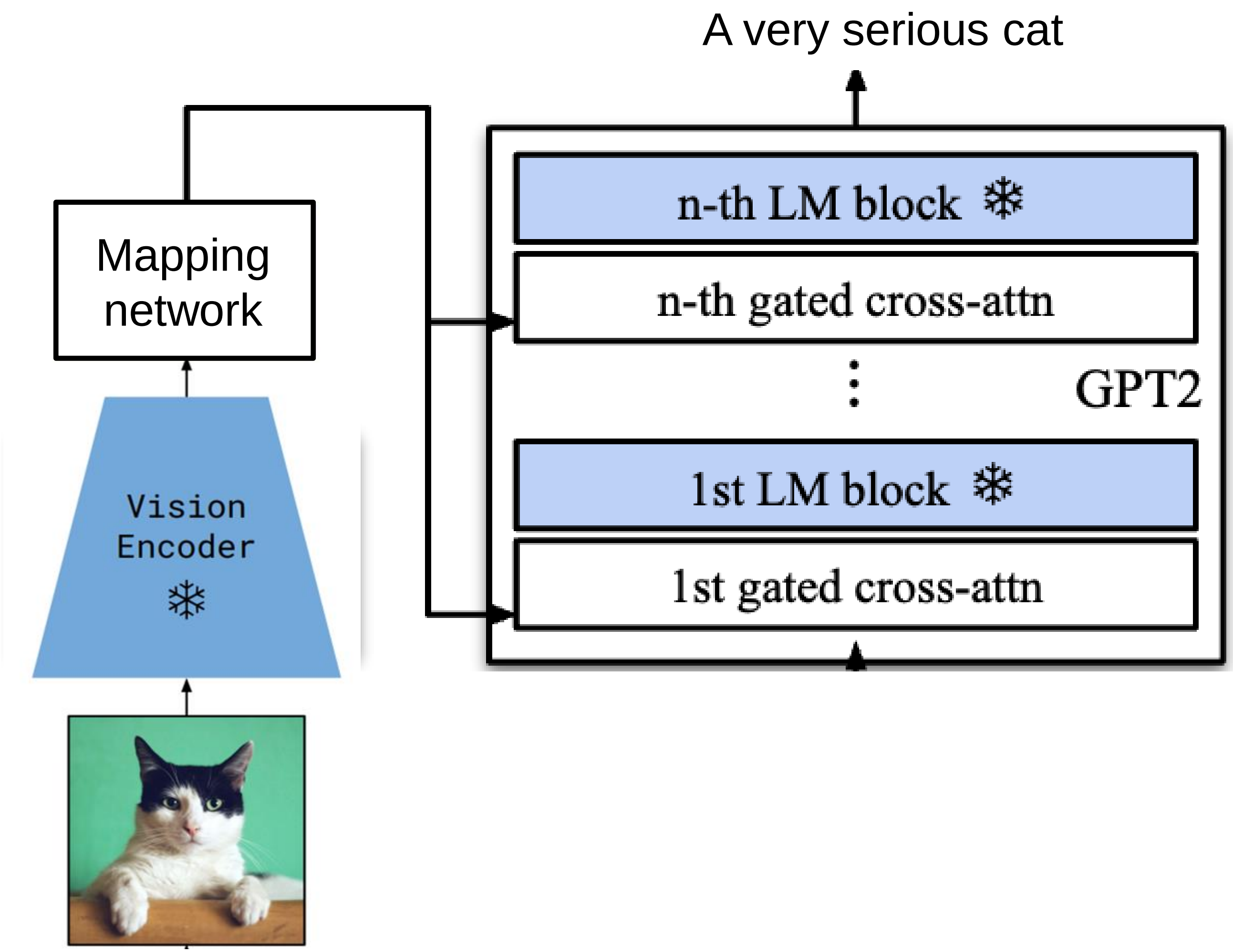


- **Input**: visual data
- **Output**: free form text
- Visual encoder: pretrained and frozen
- Language model (GPT2): pretrained and frozen
- Only the mapping network (adapter) is trained

"ClipCap: CLIP Prefix for Image Captioning", Ron Mokady, Amir Hertz, Amit H. Bermano, arXiv, 2021

"BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models", Junnan Li, Dongxu Li, Silvio Savarese, Steven Hoi, arXiv, 2023

Overview of Architecture 2: Cross-attention (X-Attn) GPT

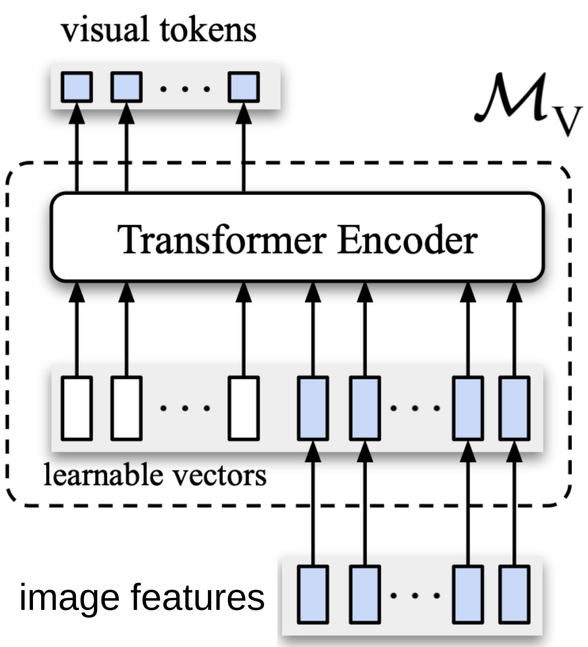


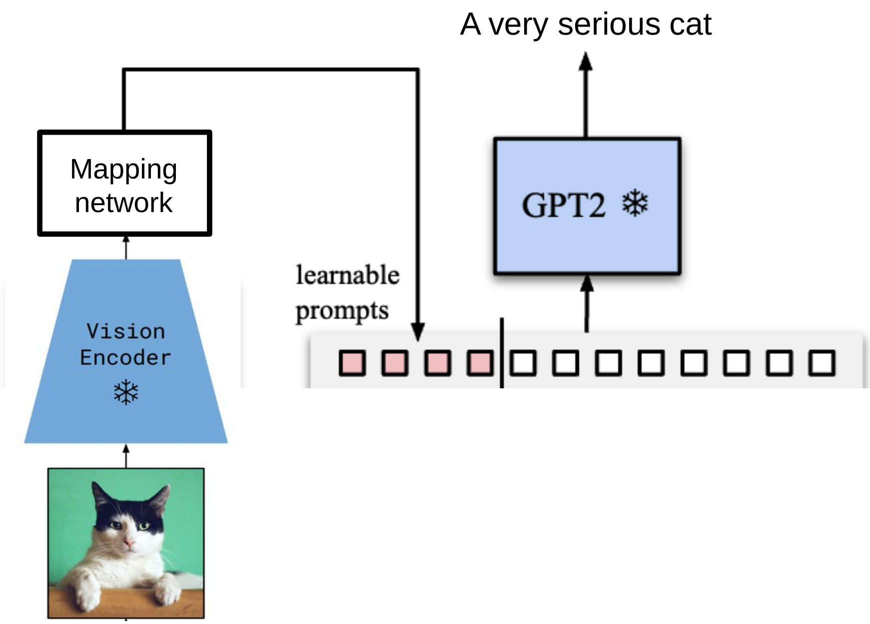
- **Input**: visual data
- **Output**: free form text
- Visual encoder: pretrained and frozen
- Language model (GPT2): pretrained and frozen
- Only the mapping network and X-Attn layers (adapters) are trained

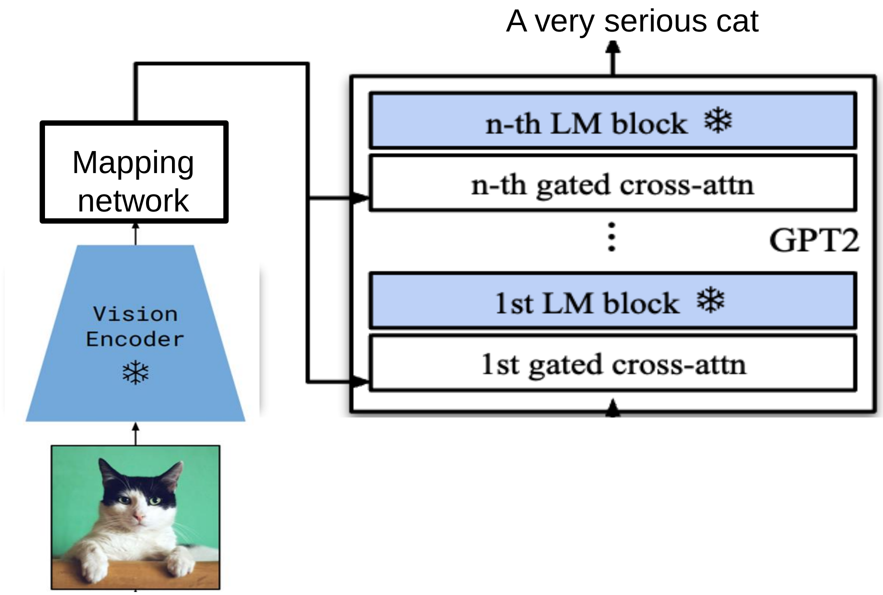
Flamingo: a Visual Language Model for Few-Shot Learning, Jean-Baptiste Alayrac et al, NeurIPS 2022

Training choices …

- Use discrete (VQ) or soft tokens for the LLM
- Architecture for mapping network, e.g.
  - Simple linear mapping
  - Transformer encoder





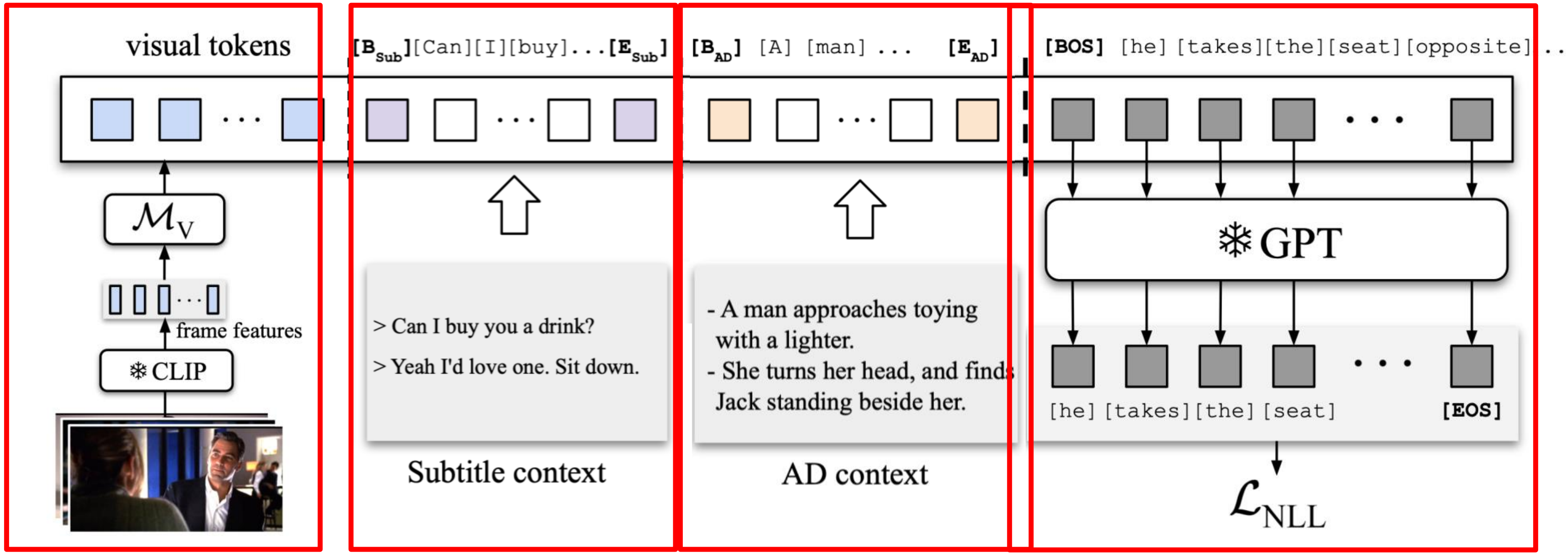


- Training end-to-end

2. A basic Audio Description model

Architecture – Prompt tuning GPT

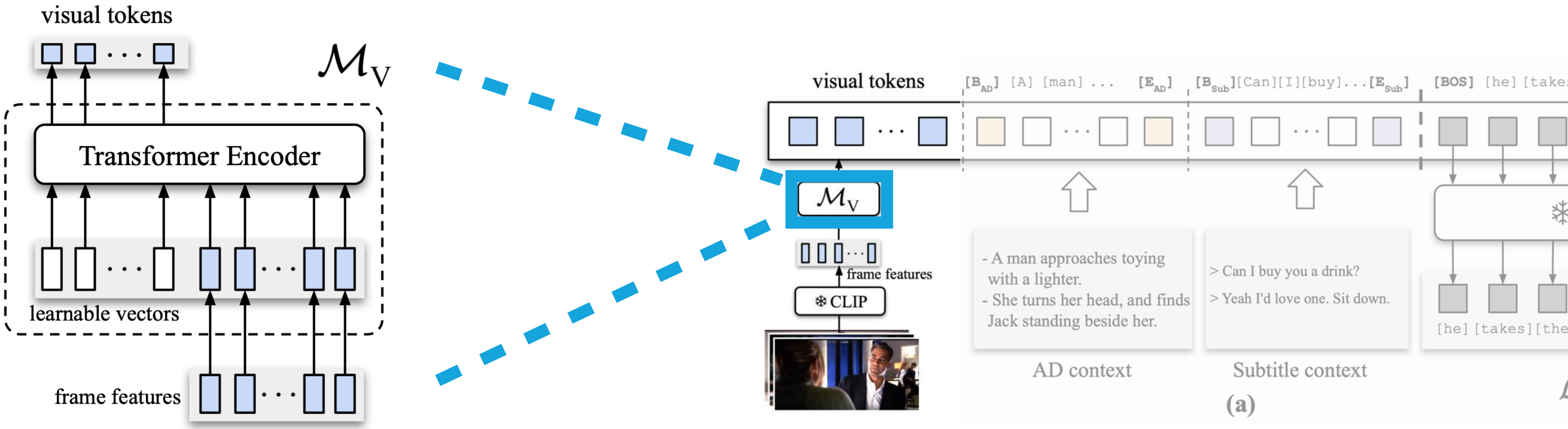
Video Captioning with Long Multimodal Context



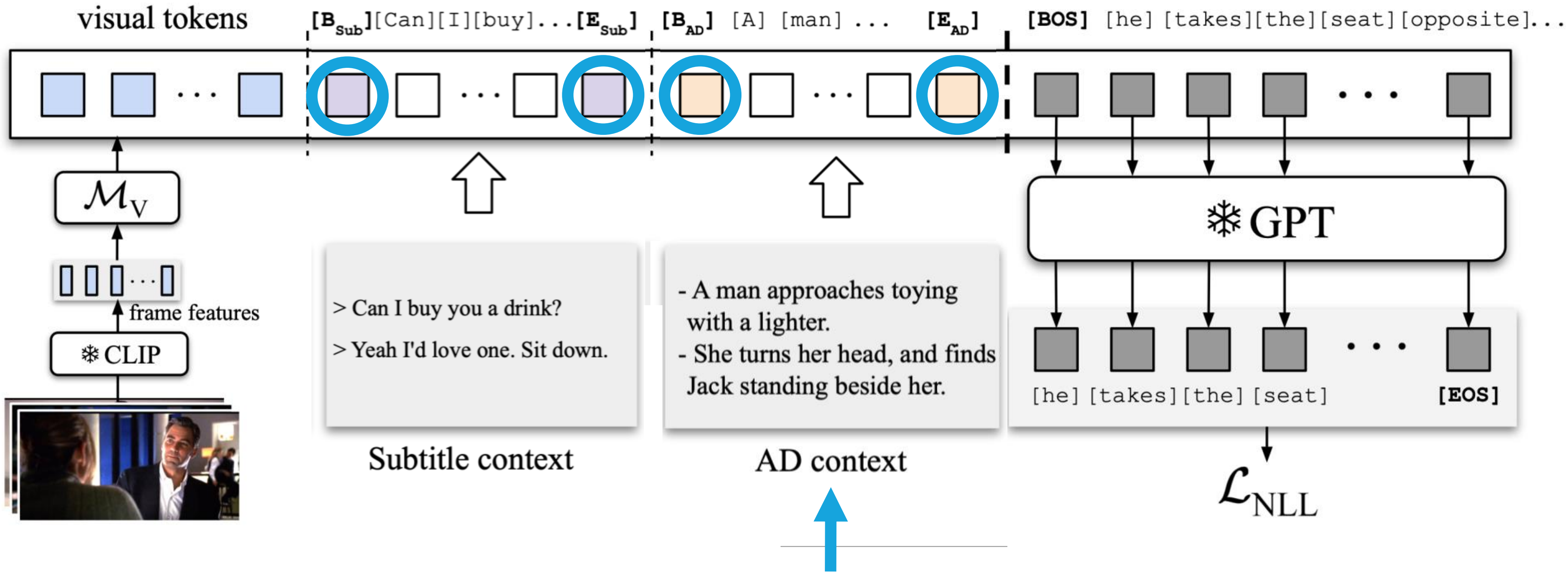
- Pretrained GPT for text generation
- All conditioning added as prompting vectors
  - Visual features (CLIP), movie subtitles, contextual AD

ClipCap, BLIP

Architecture – Prompt tuning GPT



Architecture – Prompt tuning GPT

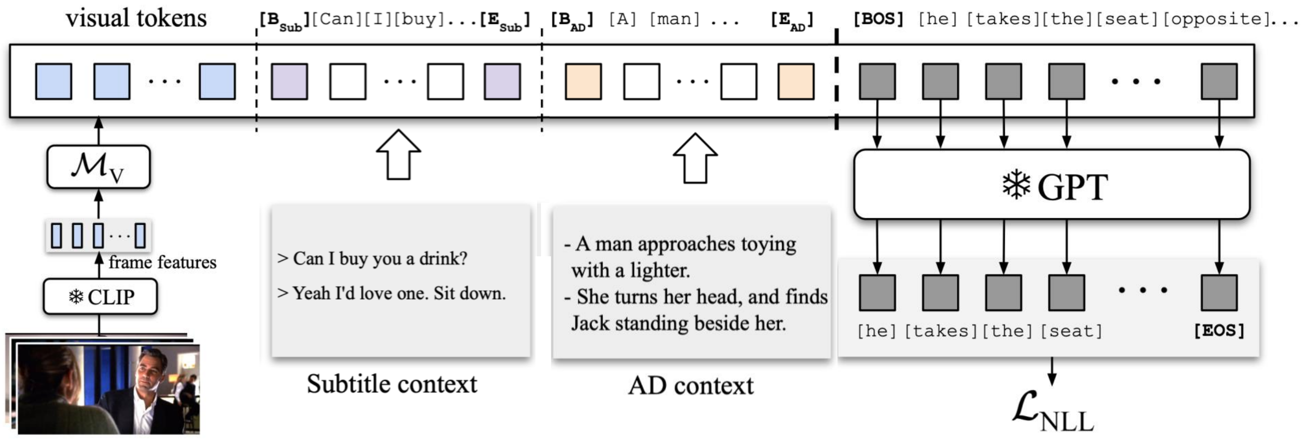


N previous captions

Training

Labelled Data (MADv2)

Model



- 488 Movies
- 316K AD captions
- 900 hrs video provided as CLIP features at 5 fps

`Complete’ movie dataset: video features, AD, subtitles

MAD: A Scalable Dataset for Language Grounding in Videos From Movie Audio Descriptions, Soldan et al, CVPR 2022

Challenge: the lack of training data

- Movie data with corresponding visual, subtitles and description elements are very limited in size
- MAD dataset has only 316k clips with AD
- Use this for an evaluation (testing) split of 10 films, but too small to adapt foundation model to AD task
- But there are several large datasets available …

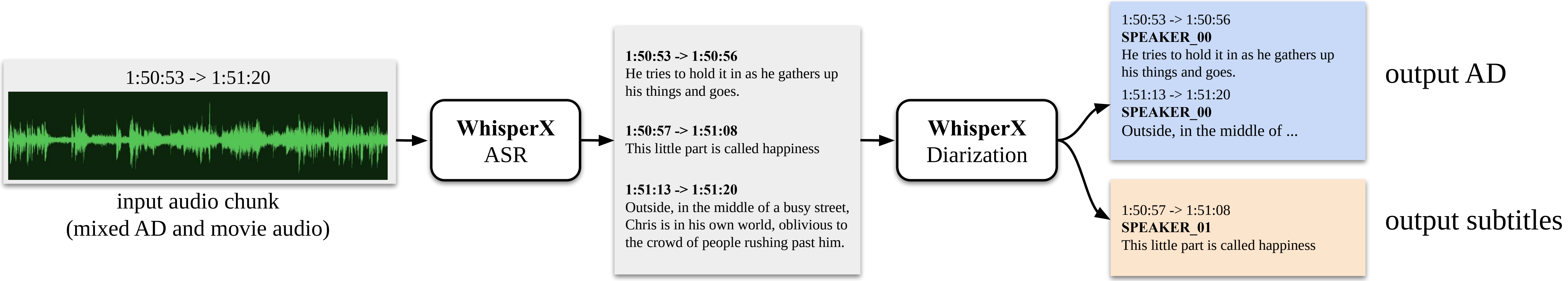
Pretrain with Partial Data

- Available large scale datasets:
  - Paired visual-textual data (without temporal context): CC3M, WebVid
  - Movie AD data (without visual information): downloaded from AudioVault
- Use partial data to pretrain particular modules from large-scale datasets
- And then finally finetune the entire architecture with the complete movie dataset (MAD training)

AudioVault Dataset

Audio soundtracks for 7000+ movies containing the combined original audio and AD audio

- Process this dataset to obtain separate subtitles and AD as text
- How? WhisperX & Diarization

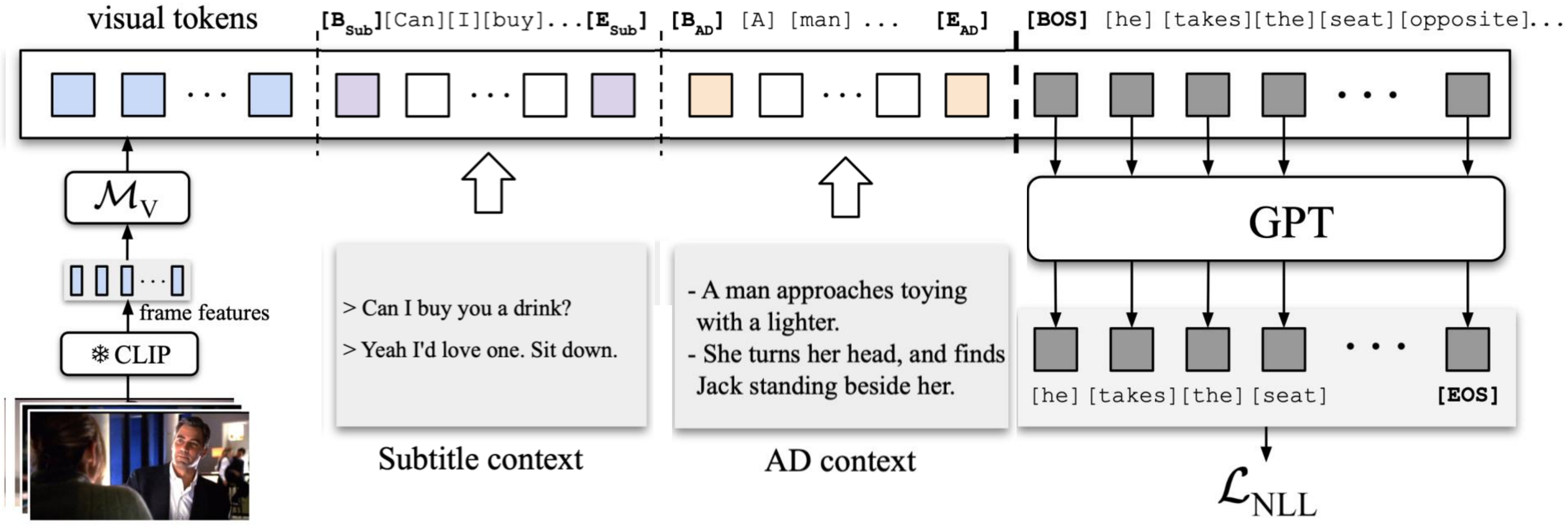


12,000+ hours transcribed

3.3M AD captions

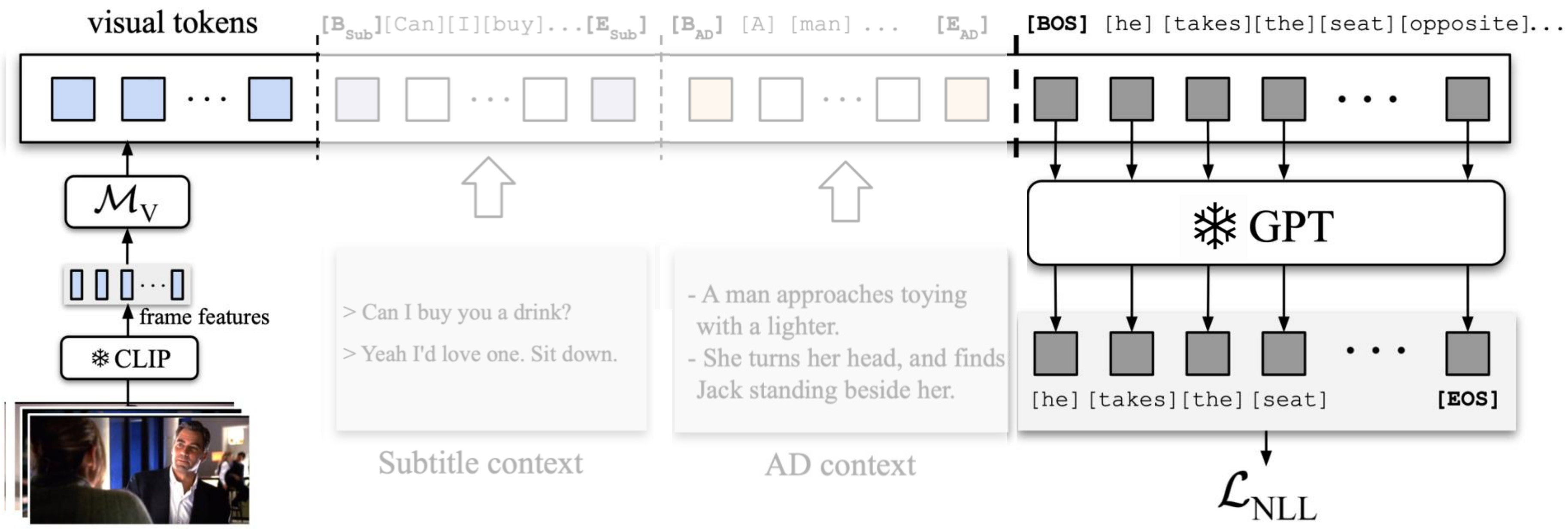
MAD-v2 and AudioVault are publicly released

Partial Pre-training



Partial Pre-training

- Visual captioning

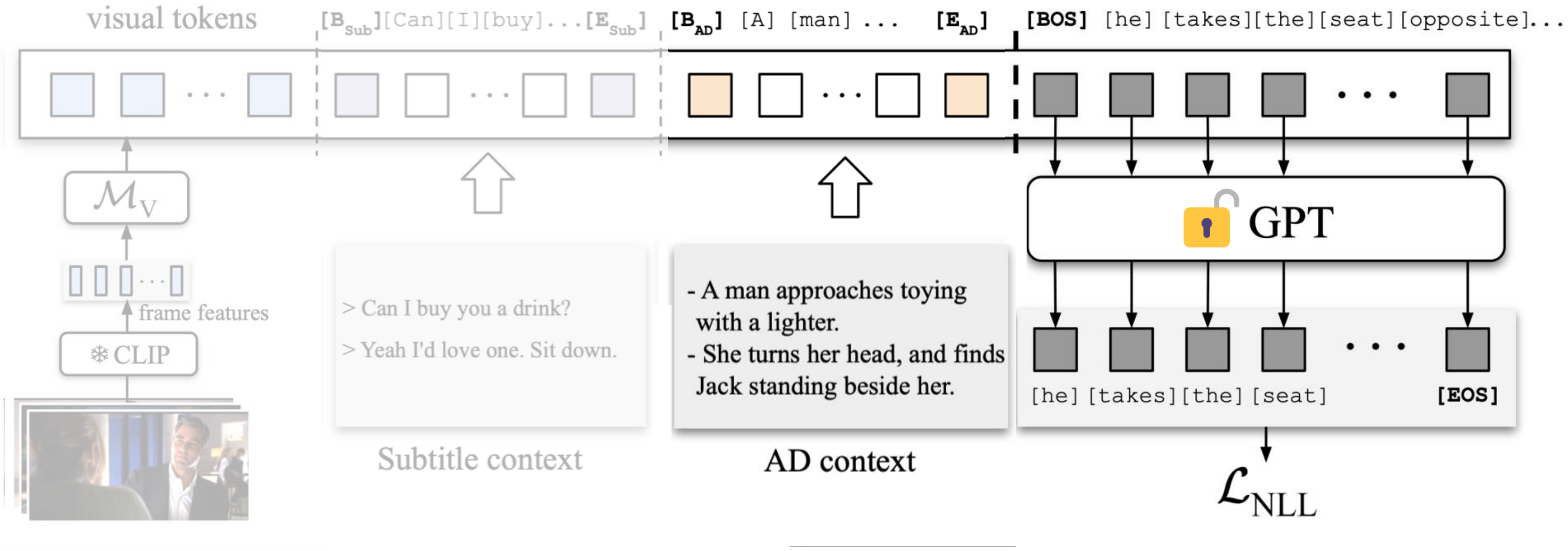


Use paired visual-textual data (without temporal context):

CC3M, WebVid (image-caption, video-caption datasets)

Partial Pre-training

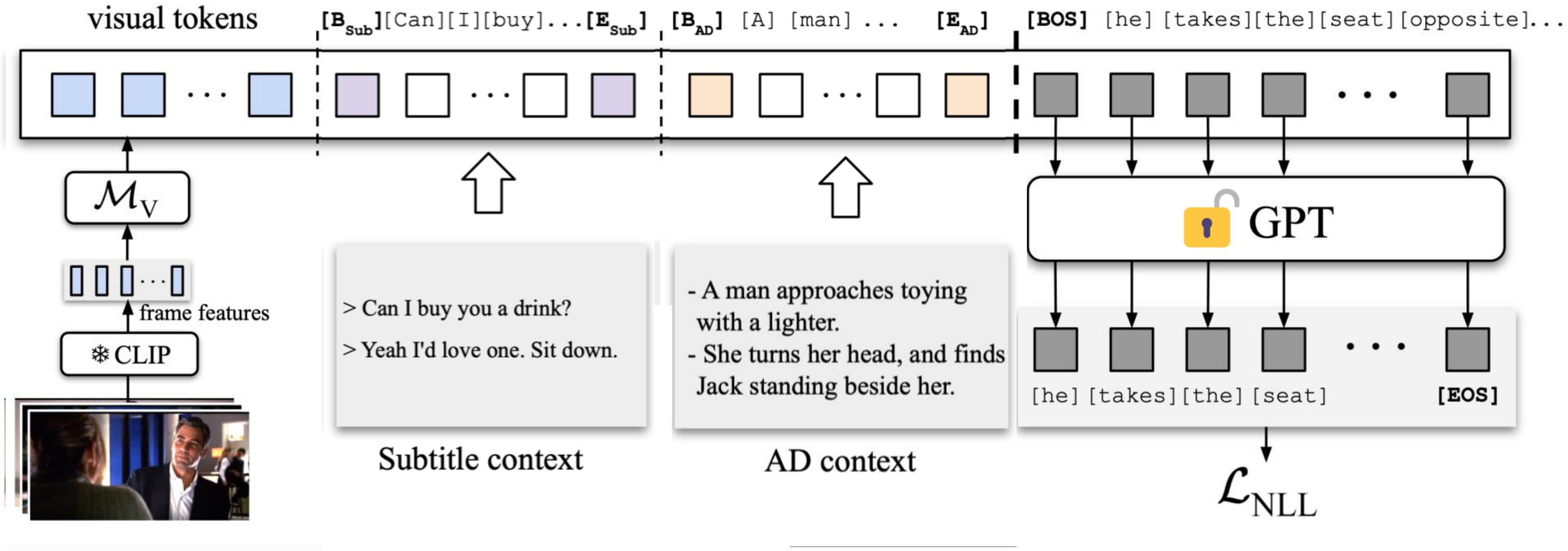
- Text only



Use text Movie AD data (without visual information) from AudioVault

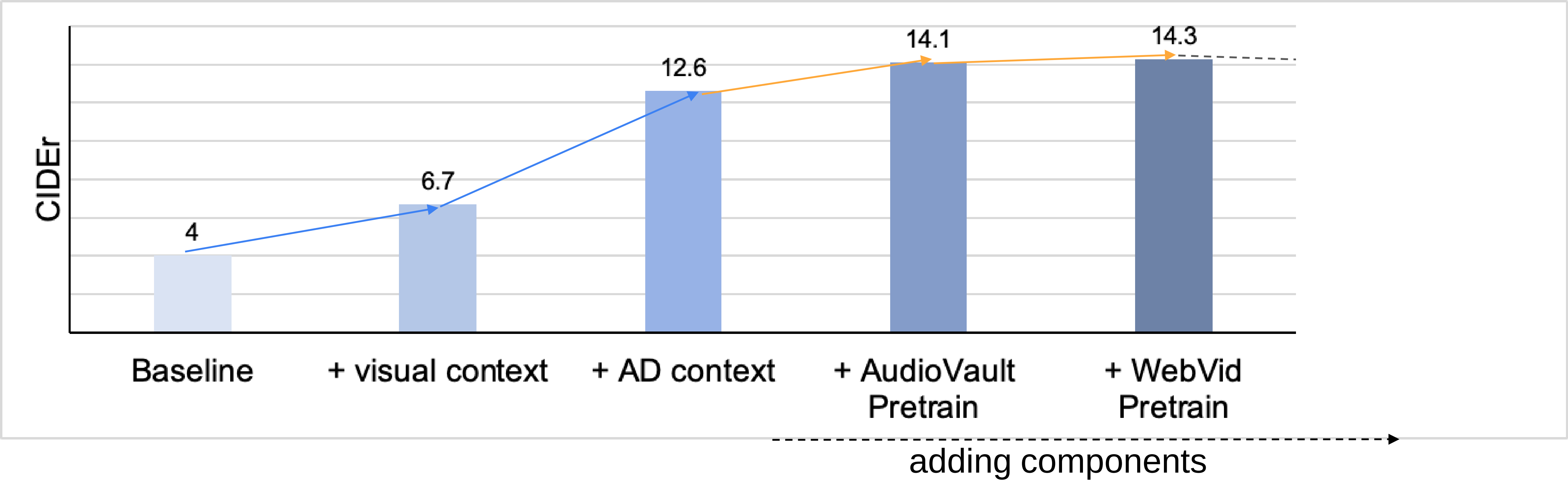
Final fine-tuning

- Complete movie data



Visual features, subtitles, and AD data from MADv2

Results: context and pretraining



CIDEr performance measure:

Scores the similarity between the predicted and actual AD for a clip, over the MAD-Eval test set

Visual context, AD context is helpful

Partial-data pretraining is helpful

However, subtitle input does not help

Qualitative Results

The Great Gatsby (2013)

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.

Harry Potter and the Order of the Phoenix (2007)

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.

Summary point & Limitations

- Developed a Prompt-tuning GPT from pre-trained foundation models using partial training

- Produces AD conditioned on visual frames of current clip and previous AD

- Does not reference character names

- Action and scene descriptions incomplete

The Great Gatsby (2013)

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.

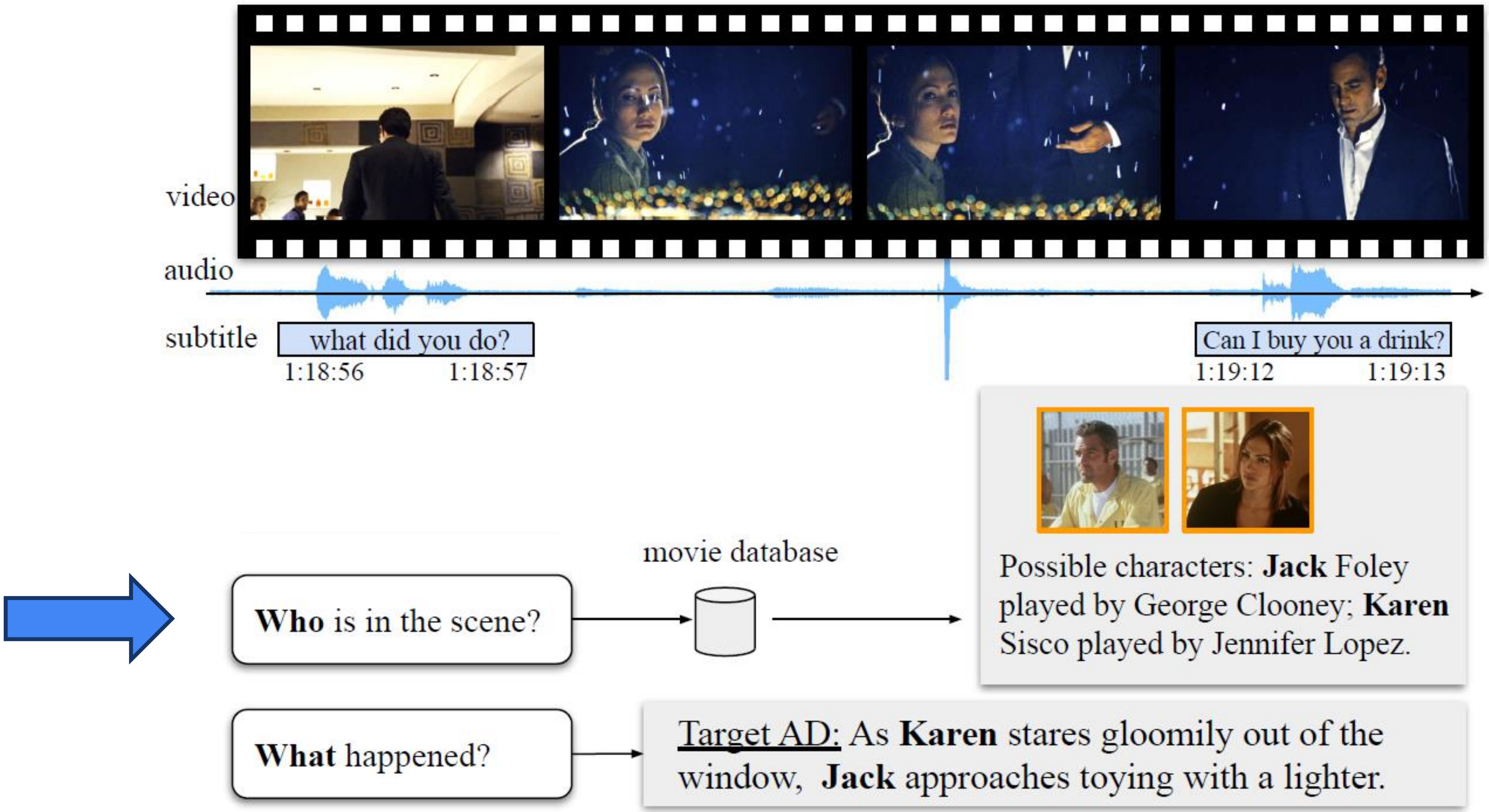
Ground-truth AD: Daisy Gatsby and Nick swim on his private beach.

Prediction: A man swims in the pool.

Outline

1. Background on visual language models
    - Two types of network architecture using adapters
2. A basic AD model, data, and training
    - Adapting pre-trained vision and language models to this task
3. Improving the `who’ in generated AD
    - Supplying supplementary information on characters
4. Improving the `what’ in generated AD
    - Adapting pre-trained video-language models to this task
    - Evaluating performance

[Who] is in the scene?



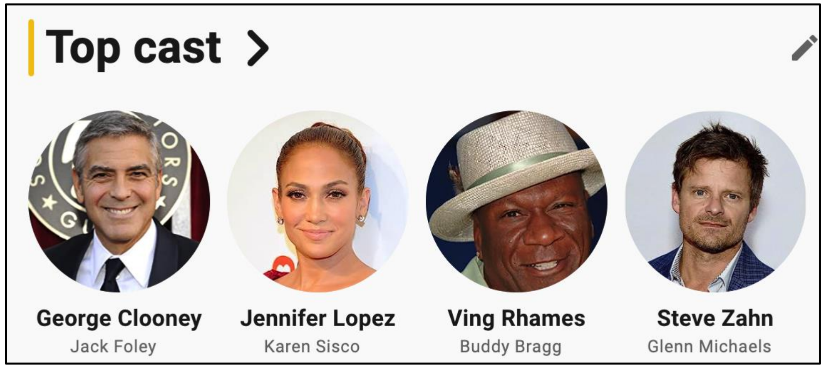
[Who] is in the scene?

- Objective: identify the active characters in a clip using face recognition
- Provide their names and example face images as prompts to aid AD naming
- How to achieve this objective?
  - Supply a "character bank" dataset for each film with names and face images
  - Use the character bank to identify the active characters in a clip

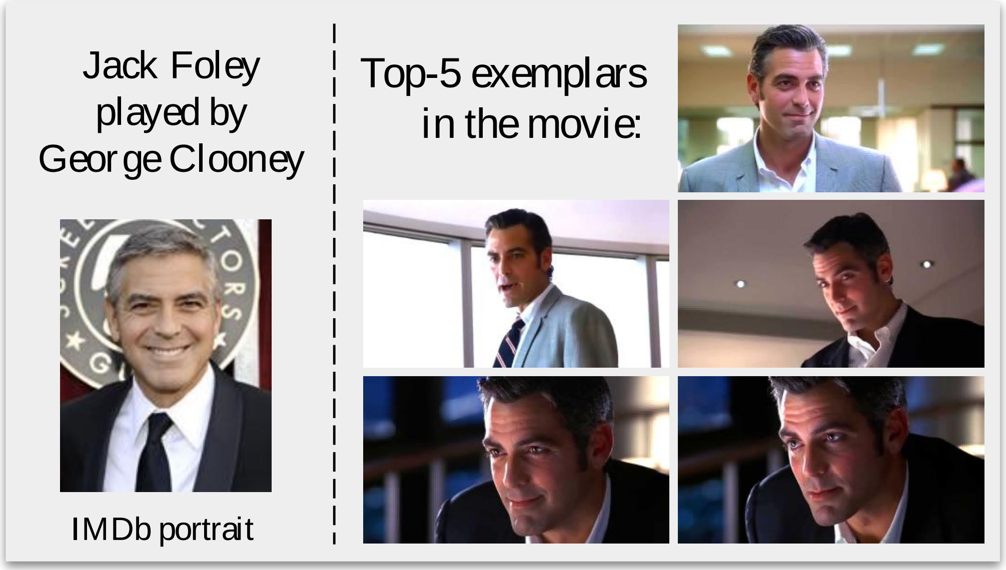
[Who] is in the scene?   Character Bank

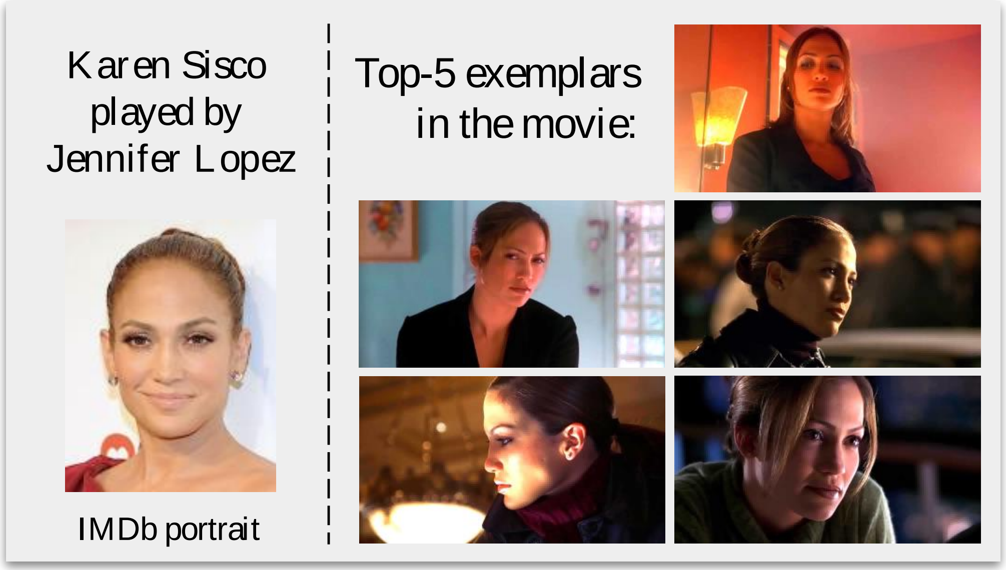
- Use the Internet Movie Database (**IMDb**)
  - provides mapping between characters and **cast** names
  - also provides cast photos

Out of Sight (1998)



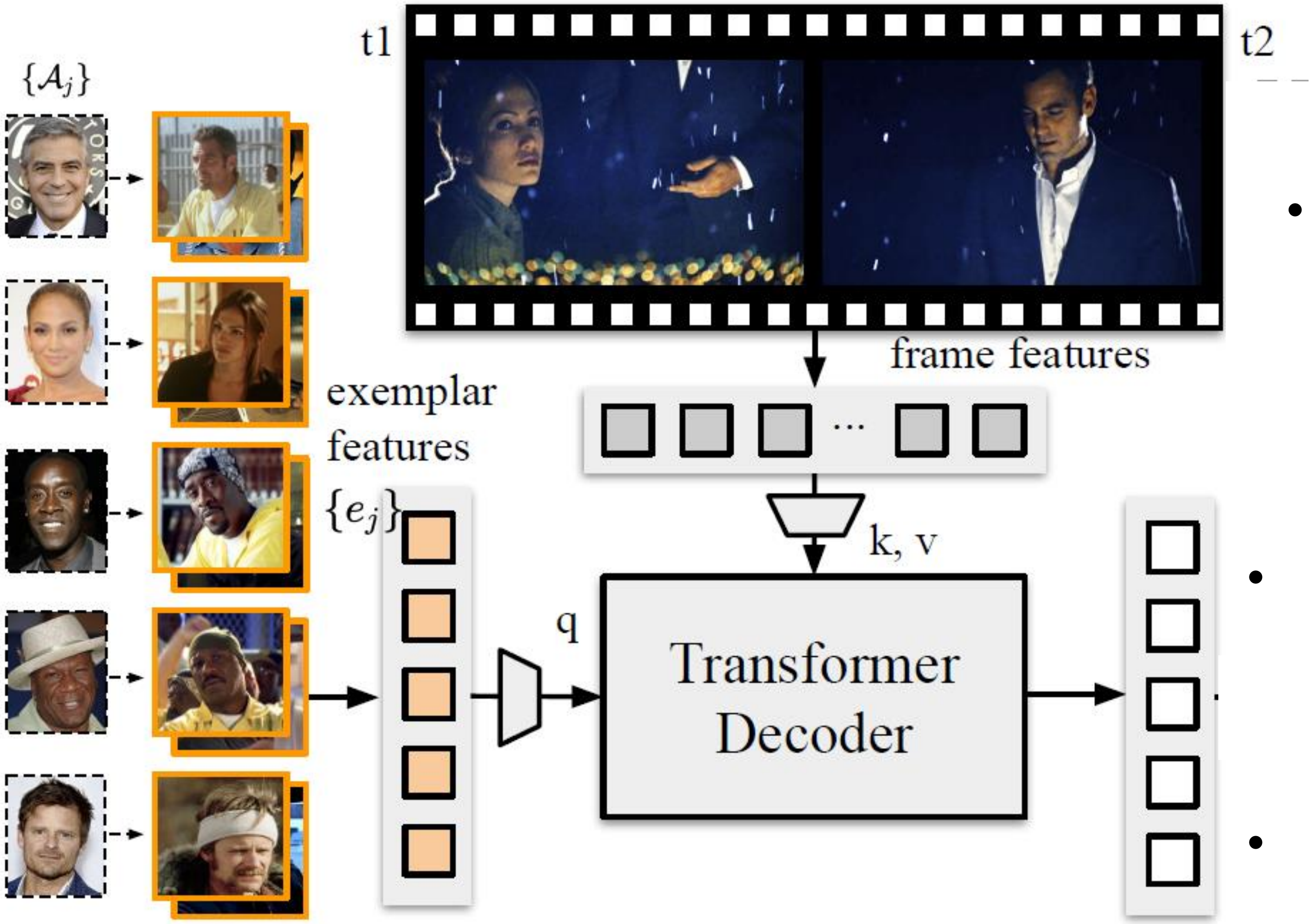
- Use cosine similarity between CLIP features of photos and frames to select visual exemplars for each actor





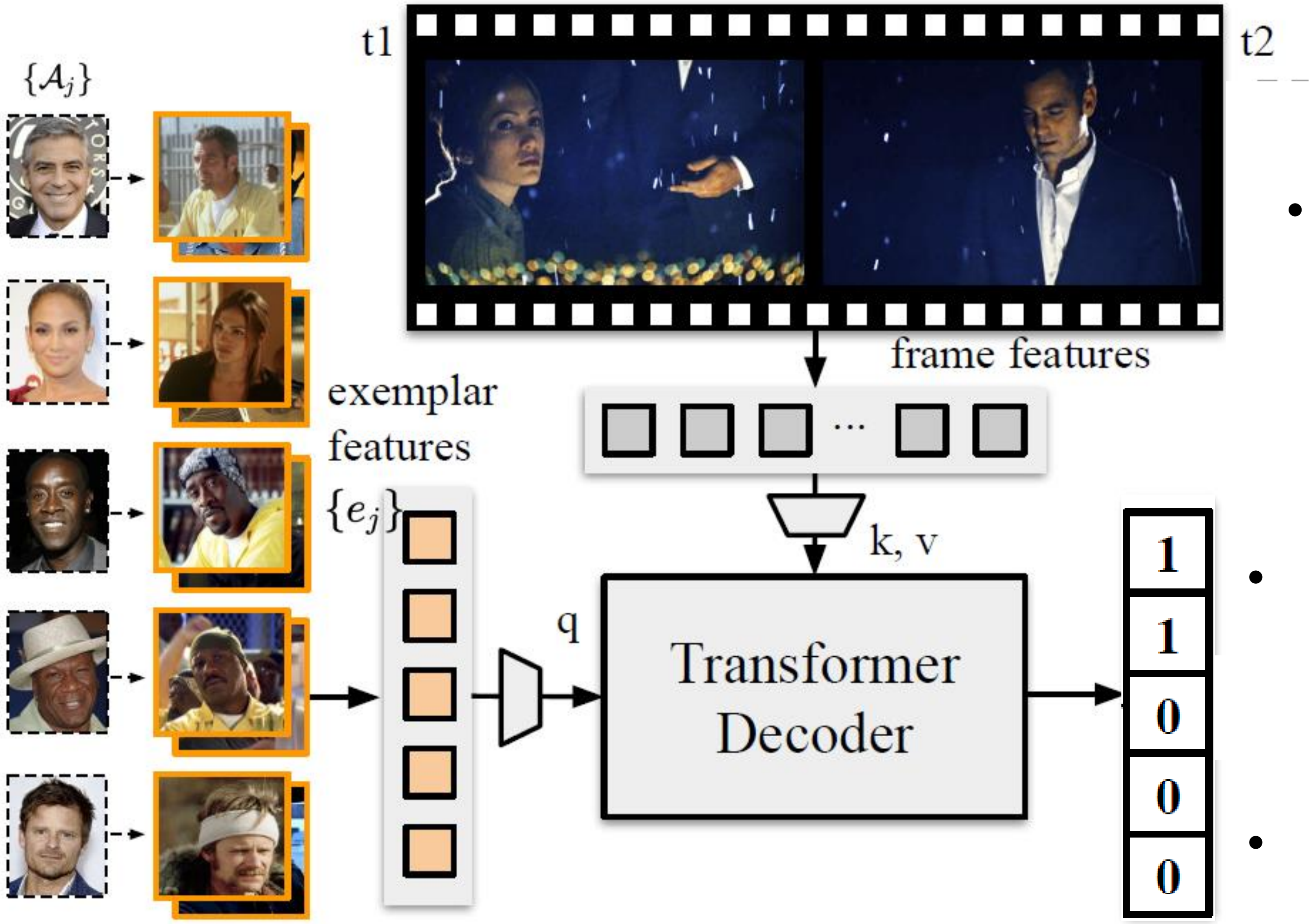
Character recognition module

Classify active characters in video clip



- Allow “face exemplars” to interact with movie frames
- Binary classification whether each character present in clip or not
- All features computed using frame-level CLIP

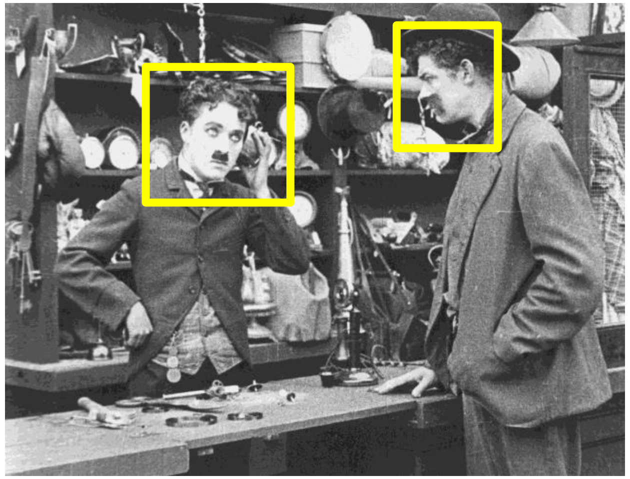
Character recognition module



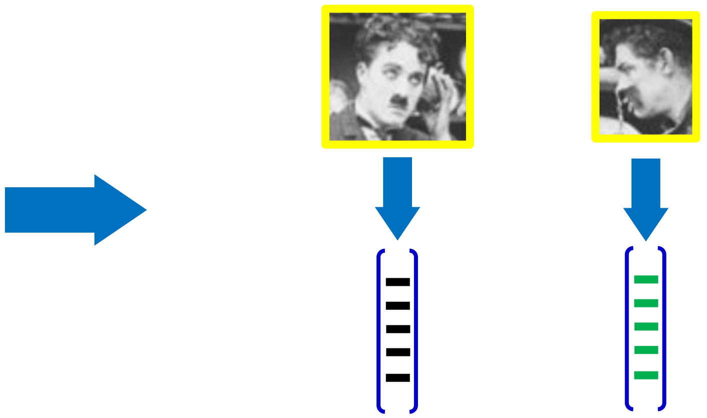
- Allow “face exemplars” to interact with movie frames
- Binary classification whether each character present in clip or not
- All features computed using frame-level CLIP

An aside: standard face recognition approach

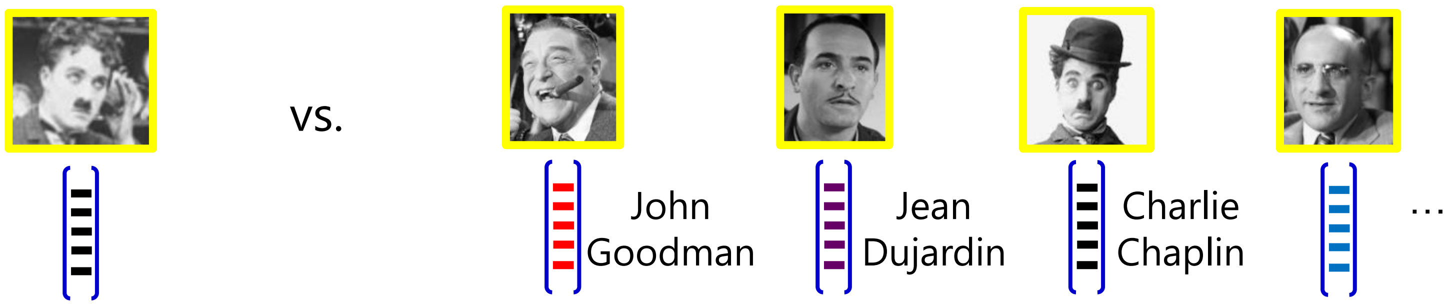
1. Detect Faces



2. Represent each face by a vector

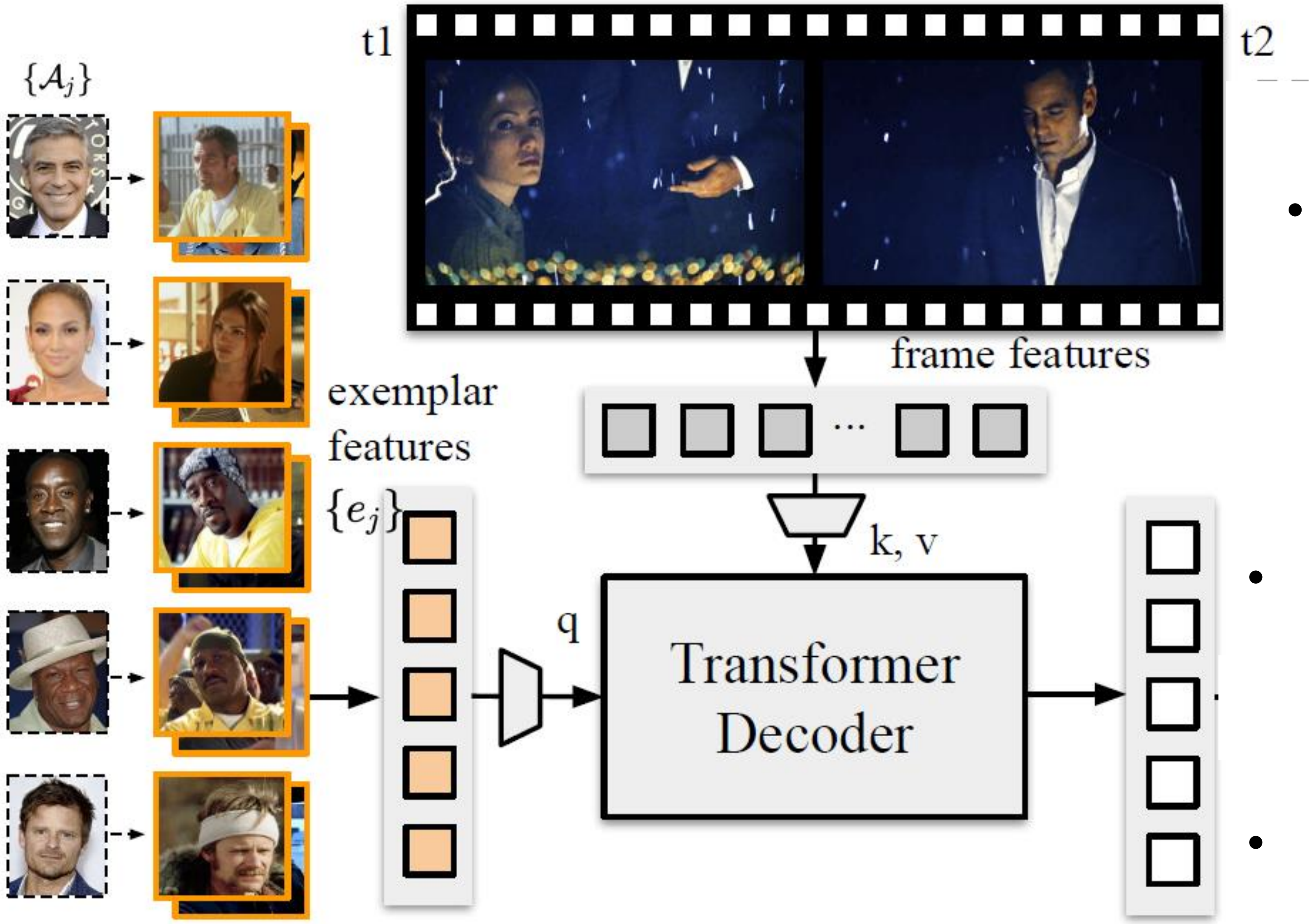


3. Recognize a face from a gallery using closest distance between vectors



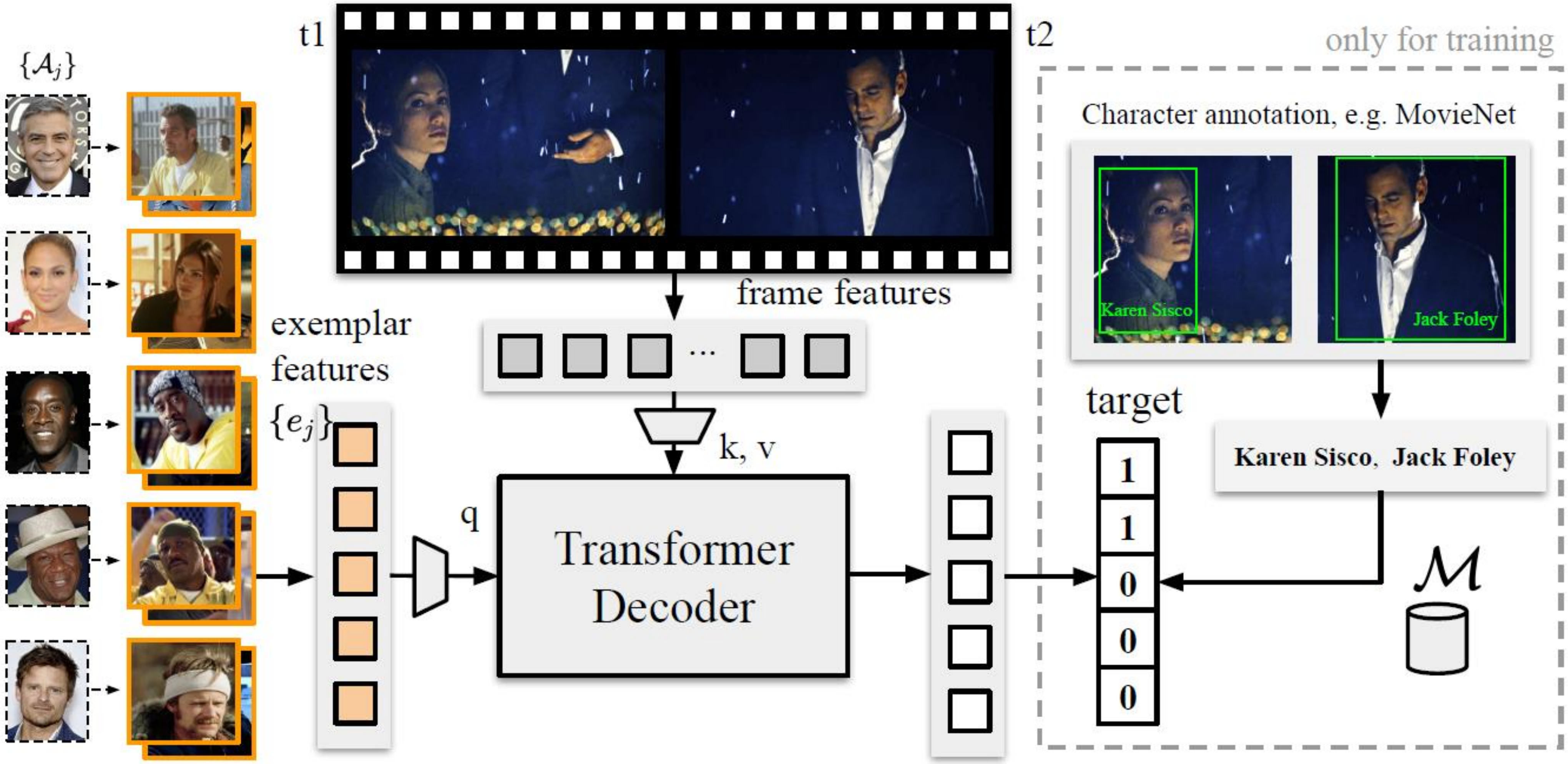
For this to work, need vectors to only represent **identity**, and not be affected by expression, pose, lighting, age, etc. Vectors obtained by deep network trained for identity

Character recognition module



- Allow “face exemplars” to interact with movie frames
- Binary classification whether each character present in clip or not
- All features computed using frame-level CLIP

Character recognition module

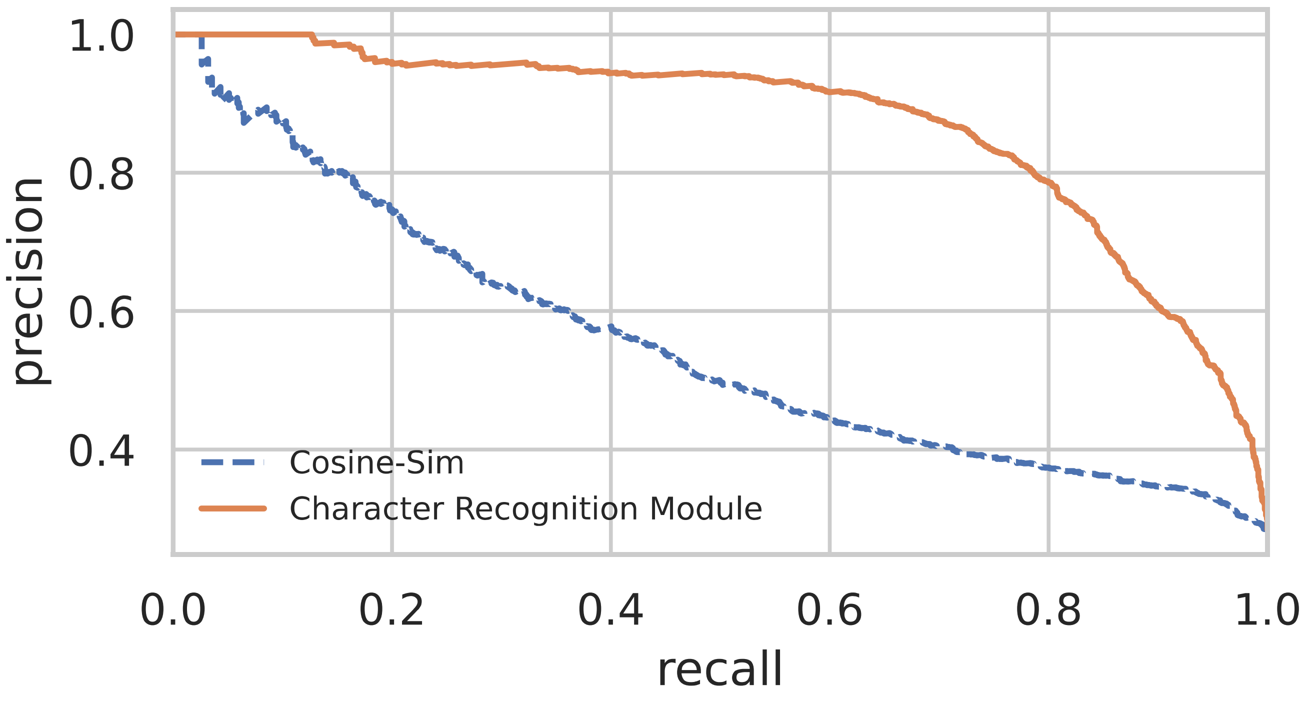


- Train <u>character recognition module</u> using MovieNet annotations

Results of [Who]

Task: predict active characters in movie clip

- Baseline: CLIP feature cosine-similarity between face exemplars and frames

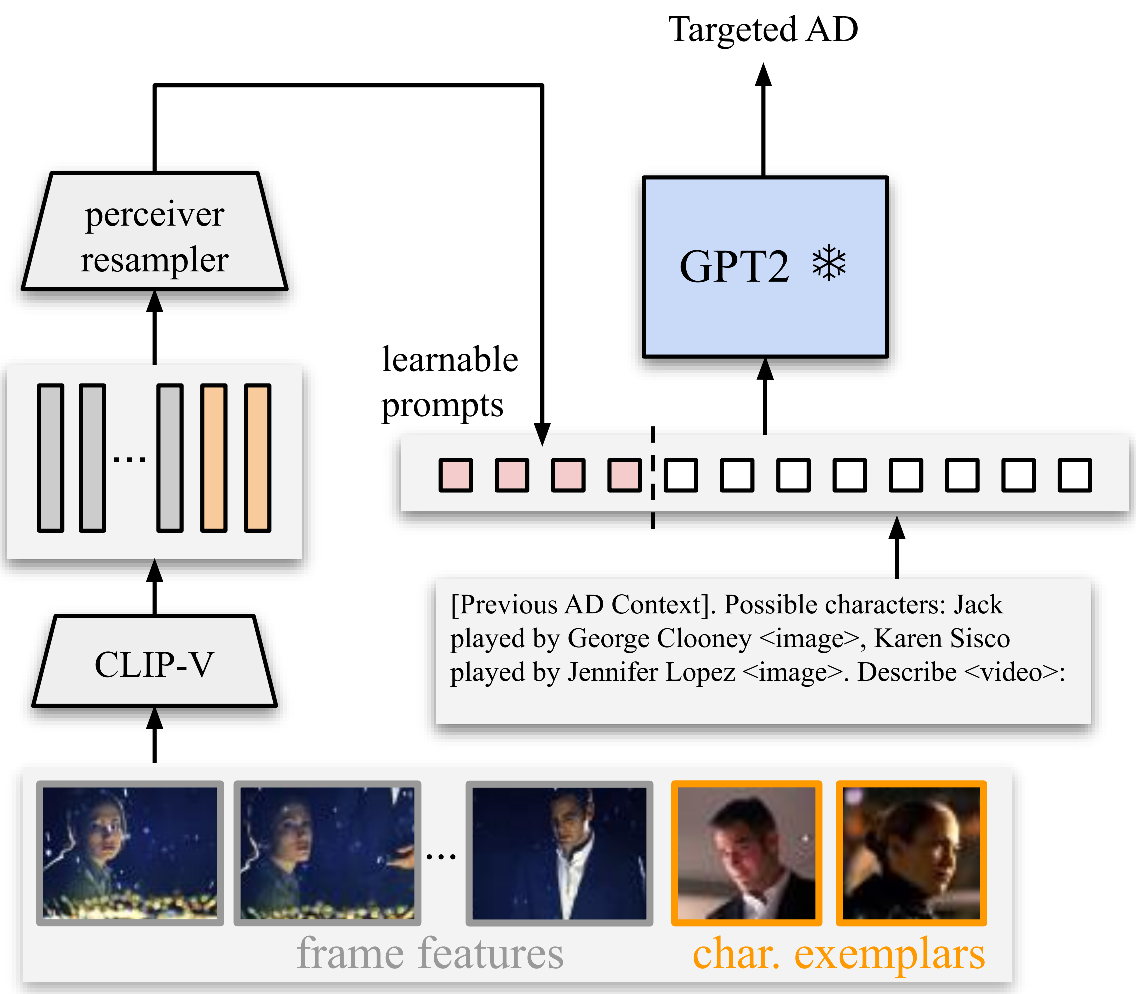


Character Bank Use for AD

For each new film:

- Download principal character data from IMDb (character names, actor names, portrait images)
- Obtain in-domain exemplars for each character as the character bank for that film
- Process all clips to determine “active characters”
- Provide active character names and visual exemplars as prompts for the AD for that clip
- No further training required

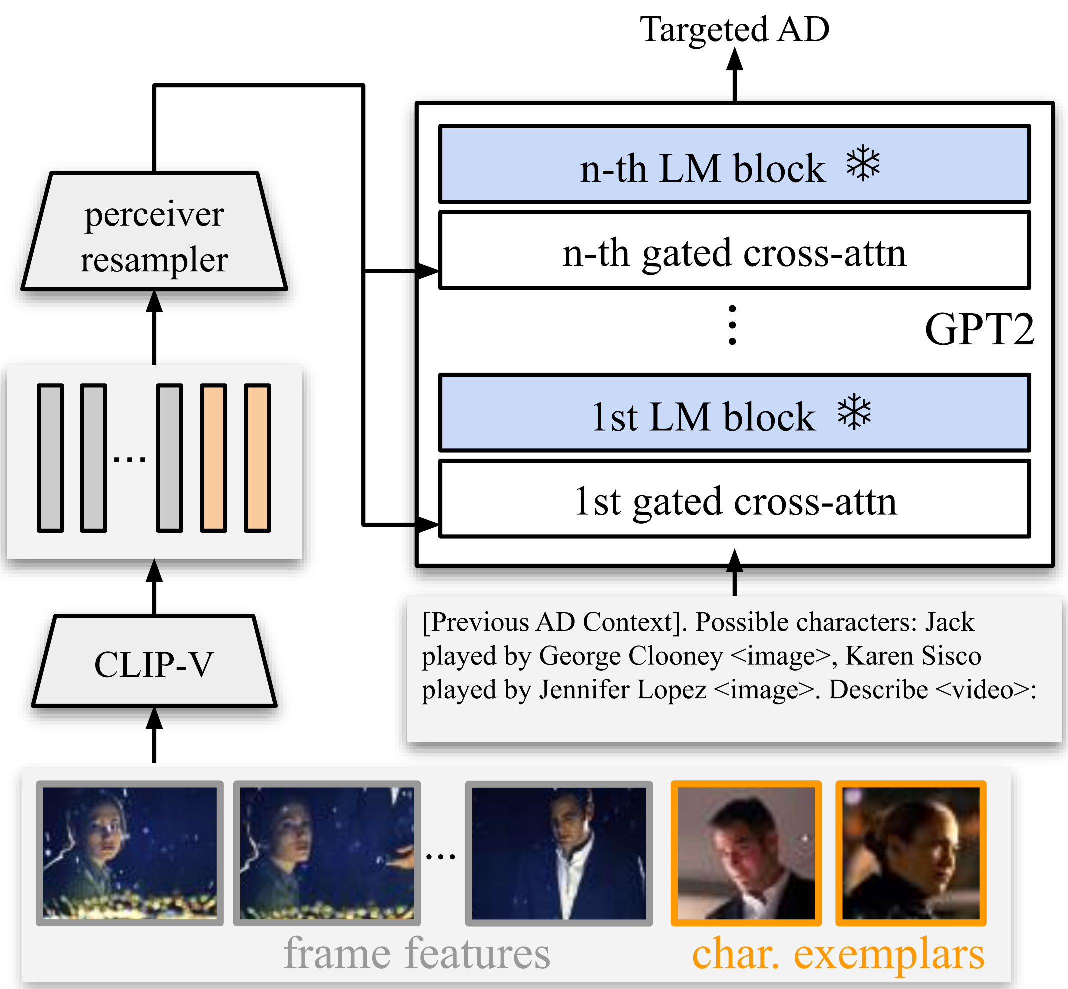
Generate AD with names – Prompt-tuned GPT



- Active **character names** are fed into the model as <u>text prompts</u>
- Also provides the **actor names**
- Also provides the **visual exemplars**

ClipCap, BLIP

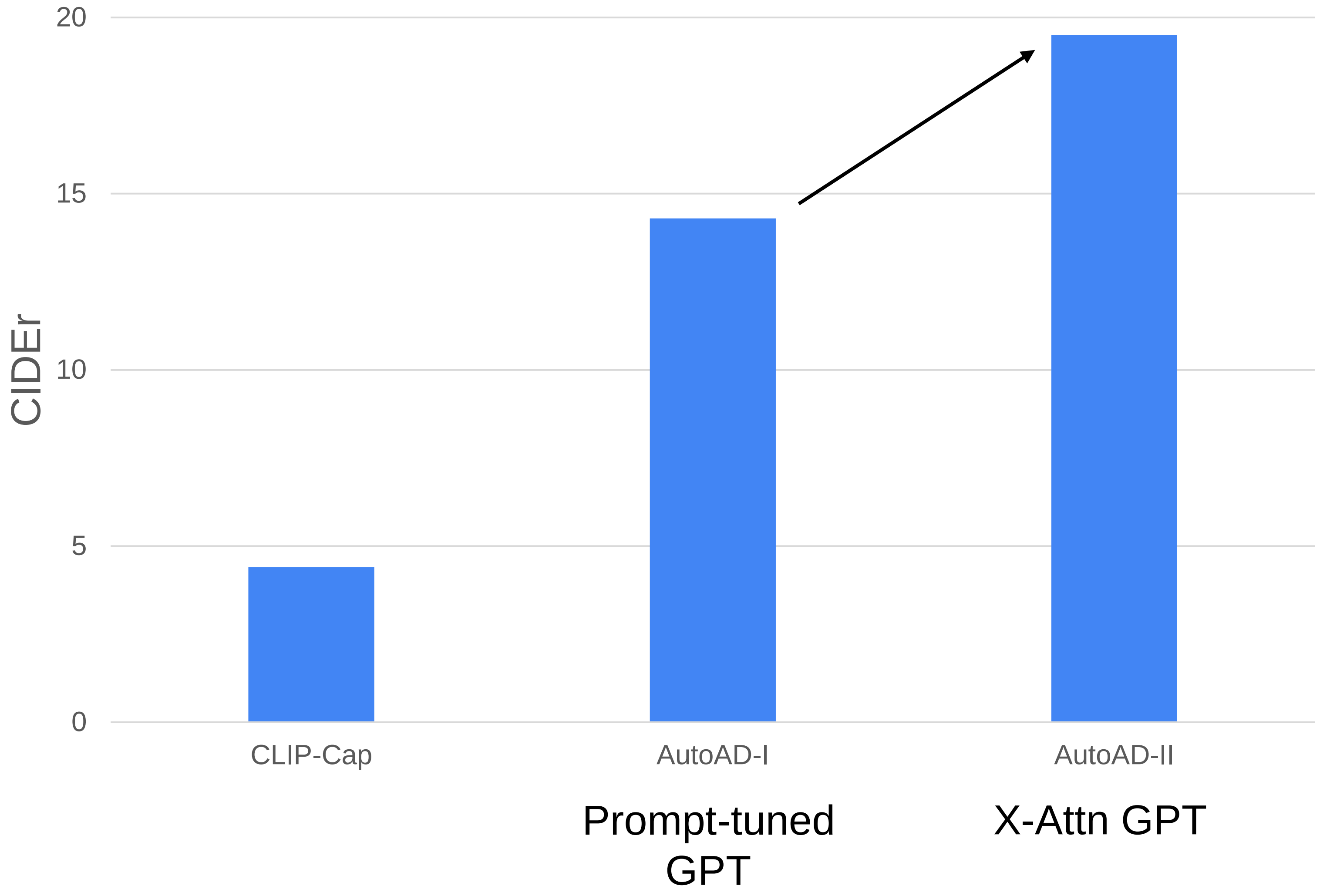
Generate AD with names – X-Attn GPT



- Active **character names** are fed into the model as <u>text prompts</u>
- Also provides the **actor names**
- Also provides the **visual exemplars**
- Use X-Attn to bring in more interactions

Flamingo

Architecture performance comparison



MAD-Eval test set

Example 1: result on `Harry Potter and the Order of the Phoenix’ (not part of training)

Predicted Audio Description: “Snape points at Harry. Harry’s eyes close in horror”

Example 2: result on `Harry Potter and the Order of the Phoenix’ (not part of training)

Predicted Audio Description:

“Hermione, Ron and Luna’s eyes are fixed on Harry, who is standing in the doorway. Harry rides on the horse's back as the horse rears up in the air.”

Summary point & Limitations

- \`Who’ performance improved significantly by introducing a Character Bank

Ides of March (2011)

Prediction: Later, Stephen walks down the street with his hands in his pockets

- \`What’ performance limited by CLIP frame descriptors

- Evaluation measures (e.g CIDEr, Bleu) not fit for purpose

Outline

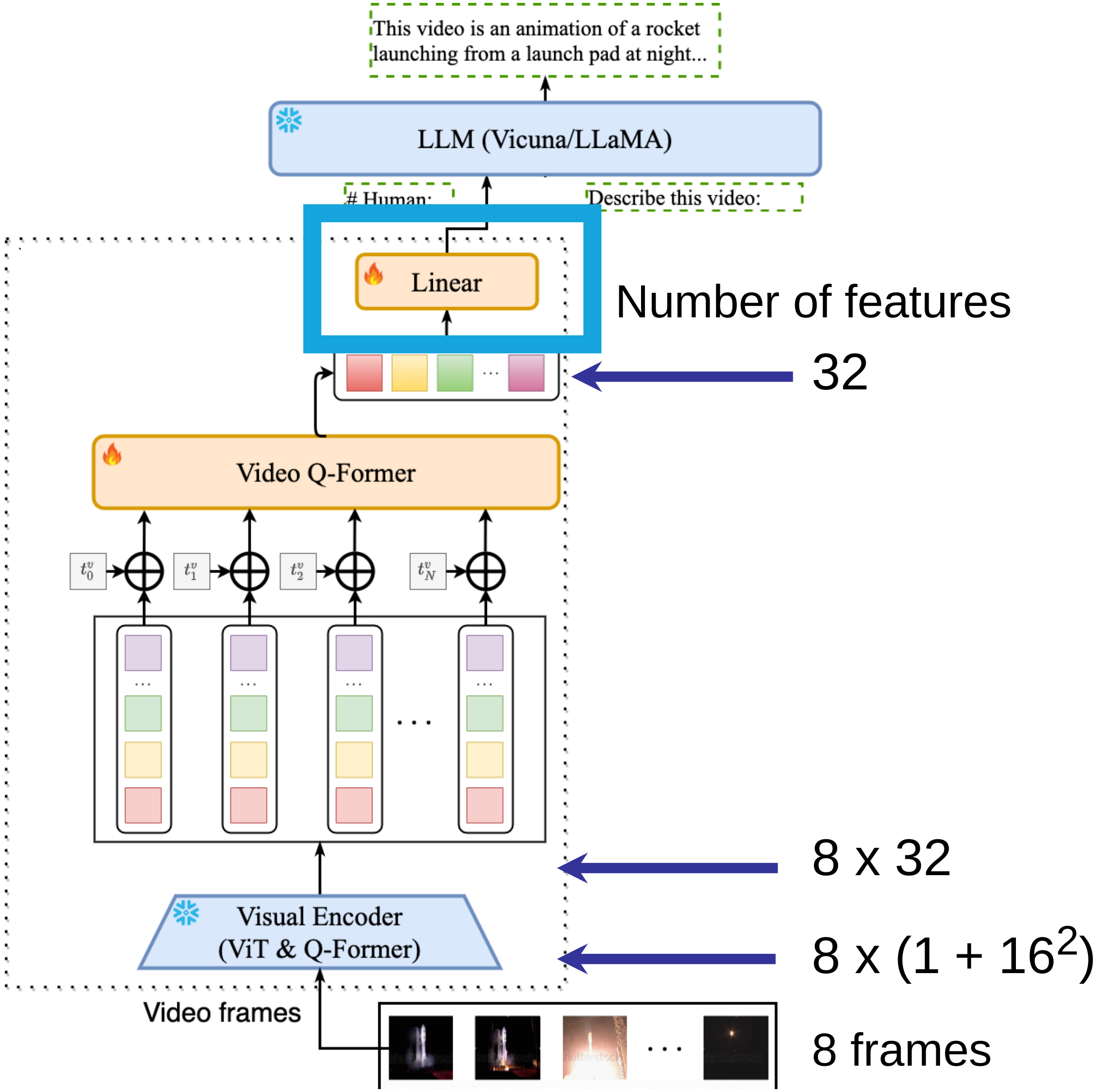
1. Background on visual language models
    - Two types of network architecture using adapters
2. A basic AD model, data, and training
    - Adapting pre-trained vision and language models to this task
3. Improving the `who' in generated AD
    - Supplying supplementary information on characters
4. Improving the `what' in generated AD
    - Adapting pre-trained video-language models to this task
    - Evaluating performance

Strong Vision-Language Models – Video-Llama

Start from pre-trained **video-language** model

- Model ingests 8 frames
- ViT spatial feature map for each frame
- Larger LM – Llama2-7B
- Video Q-Former trained on Webvid-2M

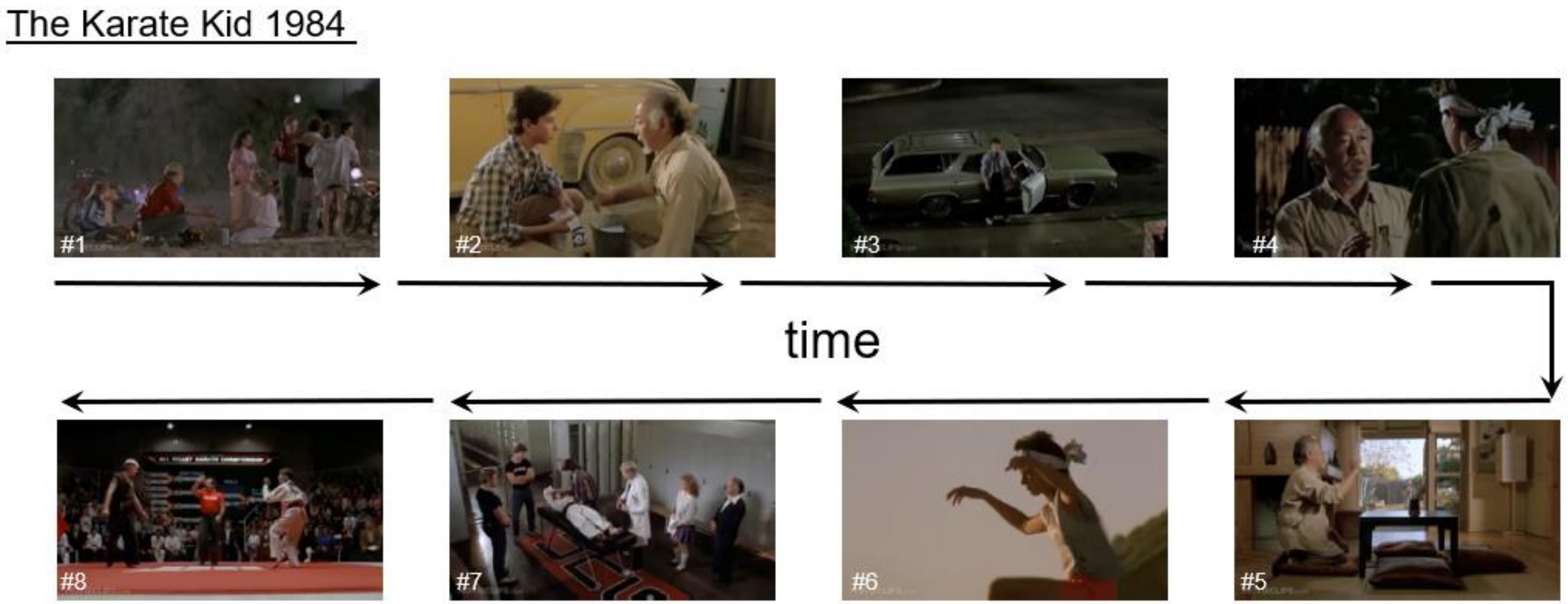
- Only train linear projection layer
- Need pixel level data to train the model …



Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding,
Hang Zhang, Xin Li, Lidong Bing, https://github.com/DAMO-NLP-SG/Video-LLaMA

The Condensed Movies Dataset

- 34,000 movie scenes from 3,600 movies
- ~10 ordered key scenes per movie, each about 2 minutes long
- Provides condensed snapshots into full-length stories



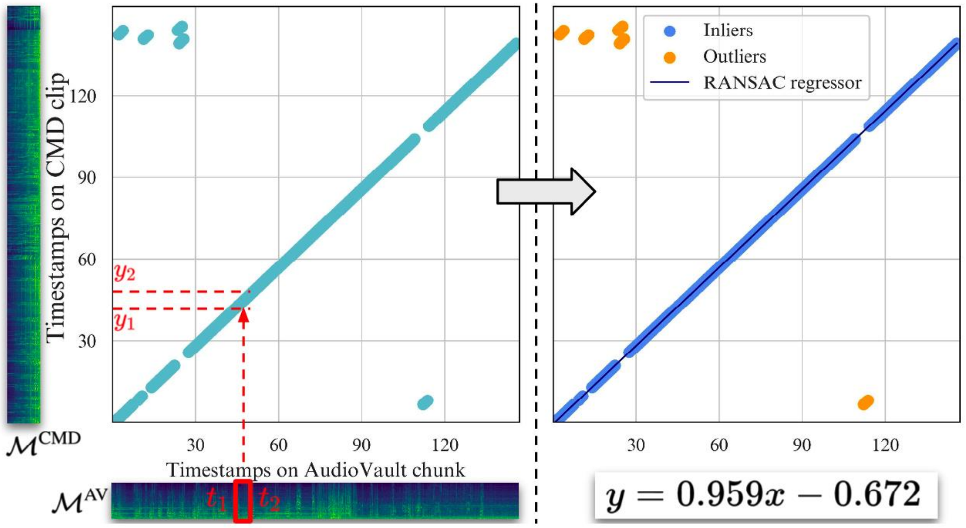
M. Bain, A. Nagrani, A. Brown, A. Zisserman, ACCV 2020

Movie datasets with pixels: CMD-AD

CONDENSED MOVIES

movie clips from YouTube

e.g. 2 minutes, with unknow [t_1,t_2]



precise temporal alignment

AudioVault

full-movie audio with AD

e.g. 1.5 hours

CMD-AD

frames

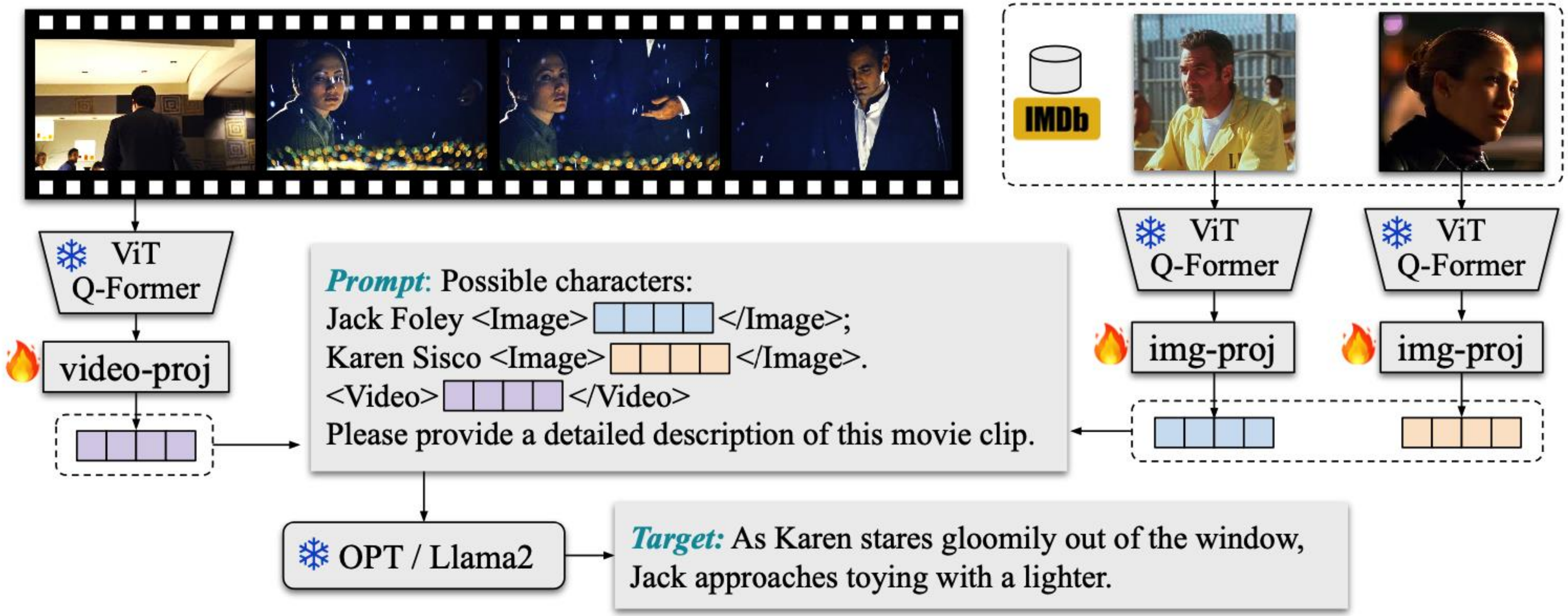
movie soundtrack

audio description

Statistics:

Train:	1332 movies
Test:	100 movies
Duration:	477 hrs
Number of AD:	101k

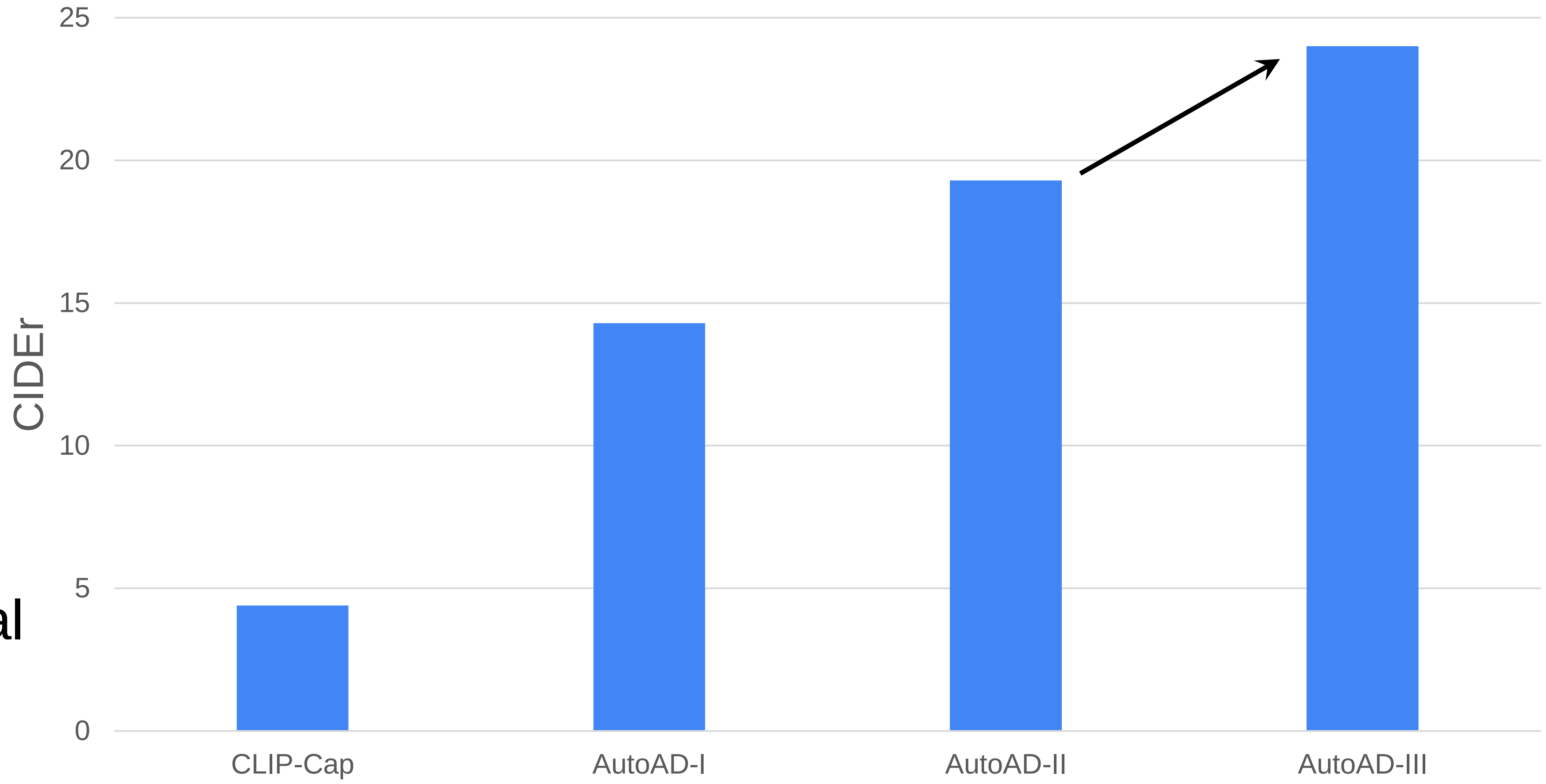
Strong Vision-Language Models (VLM) with pixel inputs



Architecture details:

- Visual feature extractor: EvaCLIP-L14
- Q-Former: 12-layer transformer
- LLM: OPT-2.7B or Llama2-7B

Architecture performance comparison



MAD-Eval
test set

Prompt-tuned GPT-2

X-Attn GPT-2

Movie-Llama

Problem of classical captioning metrics

	AD Sentence	CIDEr*	BLEU-4	METEOR	ROUGE-L
Reference	**The donkeys make a smoke message in the sky which reads, we love you, Daddy.**	-	-	-	-
Prediction 1	Donkeys use smoke to write ‘We love you daddy’ in the sky.	302.9	25.4	30.8	43.6
Prediction 2	Donkey's children write, we love you daddy, in the pale sky with their smoky breath.	226.5	18.9	20.3	25.9
Prediction 3	The young donkeys write 'Love you, daddy' in the sky.	187.0	0.0	25.7	38.6
Prediction 4	The donkey's kids use their breath to write 'We love you, Dad' in the light-colored sky.	112.3	0.0	23.1	25.2

*: to compute CIDEr for one sample, we use tf-idf from coco-eval

Source: Shrek The Third (2007)

New metric 1: LLM-AD-Eval

Prompts:

You are an intelligent chatbot designed
for evaluating the quality of generative outputs for movie audio descriptions.

...

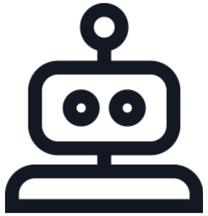
Please evaluate the following movie audio description pair:

Correct Audio Description: **{text_gt}**

Predicted Audio Description: **{text_pred}**

...

Provide your evaluation only as a matching score where the matching score is an integer value between 0 and 5, with 5 indicating the highest level of match.



gpt-3.5-turbo

Answer: 4

Perfect score = 5 for each AD

- "CLAIR: Evaluating image captions with large language models“, David Chan, Suzanne Petryk, Joseph E Gonzalez, Trevor Darrell, and John Canny. arXiv:2310.12971, 2023
- "Can large language models be an alternative to human evaluations?", Cheng-Han Chiang, Hung-yi Lee, arXiv:2305.01937, 2023
- "Judging LLM-as-a-judge with mt-bench and chatbot arena“, Lianmin Zheng et al. arXiv:2306.05685, 2023

Problem of classical captioning metrics

	AD Sentence	CIDEr*	BLEU-4	METEOR	ROUGE-L	LLM-AD-Eval
Reference	**The donkeys make a smoke message in the sky which reads, we love you, Daddy.**	-	-	-	-	-
Prediction 1	Donkeys use smoke to write ‘We love you daddy’ in the sky.	302.9	25.4	30.8	43.6	**5**
Prediction 2	Donkey's children write, we love you daddy, in the pale sky with their smoky breath.	226.5	18.9	20.3	25.9	**5**
Prediction 3	The young donkeys write 'Love you, daddy' in the sky.	187.0	0.0	25.7	38.6	**4**
Prediction 4	The donkey's kids use their breath to write 'We love you, Dad' in the light-colored sky.	112.3	0.0	23.1	25.2	**4**

*: to compute CIDEr for one sample, we use tf-idf from coco-eval

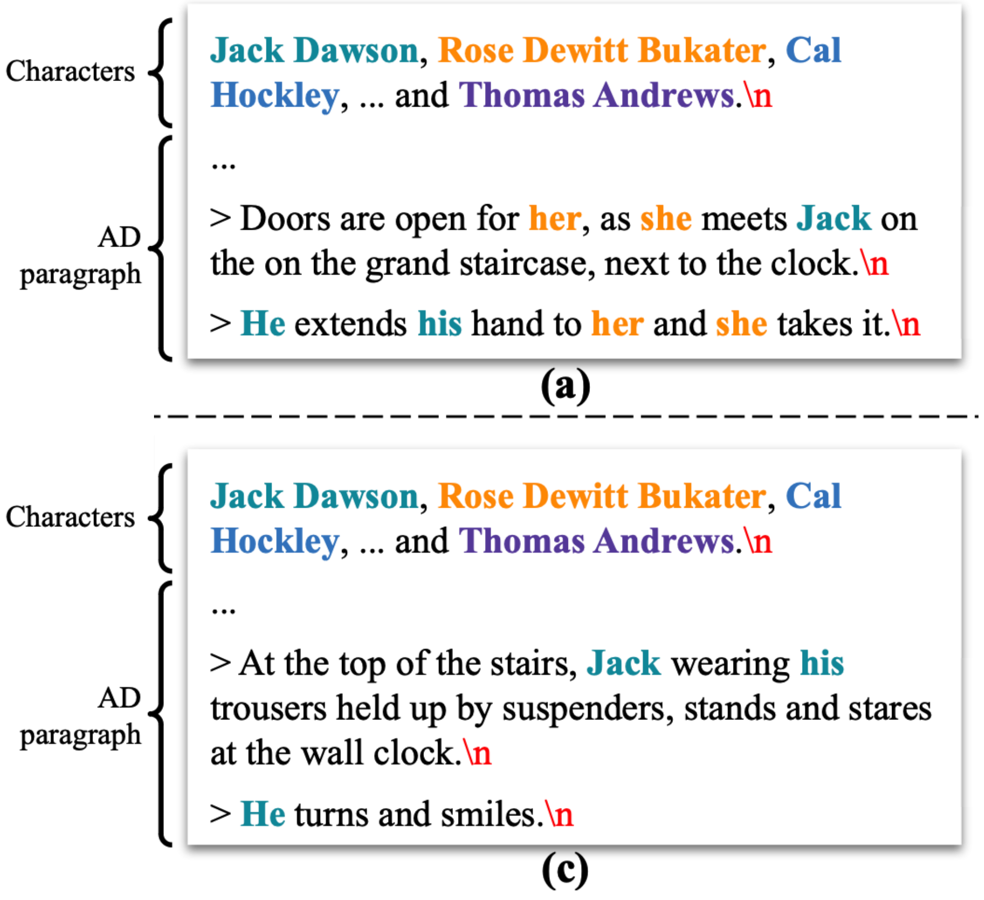
Source: Shrek The Third (2007)

New metric 2: CRITIC
-- Co-Referencing In Text for Identifying Characters

Objective: evaluate whether the characters are referred correctly

Ground truth
AD reference

Predicted AD

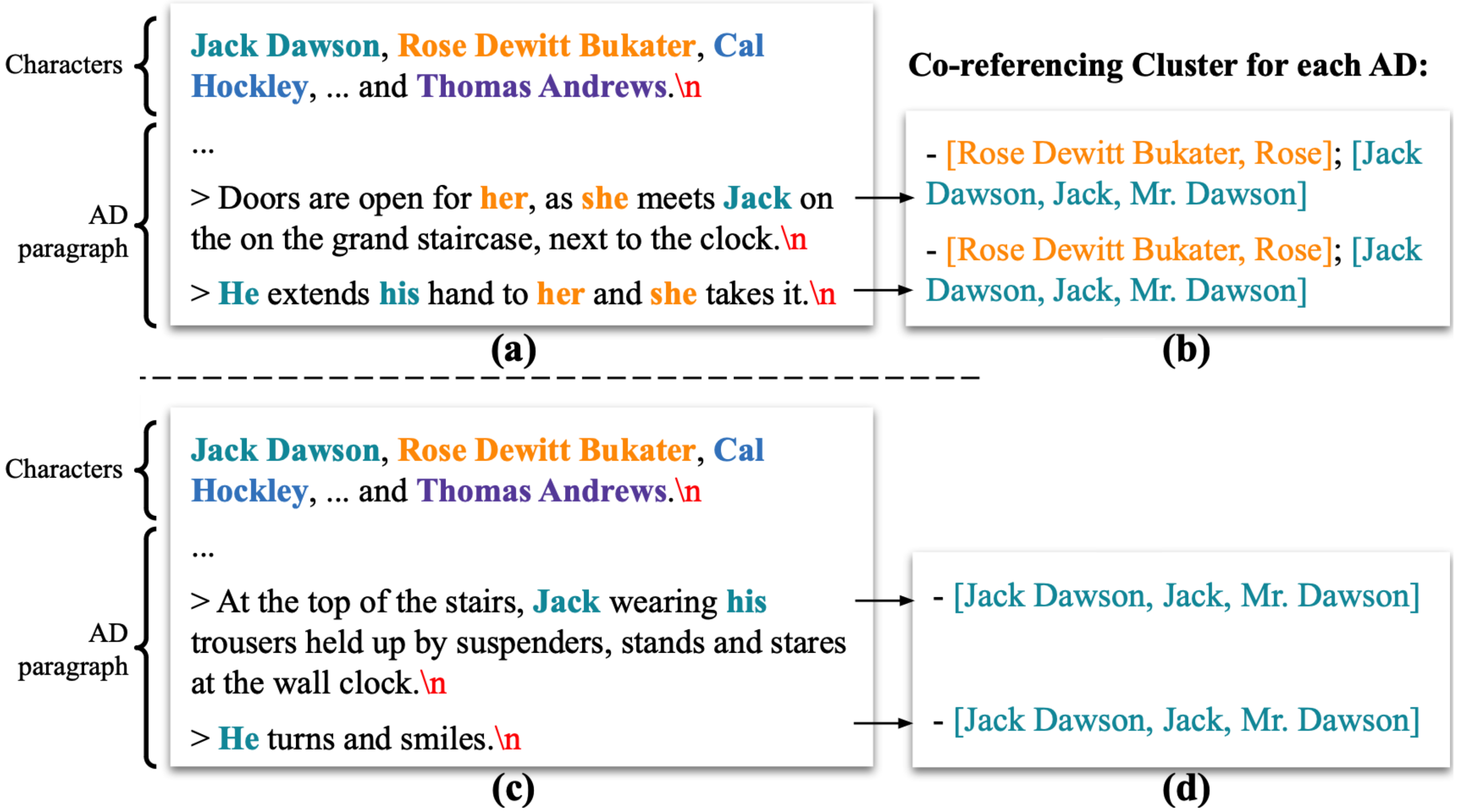


New metric 2: CRITIC
-- Co-Referencing In Text for Identifying Characters

Objective: evaluate whether the characters are referred correctly

Ground truth
AD reference

Predicted AD



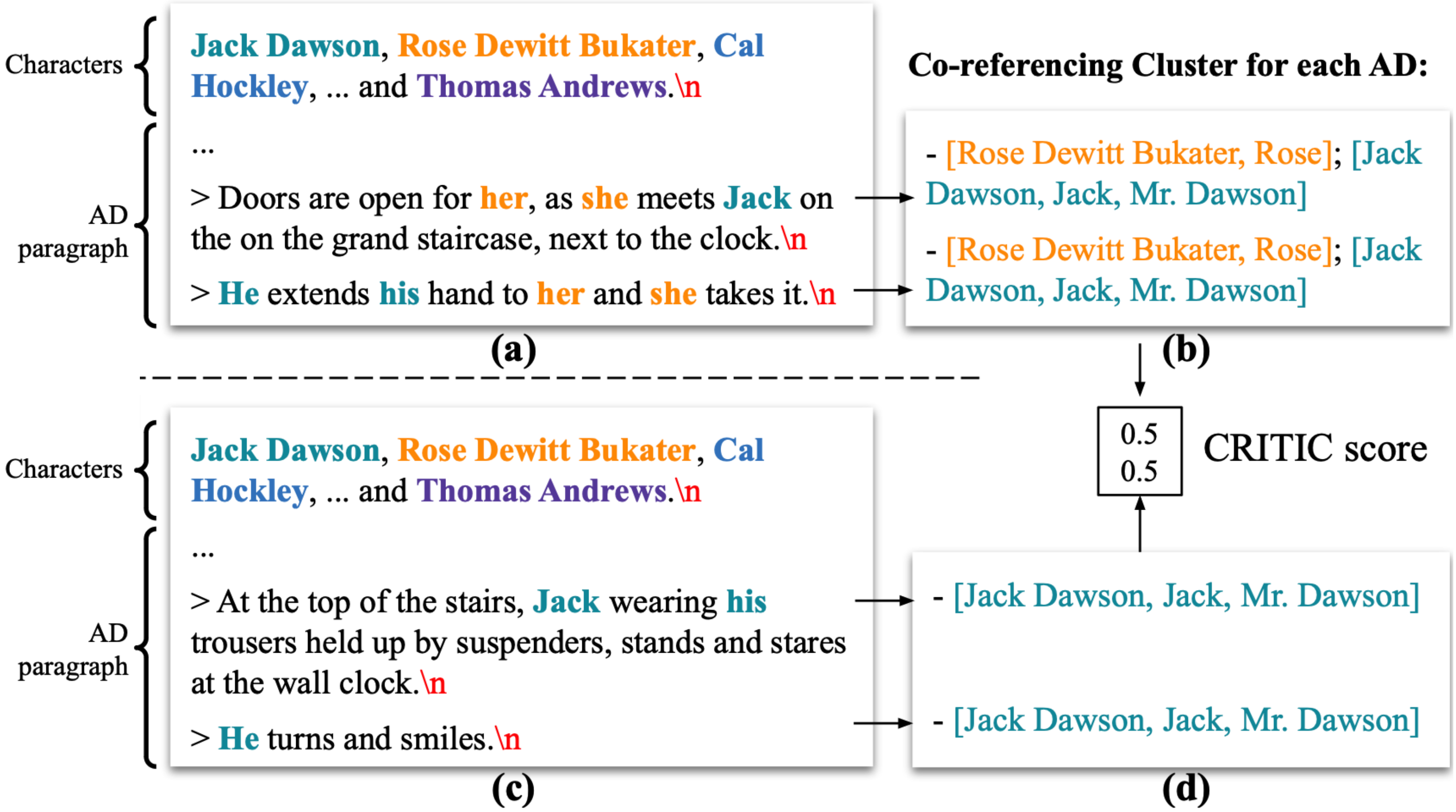
New metric 2: CRITIC

-- Co-Referencing In Text for Identifying Characters

Objective: evaluate whether the characters are referred correctly

Ground truth AD reference

Predicted AD



$$M_{\text{CRITIC}} = \frac{|E_{\text{pred}} \cap E_{\text{ref}}|}{|E_{\text{ref}}|}$$

Perfect score = 1.0 for each AD

Quantitative results

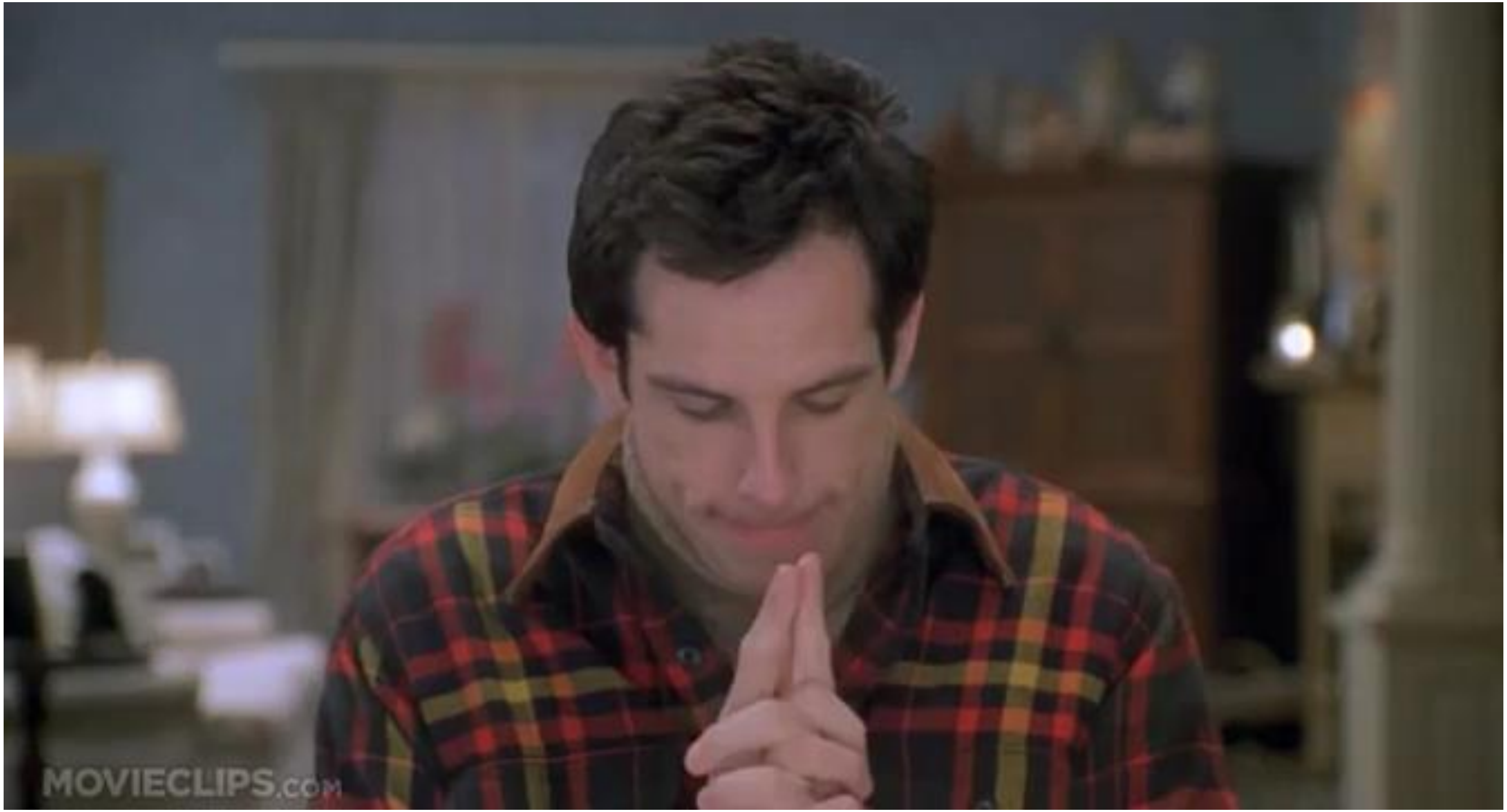
Method	CMD-AD-Eval				MAD-Eval	
	CIDEr	R@1/5	CRITIC	LLM-AD-Eval	CIDEr	R@5/16
Video-BLIP2 [69] (no ft)	4.8	22.0	12.5	1.40	5.0	35.2
Video-Llama2 [73] (no ft)	5.2	23.6	12.3	1.43	4.8	33.8
AutoAD-I [20]	-	-	-	-	14.3	42.1
AutoAD-II [21]	13.5	26.1	35.7	1.53	19.2	51.3
Movie-BLIP2 **(ours)**	22.3	29.8	62.0	2.25	22.8*	52.0*
Movie-Llama2 **(ours)**	25.0	31.2	61.1	2.29	24.0*	52.8*
MM-Narrator + GPT4	-	-	-	-	13.9 ± 0.1	-
MM-Narrator + GPT4v	-	-	-	-	9.8 ± 0.2	-

Movie-BLIP2 (from Video-BLIP)
Movie-Llama2 (from Video-Llama)

* = not trained on MAD-Train

MM-Narrator: Narrating Long-form Videos with Multimodal In-Context Learning. Zhang et al. 2023

Qualitative examples of AutoAD-III: < [title] ground-truth || prediction >



Summary

- Generating AD as a new task
  - Well defined task, so can be evaluated
  - Long form video understanding

- Appraisal
  - Have usable model for AD
  - Character bank significantly improves character naming in AD
  - General method for providing supplementary visual information as a prompt

- The future
  - Place and object banks/memory
  - Use of audio stream
  - Beyond AD: conversation and Q & A with model

Publications and resources

Authors: Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, Andrew Zisserman

- AutoAD: Movie Description in Context, CVPR 2023
- AutoAD II: The Sequel - Who, When, and What in Movie Audio Description, ICCV 2023
- AutoAD III: The Prequel - Back to the Pixels, On arXiv soon

- Datasets and models: https://www.robots.ox.ac.uk/~vgg/research/autoad/
  - MAD-v2: https://github.com/Soldelli/MAD
  - AudioVault: subtitles and AD for 7000+ movies