MM-Narrator: Narrating Long-form Videos with Multimodal In-Context Learning

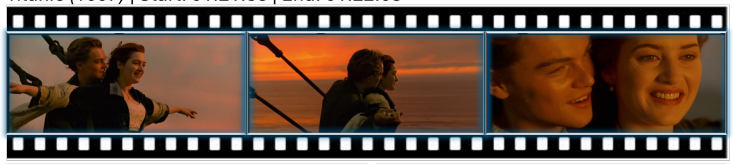
https://MM-Narrator.github.io

Chaoyi Zhang, Kevin Lin, Zhengyuan Yang, Jianfeng Wang, Linjie Li, Chung-Ching Lin, Zicheng Liu, Lijuan Wang

University of Sydney Microsoft Advanced Micro Devices

How does MM-Narrator generate AD?

Titanic (1997) | Start: 01:21:53 | End: 01:22:08



Subtitles
> All right ...
> … open your eyes.
> I'm flying.
> Jack.

Context AD
Jack and Rose stand together on the deck of the boat, enjoying their intimate moment.
Jack and Rose embrace, looking out at the ocean, feeling like they're flying.

AD Prediction (via MM-Narrator)
Jack and Rose hold hands, admiring the sunset together.

The Matrix (1999) | Start: 00:29:28 | End: 00:29:33

Subtitles
> ...you stay in Wonderland... and I show you how deep the rabbit hole goes.
> Remember … all I'm offering is the truth. Nothing more.

Context AD
Neo hesitates before choosing between the red and blue pills.
He closes his eyes, taking a moment to reflect before making his decision.

AD Prediction (via MM-Narrator)
Neo finally makes his decision and picks up the red pill.

Friends S1 E1 -The One Where It All Began (1994) | Start: 00:04:06 | End: 00:04:16

Subtitles
> This is everybody. This is Chandler and Phoebe... and Joey.
> And remember my brother, Ross?
> - Sure! - Hey.
> - Hi. - Oh, God.

Context AD
Ross listens to the conversation, occasionally sipping from his blue mug.
Rachel stands before the group, introducing herself. She stands in her wedding dress, talking to Ross and the group.

AD Prediction (via MM-Narrator)
Rachel, in her wedding dress, interacts with Ross and the group in the room.

Spider-Man: Into the Spider-Verse (2018) | Start: 00:01:28 | End: 00:01:29

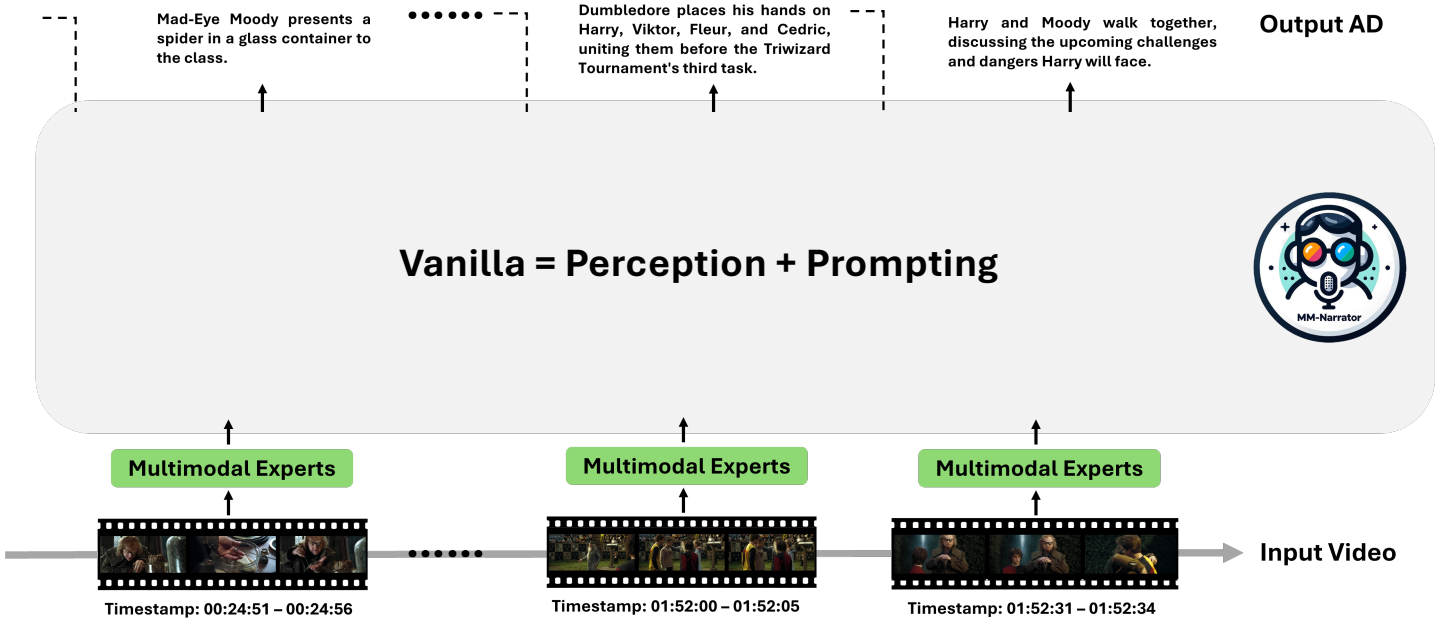
Subtitles
> All right, let's do this one last time.
> My name is Peter Parker.
> I was bitten by a radioactive spider.
> And for 10 years...

Context AD
A close-up reveals a Spider-Man comic book cover. Peter Parker's name tag is shown with a red border and white text.
Peter in his Spider-Man costume, jumps onto a metal platform.

AD Prediction (via MM-Narrator)
Spider-Man jumps off a yellow taxi and continues running on the street.

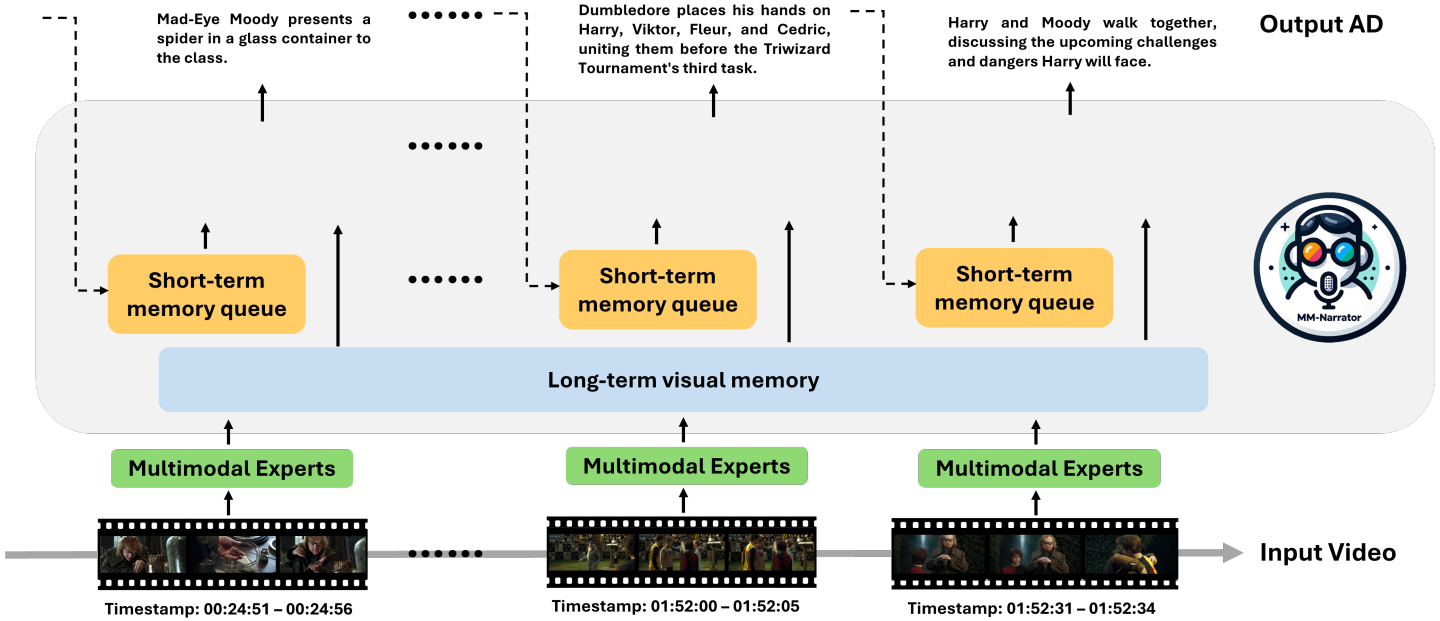
MM-Narrator is a training-free framework towards automatic **audio description (AD)** generation for **long-form videos** via iterations: for each scene, it perceives multimodal inputs (i.e., seeing visual frames and hearing character dialogues), recalls the **context AD** depicting past scenes, and infers **AD prediction** for the current scene.

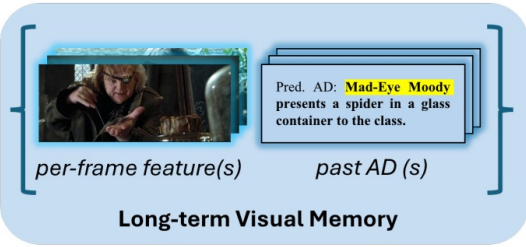
Recurrent AD Generation



- **Vanilla**: Multimodal Experts + LLM
  - Captioner (visual perception)
  - ASR (audio perception)
  - GPT-4 (prompting)

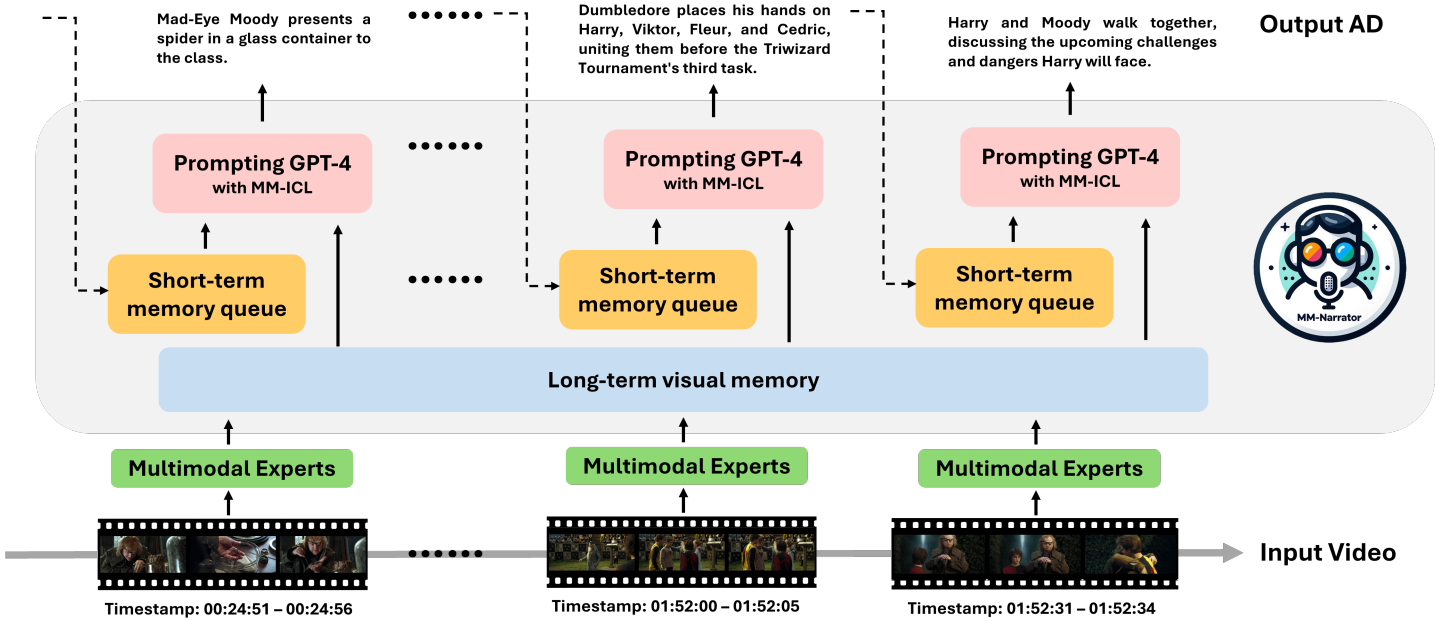
Recurrent AD Generation

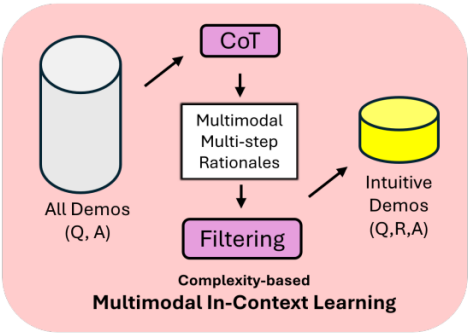




- **Vanilla**: Multimodal Experts + LLM
- **Memory Mechanism**: Short-term Memory Queue + Long-term Visual Bank

Recurrent AD Generation

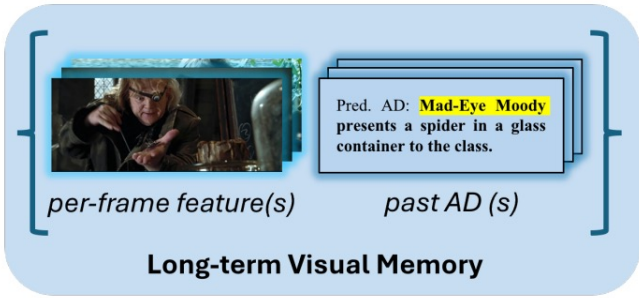




- **Vanilla**: Multimodal Experts + LLM
- **Memory Mechanism**: Short-term Memory Queue + Long-term Visual Bank
- **MM-ICL**: Complexity-based Multimodal In-Context Learning

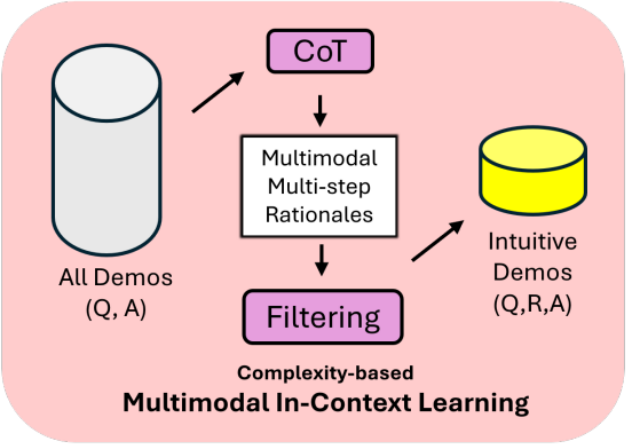
Memory Mechanism

- **Short-term Memory Queue:** past *K* AD predictions
- **Long-term Visual Memory**
  - **Visual Bank** (for frame-level character re-identification)
    - **Key**: Per-frame CLIP-ViT feature
    - **Value**: AD prediction
  - **Register-and-Recall**:
    - Turn active when only one single individual is presented in the frame (via People Detector)
    - i.e., typically in close-up shots of the character, making fra features compatible for character re-identification.
  - **Cosine similarity** (over visual features) to re-identify similar appearances.



Note: Given any AD that covers multiple frames, this frame-level visual retriever supports the MM-Narrator in re-identifying multiple characters appearing in the video clip.

Multimodal In-Context Learning



Complexity-based MM-ICL for demonstration denoising

1. Prepare MM-ICL demonstration pool of (*Q*,*A*) pairs.
2. Query LLM to articulate the CoTs as reasoning steps *R*.
3. Construct **an intuitive subset** pool of (*Q*,*R*,*A*) tuples: select the most straightforward examples, quantified by the shortest number of *R*.
4. Conduct random demonstration sampling over the subset pool (step#3).

	What makes good examples for AD task?	How to find and use them for ICL?
Random MM-ICL	No specific assumption (any example might help)	Randomly sample demonstrations
Similarity-based MM-ICL	Similar MM examples (with similar scene appearances, subtitles, character names, ...)	Retrieve similar demonstrations
Complexity-based MM-ICL ***(our proposed appro.)***	More intuitive MM examples help LLM to reason better	Denoise into a subset pool Perform random sampling

Experiment Results

- using classic captioning scores

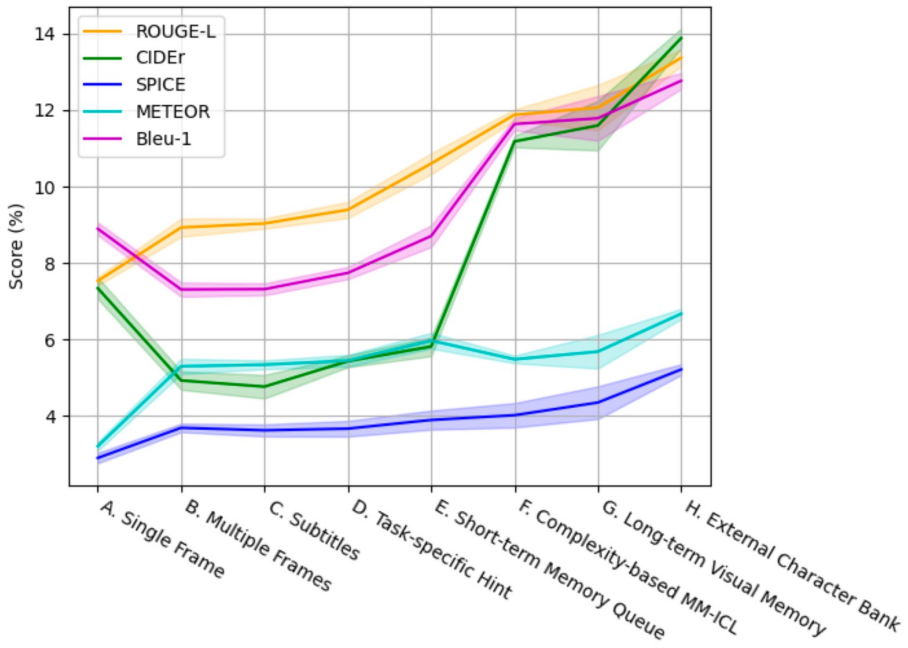
- **Comparison with fine-tuning based approaches**

Method	Training-Free	R-L (↑)	C (↑)	S (↑)	R@5/16 (↑)
ClipCap [41]	✗	8.5	4.4	1.1	36.5
ClipDec [42]	✗	8.2	6.7	1.4	-
AutoAD-I [22]	✗	11.9	**14.3**	4.4	42.1
MM-Narrator	✓	**12.1**	11.6	**4.5**	**48.0**

- **Comparison with training-free LLM/LMM baselines**

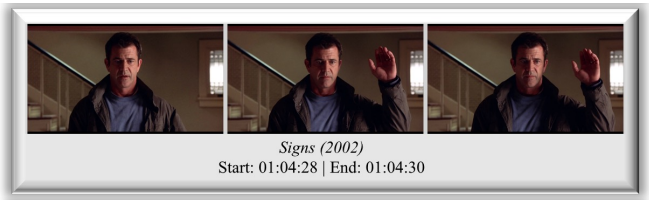
Method	LLM/LMM	R-L (↑)	C (↑)	S (↑)	R@5/16 (↑)
VLog [6]	GPT-4	7.5	1.3	2.1	42.3
VideoChat [27]	GPT-4	7.9	2.4	1.8	42.5
MM-Vid [30]	GPT-4V	9.8	6.1	3.8	46.1
MM-Narrator					
w/o MM-ICL	GPT-4	10.3	4.9	3.8	47.1
w/ MM-ICL	GPT-4	**12.1**	**11.6**	**4.5**	**48.0**

- **Building MM-Narrator from Image Captioner**



Experiment Results

- qualitative comparison

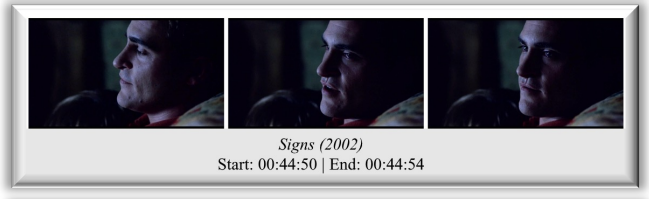


Signs (2002)
Start: 01:04:28 | End: 01:04:30

Human Annotation: GRAHAM does the same.
ClipCap: wallpaper probably with a portrait entitled actor.
VLog: A bearded man in a jacket raises his hand near a window, while someone says "Yeah. ".
MM-Vid (GPT-4V): The scene shows a man, possibly a detective, standing in the hallway of a building and raising his hand.
MM-Narrator (ours)
+ GPT-4: The man in the dark jacket raises his hand in agreement with the lake idea.
+ GPT-4V: Graham, with a stern expression, raises his hand to signal silence and attention.

Charlie St. Cloud *(2010)*
Start: 00:57:48 | End: 00:57:53

Human Annotation: A handwritten note on the back reads come find me.
ClipCap: man's hand with a pen writes on the paper.
VLog: A person holds a piece of paper, surrounded by a dark counter and a wall with peeling paint.
MM-Vid (GPT-4V): The given video clip shows a hand removing a note from a wooden door, with the note reading "Come find me".
MM-Narrator (ours)
+ GPT-4: Charlie examines a yellow note with red writing.
+ GPT-4V: Charlie finds a note saying "Come find me."



Signs (2002)
Start: 00:44:50 | End: 00:44:54

Human Annotation: MERRILL looks at GRAHAM then, nods.
ClipCap: Wallpaper probably with a portrait titled person.
AutoAD-II †: Merrill stares at him, his brow furrowed.
MM-Vid (GPT-4V): The scene shows a close-up of a man's face, who appears to be deep in thought.
MM-Narrator † (ours)
+ GPT-4: Merrill Hess looks around, deep in thought.
+ GPT-4V: Merrill's face is consumed by a mix of emotions as he reflects on a past memory, his eyes revealing a deep internal struggle.

Les Misérables (2012)
Start: 01:02:36 | End: 01:02:40

Human Annotation: Holding COSETTE VALJEAN turns and sees a man with a spade.
ClipCap: person and the child in the dark.
AutoAD-II †: the boy looks up at his father, who stares back at him with a furrowed brow.
MM-Vid (GPT-4V): A man in a top hat is carrying a young girl while looking around frantically in a dark courtyard.
MM-Narrator † (ours)
+ GPT-4: Jean Valjean carries Cosette through a dark room, seeking safety.
+ GPT-4V: Jean Valjean and Cosette, shrouded in darkness, cautiously approach the church's exit, their escape imminent.

- **ClipCap** (finetuning-based baseline)
- **AutoAD-II** (finetuning-based SOTA)
- **Vlog** (training-free LLM based on GPT-4)
- **MM-Vid** (training-free LMM based on GPT-4V)
- **MM-Narrator** (ours)

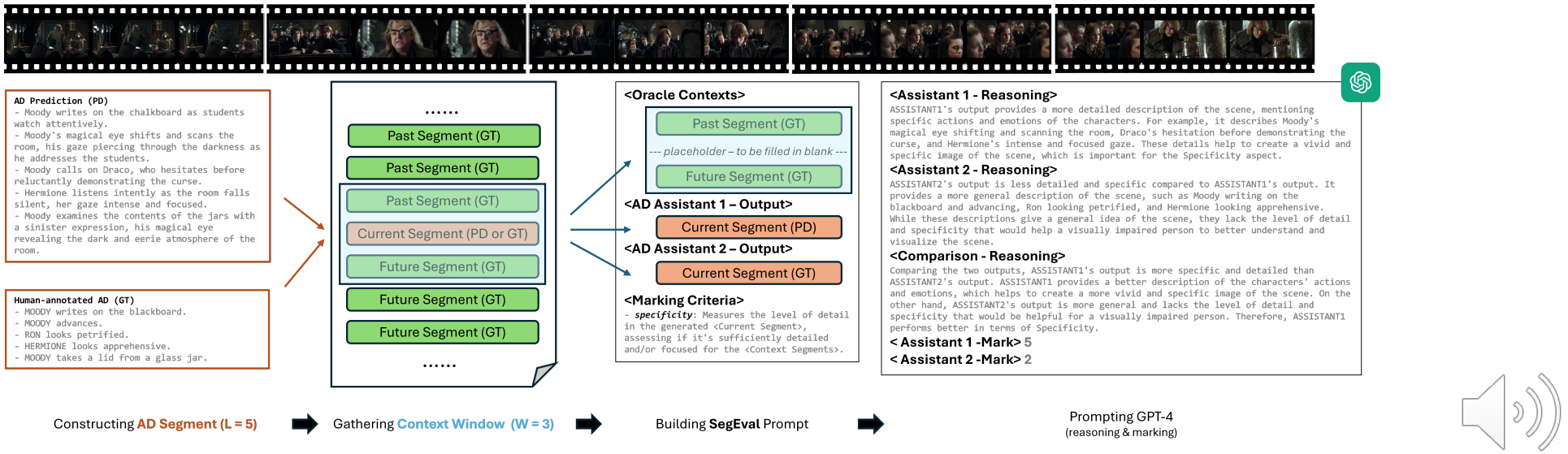
AD Evaluation with SegEval

- **Motivation:** (1) low inter-annotator agreement; (2) a few performance drop on classic reference-based captioning scores when incorporating MM-Narrator with GPT-4V (for example, R-L, C, and B-1);

Method	R-L (↑)	C (↑)	M (↑)	B-1 (↑)
MM-Narrator				
+ GPT-4	$12.1_{\pm 0.4}$	$11.6_{\pm 0.4}$	$5.7_{\pm 0.2}$	$11.8_{\pm 0.3}$
+ GPT-4V	$11.8_{\pm 0.1}$	$7.0_{\pm 0.2}$	$6.5_{\pm 0.1}$	$9.3_{\pm 0.1}$
MM-Narrator †				
+ GPT-4	$13.4_{\pm 0.0}$	$13.9_{\pm 0.1}$	$6.7_{\pm 0.0}$	$12.8_{\pm 0.0}$
+ GPT-4V	$12.8_{\pm 0.0}$	$9.8_{\pm 0.2}$	$7.1_{\pm 0.0}$	$10.9_{\pm 0.0}$

AD Evaluation with SegEval

- Motivation
- **Solution**: A segment-based GPT-4 evaluator (**SegEval**) to measure the recurrent AD generation, in terms of multi-domain qualities.



AD Evaluation with SegEval

- Motivation
- Solution
- **Result**

Method	LLM/LMM	**Text-level Quality** Context-irrelevant Scores		**Sequence-level Quality** Short-context Scores			Long-context Scores		
		Orig. ±0.02	Cons. ±0.02	Cohe. ±0.01	Dive. ±0.06	Spec. ±0.04	Cohe. ±0.01	Dive. ±0.01	Spec. ±0.03
GT	-	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ClipCap [41]	GPT-2	0.43	0.42	0.26	0.35	0.35	0.26	0.42	0.33
VLog [6]	GPT-4	1.03	0.88	0.34	0.55	0.52	0.32	0.57	0.43
MM-Vid [30]	GPT-4V	0.85	0.78	0.51	0.81	0.66	0.53	0.84	0.62
MM-Narrator	GPT-4	$1.05_{\pm 0.10}$	$1.03_{\pm 0.05}$	$0.52_{\pm 0.06}$	$0.70_{\pm 0.06}$	$0.66_{\pm 0.04}$	$0.57_{\pm 0.05}$	$0.70_{\pm 0.02}$	$0.61_{\pm 0.05}$
MM-Narrator	GPT-4V	$\mathbf{1.49}_{\pm 0.10}$	$1.45_{\pm 0.05}$	$0.94_{\pm 0.07}$	$1.01_{\pm 0.04}$	$1.13_{\pm 0.08}$	$0.87_{\pm 0.04}$	$1.05_{\pm 0.04}$	$\mathbf{1.14}_{\pm 0.05}$
MM-Narrator †	GPT-4	$0.95_{\pm 0.02}$	$1.06_{\pm 0.01}$	$0.62_{\pm 0.04}$	$0.75_{\pm 0.01}$	$0.76_{\pm 0.01}$	$0.62_{\pm 0.04}$	$0.80_{\pm 0.03}$	$0.71_{\pm 0.03}$
MM-Narrator †	GPT-4V	$1.45_{\pm 0.14}$	$\mathbf{1.46}_{\pm 0.04}$	$\mathbf{0.98}_{\pm 0.03}$	$\mathbf{1.06}_{\pm 0.04}$	$\mathbf{1.24}_{\pm 0.09}$	$\mathbf{0.94}_{\pm 0.02}$	$\mathbf{1.09}_{\pm 0.05}$	$1.1_{\pm 0.03}$