Google

MTG-AD

Mind The Gap (MTG-AD): Dataset and Fine-grained Evaluations for Inline Audio Descriptions

Subhashini Venugopalan · Yingwen Tan · Taylor Roper · Jimmy Tobin · Anton Kast · Alicia Martin · Sam Sepah · Amy Pavel

NeurIPS 2026

Why audio descriptions for YouTube?

Many YouTube videos are **inaccessible** to blind audience members.

Audio descriptions make videos accessible by narrating important visuals.

CC Closed Captions

AD Audio Description



THE VLOG IS BACK

[Music]

"Casey takes ground coffee out of the fridge and pours it into his Mr. Coffee coffee maker."



Claire Saffitz Cooks Her Ideal Thanksgiving Start to Finish | NYT Cooking

*"You're going to go **like this** all the way around" ...*

"She crimps the pie crust by pressing the edge of the crust under itself."

Problem

- **Audio Descriptions (AD)** are a **vital part** of the **watch experience** for blind and low-vision users

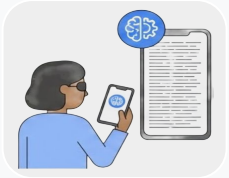
Current reality for BLV users



skip



depend on family



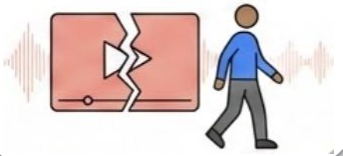
wait e.g piccybot

Inconvenient and breaks immersion

Based on Liu et. al. CHI '21 and our own participant study

79%

of survey respondents have **abandoned** a non-described video at some point.



64%

of survey respondents said that the availability of AD was **"extremely" or "very" important** when choosing something to watch.



Stats from YouTube UXR study

Challenge: AD is **time-consuming** and **expensive** to create and rarely available. **< 1%, only ~20k / 32B** videos have audio descriptions on YouTube.

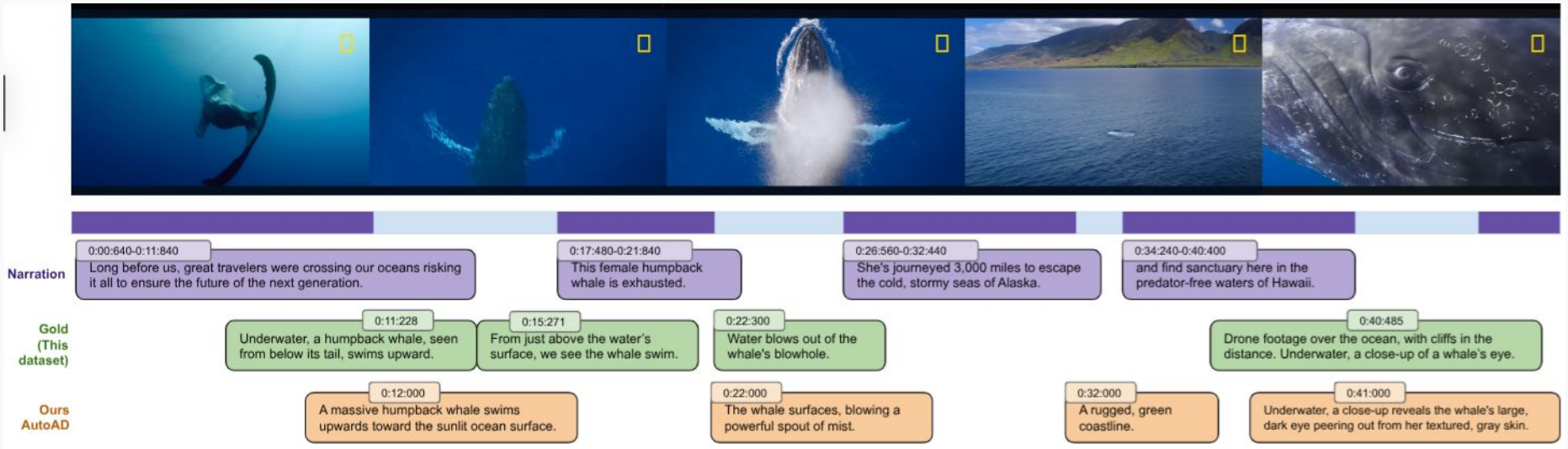
Google

AutoAD can unlock video content for blind and low vision users at scale

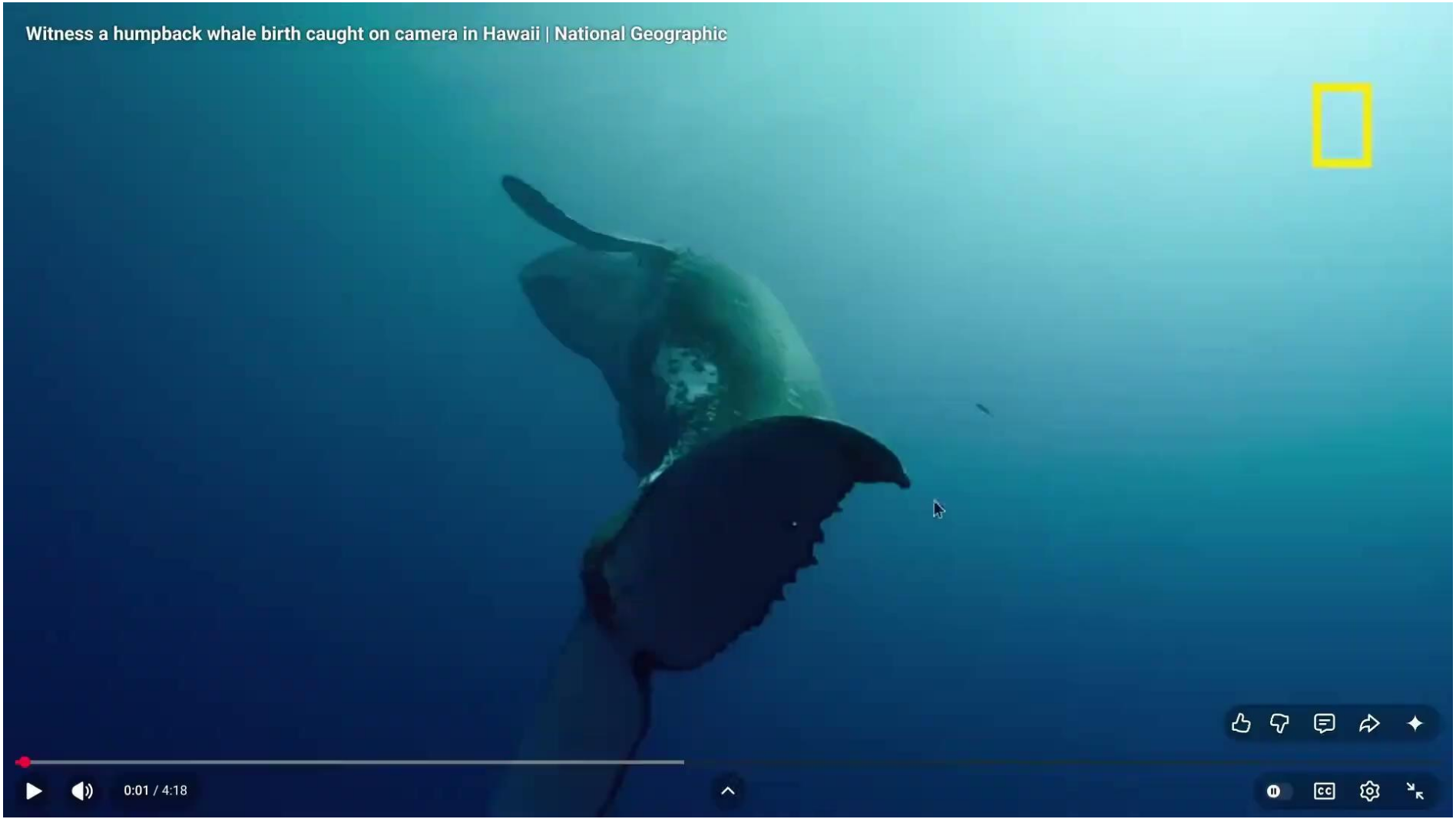


Google

Coherent, Inline, End-to-End Audio descriptions



- Capture full experience of user
- Coherent and inline with existing content
- Audio descriptions are rendered in the existing audio track


Witness a humpback whale birth caught on camera in Hawaii | National Geographic
0:01 / 4:18

What we built



Data & Framework

- Collected a high-quality dataset of professionally written AD
- Framework for making inline AD with Gemini and other models



Objective Metrics for Automated Loop

- Assessing metrics with automated evaluation to compare AD sources
- Assessing metrics with sighted human evaluation



Conducting User Studies

- Blind audience members
- YouTube creators (YouTube team)

Google

“If you cannot measure it, you cannot improve it”

~ Lord Kelvin (actual quote)

Mind The Gap Dataset

~36 hrs (434 videos)

Avg. 5mins
(2 - 9 mins long)

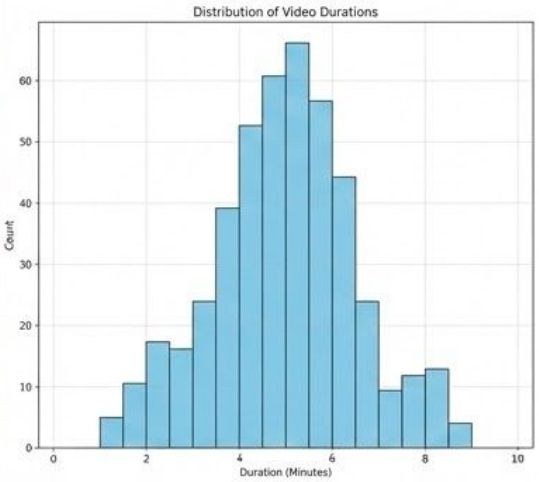
Start End

Professionally annotated

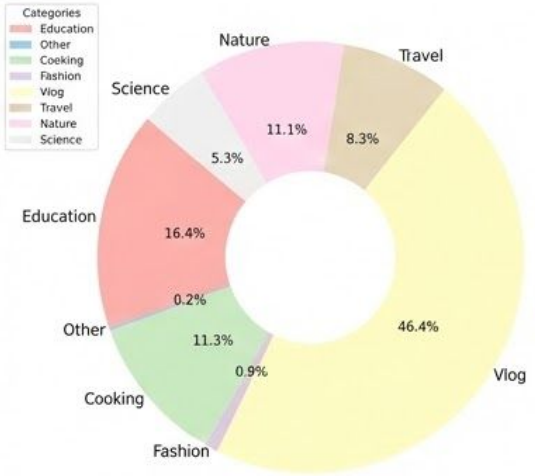
Teams of sighted and blind audio describers

~7325 AD segments

Distribution of Video Durations

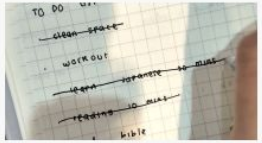


Distribution of Video Categories



Google

Samples: Captures narrative flow

Video	Simple AD Prompt	Guideline+ASR AD	AD Professionals (MTG-AD)
	A hand crosses off more items on the "TO DO LIST"	To-do list. "learn japanese 20 mins" is crossed off.	She draws a line through "learn Japanese twenty minutes" in her journal.
	A hand fills a pot with water from a tap. Overhead view of water filling a pot.	Water fills pan.	The woman fills a worn pot with water from her sink,
	A hand moves a pot to a stove. A hand turns a stove knob.	Pan on stove. Turning on Stove. Blue flame ignites.	then sets it on a gas stove burner and ignites it.
	Hands open a green instant noodle package. ...	She opens a ramen packet with ramen and seasoning.	She makes instant ramen in the pot, ...

Google

Captures important sequence of events

Video	Simple AD Prompt	Guideline+ASR AD	AD Professionals (MTG-AD)
	An elephant stands beside a road as a white truck approaches.	A large elephant stands on the grassy shoulder of a winding road, beside a lush green forest. A white truck with a blue net covering its load approaches in the distance.	A medium-sized, chubby, and tuskless elephant stands off to the side of a two-lane road hedged in by dense foliage.
		The elephant intently watches the truck as it draws nearer. Its trunk hangs loosely, almost touching the ground. Cars pass cautiously in the opposite lane.	A flatbed truck passes an oncoming cargo truck near the elephant.
	The elephant steps onto the asphalt and picks up dropped sugarcane.	The elephant carefully steps towards the stopped truck, its trunk now raised. ...	The elephant quickly steps in front of the cargo truck and the truck rapidly breaks.

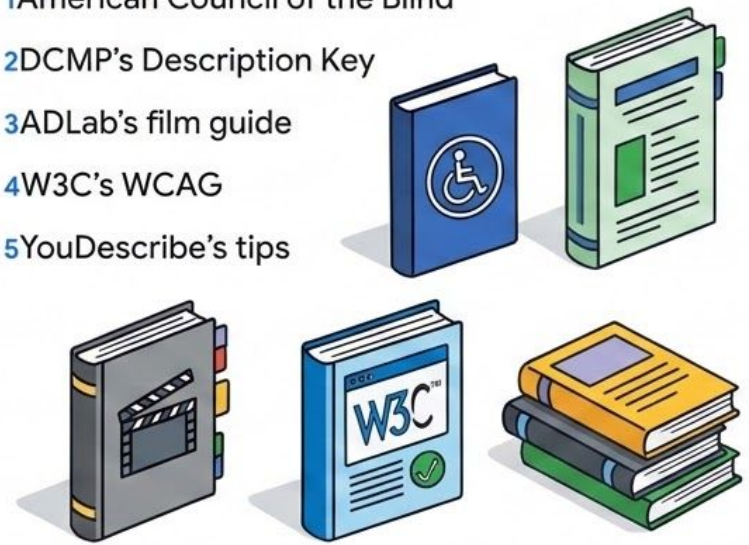
Google

AD Guidelines to Evaluation Framework

Expert AD Guidelines Reviewed

Chosen for breadth and comprehensiveness:

1. American Council of the Blind
2. DCMP's Description Key
3. ADLab's film guide
4. W3C's WCAG
5. YouDescribe's tips



Audience Perceptions & Research

We reviewed studies of **BLV audiences' perceptions** of AD to find where audience preferences depart from prescribed rules.

Together with researchers in **HCI, accessibility, and computer vision...**

SYNTHESIS OUTCOME

Ten Guidelines

covering **content, style, and audio.**

Evaluation Framework

 Higher is better

 Lower is better

Rubrics

Rubric	Description	Interpretation
Prioritization	Does AD cover important visual info to aid users.	
On-screen text	Does the AD include text on screen where available.	
Redundancy	AD accounts for information the viewer already has or is included in the original audio track. (e.g. dialog or sounds)	
Timeliness	Is the placement of the AD timely (within 5s of event)	
Audio Overlap	Overlap of audio and AD	
Silence	Is the AD missing opportunity to cover events in gaps of silence.	

Qualitative	Descriptive, Objective, Accurate

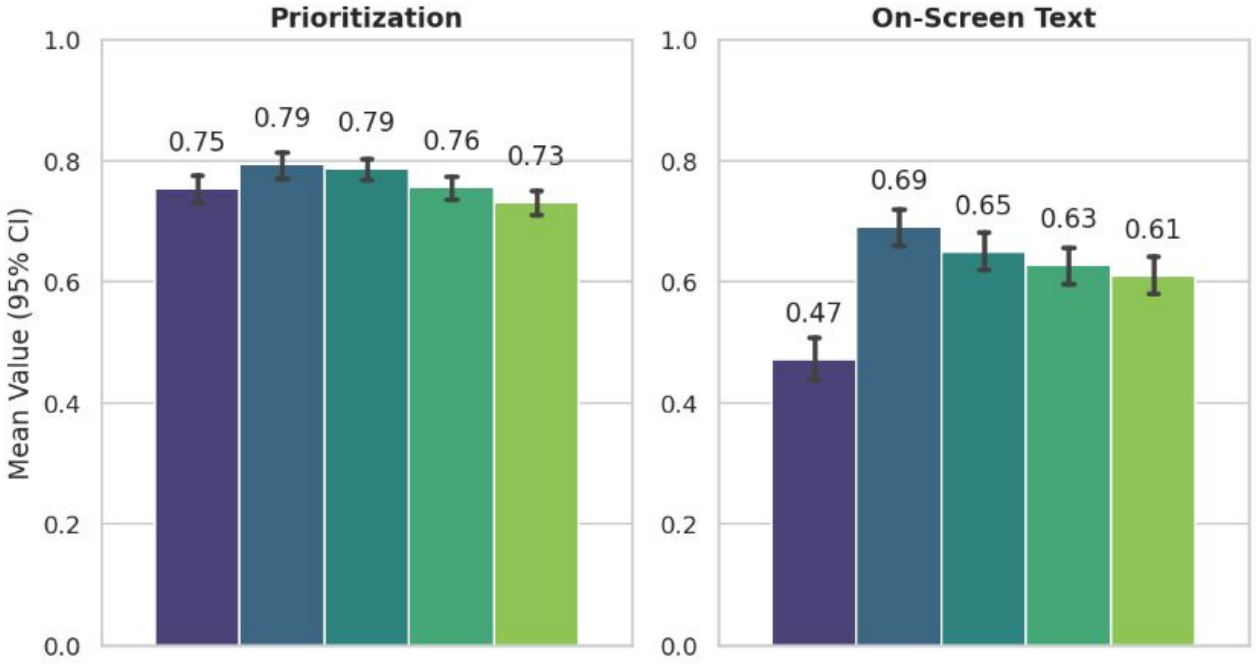
Google

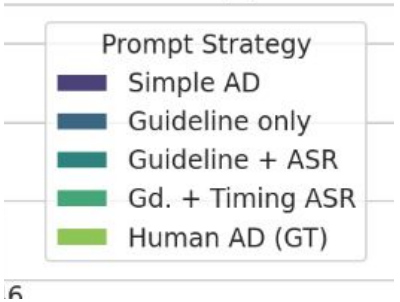
AD Generation and Evaluation



- Prompts consider inputs with different metadata sources
  - AD guidelines
  - ASR closed captions with timing information
  - CoT prompt to reason over time gaps
- Each component adds value to the overall AD

Guidelines help prioritize events e.g. on-screen text



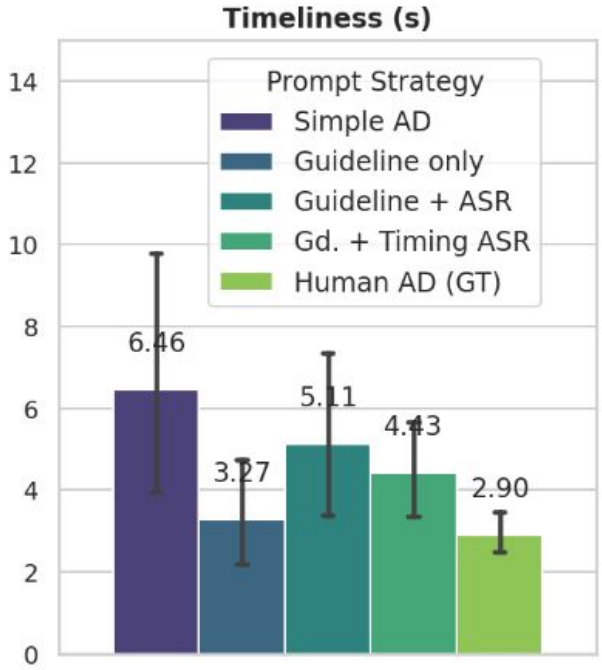


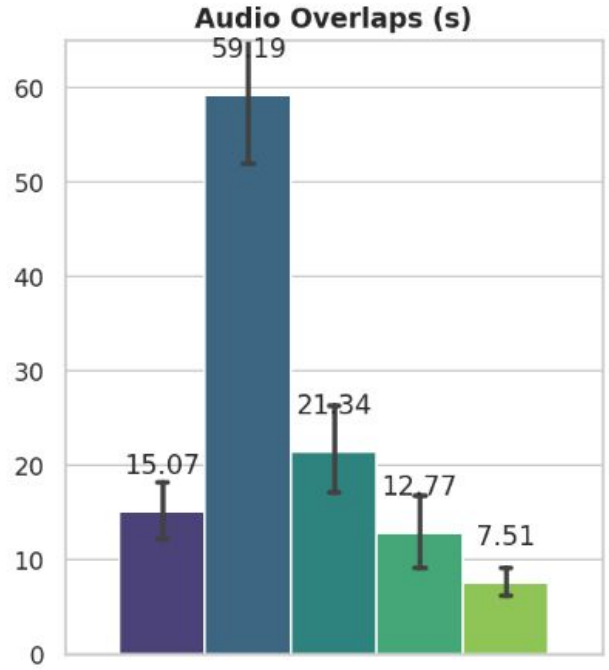


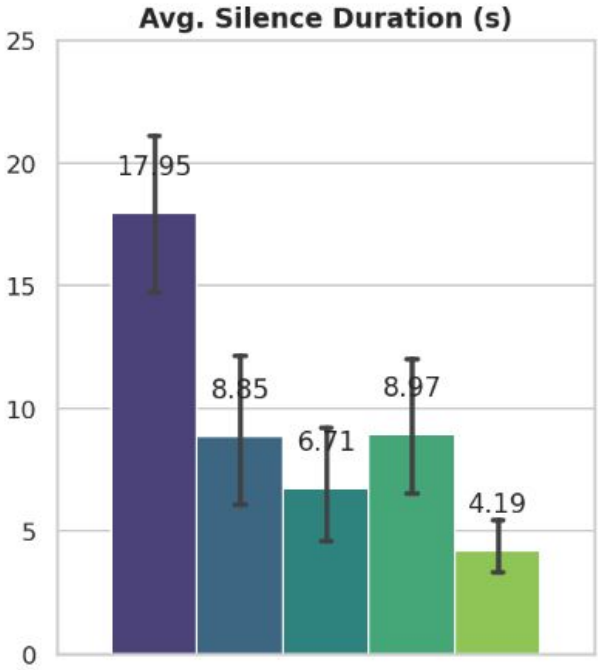
Higher is better

Google

Transcripts, CoT reasoning help placement of AD





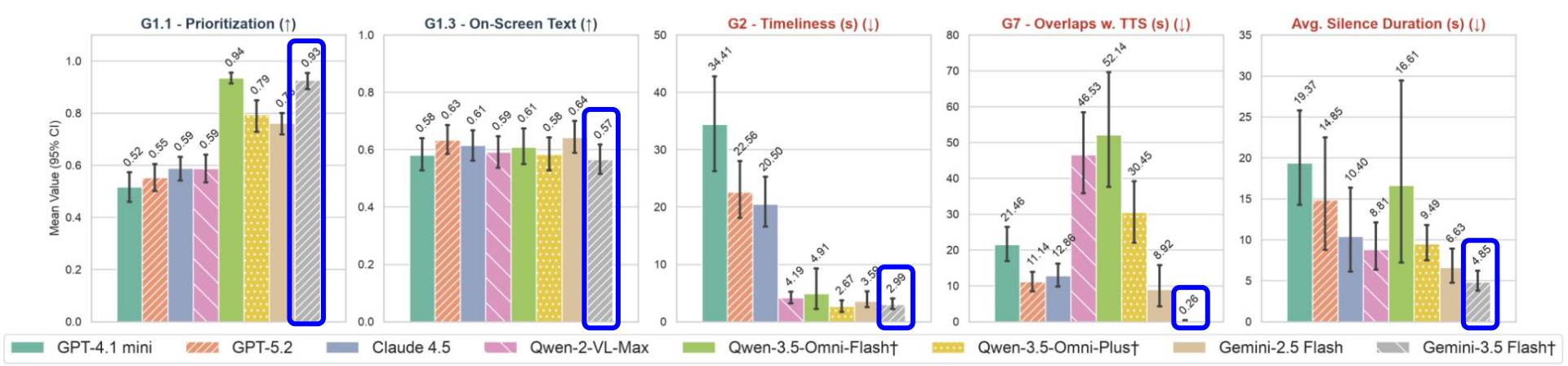




Lower is better

Google

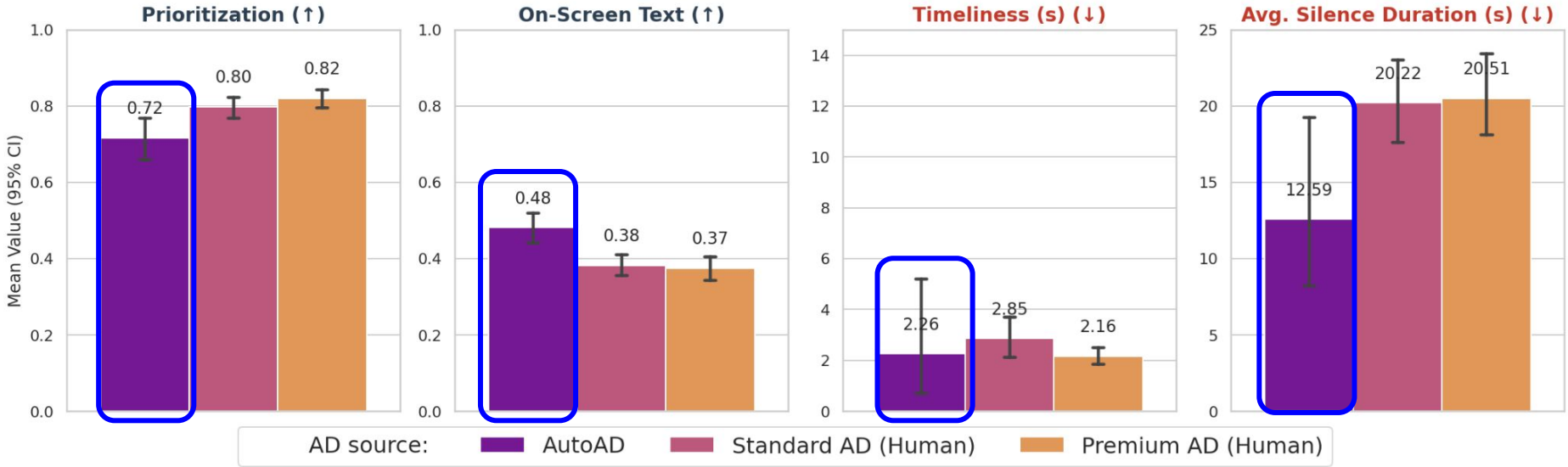
Gemini 3.5 Flash performed better than other models



Google

Metrics generalize to YouDescribe dataset

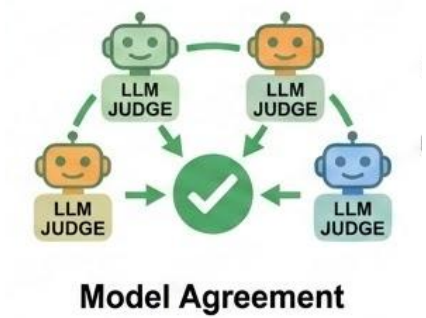
- 200 videos, also avg. 5 mins duration
- **2 sets of Human annotated AD** (premium, standard)
- AutoAD covers silences better! (Most videos silent/music-only)



Google

Evaluating LLM-Judge Quality

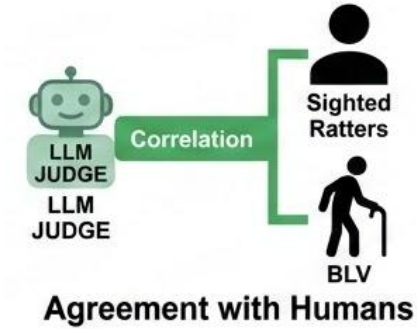
We evaluated the quality of the proposed LLM-judge based evals.



01

Inter-Model Agreement

How well different models (as judges) agree with each other, and

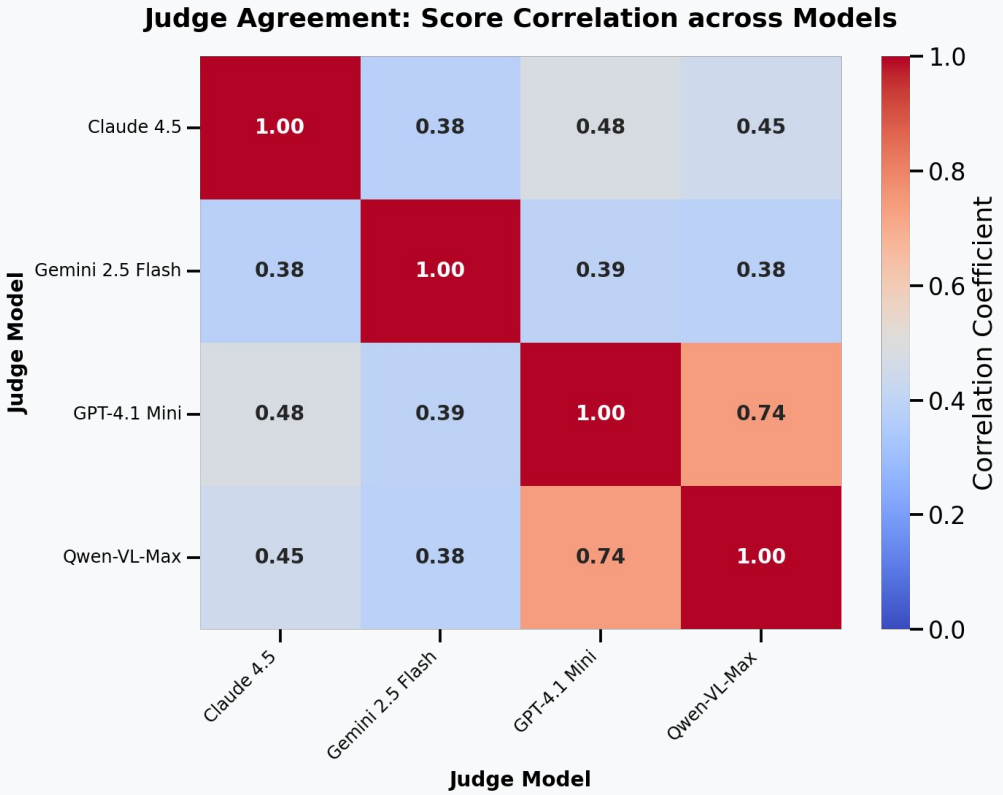


02

Human-Model Alignment

LLM judge models agreeing with sighted raters and BLV raters.

Some models agree more with each other



- GPT-4.1 mini and Qwen 2.5 VL are more correlated

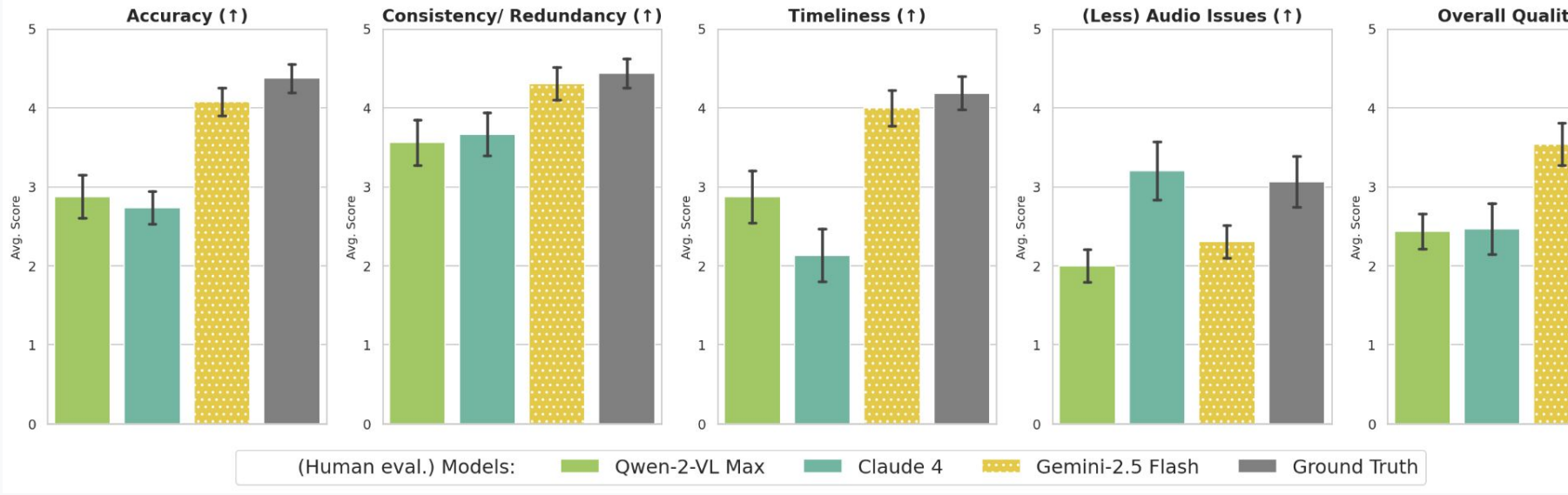
"prioritization",
"videoa11y_descriptive",
"videoa11y_objective",
"videoa11y_accurate",
"videoa11y_clear"

Google

Sighted Raters correlated with proposed metrics

- (Correlation)
- Inter-annotator agreement

Metric	Krippendorff's α	Avg Std Dev
Timeliness	0.721	0.581
Audio Issues	0.700	0.545
Objective	0.665	0.437



Google

Blind Low Vision Participant Study

Characteristics:

Demographics & History



Age 23-58, 4 Male, 3 Female

All participants **blind since an early age.**

Video Consumption Habits

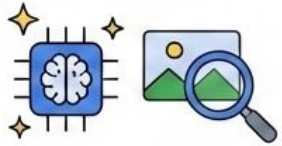


Watched videos **multiple times a day**



Watched videos with AD **less often** (e.g., weekly)

Technology & AI Use

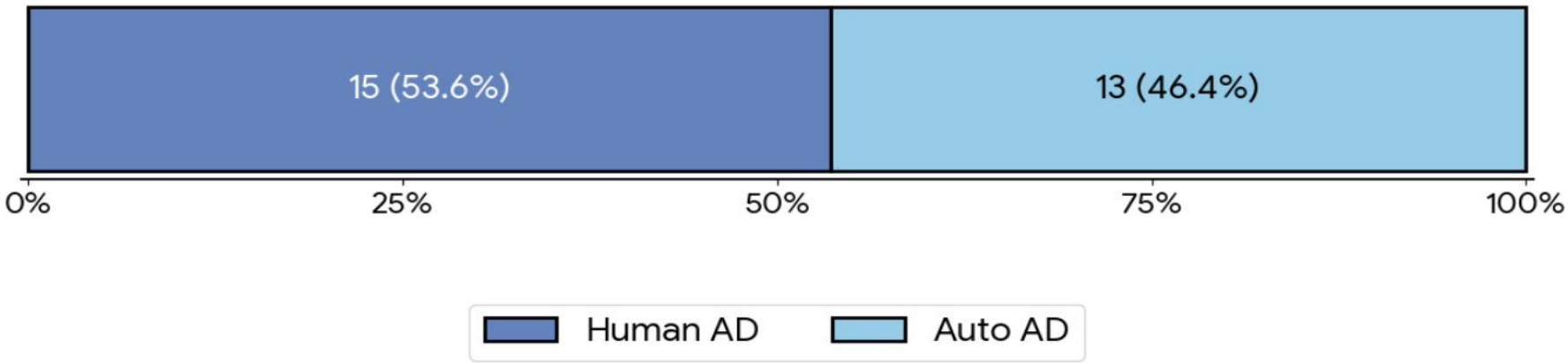


All participants **regularly used AI** to gain access to images.

Several had used it for static text descriptions of videos.

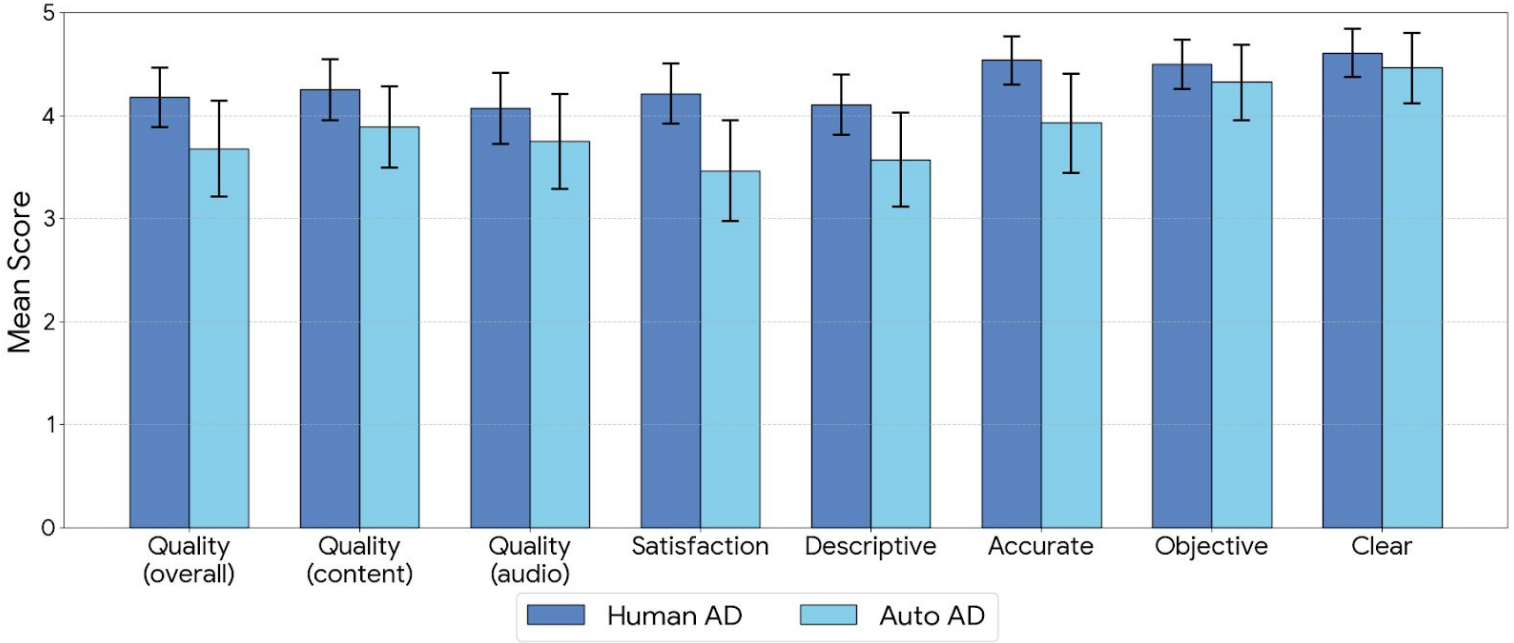
Google

Blind Low Vision Participant Study – Results



Check out the data here (See Sheet: Condition Preferences).

Overall, ratings for Auto AD closely behind Human AD

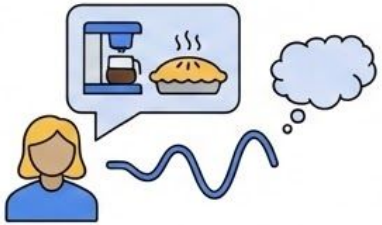


* Mean Score on a scale from 1 - low to 5 - high, error bars are 95% confidence intervals

Google

BLV Study: When did Auto AD succeed?

Complemented Narration



"it seemed to usually be adding uh adding something new to what the narrator say ... I think that's the goal of what audio description should be is add the visual context that a reasonably intelligent person would not be able to infer from listening" – **P1 (Auto AD Clip)**

Rich details



"There was more detail and they had, you know, like knowing the geyser. The swirls of the thermal... The waterfall was another detail." – **P6 (Auto AD Clip)**

"I actually figured out that there were three people. In the first one, I didn't." – **P5 (Auto AD Clip)**

Synchronization



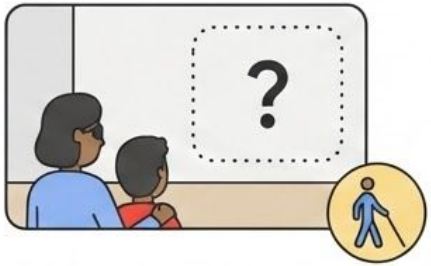
"I liked that it it seemed really synced. It didn't seem like it was trying to go over the dialogue or anything like that." – **P1 (Auto AD Clip)**

"I think the timing was good... it didn't overlap much with other audios and stuff from the video." – **P5 (Auto AD Clip)**

Google

BLV Study: What caused ratings < 4?

Missing AD



"There were, if possible, even more pauses [than in the Auto AD]... it did not describe everything and that's what you need in audio description."
– P7 (Human AD Clip)

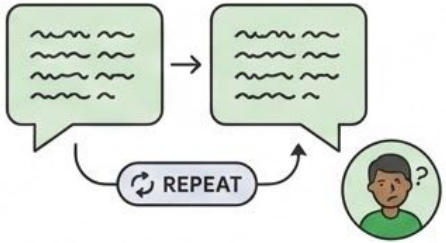
Too Much AD



"Definitely the description is overlapping the dialogue.

Which made it hard to understand both of them."
– P6 (Auto AD Clip)

Redundancy



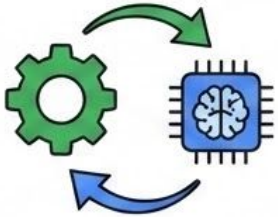
"Holy redundancy Batman... the description would say something and then the narrator would immediately say almost verbatim the exact same thing."
— P1 (Auto AD Clip)

Google

Key Takeaways



Data and Evaluation framework:
We have a professionally annotated benchmark and evaluations to measure progress.



Tech. and agentic loop:
Using mLLM is viable, and we can improve further based on our evaluation framework.



Users:
BLV users find the technology “highly beneficial” for comprehension despite the human performance gap.

Google

Are the descriptions any good?





Objective Quality Standards

AD is accurate, does not overlap, and passes rigorous objective metrics.



Consistent with Human Judgement

Objective metrics are consistent with qualitative feedback from human evaluators.



Valued by Blind Audience

Blind and Low-Vision users find them useful and enjoy the listening experience.

Mind The Gap (MTG-AD): Dataset and Fine-grained Evaluation for End-to-End Inline Audio Descriptions

Subhashini Venugopalan[1], Jenny Tan[1], Taylor Roper[1], Jimmy Tobin[1], Anton Kast[1], Alicia Martin[1], Sam Sepah[1], Amy Pavel[1,2]
[1]Google Research, [2]UC Berkeley

NeurIPS 2026

ArXiv: https://arxiv.org/abs/2610.39955

Code: https://github.com/google/mtg-ad

Dataset: hf.co/datasets/google/mind-the-gap-inline-ad