2026 SIGdial Atlanta, GA, USA

MemeBuddy

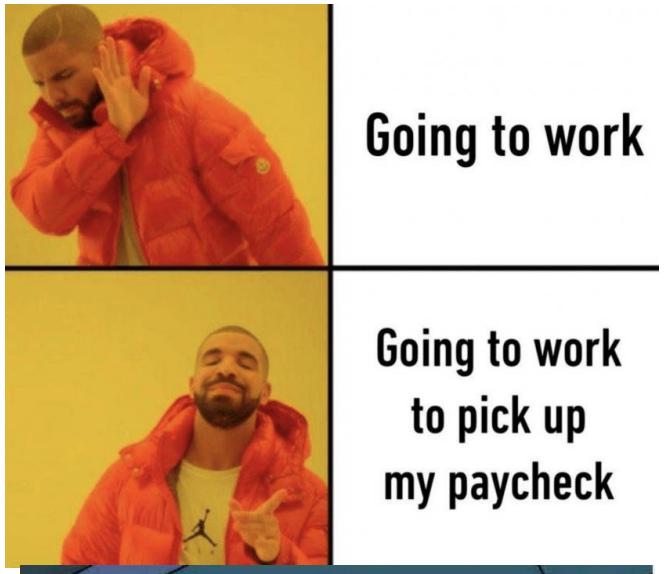
Dialog-Style Audio Representations for Engaging Non-Visual Meme Experiences

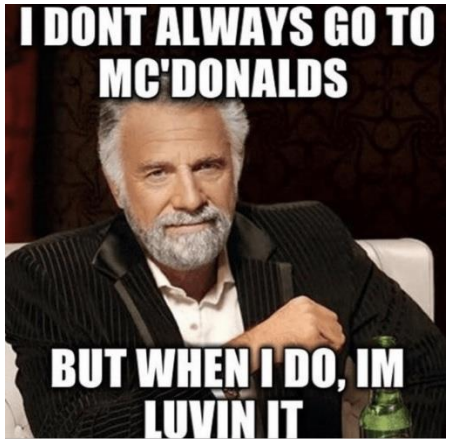
Chirag Bhansali[1] Vikas Ashok[2] Hae-Na Lee[1]

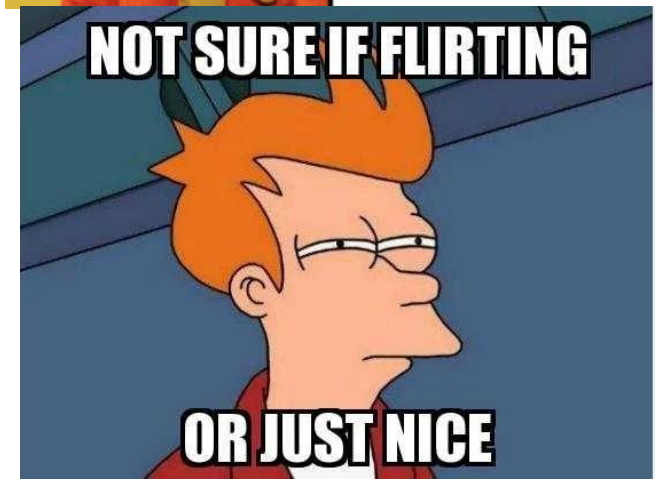
[1]Department of Computer Science and Engineering
Michigan State University

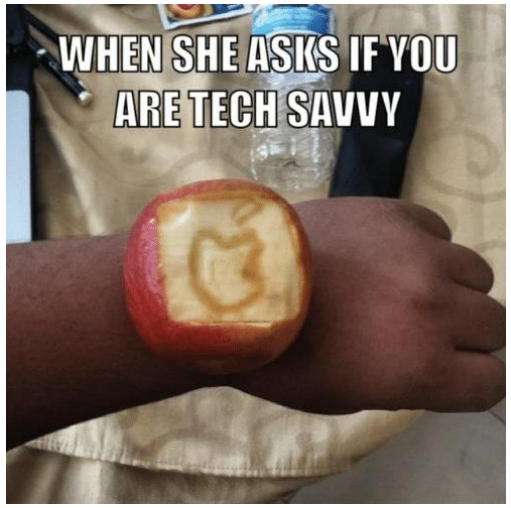
[2]Department of Computer Science
Old Dominion University

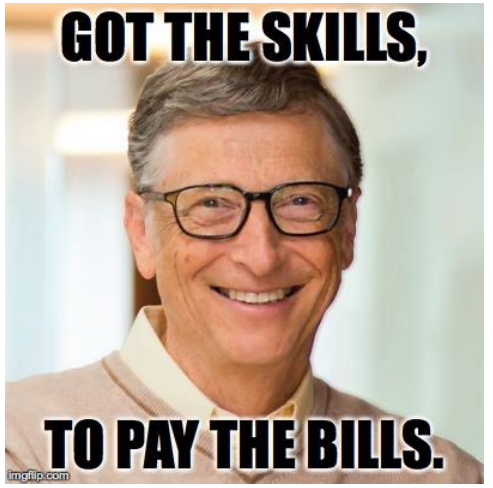
Motivation – Memes are Everywhere











Motivation

- Meme accessibility stalls at comprehension for blind people

Screen readers fall short

Blind users rely on **alt text**, which is often missing, incomplete, or insufficient for visual content

AI tools partially help

AI tools (e.g., ChatGPT, Be My AI, Seeing AI) improve comprehension, but provide caption-style, static descriptions

The engagement gap

Caption-style output misses the humor, timing, and narrative structure that make memes engaging

Key Idea

Model memes as **dialog**, not static description

- Prior work largely retains a single-speaker, descriptive paradigm

- Memes are inherently social and interpretative – meaning unfolds through reactions, timing, and shared understanding

Meaning lies not only in ***what*** is conveyed,
but in ***how*** it is structured

Contributions

Dialog-based audio representations

Generate multi-turn, role-based conversational representations of memes using a multimodal LLM

Two dialog variants + a baseline

Commentator–Commentator (CO), Commentator–Character (CH), and a caption-style baseline (BL) for comparison

Offline evaluation

Meme dataset assessed across multiple LLMs on intent clarity, fidelity, text accuracy, and speculation

User study with blind participants

Dialog-style representations improve engagement, immersion, and enjoyment while preserving comprehension

Related Work

- Prior work stays within a descriptive paradigm

- Prior work focuses primarily on **semantic access** (i.e., what is depicted in memes) with little attention to the social and experiential qualities of content
  - OCR
  - Web context
  - Automatically generated alt text
  - Facial-emotion cues
  - Emotional speech augmentation

Research Questions

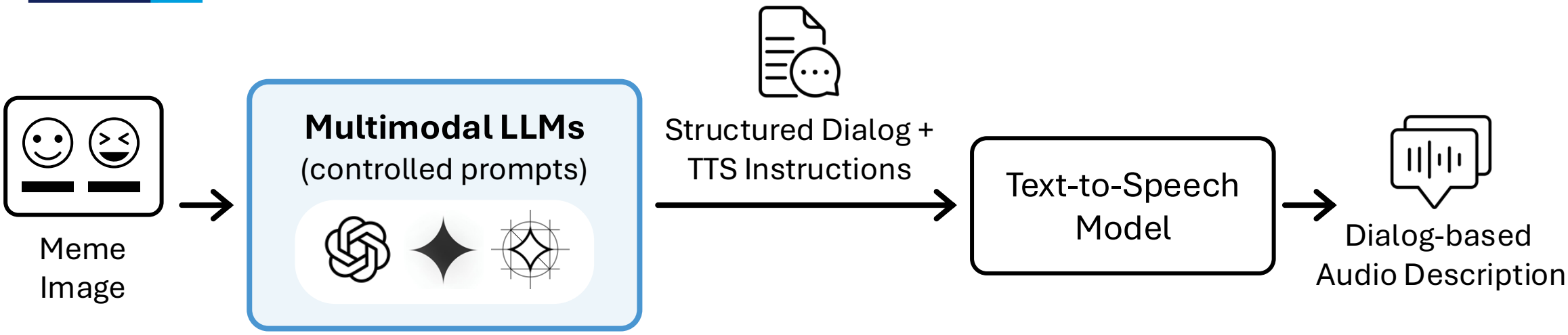
- **RQ1:** How well do the audio representations capture the intent, cultural nuances, and contextual details of a meme?

  ➔ Offline evaluation

- **RQ2:** What are the strengths and limitations of dialog-style meme representations compared to caption-style descriptions?
- **RQ3:** How can dialog-style representations be improved to enhance engagement and enjoyment?

  ➔ User study with blind participants

MemeBuddy System



- Four multimodal LLMs were compared
  - Gemini 3 Flash
  - Gemini 3 Pro
  - GPT-5.1
  - Gemma 4

Three Dialog-based Representation Strategies

Caption-style Design (Baseline)

One speaker
Narrator

- Caption-style, structured narration
- Prioritizes semantic completeness for maximal clarity

✓ Reliability

Commentator–Commentator (CO) Design

Two speakers
Host1 + Host2

- Concise two-speaker conversation, meaning distributed across turns, ends on a punchline

✓ Clarity
✓ Conversational coherence

Commentator–Character (CH) Design

Two speakers
Host + Meme Character

- The meme character speaks in first person, contributing to the narrative and punchline

✓ Immersion
✓ Expressiveness

Example – Caption-style Design

Meme Image

Narrator

Caption-style Dialog

Alright, **I've got a fun three-panel meme here featuring the actress Anna Kendrick.**
In the top panel, she's looking right at the camera with a casual, kind of amused expression. The text over her in bright yellow says, 'A CANDLE MAKER IS CALLED A "CHANDLER"'.
Then **in the middle panel,** she leans in a bit with this mischievous, excited smile, baring her teeth just a little. The text here changes to blue and reads, 'I FOUND THAT OUT ON BING...'.
Finally, **in the bottom panel,** there's no text at all. It's just Anna bursting into a big, playful laugh, looking completely proud of herself. So, **the joke is basically a giant pun.** If you put the keywords from the top and middle panels together – 'Chandler' and 'Bing' - you get Chandler Bing, the famous character from the TV show Friends. Her totally cracking up at her own dad joke is what really sells it!

Example – Dialog-based Design

Meme Image

Commentator–Commentator (CO) Dialog

Host1: This is a three-panel meme featuring actress Anna Kendrick, setting up a clever little pun.

Host2: She's got great facial expressions. **What's the setup?**

Host1: In the first panel, she has a pleasant, normal expression with text that says, 'A candle maker is called a Chandler.'

Host2: Alright, **fun little trivia fact.** What's next?

⋮

Commentator–Character (CH) Dialog

Host: Today we're looking at a three-panel meme featuring a famous actress dropping some unexpected trivia.

Character: Hi, **I'm Anna Kendrick!** In the top panel, I'm just looking perfectly normal and polite, sharing a fun fact: 'A candle maker is called a Chandler'.

Host: Then in the middle panel, you're making this hilariously forced, awkward smile with the text, 'I found that out on Bing...'

⋮

Offline Evaluation

RQ1: How well do audio representations capture the intent, cultural nuances, and contextual details of a meme?

- Evaluation setup:
  - Dataset with 100 meme images with human annotations
  - Four multimodal LLMs compared
  - **Four metrics** tailored to meme understanding

Intent Clarity

How easily setup, punchline, and tone come through (5-point Likert)

Representation Fidelity

Whether humor-critical visual elements are correctly captured

Text Accuracy

Whether on-image text is reproduced or paraphrased accurately

Speculation

Presence of unsupported, irrelevant, or hallucinated content

Offline Evaluation – Intent Clarity

Intent Clarity

How easily the meme's setup, punchline, and tone come through (5-point Likert scale)

- Gemini 3 Pro and GPT-5.1 performed best
- BL and CO conditions led, and CH trailed
  - CO closely matched BL, especially for stronger models
  - CH condition showed lower clarity ⟶ Added complexity from character/turn alignment

Offline Evaluation – Fidelity and Text Accuracy

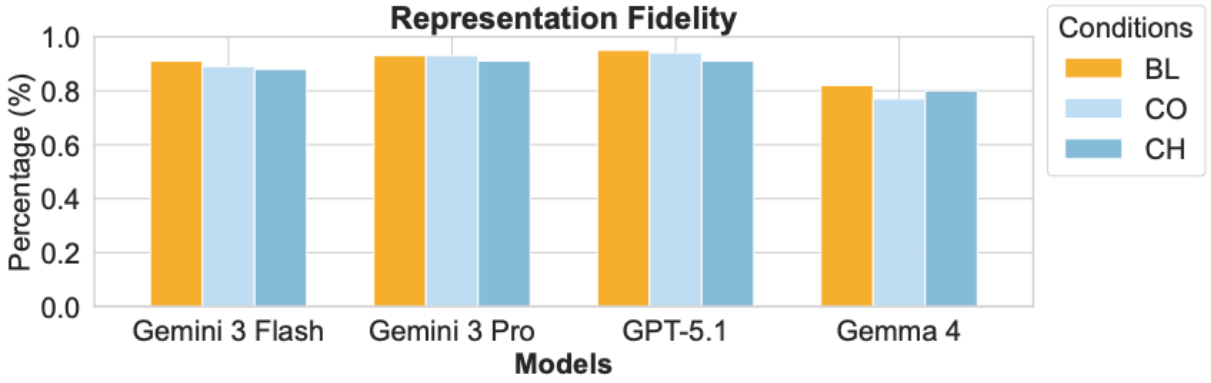
Representation Fidelity

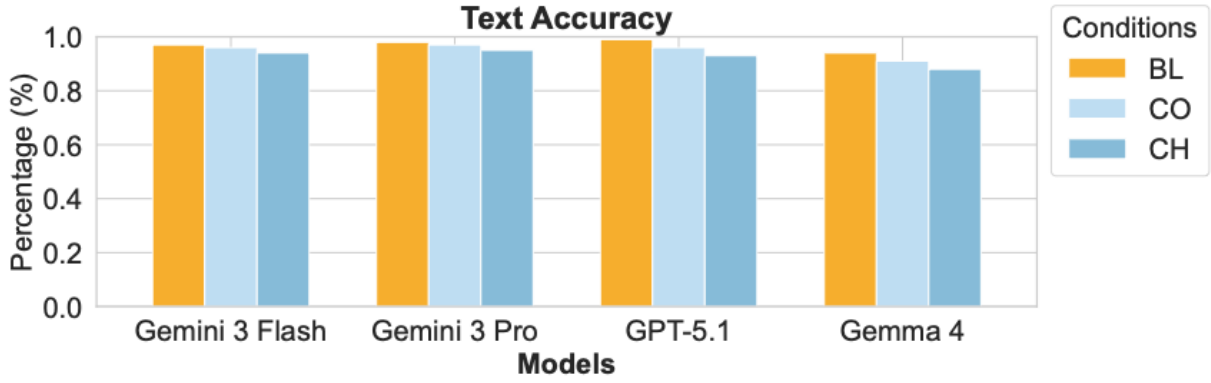
Whether humor-critical visual elements are correctly captured (incorrect/correct)

Text Accuracy

Whether on-image text is reproduced or paraphrased accurately (incorrect/correct)

- Fidelity and Text Accuracy stayed high across all three conditions





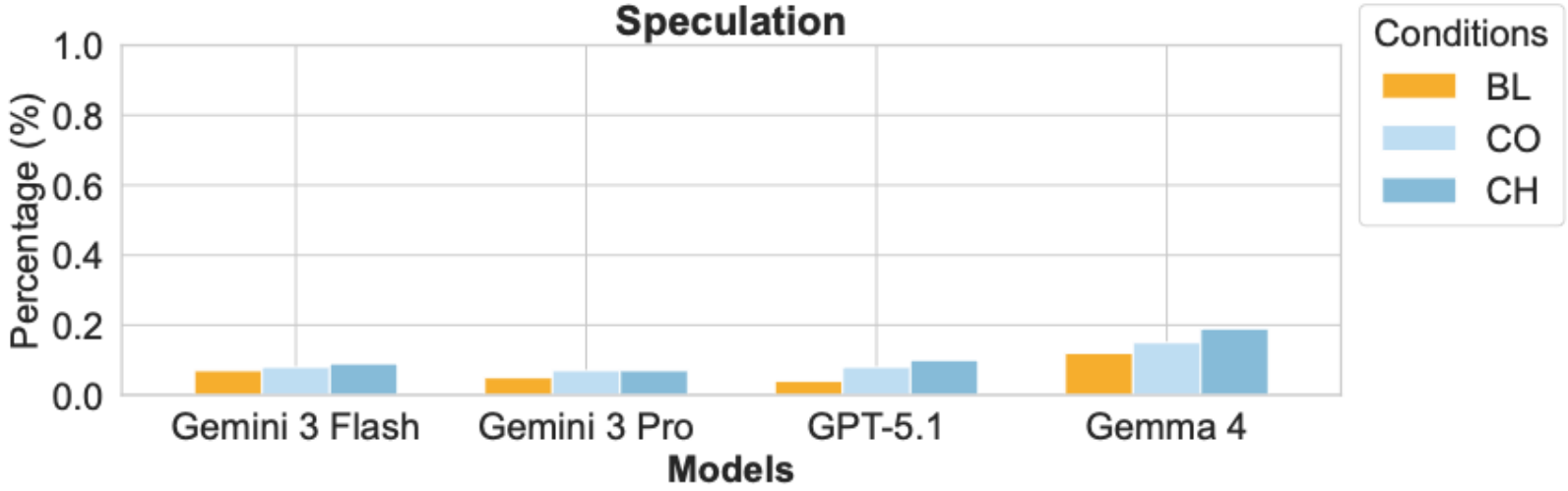
Offline Evaluation – Speculation

Speculation

Presence of unsupported, irrelevant, or hallucinated content (absent/present)

- CH condition showed higher Speculation

  ⟶ Added complexity from character/turn alignment



User Study

RQ2: What are the strengths and limitations of dialog-style meme representations compared to caption-style descriptions?

RQ3: How can dialog-style representations be improved to enhance engagement and enjoyment?

- **Task:** listen to a meme's audio representation (generated with Gemini 3 Pro) ⟶ questionnaire on comprehension, emotional response, immersion, entertainment, funniness, and willingness to engage further

14 blind participants

3 study conditions (BL, CO, CH)

6 tasks performed per participant

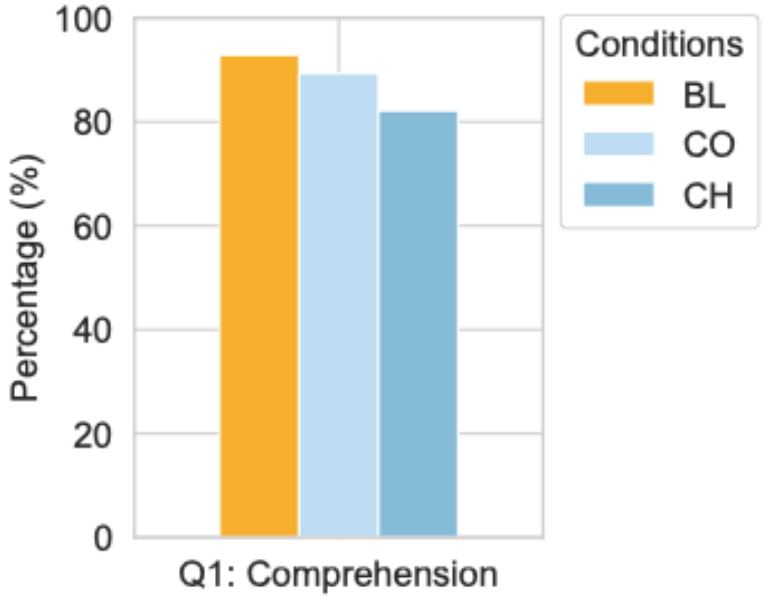
Counterbalanced memes (Latin square design)

User Study Questionnaire

Questions	
Q1: What is the author trying to convey in the meme?	Comprehension
Q2: Listening to this audio, my emotions frequently fluctuated (such as feeling happy, sad, angry, etc.).	User Experience
Q3: I was deeply immersed in this audio narrative experience.	
Q4: I found the meme audio entertaining.	
Q5: The audio format enhanced the funniness of the meme.	
Q6: I would like to listen to more memes like this.	

User Study – Comprehension

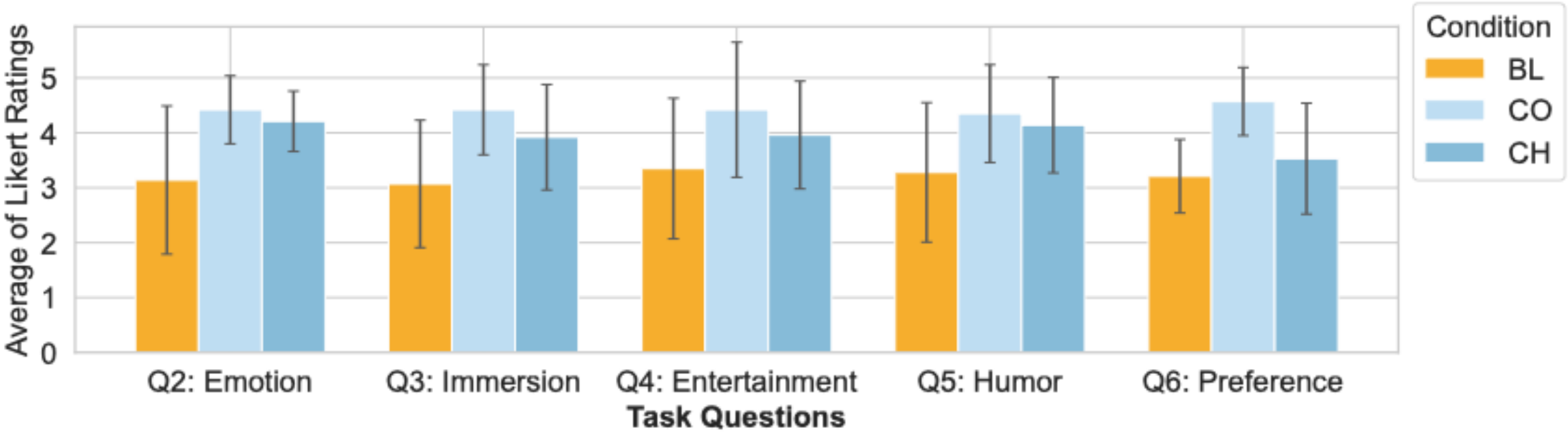
- Q1: What is the author trying to convey in the meme?
- Comprehension stayed high across all conditions
  - BL performed the best (92.85%)
    ⟶ Likely due to explicit, coverage-oriented narration
  - CO remained close to BL (89.28%)
    ⟶ Distributing information across dialog turns did not substantially hurt interpretability
  - CH showed a modest decrease (82.14%)
    ⟶ May reflect increased ambiguity in speaker roles and turn coherence



Comprehension remained above 80% in every condition

User Study – Engagement, Affect, and Preference

CO consistently yielded the highest ratings and the most robust improvements



User Study – Qualitative Insights

- Six themes from exit interviews

Two speakers improve engagement

Reduced monotony, clearer segmentation, anticipation of the next turn

Expressiveness + coherence together

Two voices expressively and coherently interacting

Dialog encourages an interactive mental model

Users wanted to replay turns, ask follow-ups, or probe unclear parts

Applicability beyond memes

Participants wanted dialog-style narration for other forms of visual content

Avoid over-explanation

Fully explicit description removed the fun of “interpreting the humor”

Need for customization

Switch between BL/CO/CH, voice styles, and turn-level navigation

Discussion

- A shift from equivalence (same information) to **experiential parity** (comparable experience), particularly for socially and culturally rich media, such as memes

- Engagement arises not from expressiveness alone, but from **structured and coherent interaction**

- Simply introducing expressive elements (e.g., meme characters) is insufficient — CO's advantage over CH shows that role clarity and turn coherence matter as much as voice variety

Limitations

- Static voice configurations likely constrained the effectiveness of CH
- English-only memes
- Static images only, excluding GIFs and videos
- 6 memes per session, limited by the 2-hour study window
- Modest sample size (14 participants) — consistent with prior blind-user studies
- Did not isolate the individual contribution of “multiple voices” vs. “temporal unfolding” to user engagement
- Cultural context inferred implicitly by the LLM, not externally grounded

Conclusion – Toward Experiential Accessibility

- **MemeBuddy** rethinks meme accessibility through dialog-style audio representation that model memes as structured two-speaker conversations

- Dialog-based representations maintain high semantic fidelity
- They improve engagement, immersion, and perceived humor over caption-style descriptions
- Optimizing for comprehension alone is insufficient for socially & culturally rich media
- Conversational structure and temporal unfolding play a central role in shaping user experience

Thank You!

Questions & discussion welcome

Chirag Bhansali Vikas Ashok Hae-Na Lee
Michigan State University Old Dominion University