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
East Lansing, MI, USA
{bhansal4, leehaena}@msu.edu

[2]Department of Computer Science
Old Dominion University
Norfolk, VA, USA
vganjigu@odu.edu

Abstract

Image memes are a pervasive form of online communication, widely used to convey humor, opinions, and cultural references. Prior work has explored making memes accessible to blind users, primarily through auto-generated descriptive captions. While these approaches improve comprehensibility and sometimes incorporate prosodic or emotional cues, they often fail to capture the humor, narrative structure, and contextual nuances that make memes engaging. We present MemeBuddy, a system that models memes as dialog, generating structured, multi-turn audio representations using role-based speakers. MemeBuddy reinterprets a meme as a conversation between two speakers, integrating extracted meme text with contextual knowledge implicitly inferred by a multimodal LLM (e.g., recognition of common meme templates and cultural references) to convey intent, timing, and implicit meaning through conversational interaction. We evaluate MemeBuddy in a user study with 14 blind participants. Results show that dialog-style meme representations consistently improve engagement and user satisfaction compared to caption-style descriptions, while maintaining comparable comprehension.

1 Introduction

Image *memes* are humorous and culturally referential visual artifacts that have become a dominant form of online communication across social media, forums, and messaging platforms (Davison, 2012; Molina, 2020; Bauckhage, 2011; Kostadinovska-Stojchevska and Shalevska, 2018; Morina and Bernstein, 2022). Users rely on memes to express opinions, share experiences, and participate in collective cultural discourse (Chen, 2012; Blommaert and Varis, 2017). Ensuring equitable access to memes is therefore critical, particularly for blind users who interact with content through synthesized speech via screen readers (e.g., JAWS, VoiceOver, NVDA).

However, meme accessibility remains limited. Screen readers primarily depend on *alt text*, which is often missing, incomplete, or insufficient for conveying meaning of images to the blind users (Voykinska et al., 2016; Gleason et al., 2019a; WebAIM, 2026). To compensate, blind users increasingly rely on AI-based tools (e.g., ChatGPT, Seeing AI, Be My AI) that generate image descriptions on demand (Sharma et al., 2025; Adnin and Das, 2024). While these systems improve *comprehensibility*, they typically produce static, caption-style outputs that fail to capture the humor, timing, and contextual nuance central to memes (Dynel, 2016; Vásquez and Aslan, 2021). As a result, the users often receive functionally correct but experientially diminished representations.

Recent work has explored enhancing engagement through emotional narration (Chen et al., 2025). However, the approach largely retains a single-speaker, descriptive paradigm. We argue that the limitation lies not only in *what* is conveyed, but in *how* it is structured. Memes are inherently social and interpretative; their meaning often unfolds through reactions, timing, and shared understanding. This motivates modeling memes as *dialogs* rather than static descriptions.

Towards this, we present MemeBuddy, a system that generates dialog-style audio representations of memes using role-based speakers. Specifically, MemeBuddy reimagines a meme as a structured, multi-turn conversation between two speakers, enabling the gradual unfolding of humor, intent, and cultural context. As illustrated in Table 1, MemeBuddy supports two dialog variants: (i) commentator–commentator (CO), where two speakers collaboratively interpret a given meme, and (ii) commentator–character (CH), where a speaker interacts with a character from the meme. A caption-style baseline (BL) is also supported for comparison.

We evaluate MemeBuddy through a user study with 14 blind participants. Results show that dialog-

Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 649–667
August 2–5, 2026.

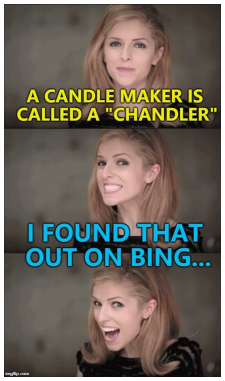
Meme Image	Baseline (BL) One speaker	Commentator–Commentator (CO) Two speakers	Commentator–Character (CH) One speaker and a meme character
	**Narrator:** Alright, I've got a fun three-panel meme here featuring the actress Anna Kendrick. In the top panel, she's looking right at the camera with a casual, 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!	**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? **Host1:** The second panel shows her making a cheesy, awkward grin. The text adds, 'I found that out on Bing...' **Host2:** Wait... Chandler... Bing. Oh, wow. **Host1:** Exactly. And the final panel is just a picture of her laughing enthusiastically at her own joke. **Host2:** So basically, it's just an elaborate dad joke hiding a 'Friends' reference, and she is absolutely thrilled with her own genius.	**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...' **Character:** Exactly. And by the bottom panel, I'm just flashing a massive, exaggerated, cheesy grin with no text at all to let the joke land. **Host:** Chandler Bing. The ultimate 'Friends' dad-joke, delivered with the perfect sarcastic smirk.

Table 1: Illustration of MemeBuddy-generated audio representations for a given meme image.

style representations consistently improve engagement, immersion, and enjoyment compared to caption-style descriptions, while maintaining comparable comprehension. Our findings further reveal trade-offs between conversational coherence and expressiveness, offering insights into how non-visual content representations can be made both accessible and engaging for blind users.

In summary, we contribute: (i) Novel dialog-based audio representations using role-based speakers for engaging meme consumption; and (ii) An empirical evaluation demonstrating that dialog structure consistently improves meme engagement while preserving comprehension.

2 Related Work

Our work closely relates to prior efforts in making visual media accessible, as well as research on meme understanding and accessibility.

2.1 Accessibility of Visual Web Content

Prior work has explored improving blind and low vision users' access to visual web content, including images, memes, GIFs, and videos (Gleason et al., 2019a; Prakash et al., 2024; Lee and Ashok, 2021). Early systems, such as *WebInSight*, generated alt text using web context and OCR, highlighting both the prevalence of missing descriptions and the potential of automation (Bigham et al., 2006). Later approaches introduced automatic image descriptions (e.g., Automatic Alt-text (Wu et al., 2017)), yet studies show that both user-provided and AI-generated descriptions are often sparse or misaligned with users' needs (Morris et al., 2016; Voykinska et al., 2016; Gleason et al., 2019a). Large-scale audits confirm that many images still lack meaningful alt text (WebAIM, 2026).

Beyond basic descriptions, research has examined richer representations of visual content. Previous works highlight the need for flexible, context-sensitive descriptions and interactive access to visual structure (Morris et al., 2018; Stangl et al., 2021; Lee et al., 2021). Accessibility efforts have also extended to dynamic media, including memes, GIFs, and videos, through approaches such as emotion-aware captions, expressive narration, and

hierarchical summaries (Gleason et al., 2019b; Prajwal et al., 2019; Chen et al., 2025; Van Daele et al., 2024; Zhang et al., 2022). However, AI-based tools, while increasingly used in practice, still exhibit limitations in usability and experiential quality (Gonzalez Penuela et al., 2024; Adnin and Das, 2024; Sharma et al., 2025).

Overall, prior work primarily focuses on *semantic access*, i.e., helping users understand what is depicted, while retaining caption-style paradigms. Less attention has been given to preserving *social and experiential qualities* of content. Our work addresses this gap by modeling memes as dialog, enabling engagement as well as comprehension.

2.2 Memes and their Accessibility

Internet memes are a multimodal and culturally-grounded form of communication that combine visual and textual elements to convey humor, opinions, and social meaning (Davison, 2012; Molina, 2020; Bauckhage, 2011; Blommaert and Varis, 2017; Kostadinovska-Stojchevska and Shalevska, 2018; Morina and Bernstein, 2022; Chen, 2012). Understanding memes requires reasoning over visual metaphor, inter-textual references (e.g., celebrities, popular culture), and contextual knowledge (Vásquez and Aslan, 2021; Hwang and Shwartz, 2023; Joshi et al., 2024), making them more complex than standard images. Prior works have addressed meme understanding through captioning, explanation, and multimodal reasoning (Hwang and Shwartz, 2023; Sharma et al., 2023; Zhong and Baghel, 2024), as well as emotion and metaphor analysis (Xu et al., 2022; Sharma et al., 2024). These efforts highlight the inherently inferential nature of memes, where meaning extends beyond literal content.

Prior accessibility work has focused on enriching meme descriptions. Gleason et al. (Gleason et al., 2019b) proposed structured descriptions beyond OCR text, while Prajwal et al. (Prajwal et al., 2019) incorporated facial-emotion cues. More recently, *MemeSpeak* (Chen et al., 2025) demonstrated that emotional speech can enhance meme narration. However, existing approaches largely remain within a *descriptive paradigm*, typically relying on single-speaker narration. This overlooks the inherently social and temporal nature of memes, where meaning unfolds through timing, interaction, and interpretation. As a result, prior work improves *understandability* but does not fully address *engagement*. Our work addresses this gap by modeling memes as dialogs through *structural re-representation*. MemeBuddy generates multi-turn, role-based conversational representations that enable meaning to gradually unfold through interaction, providing a more engaging non-visual meme experience while complementing prior approaches in this area.

3 MemeBuddy Design

The design of MemeBuddy was informed by both prior research on dialog as an effective medium for human-information interaction (Ainsworth and VanLabeke, 2004; Chi et al., 2017; Ghazarian et al., 2021; Mizukami et al., 2016; Shuster et al., 2020) and insights from an IRB-approved focus group study. The study involved seven participants with complementary expertise: two authors, a psycholinguist, a media researcher, a social computing researcher, and two blind university students with extensive experience in meme consumption and creation on social media platforms.

The focus group was conducted as a structured design elicitation session. The participants were guided using seed themes derived from prior work on accessible content representation and conversational interfaces, and were asked to collaboratively explore how visual memes could be translated into engaging non-visual audio formats. Discussions centered on key design dimensions, including number of speakers, distribution and timing of information across turns, mechanisms for conveying humor and sarcasm, and desirable prosodic characteristics of voices. All sessions were audio-recorded and transcribed for subsequent analysis.

We analyzed the transcripts using a qualitative coding process following open and axial coding procedures (Saldana, 2011). In the open coding phase, recurring concepts (e.g., "turn-taking improves clarity", "characters increase immersion but may confuse roles") were identified. These codes were then grouped during axial coding into higher-level design principles. Two dominant dialog structures emerged: (i) a commentator–commentator format emphasizing clarity and conversational coherence, and (ii) a commentator–character format emphasizing immersion and expressiveness. Additionally, participants consistently requested a caption-style fallback to ensure reliability when dialog representations are insufficient. We operationalized these insights as prompt-level constraints in MemeBuddy (Appendix A).

At a system level, MemeBuddy uses controlled prompts to a multimodal LLM to transform an input meme image into structured outputs containing both dialog and corresponding text-to-speech (TTS) instructions. These outputs are rendered as audio using expressive TTS engines. The strategies are described next.

3.1 Commentator-Commentator (CO) Design

The group strongly preferred two-speaker dialogs over multi-speaker alternatives, citing reduced cognitive load, improved coherence, and avoidance of unnecessary complexity. We encoded these preferences directly into the LLM prompt used for CO dialog generation (Table 5 in Appendix A).

In this design, the model produces a structured conversation between two speakers: *Host1*, who provides explanatory content, and *Host2*, who offers reactions, questions, or informal interpretations. The dialog is intentionally concise, with meaning distributed across turns to support gradual understanding and anticipation, and concludes with a punchline-style summary that conveys the meme's intent. The prompt also specifies TTS constraints (e.g., distinct voice styles and genders) to ensure clear speaker separation. Outputs are generated in a structured JSON format containing both dialog content and TTS specifications, enabling consistent downstream rendering.

3.2 Commentator-Character (CH) Design

To enhance immersion, participants proposed incorporating a character from the meme as an active speaker in the conversation. We operationalized this idea through the CH design, implemented via prompt constraints (Table 6 in Appendix A).

In this configuration, the dialog occurs between a *Host* (commentator) and a *Guest* (a character from the meme). The Guest may adopt a first-person perspective or respond in a manner consistent with the character's role, identity, or persona. The prompt enforces explicit self-introduction of the Guest to maintain role clarity, and guides the interaction such that the character meaningfully contributes to the narrative and punchline. Similar to CO, the dialog remains concise and structured, and includes TTS instructions specifying distinct and contextually appropriate voice characteristics. Compared to CO, which prioritizes interpretability through neutral commentary, CH emphasizes *expressiveness and narrative immersion*.

3.3 Caption-style Design (Baseline)

We also include a caption-style baseline (BL) in MemeBuddy to reflect current practices and provide a reliable fallback. This design was strongly recommended by participants for scenarios requiring maximal clarity and explicitness.

The BL prompt (Table 4 in Appendix A) generates a single-speaker narration delivered by a *Narrator*. The narration follows a structured progression: identifying the meme template (when applicable), describing visual elements in detail (e.g., layout, expressions, embedded text), and concluding with an explicit explanation of the meme's intent. The prompt encourages a natural, conversational tone while maintaining completeness and self-containment. It also specifies a consistent TTS configuration (warm, descriptive voice) to ensure intelligibility. In contrast to dialog-based designs, BL prioritizes *semantic completeness* over interaction and interpretive engagement.

3.4 Offline MemeBuddy Evaluation

Following standard practice in dialog and LLM-output evaluation, we conducted an offline human assessment of MemeBuddy-generated meme representations. The evaluation was designed to address the following research question: **RQ1:** *How well do the audio representations capture the intent, cultural nuances, and contextual details of a meme?*

3.4.1 Dataset Collection and Annotation

We constructed a dataset of 100 meme images to support the offline evaluation, ensuring diversity in structure, content, and style. The dataset was compiled from two sources: the Kaggle Meme Images Dataset (Kaggle, 2026) and the Imgflip platform (Imgflip LLC, 2026).

The Kaggle dataset, originally curated for sentiment analysis, contains a mixture of general images and memes. To ensure relevance, we filtered this dataset to retain only images that conform to common meme characteristics (e.g., presence of overlaid text, recognizable templates, or multimodal humor). From this filtered subset, we randomly sampled 62 memes. To complement this, we additionally collected 38 popular meme instances from Imgflip, prioritizing templates with high usage frequency and cultural familiarity.

We ensured that all selected memes contained both visual and textual components, as such multimodal interplay is central to meme interpretation and poses known challenges for blind users. We

excluded memes containing hateful, abusive, or politically sensitive content to avoid confounds.

Each meme was manually annotated using a structured schema (Table 7 in Appendix B) designed to capture both low-level visual details and high-level semantic intent. Specifically, annotations included template identity, panel structure, key visual elements, verbatim textual content, and a concise intent summary describing the humor mechanism. Template names were validated against popular meme repositories (e.g., Imgflip, Kapwing) to ensure consistency and correctness.

To improve annotation reliability, two annotators independently labeled each meme, followed by a reconciliation process to resolve disagreements through discussion. This process ensured consistency in interpreting subjective fields, such as `intent summary`, which are critical for evaluating generated outputs. The resulting dataset provides a grounded reference for assessing the quality and fidelity of MemeBuddy representations.

3.4.2 Models and Metrics

We evaluated MemeBuddy representations generated using multiple LLMs (with temperature set to 0 to minimize random results). Specifically, we experimented with: (i) Gemini 3 Flash, (ii) Gemini 3 Pro, (iii) GPT-5.1, and (iv) Gemma 4. For each meme, all models were prompted using the same instructions (Tables 4, 5, and 6) for each representation type (BL, CO, CH) to ensure consistency.

Evaluating dialog-like and multimodal outputs remains challenging, as standard n-gram metrics (e.g., BLEU) fail to capture semantic and experiential quality (Lowe et al., 2016). Recent work, particularly the FED framework, has advocated for fine-grained human evaluation along multiple dimensions, such as coherence, engagement, grounding, and correctness (Mehri and Eskenazi, 2020). Guided by this framework, we define the following metrics tailored to meme understanding: (i) ***Intent Clarity***: How easily a listener can understand the meme's setup, punchline, and tone; represented as a 5-point Likert item (1 = very difficult, 5 = very easy); (ii) ***Representation Fidelity***: Whether humor-critical visual elements (e.g., expressions, relationships, scene structure) are correctly captured; binary representation (0 = incorrect/missing, 1 = correct); (iii) ***Text Accuracy***: Whether on-image text is reproduced or paraphrased accurately; binary representation (0 = incorrect, 1 = correct); and (iv) ***Speculation***: Presence of unsupported, irrelevant, or hallucinated content; binary representation (0 = absent, 1 = present).

Each <meme, model, representation> triplet was independently evaluated by two annotators who were not part of the focus group, improving external validity. The annotators were provided with the original meme, generated outputs, and the dataset annotations (Table 7) as grounding reference. Items were presented in randomized order via a custom web interface to mitigate ordering effects.

We computed inter-annotator agreement for each metric (Cohen's κ for binary measures and weighted κ for Likert ratings). Rather than resolving disagreements, we retained annotator variability by aggregating ratings across annotators, following prior work (Mehri and Eskenazi, 2020; Deriu et al., 2021). Specifically, Likert ratings were averaged and binary metrics were reported as proportions. All subsequent analyses were conducted on these aggregated scores.

3.4.3 Results

Table 2 summarizes results across 100 memes and two annotators. Inter-annotator agreement was strong (Intent Clarity: $\kappa = 0.78$; Representation Fidelity: 0.71; Text Accuracy: 0.86; Speculation: 0.82), supporting reliability (variance statistics in Appendix C). Gemini Pro and GPT performed the best, followed by Gemini Flash, with Gemma trailing. BL and CO exhibited high *Intent Clarity*, while CH was lower across all models. In contrast, *Fidelity* and *Text Accuracy* remained high across all conditions.

Friedman tests showed a significant effect of condition on *Intent Clarity* for all models (Gemini Flash: $\chi^2(2) = 28.34$, $p < .001$, $W = 0.142$; Gemini Pro: $\chi^2(2) = 9.64$, $p = .008$, $W = 0.048$; GPT: $\chi^2(2) = 19.18$, $p < .001$, $W = 0.096$; Gemma: $\chi^2(2) = 29.49$, $p < .001$, $W = 0.147$), with small effect sizes. Post-hoc Wilcoxon signed-rank tests (with Holm's correction) showed CH was significantly worse than BL and CO for Gemini Pro, Gemini Flash, and GPT; for Gemma, all pairwise differences were significant (BL $>$ CO $>$ CH). No significant effects were observed for *Fidelity* and *Speculation*; for *Text Accuracy*, marginal omnibus effects were observed for GPT and Gemma, but these did not yield significant pairwise differences.

Absolute differences in *Intent Clarity* were small (typically ≤ 0.3), indicating limited practical degradation. CO closely matched BL, especially for stronger models, showing that dialog structuring

Model	Intent Clarity ↑			Representation Fidelity ↑			Text Accuracy ↑			Speculation ↓		
	BL	CO	CH	BL	CO	CH	BL	CO	CH	BL	CO	CH
Gemini 3 Flash	**4.20**	4.12	3.94	**0.91**	0.89	0.88	**0.97**	0.96	0.94	**0.07**	0.08	0.09
Gemini 3 Pro	**4.51**	4.42	4.27	**0.93**	**0.93**	0.91	**0.98**	0.97	0.95	**0.05**	0.07	0.07
GPT-5.1	**4.43**	4.32	4.10	**0.95**	0.94	0.91	**0.99**	0.96	0.93	**0.04**	0.08	0.10
Gemma 4	**3.68**	3.46	3.23	**0.82**	0.77	0.80	**0.94**	0.91	0.88	**0.12**	0.15	0.19

Table 2: Offline evaluation results aggregated across 100 memes and two annotators. Intent Clarity is a mean Likert score; Representation Fidelity, Text Accuracy, and Speculation are proportions (higher is better except Speculation).

preserves semantic quality when coherence is maintained. In contrast, CH exhibited lower clarity and higher speculation, suggesting added complexity from character role modeling and turn alignment.

Overall, results indicate that dialog representations largely preserve comprehensibility, with modest degradation mainly in CH. We next examine if these structural differences impact *engagement and enjoyment* via a user study with blind users.

4 User Study with Blind Users

We conducted an IRB-approved user study with blind participants to evaluate MemeBuddy and address the following research questions: **RQ2:** *What are the strengths and limitations of dialog-style meme representations compared to caption-style descriptions?* and **RQ3:** *How can dialog-style representations be improved to enhance engagement and enjoyment?*

4.1 Study Design

We recruited 14 blind participants through email lists and word-of-mouth. All participants were familiar with AI-based image description tools and reported regular exposure to memes. Table 11 (Appendix D) shows the demographics of participants.

We used a within-subject design with three conditions:

- **Baseline (BL):** caption-style audio description by a single speaker.
- **Commentator–Commentator (CO):** dialog between two speakers.
- **Commentator–Character (CH):** dialog between a speaker and a meme character.

The BL condition reflects the dominant paradigm used by current AI-based accessibility tools, where images are described through single-speaker, caption-style narration. As such, it provides an ecologically valid and deployment-relevant baseline for comparison. Each participant completed 6 tasks (2 per condition), given the session time limit of 2 hours. In each task, the participant listened to a meme's audio representation and completed a questionnaire measuring comprehension, emotional response, immersion, entertainment, funniness, and willingness to engage further (Table 3). Q1 was open-ended; Q2–Q6 used 5-point Likert scales adapted from prior work (Chen et al., 2025; Xu et al., 2024; O'Brien et al., 2018; Richardson et al., 2018; Nabi et al., 2006; Ryan et al., 1983). These measures capture dialog effectiveness from a user perspective, reflecting perceived coherence, pacing, and conversational engagement rather than purely structural dialog metrics. We used Gemini 3 Pro to generate the representations, given its best performance in the offline evaluation.

To control for ordering and learning effects, we used six distinct memes (Appendix E) and counterbalanced study conditions using a Latin square design (Bradley, 1958). The selected memes were self-contained and required only common background knowledge, minimizing the influence of participants' prior familiarity with specific topics and enabling a fair comparison across conditions. The study was conducted in person using a Windows laptop with JAWS and NVDA installed. Following a brief practice session, participants completed all six tasks and then participated in a semi-structured exit interview. Each session lasted approximately two hours, and participants received an $80 gift

Questions
Q1: What is the author trying to convey in the meme?
Q2: Listening to this audio, my emotions frequently fluctuated (such as feeling happy, sad, angry, etc.).
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.

Table 3: User study questions.

card as compensation.

4.2 Measures and Analysis

With participants' written consent, we collected questionnaire responses, think-aloud observations, and interview data. For Q1, responses were manually coded for correctness against the intended meme meaning. For Q2–Q6 (ordinal repeated measures), we used Friedman tests followed by Wilcoxon signed-rank post-hoc tests with Holm correction. Qualitative data were analyzed using open and axial coding (Saldana, 2011). Two authors independently conducted initial coding, resolved discrepancies through discussion, and collaboratively derived higher-level themes.

4.3 Results

Comprehension (Q1). We coded open-ended responses against the intended meme meaning (binary correct/incorrect). Two annotators achieved high agreement (Cohen's $\kappa = 0.87$), with disagreements resolved by a third annotator.

All conditions achieved high comprehension: BL (92.85%), CO (89.28%), and CH (82.14%). BL performed best, likely due to its explicit and coverage-oriented narration. CO remained close to BL, suggesting that distributing information across dialog turns does not substantially affect interpretability. CH showed a modest decrease, which may reflect increased ambiguity in speaker roles and turn coherence. Importantly, comprehension remained above 80% across all conditions.

Engagement, Affect, and Preference (Q2–Q6). Results are shown in Figure 1. We analyzed Likert responses using Friedman tests with Wilcoxon signed-rank post-hoc tests (Holm-corrected); full statistics are reported in Appendix F. Friedman tests indicated a significant effect of condition for all measures ($p < .001$). However, post-hoc comparisons revealed more nuanced patterns.

Key patterns: (i) **Emotional response (Q2)** increased significantly under CO and CH compared to BL; (ii) **Immersion (Q3)** improved under CO and CH, with CO outperforming BL and CH; (iii) **Entertainment and funniness (Q4–Q5)** showed selective improvements: CO significantly outperformed CH (Q4) and BL (Q5), while other pairwise differences were not consistently significant; (iv) **Preference (Q6)** showed the strongest effect, with CO significantly outperforming both BL and CH, while CH did not significantly differ from BL.

Overall, CO consistently yielded the highest ratings and most robust improvements across measures, whereas CH showed more variable gains. These results suggest that engagement benefits arise not only from multi-speaker dialog, but also from maintaining conversational coherence. Collectively, the findings indicate that structural re-representation plays a central role in shaping affective and experiential aspects of meme consumption.

Repeated Listening and Interaction Effort. Participants replayed memes under the CO condition more frequently for enjoyment, whereas replays under BL were primarily for clarification, based on think-aloud observations. Informally, CO appeared to reduce perceived effort by segmenting information into conversational turns. CH exhibited mixed patterns, with replays occurring both for enjoyment and for disambiguation when speaker roles were unclear. These patterns align with Q3–Q6 outcomes and suggest that *turn-level segmentation* contributes to engagement.

Qualitative Insights. Six themes emerged from interview data analysis.

(i) Two speakers improve engagement. Participants consistently preferred dual-speaker dialog over single-speaker narration, citing reduced monotony, improved pacing, and clearer segmentation of information. Turn-taking helped distribute content into manageable units, making it easier to follow and less cognitively demanding. Participants also noted that dialog introduced anticipation, as they expected reactions or follow-ups from the second speaker. As one participant remarked, *"It feels like when two people are talking, it's not boring anymore."* These benefits were most consistently observed in the CO condition, where conversational flow was clear and predictable; CH also improved engagement over BL, but less reliably.

(ii) Emotional expressiveness and coherence jointly shape experience. Participants emphasized that engagement depends not only on having multiple voices, but also on how expressively and coherently those voices interact. While CH introduced character-driven dialog, its effectiveness was sometimes limited by flat prosody or unclear conversational roles. Participants suggested richer emotional cues (e.g., sarcasm, emphasis, pacing) and more consistent voice behavior to make interactions feel natural. For example, one participant noted, *"If the character sounded more expressive, it would feel like a real clip."* These observations

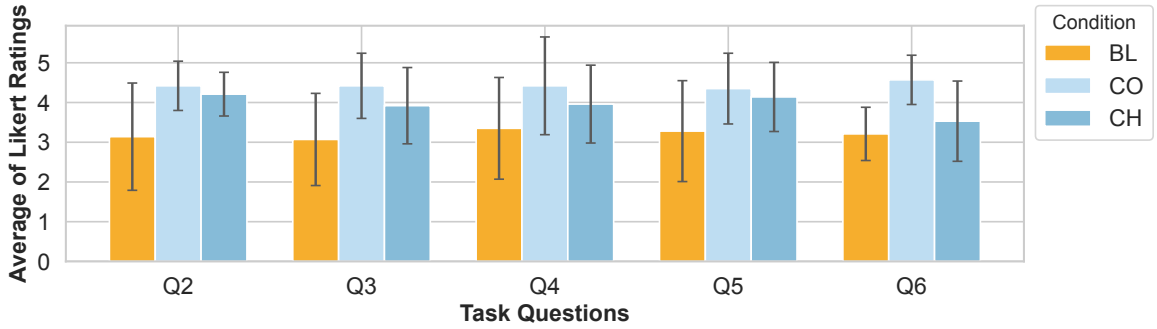


Figure 1: Average Likert ratings (Q2 – Q6) in the questionnaire. All effects significant ($p < .001$). CO consistently outperforms BL, and CH shows intermediate gains.

help explain why CH condition improved engagement over BL condition but did not match the consistency of CO condition, where speaker roles were clearer and interactions more coherent.

(iii) Dialog encourages an interactive mental model. Participants desired to engage more actively with dialog-style representations, such as replaying specific turns, asking follow-up questions, or probing unclear parts. This suggests that dialog naturally shifts users from passive consumption toward a more interactive interpretation process. As one participant put it, *"I wish I could jump in and ask what they meant."* This expectation of interaction was less evident in the BL condition, which was perceived as more static and explanatory.

(iv) Applicability beyond memes. Participants noted that dialog-style representations could extend beyond memes to other forms of visual content, particularly personal photos. They felt that dialog could better convey emotional context and lived experience compared to static descriptions. One participant explained, *"This would be great for my photos ... it would bring back memories better than just descriptions."* This suggests broader applicability of dialog-based representations for experiential accessibility.

(v) Avoid over-explanation. Participants preferred representations that preserve some level of inference rather than fully explaining the meme. BL descriptions were often perceived as overly explicit, reducing enjoyment by eliminating the need to interpret the humor. As one participant noted, *"It tells me everything, so there's nothing left to figure out."* In contrast, dialog, particularly in CO, allowed meaning to unfold across turns, supporting humor through timing and partial revelation.

(vi) Need for customization. Participants highlighted the importance of flexibility in representation. They expressed interest in switching between BL, CO, and CH depending on context, as well as customizing voice styles and navigating content at finer granularity (e.g., replaying individual turns). One participant suggested, *"I'd like to go back to just one part if I miss something."* These preferences indicate that no single representation is universally optimal, and that adaptable interfaces may better support diverse user needs.

5 Discussion

Our results show that dialog-style representations can improve the non-visual experience of memes for blind users, while maintaining high comprehension (Q1). However, these improvements are not uniform across conditions or metrics, with the commentator–commentator (CO) design consistently yielding the most robust gains.

RQ2: Strengths and limitations. Dialog-style representations shift meme access from an *informational* toward a more *experiential* paradigm. While caption-style descriptions (BL) ensure semantic clarity (Gonzalez Penuela et al., 2024; Gleason et al., 2019b), they often fail to capture humor, timing, and cultural nuance. In contrast, dialog representation distributes meaning across turns, enabling users to infer intent through temporal unfolding. This is reflected in significant improvements in emotional response (Q2), immersion (Q3), and user preference (Q6), with more selective gains for entertainment and funniness (Q4–Q5).

These findings align with prior work on narrative engagement and experiential design, which emphasize temporal structure and anticipation (Ryan et al., 1983; Ghazarian et al., 2021). Dialog introduces rhythm and contrast, allowing humor to emerge through interaction rather than explicit explanation.

At the same time, effectiveness depends critically on conversational coherence. The consistent

advantage of CO over CH suggests that simply introducing additional expressive elements (e.g., meme characters) is insufficient. In CH, role ambiguity and breakdowns in turn coherence occasionally increased cognitive effort and reduced immersion. This highlights an important design principle: *engagement arises not from expressiveness alone, but from structured and coherent interaction*; expressive elements that disrupt role clarity can reduce both interpretability and immersion.

RQ3: Improving engagement. Our findings suggest three directions for improvement. First, *emotional expressiveness* remains critical. Participants noted that richer prosody (e.g., sarcasm, emphasis) would improve realism and engagement. While prior work demonstrates benefits of emotional narration (Chen et al., 2025), our results suggest these gains may be amplified when applied at the level of dialog. Second, dialog representations naturally invite *interaction*. Participants expressed interest in asking follow-up questions and revisiting specific turns, suggesting that dialog-based systems should evolve toward interactive conversational interfaces. This aligns with emerging reliance on conversational visual assistants (Pal et al., 2026; Sindhu et al., 2025). Third, *customization* is essential. Participants requested flexibility in representation style, voice characteristics, and navigation granularity. This indicates that future systems should support adaptive, user-controlled configurations rather than a single fixed representation.

Implications for Accessible Content Design. Our work extends prior research on accessible image descriptions (Chen et al., 2025; Xu et al., 2024), which has largely focused on semantic comprehension. We show that *structural re-representation*, specifically dialog-style rendering, plays a central role in enhancing engagement.

This perspective aligns with prior work emphasizing that user experience depends on both content and presentation (Chen et al., 2025). By modeling memes as conversations, MemeBuddy better reflects their inherently social nature. More broadly, our findings suggest a shift from *equivalence* (same information) to *experiential parity* (comparable experience), particularly for socially and culturally rich media such as memes.

Limitations. Our work has several limitations. First, static voice configurations likely constrained the effectiveness of CH; future work should explore context-aware expressive speech. Second, we only considered English memes, limiting generalizability. Third, we focus on static images, excluding GIFs and videos where dialog representations may be even more effective. Fourth, each study session was limited to six meme tasks to fit within the two-hour session duration. Future longitudinal in-the-wild studies should evaluate dialog-style representations using substantially larger and more diverse meme collections spanning a broader range of topics, formats, and cultural contexts. Fifth, our user study sample size is modest, however, it is consistent with prior user studies involving blind participants, where recruitment is challenging and within-subject designs provide sufficient sensitivity for comparative evaluation. Sixth, we did not isolate the individual contributions of MemeBuddy's two key design elements, i.e., multiple voices and temporal unfolding, to user engagement. Although participants consistently identified both as important during the exit interviews, future work should disentangle their effects through controlled experiments to quantify their respective contributions to engagement, immersion, and enjoyment. Lastly, we note that contextual interpretation (e.g., cultural references) is inferred implicitly by the LLM and not externally grounded, which may introduce errors for niche memes.

6 Conclusion

We presented MemeBuddy, a system that rethinks meme accessibility through dialog-style audio representations that model memes as structured two-speaker conversations. Across an offline evaluation and a user study with blind participants, we show that dialog-based representations maintain high semantic fidelity while improving engagement, immersion, and perceived humor over caption-style descriptions. These findings highlight a key limitation of current approaches: optimizing for comprehension alone is insufficient for socially and culturally rich media such as memes. Instead, conversational structure and temporal unfolding play a central role in shaping user experience. By moving toward *experiential accessibility*, MemeBuddy points to a promising direction for assistive systems that support both understanding and engagement.

Acknowledgments

We thank anonymous reviewers for their insightful feedback. We also thank the participants and Dr. William Seiple for their support with the user study.

References

Rudaiba Adnin and Maitraye Das. 2024. "i look at it as the king of knowledge": How blind people use and understand generative ai tools. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '24, New York, NY, USA. Association for Computing Machinery.

Shaaron Ainsworth and Nicolas VanLabeke. 2004. Multiple forms of dynamic representation. *Learning and Instruction*, 14(3):241–255. Dynamic Visualisations and Learning.

Christian Bauckhage. 2011. Insights into internet memes. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, pages 42–49.

Jeffrey P. Bigham, Ryan S. Kaminsky, Richard E. Ladner, Oscar M. Danielsson, and Gordon L. Hempton. 2006. Webinsight: making web images accessible. In *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility*, Assets '06, page 181–188, New York, NY, USA. Association for Computing Machinery.

Jan Blommaert and Piia Varis. 2017. Conviviality and collectives on social media: Virality, memes, and new social structures. *Multilingual Margins: A journal of multilingualism from the periphery*, 2(1):31.

James V Bradley. 1958. Complete counterbalancing of immediate sequential effects in a latin square design. *Journal of the American Statistical Association*, 53(282):525–528.

Carl Chen. 2012. The creation and meaning of internet memes in 4chan: Popular internet culture in the age of online digital reproduction. *Habitus*, 3(1):6–19.

Gengyu Chen, Chiqian Xu, Junru Jiang, Yuting Lu, Xiao Wang, Huamin Qu, Shuchang Xu, and Guanhong Liu. 2025. Memespeak: Making image memes blind-accessible through emotional speech augmentation. In *Companion Publication of the 2025 Conference on Computer-Supported Cooperative Work and Social Computing*, CSCW Companion '25, page 440–445, New York, NY, USA. Association for Computing Machinery.

Michelene T. H. Chi, Seokmin Kang, and David L. Yaghmourian. 2017. Why students learn more from dialogue- than monologue-videos: Analyses of peer interactions. *Journal of the Learning Sciences*, 26(1):10–50.

Patrick Davison. 2012. *9. The Language of Internet Memes*, pages 120–134. New York University Press, New York, USA.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54(1):755–810.

Marta Dynel. 2016. "i has seen image macros!" advice animals memes as visual-verbal jokes. *International Journal of Communication*, 10:29–29.

Sarik Ghazarian, Zixi Liu, Tuhin Chakrabarty, Xuezhe Ma, Aram Galstyan, and Nanyun Peng. 2021. DiSCoL: Toward engaging dialogue systems through conversational line guided response generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 26–34, Online. Association for Computational Linguistics.

Cole Gleason, Patrick Carrington, Cameron Cassidy, Meredith Ringel Morris, Kris M. Kitani, and Jeffrey P. Bigham. 2019a. "it's almost like they're trying to hide it": How user-provided image descriptions have failed to make twitter accessible. In *The World Wide Web Conference*, WWW '19, page 549–559, New York, NY, USA. Association for Computing Machinery.

Cole Gleason, Amy Pavel, Xingyu Liu, Patrick Carrington, Lydia B. Chilton, and Jeffrey P. Bigham. 2019b. Making memes accessible. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, page 367–376, New York, NY, USA. Association for Computing Machinery.

Ricardo E Gonzalez Penuela, Jazmin Collins, Cynthia Bennett, and Shiri Azenkot. 2024. Investigating use cases of ai-powered scene description applications for blind and low vision people. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

EunJeong Hwang and Vered Shwartz. 2023. MemeCap: A dataset for captioning and interpreting memes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1445, Singapore. Association for Computational Linguistics.

Imgflip LLC. 2026. Imgflip - create and share awesome images. `https://imgflip.com/`.

Saurav Joshi, Filip Ilievski, and Luca Luceri. 2024. Contextualizing internet memes across social media platforms. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 1831–1840, New York, NY, USA. Association for Computing Machinery.

Kaggle. 2026. 6992 meme images dataset with labels. `https://www.kaggle.com/datasets/hammadjavaid/6992-labeled-meme-images-dataset`.

Bisera Kostadinovska-Stojchevska and Elena Shalevska. 2018. Internet memes and their socio-linguistic features. *European Journal of Literature, Language and Linguistics Studies*, 2(4).

Hae-Na Lee and Vikas Ashok. 2021. Towards enhancing blind users' interaction experience with online videos via motion gestures. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, HT '21, page 231–236, New York, NY, USA. Association for Computing Machinery.

Jaewook Lee, Yi-Hao Peng, Jaylin Herskovitz, and Anhong Guo. 2021. Image explorer: Multi-layered touch exploration to make images accessible. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '21, New York, NY, USA. Association for Computing Machinery.

Ryan Lowe, Iulian Vlad Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. On the evaluation of dialogue systems with next utterance classification. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 264–269, Los Angeles. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting. Association for Computational Linguistics.

Masahiro Mizukami, Koichiro Yoshino, Graham Neubig, David Traum, and Satoshi Nakamura. 2016. Analyzing the effect of entrainment on dialogue acts. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 310–318, Los Angeles. Association for Computational Linguistics.

Maria D Molina. 2020. What makes an internet meme a meme? six essential characteristics. *Handbook of visual communication*, pages 380–394.

Durim Morina and Michael S. Bernstein. 2022. A web-scale analysis of the community origins of image memes. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW1).

Meredith Ringel Morris, Jazette Johnson, Cynthia L. Bennett, and Edward Cutrell. 2018. Rich representations of visual content for screen reader users. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–11, New York, NY, USA. Association for Computing Machinery.

Meredith Ringel Morris, Annuska Zolyomi, Catherine Yao, Sina Bahram, Jeffrey P. Bigham, and Shaun K. Kane. 2016. "with most of it being pictures now, i rarely use it": Understanding twitter's evolving accessibility to blind users. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 5506–5516, New York, NY, USA. Association for Computing Machinery.

Robin L Nabi, Carmen R Stitt, Jeff Halford, and Keli L Finnerty. 2006. Emotional and cognitive predictors of the enjoyment of reality-based and fictional television programming: An elaboration of the uses and gratifications perspective. *Media psychology*, 8(4):421–447.

Heather L. O'Brien, Paul Cairns, and Mark Hall. 2018. A practical approach to measuring user engagement with the refined user engagement scale (ues) and new ues short form. *International Journal of Human-Computer Studies*, 112:28–39.

Ratnabali Pal, Samarjit Kar, Dilip K. Prasad, and Arif Ahmed Sekh. 2026. Visionaidqa: Advancing visual question answering for the visually impaired. *ACM Comput. Surv.*, 58(9).

K R Prajwal, C V Jawahar, and Ponnurangam Kumaraguru. 2019. Towards increased accessibility of meme images with the help of rich face emotion captions. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 202–210, New York, NY, USA. Association for Computing Machinery.

Yash Prakash, Akshay Kolgar Nayak, Shoaib Mohammed Alyaan, Pathan Aseef Khan, Hae-Na Lee, and Vikas Ashok. 2024. Improving usability of data charts in multimodal documents for low vision users. In *Proceedings of the 26th International Conference on Multimodal Interaction*, ICMI '24, page 498–507, New York, NY, USA. Association for Computing Machinery.

Daniel C. Richardson, Nicole K. Griffin, Lara Zaki, Auburn Stephenson, Jiachen Yan, Thomas Curry, Richard Noble, John Hogan, Jeremy I. Skipper, and Joseph T. Devlin. 2018. Measuring narrative engagement: The heart tells the story. *bioRxiv*.

Richard M Ryan, Valerie Mims, and Richard Koestner. 1983. Relation of reward contingency and interpersonal context to intrinsic motivation: A review and test using cognitive evaluation theory. *Journal of personality and Social Psychology*, 45(4):736.

Johnny Saldana. 2011. *Fundamentals of qualitative research*. Oxford university press.

Shivam Sharma, Siddhant Agarwal, Tharun Suresh, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. 2023. What do you meme? generating explanations for visual semantic role labelling in memes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8):9763–9771.

Shivam Sharma, Ramaneswaran S, Md. Shad Akhtar, and Tanmoy Chakraborty. 2024. Emotion-aware multimodal fusion for meme emotion detection. *IEEE Transactions on Affective Computing*, 15(3):1800–1811.

Tanusree Sharma, Yu-Yun Tseng, Lotus Zhang, Ayae Ide, Kelly Avery Mack, Leah Findlater, Danna Gurari, and Yang Wang. 2025. "before, i asked my mom, now i ask chatgpt": Visual privacy management with generative ai for blind and low-vision people. In

Proceedings of the 27th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS '25, New York, NY, USA. Association for Computing Machinery.

Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2020. Image-chat: Engaging grounded conversations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2414–2429, Online. Association for Computational Linguistics.

B Sindhu, B Bhagya Preethi, S Leela Venkata Lakshmi, B Praveen Reddy, and S Srinivas Kiran. 2025. Voice-assisted artificial intelligence based question answering system for the visually impaired. In *Algorithms in Advanced Artificial Intelligence*, pages 135–140. CRC Press.

Abigale Stangl, Nitin Verma, Kenneth R. Fleischmann, Meredith Ringel Morris, and Danna Gurari. 2021. Going beyond one-size-fits-all image descriptions to satisfy the information wants of people who are blind or have low vision. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '21, New York, NY, USA. Association for Computing Machinery.

Tess Van Daele, Akhil Iyer, Yuning Zhang, Jalyn C Derry, Mina Huh, and Amy Pavel. 2024. Making short-form videos accessible with hierarchical video summaries. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Violeta Voykinska, Shiri Azenkot, Shaomei Wu, and Gilly Leshed. 2016. How blind people interact with visual content on social networking services. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, page 1584–1595, New York, NY, USA. Association for Computing Machinery.

Camilla Vásquez and Erhan Aslan. 2021. "cats be outside, how about meow": Multimodal humor and creativity in an internet meme. *Journal of Pragmatics*, 171:101–117.

WebAIM. 2026. The webaim million - the 2025 report on the accessibility of the top 1,000,000 home pages. `https://webaim.org/projects/million/`.

Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. 2017. Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, page 1180–1192, New York, NY, USA. Association for Computing Machinery.

Bo Xu, Tingting Li, Junzhe Zheng, Mehdi Naseriparsa, Zhehuan Zhao, Hongfei Lin, and Feng Xia. 2022. Met-meme: A multimodal meme dataset rich in metaphors. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2887–2899, New York, NY, USA. Association for Computing Machinery.

Shuchang Xu, Chang Chen, Zichen Liu, Xiaofu Jin, Lin-Ping Yuan, Yukang Yan, and Huamin Qu. 2024. Memory reviver: Supporting photo-collection reminiscence for people with visual impairment via a proactive chatbot. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, UIST '24, New York, NY, USA. Association for Computing Machinery.

Mingrui Ray Zhang, Mingyuan Zhong, and Jacob O. Wobbrock. 2022. Ga11y: An automated gif annotation system for visually impaired users. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Yang Zhong and Bhiman Kumar Baghel. 2024. Multimodal understanding of memes with fair explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2007–2017.

A Meme Representation Generation Prompts

Caption-style Description (Baseline) Generation Prompt

```
You are tasked with explaining memes to a blind person in an engaging monologue as a single
narrator, and also generating TTS voice style instructions.

## Rules for Output
- Output strictly in JSON with two top-level keys:
  - "dialogues" → array of objects with "speaker": "Narrator" and "text".
  - "tts_instructions" → array with one object describing how the narrator's voice should
  sound.
- No extra narration, no text outside the JSON.

## Rules for Dialogue
- Always use "speaker": "Narrator".
- Start by recognizing or naming the meme template (if known).
- Describe what's happening visually so a blind listener can clearly imagine it (panels,
labels, expressions, text placement).
- Keep the tone casual, cohesive, and natural — like a friend explaining something funny. Add
light filler words or small reactions so it doesn't sound robotic.
- End with a short summary punchline that explains the joke in plain language.
- Keep total length around 25-75 seconds depending on meme complexity. Don't pad unnecessarily.

## Rules for TTS Instructions
- Provide "tts_instructions" with one object:
  - "speaker": "Narrator"
  - "style": Warm, conversational, descriptive voice. Clear pacing, friendly tone, with a hint
  of humor.

---

# Example Input → Output

Input: Meme: Distracted Boyfriend with labels "me," "responsibilities," and "new hobby."

Output:
{
  "dialogues": [
    {"speaker": "Narrator", "text": "Oh hey, this one's the distracted boyfriend meme."},
    {"speaker": "Narrator", "text": "You've got the guy walking with his girlfriend, but his
    head's turned to check out another woman passing by."},
    {"speaker": "Narrator", "text": "In this version, the girlfriend is labeled
    'responsibilities,' the guy is 'me,' and the other woman is 'new hobby.'"},
    {"speaker": "Narrator", "text": "So yeah, the joke is basically about ignoring your duties
    just because something fun comes along."}
  ],
  "tts_instructions": [
    {"speaker": "Narrator", "style": "Warm, conversational, descriptive voice. Clear pacing,
    friendly tone, with a hint of humor."}
  ]
}
```

Table 4: A prompt for generating caption-style description.

Commentator-Commentator (CO) Dialog Generation Prompt

```
Your task is to describe a meme to a blind person in an engaging dialog format between two
commentators, and also generate text-to-speech (TTS) voice style instructions. While creating
the dialog and TTS instructions, you should strictly follow these rules:

Rules for constructing the dialog:

- The two commentators must be labeled "Host1" and "Host2".
- Host1 should start the conversation.
- Host1's utterances should be more explanatory and Host2's utterances should be more
casual/reactive.
- The conversation should not be overly detailed. There should be room for self-interpretation
for the listener.
- The conversation should be natural and concise: 25-75 seconds if narrated with a TTS. Don't
pad content unnecessarily.
- The conversation should end with a plain punchline-style summary from Host2 to convey the
meme intent such as humor or sarcasm.

Rules for generating TTS instructions:

- Host1's voice should have a clear, explanatory style.
- Host2's voice should be more casual and reactive, complementing Host1's voice.
- Host 1 and Host 2 should not have the same gender.
- The voices of Host1 and Host2 should be distinct so the listener can easily tell them apart.
- Gender assignment should be either "male" or "female".

Rules for Output format:

- Output must be strictly a JSON object with two keys: "dialogs" and "tts_instructions".
- "dialogs" must be an array of objects, with each object having "speaker" and "text" fields
along with their values.
- "tts_instructions" must be an array of objects, with each object having "speaker", "style",
and "gender" fields along with their values.
- An illustrative example showing an input-output pair is given next.

Example Input-Output pair:

Input: [Drake "Yes/No" meme image with labels "Working overtime" and "Taking a nap."]

Output:
{
  "dialogs": [
    {"speaker": "Host1", "text": "This is the classic Drake meme—two stacked panels showing
    preference."},
    {"speaker": "Host2", "text": "Oh yeah, the top panel has Drake waving his hand dismissively
    at 'working overtime.'"},
    {"speaker": "Host1", "text": "And the bottom shows Drake pointing approvingly at 'taking a
    nap.'"},
    {"speaker": "Host2", "text": "So basically: work bad, sleep good. The eternal struggle!"}
  ],
  "tts_instructions": [
    {"speaker": "Host1", "style": "Clear, explanatory tone. Professional but friendly, like a
    radio host.", "gender": "female"},
    {"speaker": "Host2", "style": "Casual, laid-back voice with good energy. More
    conversational and reactive.", "gender": "male"}
  ]
}
```

Table 5: A prompt for generating Commentator-Commentator (CO) dialog.

Commentator-Character (CH) Dialog Generation Prompt

```
Your task is to describe a meme to a blind person in an engaging dialog format between a commentator
(i.e., Host) and a character (i.e., Guest) from the meme, and also generate text-to-speech (TTS)
voice style instructions. While creating the dialog and TTS instructions, you should strictly follow
these rules:

Rules for Dialog:

- The speakers must be labeled "Host" and "Guest".
- The Host should start the conversation.
- For Guest, if there are more than one characters in the meme, select the character that is
associated with the punch-line that conveys intent like humor or sarcasm.
- The Guest must introduce themselves in their first utterance in the dialog. If the selected
character/person from the meme has a known name, use it (e.g., "Hi, I'm Drake"). If the name is
unknown, have the Guest describe themselves by relating to how the Host refers to them (e.g., if
Host says "the person in the meme," Guest says "Yeah, I'm the person in the meme"). This helps blind
listeners distinguish between speakers.
- If the meme suits first-person narration (e.g., character/person style meme), then the Guest
should speak as if they are the character. If not, the guest should speak like a commentator.
- The conversation should not be overly detailed. There should be room for self-interpretation for
the listener.
- The conversation should be natural and concise: 25-75 seconds if narrated with a TTS. Don't pad
content unnecessarily.
- The conversation should end with a plain punchline-style summary conveying the meme intent such as
humor or sarcasm.

Rules for TTS Instructions:

- First decide the Guest's gender based on the meme character or context. If the gender of the
character is not clear (e.g., a bird), choose the gender that creates better contrast between the
two speakers.
- The gender of the Host should be the opposite of Guest to maximize voice distinction.
- The gender should be either “male” or “female”.
- The voice style for the Host should be warm and clear with an explanatory tone. The pacing should
be moderate like a professional podcast narrator.
- The voice style for the Guest will depend on the meme character. If the selected meme character
for the Guest is well known (e.g., a celebrity), then configure the style to match their
publicly-known speech attributes (tone, personality, pacing). If the selected character is not a
public figure, then use a friendly and conversational style that best complements the Host’s style.
- Try your best to make the voice profiles of Host and Guest distinct so the listener can easily
tell them apart.

Rules for Output format:

- Output must be strictly a JSON object with two keys: "dialogs" and "tts_instructions".
- "dialogs" must be an array of objects, with each object having “speaker” and “text” fields along
with their values.
- "tts_instructions" must be an array of objects, with each object having “speaker”, “style”, and
“gender” fields along with their values.
- An illustrative example showing an input-output pair is given next.

Example Input-Output pair:

Input: [Meme: Drake “Yes/No” with labels “Working overtime” and “Taking a nap.”]

Output: {
  "dialogs": [
    {"speaker": "Host", "text": "This is the Drake meme—the two stacked panels."},
    {"speaker": "Guest", "text": "Yeah, in the top I’m waving my hand like, ‘nope,’ at the words
    ‘working overtime.’"},
    {"speaker": "Host", "text": "And then in the bottom, you’re pointing happily at ‘taking a
    nap.’"},
    {"speaker": "Guest", "text": "Exactly—work is out, sleep is in."}
  ],
  "tts_instructions": [
    {"speaker": "Host", "style": "Warm, clear, explanatory tone. Moderate pace, like a friendly
    podcast narrator.", "gender": "female"},
    {"speaker": "Guest", "style": "Casual, smug, playful voice—try to sound like Drake, with a
    relaxed pace and confident energy.", "gender": "male"}
  ]
}
```


Table 6: A prompt for generating Commentator-Character (CH) dialog.

B Meme Annotation Schema

Field	Description
`template name`	Meme template identity ("unknown" if a template is unrecognizable)
`panels`	Panel count and layout order
`key visuals`	Per-panel visual elements (e.g., characters, poses, expressions, objects)
`labels`	Exacted on-image text (case-sensitive, verbatim)
`intent summary`	Explanation of humor mechanism in a meme in 1–2 sentences

Table 7: Meme image annotation schema.

C Additional Offline Evaluation Details

C.1 Variance Statistics

Table 8 reports the standard deviations (SD) for *Intent Clarity* and binomial standard errors (SE) for the binary metrics, computed across 100 memes and two annotators. Variance is moderate overall, with slightly higher variability in CH across models, indicating less consistent performance for character-based dialog generation.

Model	Intent Clarity (SD)			Representation Fidelity (SE)			Text Accuracy (SE)			Speculation (SE)		
	BL	CO	CH	BL	CO	CH	BL	CO	CH	BL	CO	CH
Gemini 3 Flash	0.61	0.66	0.72	0.029	0.031	0.032	0.017	0.020	0.024	0.026	0.028	0.029
Gemini 3 Pro	0.52	0.58	0.64	0.026	0.026	0.029	0.014	0.017	0.022	0.022	0.026	0.026
GPT-5.1	0.54	0.60	0.68	0.022	0.024	0.029	0.010	0.020	0.025	0.020	0.028	0.030
Gemma 4	0.78	0.82	0.89	0.038	0.042	0.040	0.024	0.029	0.033	0.032	0.035	0.039

Table 8: Variance statistics for offline evaluation metrics. Standard deviations (SD) are reported for *Intent Clarity*; binomial standard errors (SE) are reported for binary metrics (*Representation Fidelity*, *Text Accuracy*, *Speculation*).

C.2 Effect Sizes

Table 9 shows Kendall's W for Friedman tests on *Intent Clarity*.

Model	Kendall's W
Gemini 3 Flash	0.142
Gemini 3 Pro	0.048
GPT-5.1	0.096
Gemma 4	0.147

Table 9: Effect sizes for condition differences in *Intent Clarity*.

All effect sizes are **small** ($W \leq 0.15$), indicating that although condition differences are statistically significant, their practical impact on comprehensibility is limited.

C.3 Pairwise Effect Sizes

Table 10 reports representative Wilcoxon signed-rank effect sizes (r).

Comparison	Effect Size (r)
BL vs. CO	0.10–0.18 (small)
BL vs. CH	0.25–0.38 (small–moderate)
CO vs. CH	0.20–0.32 (small–moderate)

Table 10: Representative pairwise effect sizes for *Intent Clarity*.

Pairwise comparisons show larger (but still modest) effects involving CH, consistent with reduced clarity in character-driven dialog.

D User Study Participant Demographics

ID	Age / Gender	Visual Condition	Web Browsing Experience	Preferred Screen Reader	Social Forums Experience	Meme Exposure	AI Assistive Tools
P1	44/M	Blind	Daily	JAWS, NVDA	2-3 times/week	X, Reddit	Be My AI, ChatGPT
P2	32/M	Blind	Daily	JAWS	Daily	Yahoo, Reddit	ChatGPT
P3	25/F	Blind	Daily	JAWS, NVDA	4-5 times/week	WhatsApp	ChatGPT
P4	56/F	Low Vision	2-3 times/week	JAWS	0-1 time/week	Facebook	JAWS Picture Smart AI
P5	66/M	Blind	1-2 times/week	JAWS	1 time/week	WhatsApp	Be My AI
P6	52/F	Low Vision	Daily	JAWS	2-4 times/week	Messages, Reddit	ChatGPT
P7	41/F	Blind	Daily	JAWS	Daily	Facebook	Be My AI
P8	38/F	Blind	Daily	JAWS, VoiceOver	4 times/week	X, Bluesky	Be My AI
P9	29/M	Light Perception	Daily	JAWS, NVDA	Daily	Messages, WhatsApp	Be My AI, JAWS Picture Smart AI
P10	22/F	Blind	Daily	NVDA	Daily	Facebook, Yahoo	Be My AI, ChatGPT
P11	46/M	Blind	4-5 times/week	JAWS	2-3 times/week	Facebook	Be My AI, JAWS Picture Smart AI
P12	30/M	Light Perception	Daily	JAWS	Daily	Facebook, Reddit	ChatGPT
P13	23/F	Blind	Daily	NVDA	Daily	Messages, Facebook	Be My AI, ChatGPT
P14	27/M	Blind	Daily	JAWS	Daily	X, Yahoo	Be My AI

Table 11: Demographics of blind participants in the user study. All information was self-reported by the participants.

E Memes in User Study Evaluation

The six memes shown in Figure 2 were used in the user study.

Figure 2: Image memes used in the user study.

F Full Inferential Statistical Results (Q2–Q6)

We analyzed Likert-scale responses (Q2–Q6) using Friedman tests followed by pairwise Wilcoxon signed-rank tests with Holm correction for multiple comparisons. Effect sizes are reported as Kendall's W for omnibus tests and r for pairwise comparisons. The inferential statistical results for Q2–Q6 are shown in Table 12.

Q2: Emotional fluctuation			
Friedman test			
$\chi^2(2) = 21.23, p < .001, W = 0.38$			
Post-hoc (Wilcoxon signed-rank test)			
Comparison	W	p**-value**	**Effect Size** (r)
BL vs. CO	0.0	.0012	.89
BL vs. CH	13.5	.0030	.75
CO vs. CH	3.5	.0578	.72
Q3: Immersion			
Friedman test			
$\chi^2(2) = 16.24, p < .001, W = 0.29$			
Post-hoc (Wilcoxon signed-rank test)			
Comparison	W	p**-value**	**Effect Size** (r)
BL vs. CO	35.0	.0027	.68
BL vs. CH	90.0	.0414	.43
CO vs. CH	60.0	.0414	.49
Q4: Entertainment			
Friedman test			
$\chi^2(2) = 16.48, p < .001, W = 0.29$			
Post-hoc (Wilcoxon signed-rank test)			
Comparison	W	p**-value**	**Effect Size** (r)
BL vs. CO	70.0	.1268	.40
BL vs. CH	74.0	.2311	.27
CO vs. CH	14.0	.0140	.73
Q5: Funniness			
Friedman test			
$\chi^2(2) = 18.00, p < .001, W = 0.32$			
Post-hoc (Wilcoxon signed-rank test)			
Comparison	W	p**-value**	**Effect Size** (r)
BL vs. CO	64.0	.0391	.51
BL vs. CH	61.0	.0623	.46
CO vs. CH	9.0	.0833	.58
Q6: Willingness to consume more			
Friedman test			
$\chi^2(2) = 29.56, p < .001, W = 0.53$			
Post-hoc (Wilcoxon signed-rank test)			
Comparison	W	p**-value**	**Effect Size** (r)
BL vs. CO	0.0	$< .001$	.90
BL vs. CH	61.0	.1578	.32
CO vs. CH	0.0	$< .001$	.90

Table 12: Inferential statistical results for Q2–Q6.