
MemeSonic: Affective Meme Audio Generation

Yiqiao Huang
Harvard University
Cambridge, MA
yhuang1@fas.harvard.edu

Hongbee Park
Harvard University
Cambridge, MA
hongbee_park@mde.harvard.edu

Ruyi Yang
Harvard University
Cambridge, MA
ruyi_yang@gsd.harvard.edu

Abstract

Conventional multimodal affect understanding assumes that image and text provide complementary signals. Memes fundamentally violate this assumption: their meaning often emerges from cross-modal incongruity, such as sarcastic text paired with a cheerful image. We present **MemeSonic**, an incongruity-aware framework for meme understanding and expressive speech generation. We first model image-text incongruity as a continuous signal and introduce **FusionMoE**, a mixture-of-experts architecture that adaptively selects fusion strategies for redundancy, complementarity, and irony-dominant cases. Across ten tasks on three meme benchmarks, FusionMoE outperforms conventional fusion baselines by up to **+14.5%** Macro-F1, with the largest gains on sarcasm and metaphor understanding. We further investigate whether expressive speech can serve as an auxiliary affective modality by generating voiceovers conditioned on multimodal affective cues and using them as pseudo supervision for tri-modal learning. A controlled alignment study shows that naive tri-modal fusion provides no benefit, while a learned modality projector improves sentiment accuracy to **82%**. These results suggest that expressive speech captures affective information beyond image and text alone, highlighting the value of modeling multimodal incongruity as an explicit auditory signal. Github Page: <https://github.com/emmaruyiyang/MMAI-MemeSonic>.

1 Introduction

Traditional multimodal sentiment understanding assumes that image and text are complementary—that both modalities reinforce the same semantic or emotional judgment. In memes, however, the image frequently conveys positive affect while the accompanying text carries ironic, negative, or teasing undertones; it is precisely this cross-modal conflict, contrast, and pragmatic shift that serves as the primary vehicle for meme affect. Image-text inconsistency in memes is therefore not noise but a decisive clue for deciphering underlying sentiment and intent. Building on this observation, we explicitly model image-text *incongruity* as a central variable in meme affective understanding. Concretely, we quantify the distributional divergence between image and text within an affective space to derive a continuous incongruity score that characterizes whether—and to what degree—the emotional content of the two modalities conflicts. This score enables the model to distinguish standard complementary samples from memes that rely on contrast or irony, and allows us to reframe the task itself as *incongruity-aware affective understanding*: the model must leverage not only the content of image and text but also the inconsistency between them to jointly predict sentiment and intent. Different decision modes—image-dominant, text-dominant, consistency-fusion, and incongruity-dominant—are naturally accommodated within a Mixture-of-Experts architecture [Yu et al., 2024].

Taking one step further, if image-text incongruity constitutes a form of *implicit pragmatic affect*, it should be externalizable through vocal prosody. We therefore generate expressive voiceovers tailored to each meme’s sentiment and intent, converting the affective nuance latent in the image-text contrast into audible prosodic signals. Finally, because real-world memes rarely come with human-annotated affective audio, we reintroduce these generated voiceovers into the learning process as a *privileged affective modality*—not merely an additional input channel, but a pragmatic supervisory signal encoding *how this meme ought to be spoken*. By training on this pseudo-audio alongside image, text, and the incongruity score, the model better captures the sentiments and intents that arise from image-text conflict, ultimately improving affective understanding. We present **MemeSonic**, a unified framework that integrates this incongruity-centered task formulation, expressive speech synthesis, and pseudo-GT audio supervision into a single system.

2 Related Work

2.1 Meme Understanding and Multimodal Sentiment

Meme understanding has been studied primarily as a multimodal classification problem. The Hateful Memes Challenge [Kiela et al., 2020] deliberately constructed a dataset in which neither modality alone is sufficient for classification, establishing that meme meaning requires genuine cross-modal reasoning. Cai et al. [2019] proposed a hierarchical fusion model for multimodal sarcasm detection on Twitter, and MemeCLIP [Shi et al., 2024] leveraged CLIP-based representations to analyze meme sentiment and humor. Memotion [Sharma et al., 2020, Caton et al., 2023] and MMSD 2.0 [Qin et al., 2023] provide benchmarks that broaden the task space to sentiment, sarcasm, humor, and intent. On the theoretical side, the incongruity-resolution account of humor [Suls, 1972] holds that humor arises from a mismatch between expectation and outcome, and Yang et al. [2015] and Hessel et al. [2023] have explored computational approaches to modeling this mechanism. However, these works remain confined to classification or explanation; none externalize non-literal meme affect as speech or reintroduce generated audio as a supervisory signal for understanding. Our work extends this line: Phase 1 externalizes pragmatic affect as expressive voiceover, and Phase 3 tests whether that voiceover carries affective signal beyond image and text.

2.2 Multimodal Fusion and Interaction Decomposition

Standard multimodal fusion strategies, including concatenation, late fusion, and multimodal transformers [Tsai et al., 2019], combine modality embeddings without explicit awareness of *how* modalities interact. Contrastive vision-language models [Radford et al., 2021, Li et al., 2022, Jia et al., 2021] align image and text in a shared space, providing a strong foundation for zero-shot recognition and retrieval, but this alignment relies on the assumption that paired data are semantically consistent, which breaks down in conflict-dominant domains such as memes. ImageBind [Girdhar et al., 2023] extends shared-space alignment to six modalities including audio, yet because it is trained primarily on environmental sounds, it achieves effectively zero performance on TTS speech alignment in our probing ($A \leftrightarrow T$ Top-1 ≈ 0.00). More recently, Yu et al. [2024] proposed a Multimodal Mixture-of-Experts (MMoE) that decomposes cross-modal interactions into redundancy, uniqueness, and synergy, with separate experts handling each type, yielding gains on sarcasm and humor benchmarks where synergy is theoretically expected to dominate. We adopt the R/U/S decomposition as the skeleton of our Phase 2 fusion, but instantiate the synergy expert through an explicit incongruity score derived from the divergence between unimodal sentiment distributions, providing interpretable, affect-grounded synergy tailored to the conflict-dominant nature of memes.

2.3 Controllable and Affective Speech Synthesis

Text-to-speech research has progressed from neutral voicing toward expressive, controllable synthesis. Early end-to-end systems such as Tacotron [Wang et al., 2017] and FastSpeech 2 [Ren et al., 2021] established foundations for natural speech generation, and StyleTTS 2 [Li et al., 2023] introduced diffusion-based style modeling that enables fine-grained control over prosody and speaking style. Emotion-transfer methods [Lei et al., 2022] condition synthesis on affect labels for cross-lingual settings, and emotion2vec [Ma et al., 2023] provides a self-supervised representation specifically for speech emotion. A central challenge when integrating such audio representations with vision-language systems is the modality gap [Liang et al., 2022]: embeddings from different encoders inhabit

incompatible regions of their respective spaces, rendering naive fusion ineffective. Commercial systems (e.g., ElevenLabs) expose voice catalogs and emotion-tag prompting, while open-source systems (e.g., ChatTTS) support discrete prosodic control tokens, but in all cases the conditioning signal is generic: a style prompt or reference speaker rather than the interpreted affect of a specific multimodal input. To our knowledge, MemeSonic is the first system to condition expressive synthesis directly on meme-level affective interpretation, including the incongruity score and image-text context (Phase 1), and to address the modality gap through a learned linear projector that enables generated speech to serve as a supervisory signal for affective understanding (Phase 3).

3 Problem Statement

Given a meme dataset $\mathcal{D} = \{(v_i, t_i, y_i^{(s)}, y_i^{(i)})\}_{i=1}^N$, our goal is to predict each meme’s affective sentiment $y^{(s)}$ and pragmatic intention $y^{(i)}$ from its image v and text t . Unlike standard multimodal affective understanding, meme affect often emerges from image-text conflict rather than cross-modal agreement. Let $\delta(v, t)$ denote image-text incongruity. We learn $F_\theta(v, t, \delta(v, t)) \rightarrow (y^{(s)}, y^{(i)})$ by minimizing the joint sentiment-intention loss:

$$\theta^* = \arg \min_{\theta} \mathbb{E}_{(v, t, y^{(s)}, y^{(i)}) \sim \mathcal{D}} [\mathcal{L}_s + \mathcal{L}_i]. \quad (1)$$

While image-text incongruity is crucial for meme understanding, the affective meaning induced by such incongruity is often implicit. We further externalize such implicit affect through expressive voiceover generation. Given a script $s(t)$ and affective condition c , we synthesize pseudo speech $\tilde{a} = \Psi(s(t), c)$ that conveys the meme’s pragmatic affect through prosody.

Since human-annotated meme audio is rarely available, we use $\tilde{a}$ as a pseudo affective modality and learn $G_\omega(v, t, \tilde{a}, \delta(v, t)) \rightarrow (y^{(s)}, y^{(i)})$, with the objective of improving affective prediction beyond image, text, and incongruity alone:

$$\text{Acc}(G_\omega(v, t, \tilde{a}, \delta(v, t))) > \text{Acc}(F_\theta(v, t, \delta(v, t))). \quad (2)$$

This work therefore studies how image-text incongruity can be used to understand meme affect, and how generated expressive speech can externalize such implicit pragmatic affect to further improve meme understanding.

4 Proposed Method

MemeSonic consists of three phases (Figure 1): (1) Meme-to-Speech Generation, which externalizes meme affect as expressive speech; (2) Incongruity-Aware Fusion, which models image-text conflict for affect understanding; and (3) Tri-Modal Affective Fusion, which studies whether generated speech provides non-redundant affective signal.

4.1 Phase 1: Meme-to-Speech Generation

Script Generation and Prosody Prompt Adapter Given a meme image v and text t , we first generate a voiceover script s from the image-text context. Rather than literally describing the meme, the script verbalizes its pragmatic intent, including sarcasm, irony, humor, and affective reversal.

Discrete affect labels are too coarse to capture meme-specific delivery styles. We therefore train a Qwen3-4B-based emotion-to-prosody adapter. Given the affect label $\ell = (y^{(s)}, y^{(i)})$ and image-text context $r(v, t)$, the adapter generates a prosody prompt $\pi = A_\eta(\ell, r(v, t))$, where π controls tone, rhythm, emphasis, intensity, and speaking style.

Conditioned Speech Generation We synthesize meme voiceover audio by conditioning the TTS model on the script, prosody prompt, and affect label: $a^* = \Psi(s, \pi, \ell)$. This conditioning adjusts both voice timbre and intonation style according to the meme’s visual-textual semantics and pragmatic affect. The resulting speech externalizes implicit meme emotion as an explicit prosodic signal, and serves as pseudo ground-truth audio for downstream understanding.

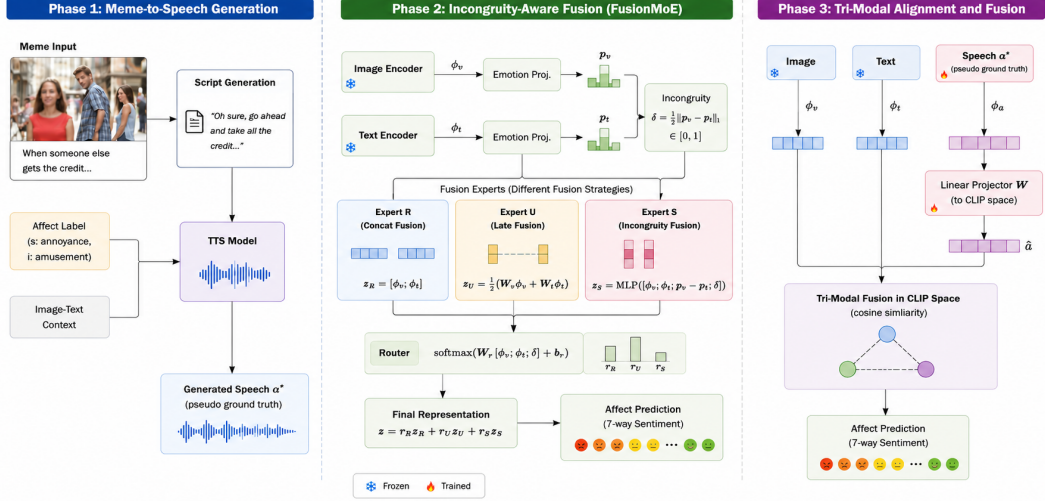


Figure 1: Overview of MemeSonic. **Phase 1** generates meme voiceovers using GPT-4.1-mini with Qwen3-based prosody prompting and incongruity-aware control. **Phase 2** computes the incongruity score δ and routes inputs through FusionMoE. **Phase 3** aligns speech embeddings into a shared CLIP/ImageBind space for multimodal classification.

4.2 Phase 2: Incongruity-Aware Multimodal Fusion

Explicit Affect Decomposition We encode the meme image v and text t using frozen CLIP encoders ϕ_v and ϕ_t . Each embedding is projected onto C emotion prototypes $\{\mathbf{e}_c\}_{c=1}^C$, producing emotion distributions $p_v, p_t \in \Delta^C$:

$$p_c^{(\cdot)} = \frac{\exp(\phi_{(\cdot)}(\cdot)^\top \mathbf{e}_c / \tau)}{\sum_{j=1}^C \exp(\phi_{(\cdot)}(\cdot)^\top \mathbf{e}_j / \tau)} \quad (3)$$

We define image-text incongruity as:

$$\delta(v, t) = \frac{1}{2} \|p_v - p_t\|_1 \in [0, 1] \quad (4)$$

where larger δ indicates stronger affective conflict.

FusionMoE Different affective tasks favor different fusion assumptions. Literal sentiment often benefits from direct semantic aggregation, while sarcasm and humor require explicit modeling of image-text conflict. We therefore model meme understanding as adaptive fusion over complementary fusion strategies.

Note. Our MoE differs from LLM-style MoE: experts correspond to explicit fusion strategies rather than learned parameter specializations. The expert set is task-dependent and can be adapted to other multimodal reasoning tasks.

Expert R (Concat Fusion). This expert performs standard representation-level fusion through direct feature concatenation:

$$z_R = [\phi_v(v); \phi_t(t)] \quad (5)$$

It captures aligned semantic signals and serves as a strong baseline for literal or weakly incongruent memes.

Expert U (Late Fusion). This expert performs decision-level fusion by averaging unimodal predictions:

$$z_U = \frac{1}{2} (W_v \phi_v(v) + W_t \phi_t(t)) \quad (6)$$

It is more robust when one modality dominates or when image-text interaction is weak.

Expert S (Incongruity Fusion). This expert explicitly models cross-modal conflict using emotion disagreement and incongruity features:

$$z_S = \text{MLP}([\phi_v(v); \phi_t(t); p_v - p_t; \delta]) \quad (7)$$

It is designed for sarcasm, irony, and other non-literal memes where affect emerges from image-text tension.

A router predicts soft expert weights:

$$\mathbf{r} = \text{softmax}(W_r[\phi_v(v); \phi_t(t); \delta] + b_r) \quad (8)$$

The final representation is computed as:

$$z = r_R z_R + r_U z_U + r_S z_S \quad (9)$$

4.3 Phase 3: Tri-Modal Affective Fusion

We next incorporate generated speech as a third affective modality for meme understanding. Naive tri-modal fusion fails because speech and vision-language embeddings lie in incompatible representation spaces. In particular, ImageBind targets general audio semantics rather than expressive speech.

To bridge this gap, we train a linear projector $W \in \mathbb{R}^{D_a \times D_e}$ that maps Emotion2Vec speech embeddings into the shared CLIP/ImageBind semantic space:

$$\hat{a} = \text{normalize}(W\phi_a(a^*) + b) \quad (10)$$

The projector is supervised using MSE loss against the sentiment prototype of the sample:

$$\mathcal{L}_{\text{proj}} = \frac{1}{N} \sum_{i=1}^N \|\hat{a}_i - \text{normalize}(\mathbf{e}_{y_i^{(s)}})\|_2^2 \quad (11)$$

After projection, image, text, and speech embeddings are fused in the shared semantic space for cosine-similarity classification. This experiment isolates modality mismatch from signal quality, showing that generated speech becomes a useful affective modality once properly aligned.

5 Experimental Methodology

5.1 Datasets

We evaluate on three benchmarks spanning ten meme understanding tasks (Table 10).

MET-Meme [Xu et al., 2022] covers intention, sentiment, hate, and metaphor classification. **Memotion 7k** [Sharma et al., 2020] covers humour, sarcasm, offensiveness, motivation, and sentiment tasks. **MMSD 2.0** [Qin et al., 2023] focuses on binary sarcasm detection.

We use the official splits for MET-Meme and MMSD 2.0, and a random 80/10/10 split for Memotion 7k (seed 42). All datasets provide meme images paired with overlaid text as input.

5.2 Baselines

Phase 1 — Meme-to-Speech Generation. We compare against: (1) **Vanilla TTS**, using default TTS generation without affect conditioning; (2) **Label-Guided TTS**, conditioned only on discrete affect labels; and (3) **MemeSonic**, which uses the proposed emotion-to-prosody adapter conditioned on affect labels and image-text context.

Phase 2 — Multimodal Fusion. We compare with standard multimodal baselines including Unimodal-Image, Unimodal-Text, Concat Fusion, Late Fusion, and MULT. We additionally evaluate multimodal LLMs including GPT-4o, Gemini-2.5-Flash, o4-mini (CoT), and Qwen3-VL-30B (Thinking). Our methods include Incongruity Fusion and FusionMoE.

Phase 3 — Tri-Modal Fusion. To evaluate whether generated speech provides non-redundant affective information, we synthesize multiple speech variants for each meme using different prosody settings. Speech embeddings are projected into a shared multimodal semantic space using a learned linear projector, enabling joint image-text-speech fusion for affect classification.

5.3 Evaluation Protocol

Audio Generation. We evaluate generated speech using both automatic and human evaluation. Automatic evaluation uses Emotion2Vec cosine similarity against pseudo ground-truth speech. Human evaluation uses Mean Opinion Score (MOS) on naturalness (NMOS), expressiveness (EMOS), and pragmatic appropriateness (AMOS). We additionally conduct pairwise A/B preference tests against baseline systems.

Affective Understanding. For meme understanding tasks, we report Accuracy and Macro-F1. The tri-modal study additionally evaluates cosine-similarity classification after speech alignment.

5.4 Implementation Details

We use frozen vision-language and speech encoders for multimodal representation learning. FusionMoE is optimized with AdamW and early stopping based on validation Macro-F1. Additional implementation details are provided in Appendix E.

6 Results and Discussion

6.1 Phase 1: Expressive Meme Voiceover

Key findings.

Table 1: Subjective evaluation of meme voiceover generation. MemeSonic improves expressiveness and affective appropriateness while maintaining comparable naturalness.

Method	NMOS $\uparrow$	EMOS $\uparrow$	AMOS $\uparrow$
Vanilla TTS	3.92 ± 0.16	2.35 ± 0.31	2.42 ± 0.38
Label-Guided TTS	3.85 ± 0.18	3.12 ± 0.28	3.05 ± 0.35
MemeSonic	3.88 ± 0.15	3.86 ± 0.24	3.95 ± 0.22

Comparison	Prefer MemeSonic (%)	Prefer Baseline (%)	No Pref. (%)
MemeSonic vs. Vanilla TTS	84.7%	9.3%	6.0%
MemeSonic vs. Label-Guided	65.3%	27.3%	7.3%

Expressive voiceover generation. MemeSonic substantially improves expressive and pragmatic quality while preserving comparable naturalness. The gain over Label-Guided TTS suggests that coarse affect labels are insufficient for meme delivery, which often depends on subtle prosodic cues such as irony, exaggeration, and affective reversal. Incorporating image-text context through the prosody adapter produces more meme-aligned voiceover behavior.

6.2 Phase 2: Incongruity-Aware Meme Understanding

Table 2 reports representative results on literal and non-literal meme understanding tasks. Full benchmark results are provided in Appendix B.

Incongruity-aware fusion. Incongruity-aware methods perform strongly on sarcasm and metaphor tasks, where meme meaning depends on image-text contradiction rather than literal semantics.

FusionMoE provides more consistent performance across heterogeneous tasks by adaptively combining complementary fusion assumptions rather than relying on a single fusion strategy.

LLMs perform strongly on commonsense-heavy tasks, while trained multimodal models remain more reliable on structured affect classification.

Table 2: Meme understanding results (Macro-F1). Full results in Appendix B. **Bold** = best trained model.

Method	Sent.	Sarcasm	Metaph.	MMSD
Concat Fusion	31.0	30.2	82.9	82.6
Late Fusion	27.8	32.1	81.8	82.0
MULT	24.3	28.4	84.7	83.0
GPT-4o	28.3	26.8	36.7	80.0
Gemini-2.5-Flash	32.7	38.8	71.3	85.9
o4-mini (CoT)	33.3	44.0	73.1	84.0
Qwen3-VL-30B	30.2	49.0	57.1	57.2
Incongruity Fusion	25.3	24.4	87.6	84.0
FusionMoE	26.8	25.3	87.6	84.0

6.3 Phase 3: Tri-Modal Fusion

Table 3 evaluates whether generated speech provides non-redundant affective information for meme understanding.

Table 3: Effect of speech alignment on tri-modal fusion.

Setting	Accuracy	Macro-F1
V only	0.15	0.38
V + T	0.27	0.45
V + T + A (naive)	0.27	0.45
V + T + A _{proj}	0.82	0.49

Tri-modal alignment. Naive tri-modal fusion provides no improvement over the image-text baseline despite the presence of affective information in speech. After alignment, performance improves substantially, indicating that modality mismatch—not lack of signal—is the primary obstacle to integrating speech into meme understanding.

6.4 Failure Analysis

Commonsense-heavy memes. Performance drops when interpretation requires external cultural or political knowledge beyond the meme itself.

Highly subjective humor and sarcasm. Some memes rely on deeply subjective interpretations or multi-layered irony, making affect understanding inherently ambiguous even for humans.

7 Conclusion

We presented **MemeSonic**, a framework for meme affective understanding and expressive voiceover generation. The central insight is that image-text incongruity is not noise, but a structural cue for understanding meme affect and pragmatic intent. MemeSonic models this cue explicitly and further externalizes implicit pragmatic affect as expressive speech.

(1) *Image-text incongruity improves meme affect understanding.* Compared with standard multimodal fusion, MemeSonic better handles cases where image and text convey conflicting signals, especially in sarcasm, humor, and non-literal affect. This shows that meme meaning often emerges not from either modality alone, but from the tension between them.

(2) *Expressive voiceover generation externalizes pragmatic affect.* MemeSonic achieves the best EMOS and AMOS scores while maintaining comparable speech naturalness. This suggests that the

learned prosody prompt adapter, conditioned on affect labels and image-text context, produces more meme-aligned delivery than direct label-guided or vanilla TTS.

(3) *Generated speech provides non-redundant affective supervision.* By using generated voiceover as a pseudo affective modality, MemeSonic introduces an additional supervisory signal for meme understanding. Our diagnostic tri-modal study suggests that affective information exists in speech.

Limitations. MemeSonic still relies on general-purpose vision and audio encoders, which may miss meme-specific visual semantics and expressive speech cues. In addition, humor, sarcasm, and pragmatic affect are inherently subjective, and existing datasets rarely provide aligned image, text, affect, intention, and audio annotations for the same memes. Our experiments should therefore be read as a focused validation of the proposed direction rather than a complete solution.

Future Work. A key next step is to build meme-native multimodal datasets with fine-grained annotations for affect, intention, humor, and paired voiceover audio. Such resources would support end-to-end training and evaluation of the full pipeline. Another direction is joint tri-modal pre-training, where image, text, and speech are aligned in a shared representation space from the start. Finally, image-text incongruity is not meme-specific: news captions, political cartoons, and short-form videos also rely on cross-modal tension. Extending MemeSonic to these domains would further test the generality of incongruity as an affective signal.

Acknowledgments and Disclosure of Funding

References

- Yitao Cai, Huiyu Cai, and Xiaojun Wan. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *ACL*, 2019.
- Matthew Caton et al. Memotion 3: Dataset for meme sentiment, emotion and intensity analysis. In *SemEval*, 2023.
- Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *CVPR*, 2023.
- Jack Hessel, Ana Marasovic, Jena D Hwang, Lillian Lee, James Davidson, Rowan Zellers, and Yejin Choi. Do androids laugh at electric sheep? humor “understanding” benchmarks from the new yorker caption contest. *ACL*, 2023.
- Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021.
- Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. In *NeurIPS*, 2020.
- Ziyue Lei et al. Msemotts: Multi-scale emotion transfer for cross-lingual speech synthesis. In *ICASSP*, 2022.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.
- Yinghao Aaron Li, Cong Han, Vinay S. Raghavan, Gavin Mischler, and Nima Mesgarani. StyleTTS 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. In *NeurIPS*, 2023.
- Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *NeurIPS*, 2022.
- Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. Emotion2Vec: Self-supervised pre-training for speech emotion representation. *arXiv preprint arXiv:2312.15185*, 2023.

- Libo Qin et al. MMSD2.0: Towards a reliable multi-modal sarcasm detection system. In *ACL Findings*, 2023.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.
- Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. FastSpeech 2: Fast and high-quality end-to-end text to speech. In *ICLR*, 2021.
- Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas Pykl, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Bjorn Gambäck. SemEval-2020 task 8: Memotion analysis — the visuo-lingual metaphor! In *SemEval*, 2020.
- Xu Shi, Mingzhe Xu, and Jiaxuan Liu. Memeclip: Leveraging clip representations for meme sentiment and humor analysis. In *EMNLP*, 2024.
- Jerry M Suls. A two-stage model for the appreciation of jokes and cartoons: An information-processing analysis. *The psychology of humor*, 1972.
- Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *ACL*, 2019.
- Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. In *Interspeech*, 2017.
- Younes Xu et al. 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*, 2022. doi: 10.1145/3477495.3532019.
- Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. Humor recognition and humor anchor extraction. In *EMNLP*, 2015.
- Haofei Yu, Zhengyang Qi, Lawrence Jang, Ruslan Salakhutdinov, Louis-Philippe Morency, and Paul Pu Liang. MMoE: Enhancing multimodal models with mixtures of multimodal interaction experts. In *EMNLP*, 2024.

This appendix provides additional implementation details, full quantitative results, and qualitative observations omitted from the main paper due to space constraints.

A Demo and Resources

A live demo of the MemeSonic pipeline—including meme input, incongruity score visualization, and synthesized audio playback—is available at:

<https://tinyurl.com/4schzjb5>

Generated audio samples from our 100-meme evaluation set are available on Google Drive, organized by sentiment class and voice setting for qualitative comparison across synthesis strategies: Google Drive folder

B Full Classification Results

Tables 4 and 5 report Accuracy and Macro-F1 across all methods and tasks. Bold green indicates the best trained model. LLM baselines are included for contextual reference only and are not directly comparable to trained models due to differing inference settings and reasoning budgets.

Table 4: Accuracy across all methods and tasks. Bold green = best trained model.

Dataset	Task	Trained Models							LLM Zero-Shot				
		Uni-I	Uni-T	Concat	Late	Incong	MULT	FusionMoE	GPT-4o	Gemini	o4-mini	Gemini-CoT	Qwen3
MET-Meme	intention	45	40	42	42	37	45	44	52	48	56	46	50
MET-Meme	sentiment	39	36	44	40	48	37	45	28	34	32	36	32
MET-Meme	hate	84	82	84	85	87	86	86	76	74	86	76	82
MET-Meme	metaphor	81	76	83	82	88	85	88	58	72	74	64	64
Memotion	humour	54	49	58	58	66	54	65	68	68	66	46	68
Memotion	sarcasm	37	38	33	34	47	44	42	52	50	56	44	54
Memotion	offensive	53	54	52	51	52	58	48	38	46	42	38	40
Memotion	motivational	57	60	61	60	68	65	62	52	54	52	46	54
Memotion	sentiment	45	46	44	49	56	50	46	30	26	28	22	24
MMSD 2.0	sarcasm	78	78	83	82	84	83	83	80	86	84	74	58

Table 5: Macro-F1 across all methods and tasks. Bold green = best trained model.

Dataset	Task	Trained Models							LLM Zero-Shot				
		Uni-I	Uni-T	Concat	Late	Incong	MULT	FusionMoE	GPT-4o	Gemini	o4-mini	Gemini-CoT	Qwen3
MET-Meme	intention	32	30	32	31	21	27	29	35	34	45	27	40
MET-Meme	sentiment	25	24	31	28	25	24	29	28	33	33	39	30
MET-Meme	hate	39	39	38	38	37	41	38	30	36	44	30	31
MET-Meme	metaphor	81	76	83	82	88	85	88	37	71	73	64	57
Memotion	humour	36	33	36	36	27	36	30	27	31	30	21	32
Memotion	sarcasm	33	33	30	32	24	28	29	27	39	44	39	49
Memotion	offensive	32	30	33	34	35	24	26	24	31	60	29	27
Memotion	motivational	49	48	47	51	41	43	53	38	39	38	37	39
Memotion	sentiment	32	31	31	31	25	37	34	27	24	27	22	19
MMSD 2.0	sarcasm	78	78	83	82	84	83	83	80	86	84	74	57

C t-SNE Visualization of Audio Embeddings

Figure 2 shows a t-SNE projection of Emotion2Vec embeddings for all 900 audio samples (100 memes $\times$ 9 voice settings), colored by ground-truth sentiment class. Points cluster by sentiment rather than by voice setting, confirming that Emotion2Vec captures semantic emotion content and is largely insensitive to surface prosodic variation (pitch, rate, stability) introduced by different ElevenLabs voice settings.

D Prosody Grid Details

The 9-point prosody grid used in the tri-modal alignment study spans:

$$\text{stability} \in \{0.2, 0.5, 0.8\} \times \text{style} \in \{0.2, 0.5, 0.8\}$$

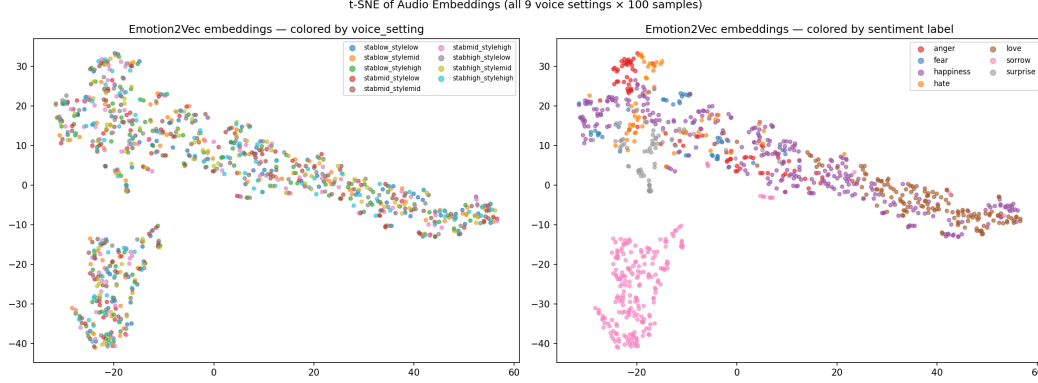


Figure 2: t-SNE of Emotion2Vec embeddings for 900 audio files (100 memes $\times$ 9 prosody settings). Colors indicate GT sentiment class. Clustering by sentiment rather than voice setting confirms Emotion2Vec’s robustness to prosodic surface variation.

using ElevenLabs voice_settings. All other parameters (similarity_boost=0.75, use_speaker_boost=True) are held constant. Table 6 summarizes the intended perceptual effect of each setting combination.

Table 6: Prosody grid: perceptual intent of each (stability, style) combination.

Stability	Style	Perceptual character
0.2	0.2	Unstable, flat — inconsistent emotion, low expressiveness
0.2	0.5	Unstable, moderate — variable rhythm, some emotional coloring
0.2	0.8	Unstable, expressive — erratic and emotionally intense
0.5	0.2	Neutral, flat — consistent but monotone (close to vanilla TTS)
0.5	0.5	Balanced — moderate stability and expressiveness
0.5	0.8	Neutral stability, high style — expressive with controlled variation
0.8	0.2	Very stable, flat — robotic, minimal affect
0.8	0.5	Stable, moderate — clear and measured delivery
0.8	0.8	Stable, expressive — polished and emotionally rich

E Implementation Details

Tables 7–9 summarize hyperparameters for each phase. All experiments were conducted on a consumer-grade MacBook Pro M2 setup without dedicated GPU acceleration.

F Error Analysis

Where Incongruity Fusion fails. Incongruity Fusion underperforms on MET/intention (−15.6% F1 vs. Unimodal-Image) and Memotion/humour (−8.9% F1 vs. Late Fusion). This suggests that incongruity is not universally beneficial: it introduces noise when tasks depend on holistic semantics rather than modality conflict. In particular, intention labels are driven by joint distributional cues rather than emotional mismatch. A more effective approach may be task-adaptive gating of incongruity signals.

LLMs and reasoning models. o4-mini (CoT) achieves strong performance on Memotion/offensive (59.9% F1) and sarcasm (44.0% F1), highlighting the benefit of explicit reasoning for socially

Table 7: Phase 2 — Multimodal Fusion hyperparameters.

Component	Hyperparameter	Value
Encoder	Backbone	CLIP ViT-B/32 (frozen)
	Output dim	512
MLP Head	Hidden layers	$512 \rightarrow 256 \rightarrow \text{num_classes}$
	Activation	ReLU
	Dropout	$p = 0.3$
Training	Optimizer	AdamW
	Learning rate	5×10^{-4}
	Weight decay	10^{-3}
	Batch size	64
	Max epochs	15
	Early stopping	patience 5 (val Macro-F1)
FusionMoE	Temperature τ	0.05
	Gating network	Single linear layer
	Gating input	$[\phi_v; \phi_t; \delta]$ (dim 1025)

Table 8: Phase 3 — Tri-Modal Alignment hyperparameters.

Component	Hyperparameter	Value
Audio Encoder	Model	Emotion2Vec (iic/emotion2vec_plus_large)
	Output dim	1,024
Target Space	Model	ImageBind
	Dim	1,024
Projector	Architecture	Linear + L2 normalize
	Optimizer	AdamW
	Learning rate	10^{-3}
	LR schedule	Cosine decay
	Batch size	32
	Epochs	300
Data	Train/Test split	80/20 (720/180 samples)
	Augmentation	9 prosody variants per meme

contextual tasks. This advantage is not captured by CLIP-based models. Integrating LLM reasoning with incongruity-aware features is a promising direction.

G Dataset Statistics

Table 10 summarizes the dataset splits and class distributions used in Phase 1 experiments.

H Additional Figures

I Qualitative Case Studies

To better illustrate the role of cross-modal incongruity in affective meme understanding, we provide several representative qualitative examples.

- Cases where image-text incongruity improves sarcasm recognition.
- Examples where MemeSonic generates stronger prosodic emphasis than vanilla TTS baselines.
- Failure cases involving political references, layered irony, or culturally specific humor.

Table 9: Phase 1 — Audio Generation hyperparameters (Incongruity-to-Prosody Adapter).

Component	Hyperparameter	Value
Prosody Mapping	stability formula	$1 - 0.6 \delta$
	style formula	0.7δ
	Training	None (fixed affine mapping)
ElevenLabs	similarity_boost	0.75
	use_speaker_boost	True
	Voice	Fixed identity across all memes

Table 10: Dataset statistics for multimodal meme understanding benchmarks.

Dataset	Task	Train	Val	Test	Classes
MET-Meme	intention	1,400	300	340	5 (interactive, expressive, entertaining, offensive, other)
MET-Meme	sentiment	1,400	300	340	7 (happiness, love, anger, sorrow, fear, hate, surprise)
MET-Meme	hate	1,400	300	340	4 (intensity levels)
MET-Meme	metaphor	1,400	300	340	2 (binary)
Memotion	humour	5,593	699	700	3
Memotion	sarcasm	5,593	699	700	3
Memotion	offensive	5,593	699	700	4
Memotion	motivational	5,593	699	700	2
Memotion	sentiment	5,593	699	700	3 (pos/neu/neg)
<i>Memotion splits are random 80/10/10 (seed 42); not official splits.</i>					
MMSD 2.0	sarcasm	16,000	3,000	3,482	2 (binary)

Example Structure. For each case study, we recommend reporting the meme caption, observed incongruity signal, generated speech behavior, and a short interpretation of why the system succeeds or fails.

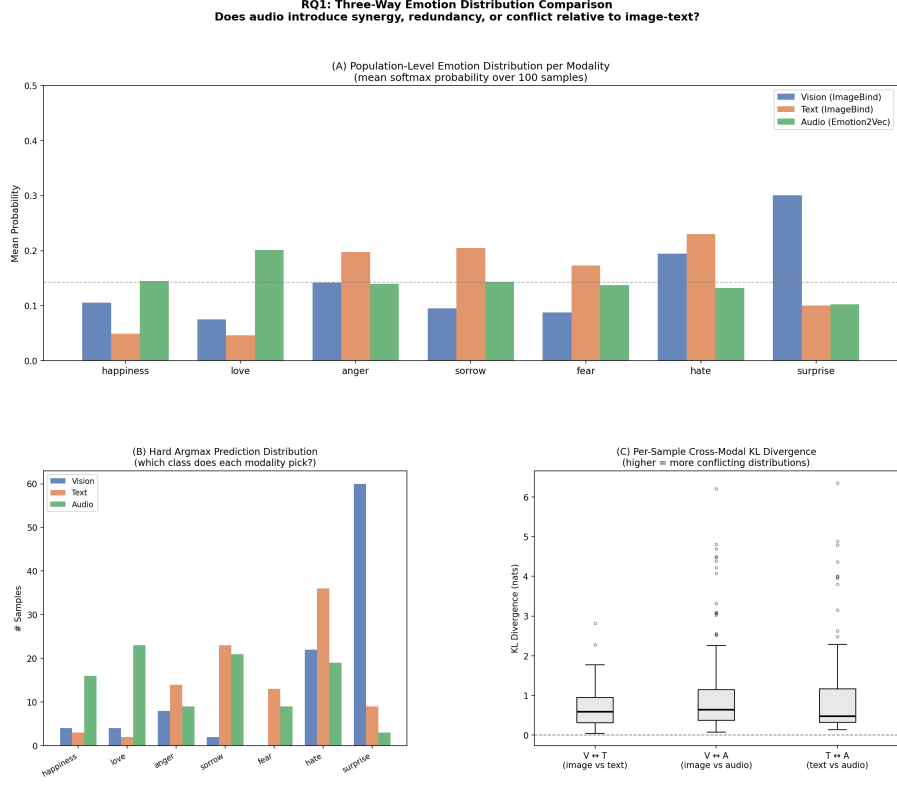


Figure 3: Three-way emotion distribution across image, text, and audio modalities. The divergence between per-modality distributions motivates the incongruity score δ as an explicit signal for cross-modal conflict.

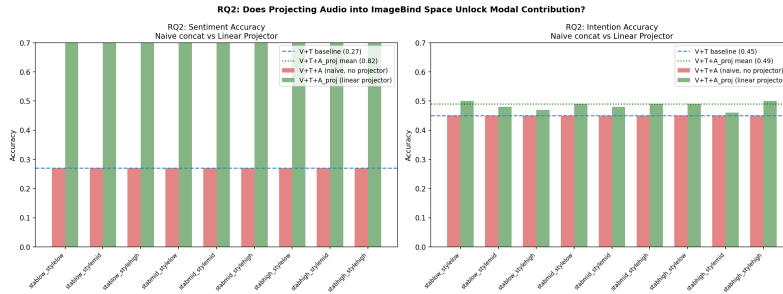


Figure 4: Tri-modal fusion accuracy with and without the linear projector. Naive fusion (all 9 prosody settings) yields no improvement over the image-text baseline; the learned projector lifts sentiment accuracy from 0.27 to 0.82.

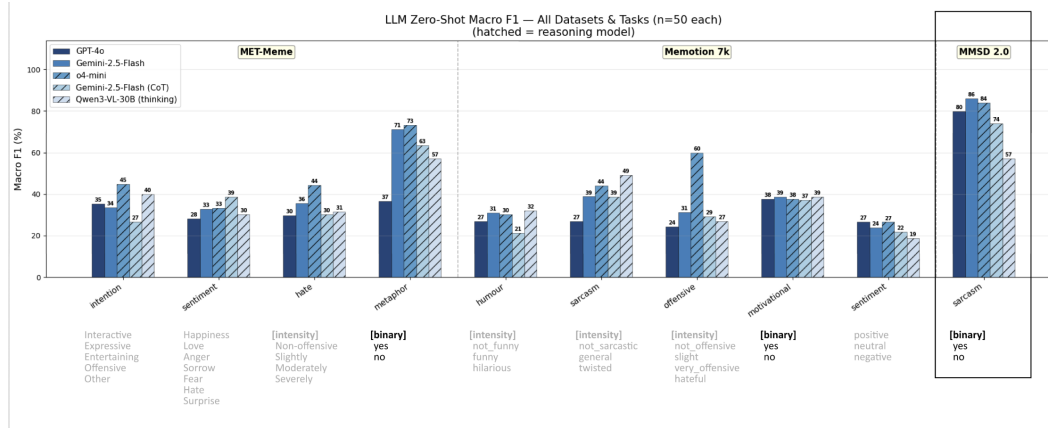
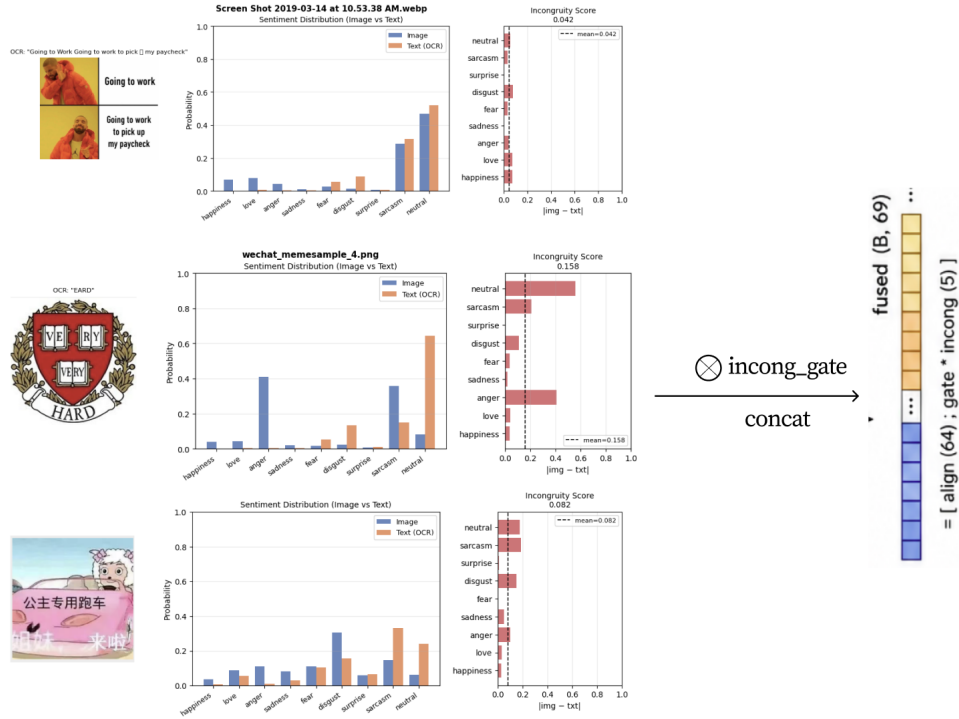


Figure 5: LLM zero-shot Macro-F1 across all datasets and tasks ($n = 50$ each). Hatched bars indicate reasoning models (CoT). LLMs perform strongly on open-ended tasks (intention, sarcasm) but underperform trained models on structured affect tasks (metaphor, motivational).



Figure 6: Sentiment–intention space of meme affect labels. Emotions cluster into four quadrants by valence (positive/negative) and intention (high/low), revealing the structured pragmatic space that meme affect models must navigate.



real implementation is not like this (deterministic)

Figure 7: Illustration of the Incongruity Fusion (Expert-S) mechanism. Per-modality emotion distributions are computed and their difference $p_v - p_t$ is concatenated with the incongruity score δ as explicit conflict features for the MLP classifier.

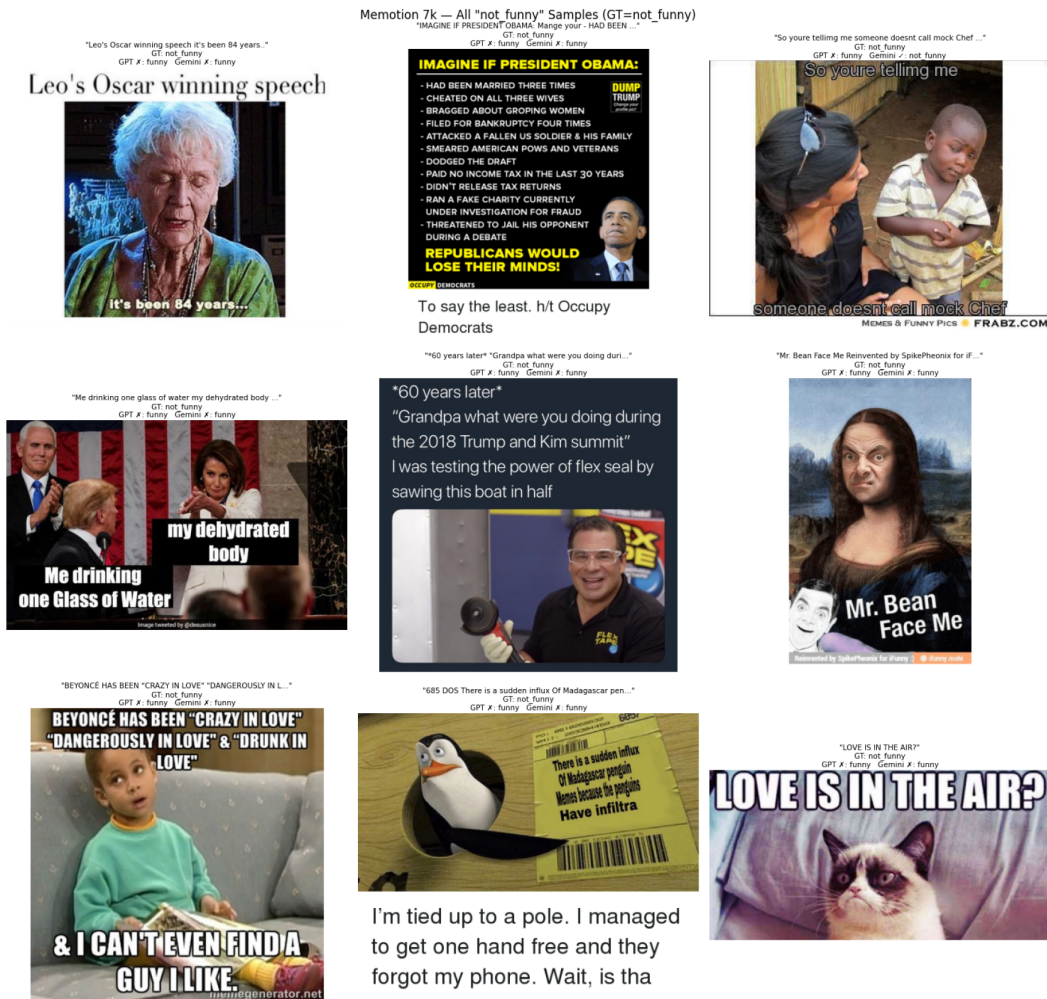


Figure 8: Error analysis: examples from Memotion 7k where LLM predictions (GPT-4o, Gemini) diverge from ground truth. Cases labeled *not_funny* by annotators are often classified as *funny* by LLMs, reflecting subjectivity in humor perception.