BanglaMemeX: Advancing Cultural Metaphoric Image Interpretation in Bangla with a Multimodal Explainable Dataset

Anonymous ACL submission

Abstract

Vision Language Models (VLMs) have achieved strong performance on multimodal benchmarks, yet their ability to reason about culturally grounded and metaphor-rich content remains insufficiently studied. Internet memes present a challenging setting where meaning emerges from implicit interactions between image, overlaid text, sarcasm, and shared socio-cultural knowledge rather than literal visual recognition. This challenge is amplified in low-resource languages such as Bangla, where code-mixing, stylized scripts, and culturally specific symbolism introduce substantial distribution shift. In this work, we introduce **BanglaMemeX**, a culturally grounded multimodal benchmark comprising 3,000 Bangla memes annotated with multi-dimensional labels (humor, sarcasm, offensiveness, motivational intent) and human-written explanations that explicitly describe textual and visual metaphors. We systematically evaluate modern VLMs on both classification and explanation generation, revealing that current models struggle to interpret implicit cultural cues despite reasonable surface-level accuracy. Our results highlight the need for culturally-aware multimodal systems capable of grounded reasoning under linguistic and cultural distribution shift. We envision BanglaMemeX as a foundational resource for advancing explainable vision–language understanding in low-resource and culturally sensitive settings.understanding.

1 Introduction

Social media platforms have become major channels for public discourse and cultural exchange. Among the most common forms of communication on these platforms are internet memes, which combine images, text, and cultural references to convey opinions, humor, or social commentary. Due to their visual and concise format, memes spread rapidly and have been studied for tasks such as hate speech detection, sentiment analysis, and offensive content classification (Xu et al., 2022). However, meme meaning rarely arises from literal image to text correspondence. Instead, interpretation often depends on metaphor, sarcasm, irony, and shared cultural context. As a result, categorical labels alone provide only a partial view of the underlying reasoning behind a meme's interpretation. Explanations can therefore improve interpretability and help analyze model decisions in socially sensitive domains (Ribeiro et al., 2016; Doshi-Velez and Kim, 2017). This challenge becomes more pronounced in multimodal settings where meaning emerges from interactions between visual and textual elements. Despite progress in Vision Language Models (Radford et al., 2021), many multimodal benchmarks focus on literal tasks such as captioning or visual question answering, which primarily evaluate cross-modal alignment rather than culturally grounded reasoning.

Several Bangla meme datasets have recently been introduced. MemoSen (Hossain et al., 2022a) studies sentiment classification, while MUTE (Hossain et al., 2022b), BanHateMeme (Mia and Fahim, 2025), and BanglaAbuseMeme (Das and Mukherjee, 2023) focus on hate or abusive content detection. BANMIME (Mia et al., 2025) introduces misogyny detection with explanations, and CMBAN (Alam et al., 2025) studies contextual classification in cartoon memes. While these datasets support important tasks, most emphasize categorical prediction and do not explicitly evaluate whether models correctly interpret metaphor and reasoning.

To address this limitation, we introduce **BanglaMemeX**, a dataset of 3,000 Bangla memes annotated with labels for humor, sarcasm, offensiveness, and motivational intent, along with overall sentiment and human-written explanations describing the textual and visual metaphors in each meme. These explanations allow evaluation beyond classification accuracy and enable analysis of model reasoning. We evaluate modern VLMs using

Dataset Name	Volume	Modality	Explored Hate Domain	Explanation
BanglaSarc3 (Biswas et al., 2025a)	12,089	Text	Sarcasm	×
CMBAN (Alam et al., 2025)	2,641	Multimodal	Cartoon-based Meme	×
BanMiMe (Mia et al., 2025)	2,000	Multimodal	Misogyny Meme	✓
BanHateMeme (Mia and Fahim, 2025)	3,819	Multimodal	Misogyny & Sarcastic Meme	×
BanglaAbuseMeme (Das and Mukherjee, 2023)	4,043	Multimodal	Abuse, Vulgarity, Sarcasm, Sentiment	×
BanHate (Raquib et al., 2025)	19,203	Text	General Hate Category	×
BanglaMemeX (Ours)	3,000	Multimodal	Humor, Sarcasm, Abuse, Motivation, Sentiment	✓

Table 1: Overview of major Bangla textual and multimodal benchmark datasets (2024-2026). The Explanation column indicates whether the dataset includes explicit human-written explanation annotations.

zero-shot prompting, chain-of-thought reasoning, and parameter-efficient fine-tuning. Explanation quality is assessed using BERTScore (Zhang et al., 2020) and an LLM-as-a-judge evaluation. We also employed a structured *LLM Council* prompting for explanation generation and explored LoRA-based adaptation to improve robustness under cultural and linguistic variation (Hu et al., 2022).

Our contributions are threefold. First, we introduce BanglaMemeX, a culturally grounded multimodal dataset with explanation-level annotations for metaphor-rich meme understanding. Second, we provide a systematic evaluation of modern vision–language models under cultural and linguistic shift, examining whether models can correctly interpret metaphorical Bangla memes and how explanation supervision influences reasoning quality.

2 Related Work

Vision Language Models Vision Language Models (VLMs) have progressed through large-scale contrastive and generative pretraining. CLIP (Radford et al., 2021) demonstrated that image–text alignment at scale enables strong zero-shot transfer across tasks. Subsequent models such as BLIP (Li et al., 2022) and Flamingo (Alayrac et al., 2022) extended this framework toward unified multimodal understanding and few-shot reasoning. These systems perform effectively on captioning, visual question answering, and instruction-following benchmarks. However, most evaluation settings emphasize literal semantic meaning and are primarily developed using high-resource languages. Their performance under linguistic nuance and cultural variation, particularly for metaphor-rich content, remains less explored.

Multimodal Meme Understanding in Bangla Memes require visual context, linguistic clues, and sociocultural expertise to comprehend. According to the Hateful Memes Challenge (Kiela et al., 2020), multimodal fusion is essential for accurate hate detection. Later research focused on misogyny detection, sentiment analysis, and multimodal classification (Xu et al., 2022). Bangladeshi datasets have addressed classification tasks. MemoSen (Hossain et al., 2022a) analyzed Bangla meme sentiment, whereas MUTE (Hossain et al., 2022b) and BanHateMeme (Mia and Fahim, 2025) detected multimodal hate. BengaliAbuseMeme (Das and Mukherjee, 2023) expanded this approach to classify abusive content. BANMIME (Mia et al., 2025) added misogyny intensity categories and metaphor localization, whereas CMBAN (Alam et al., 2025) added humor and sarcasm to contextual classification. BanglaSarc3 (Biswas et al., 2025b) investigated sarcasm recognition in code-mixed text. These materials expanded Bangla multimodal research annotation techniques and benchmark coverage. Most assessment methodologies, however, remain focused on category prediction tasks such as hatred, abuse, emotion, or sarcasm categorization. Current Bangla benchmarks prioritize label-based evaluation for sentiment, hatred, abuse, and sarcasm (see Table 1). While some datasets incorporate structured annotations such as metaphor localization (Mia et al., 2025), systematic examination of explanation alignment and reasoning consistency has received comparably little attention.

3 Dataset

We introduce **BanglaMemeX**, a curated collection of 3,000 Bangla multimodal memes annotated across four communicative dimensions—Humor, Sarcasm, Offensiveness, and Motivational intent—together with Overall Sentiment and explanation-level supervision. The dataset construction workflow is depicted in Figure 1.

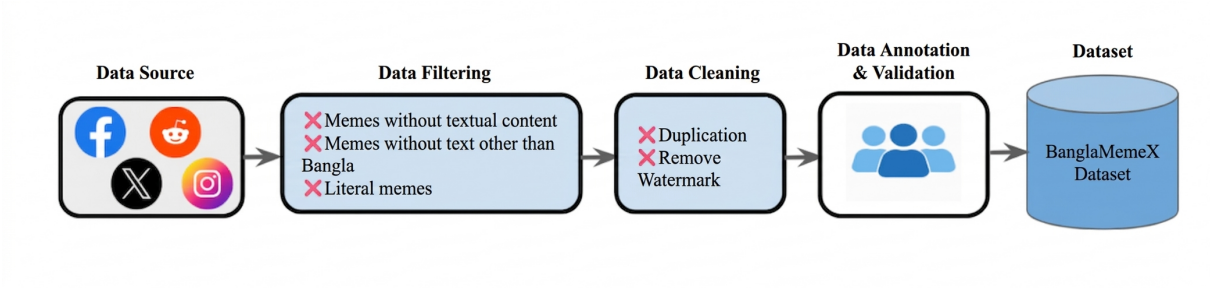


Figure 1: Overview of the BanglaMemeX dataset construction pipeline.

3.1 Data Collection and Annotation

Data Collection. We collected 105,835 memes from publicly accessible posts on Facebook, Instagram, Reddit, and X, focusing on Bangla-speaking communities where memes commonly combine visual imagery with overlaid Bangla or code-mixed Bangla–English text. Only publicly available content was considered, and no user-identifiable information or engagement metadata were retained during collection. Images are redistributed only when permitted under platform terms.

Filtering and Cleaning. The collected memes were processed through a structured filtering pipeline to ensure linguistic and visual consistency. First, duplicate and near-duplicate images were removed through manual inspection. We then discarded memes containing third-party watermarks, such as logos of social media pages or groups embedded within the image. Memes whose overlaid text was not written in Bangla or code-mixed Bangla–English were also excluded.

Automatic OCR tools often struggle to reliably extract Bangla meme text due to stylized fonts, mixed scripts, and low-resolution overlays commonly used in memes. Therefore, instead of relying on external OCR systems, all textual content was manually transcribed by annotators during the annotation process. After applying these filtering steps, the final dataset consists of 3,000 memes.

Annotation. The dataset was annotated by three native Bangla-speaking annotators who are familiar with contemporary Bangla meme culture and are undergraduate students in linguistics. The annotation process took place over a period of three months, and annotators were compensated BDT 5 per meme. Because the dataset contains instances of hateful and misogynistic content, structured annotation breaks were incorporated throughout the process to reduce cognitive fatigue and maintain annotation quality. The full annotation process required approximately 1500 annotator hours. On average, annotators spent 10 minutes per meme to identify labels, metaphor types and produce explanations.

Each meme was annotated with multiple components. Annotators first assigned communicative labels capturing humor, sarcasm, offensiveness, motivational and overall intent. They then identified the textual metaphor, describing the figurative meaning conveyed by the overlaid text, and the visual metaphor, referring to the symbolic meaning represented through elements within the image. Finally, annotators wrote a short natural-language explanation linking the visual and textual elements to the underlying cultural meaning or communicative intent.

To support visual metaphor annotation, we designed a semi-assisted annotation pipeline. Each meme image was first processed using the Segment Anything Model (SAM) (Kirillov et al., 2023), which generates region-level segmentations of the image. Annotators then selected the region corresponding to the visual metaphor and described its symbolic role before writing the final explanation. Full annotation guidelines and representative examples are provided in Appendix A.

3.2 Data Analysis

The dataset is divided into training and test sets using stratified multi-label sampling to preserve label distributions across communicative categories. Inter-annotator agreement was measured using Fleiss' κ (Fleiss, 1971) across three annotators. For multi-class labels, agreement was computed separately for each class using a one-vs-rest formulation. The resulting per-class κ values are shown in Table 6 in A. Dataset statistics, label distribution, and explanation length values are presented in Table 2. To verify the quality and consistency of the annotated explanations, we conducted both

Dataset Attribute	Value
Total Memes	3,000
Training/Test Split	2,550/450
Label Distributions	
Humor	
Funny	2,438 (81.3%)
Very Funny	341 (11.4%)
Not Funny	219 (7.3%)
Sarcastic	
Little Sarcastic	1,489 (49.6%)
Not Sarcastic	768 (25.6%)
Very Sarcastic	741 (24.7%)
Offensive	
Not Offensive	1,955 (65.2%)
Slight Offensive	714 (23.8%)
Very Offensive	228 (7.6%)
Hateful Offensive	101 (3.4%)
Motivational	
Not Motivational	2,831 (94.4%)
Motivational	167 (5.6%)
Metaphor Statistics	
Image Metaphor Mean Length	32.83 chars
Textual Metaphor Mean Length	11.49 chars
Explanation Statistics	
Explanation Mean Length	99.58 chars
Explanation Modality	Text
Maximum Explanation Length	541 chars

Table 2: Summary statistics and label distributions of the BanglaMemeX dataset. Length values are reported in characters.

Figure 2: Example of a BanglaMemeX instance with multi-dimensional labels, metaphor annotations, and explanation.

human evaluation and semantic similarity analysis. Each annotated meme contains three textual components: (i) the textual metaphor description (ii) the visual metaphor description and (iii) the final explanation. Self evaluation was performed by three independent annotators as they added a five-point Likert scale assessing correctness, clarity, and cultural appropriateness for each textual component. Annotations receiving an average score below 3.5 were flagged for revision and sent for re-annotation until satisfaction of threshold. The human quality score was computed as:

$$Q = \frac{1}{N} \sum_{i=1}^{N} s_i \quad (1)$$

where s_i represents the score assigned by annotator i and N is the total number of annotators.

We did not only rely on self evaluation. To measure semantic agreement between annotators, we computed BERTScore (Zhang et al., 2020), which evaluates similarity using contextual embeddings from pretrained transformer models. Table 3 reports the human evaluation scores and semantic similarity results for textual metaphor descriptions, visual metaphor explanations, and the final multimodal explanation.

BERTScore was computed at the instance level between explanations written by different annotators. For each meme, pairwise BERTScore F1 values were calculated across annotator explanations and averaged to obtain an instance-level agreement score. Corresponding textual component with maximum BERTScore F1 were selected. The final dataset-level BERTScore reported in Table 3 corresponds to the mean score across all accepted annotations. complementary overlap metrics also indicate strong agreement (ROUGE-1 F1: 85.9, ROUGE-2 F1: 50.8, ROUGE-L F1: 84.7) (Lin, 2004). Figure 2 provides a representative example illustrating the multimodal annotation structure, including categorical labels, textual and visual metaphors, and the corresponding human-written explanation. The dataset, annotation guidelines, and evaluation scripts will be publicly released upon acceptance to facilitate reproducibility and future research.

Annotation Component	Human Score	mean BERT-P	mean BERT-R	mean BERT-F1
Textual Metaphor Description	4.12	0.88	0.86	0.87
Visual Metaphor Explanation	4.08	0.87	0.85	0.86
Final Multimodal Explanation	4.18	0.89	0.86	0.87

Table 3: Human evaluation and semantic similarity analysis of BanMemeX metaphor annotations.

4 Experiment Setup

We evaluate multimodal reasoning and culturally grounded interpretation of memes on the BanglaMemeX dataset using three experimental settings: (i) prompt-based inference, (ii) a role-based LLM Council , and (iii) parameter-efficient fine-tuning via Low-Rank Adaptation (LoRA). The overall experimental pipeline is illustrated in Figure 3.

4.1 Problem Formulation

Each meme instance consists of an image I and optional overlaid text T. The objective is to predict a set of semantic labels

$$Y = \{y_1, y_2, \ldots, y_k\},$$

corresponding to *humor*, *sarcasm*, *offensiveness*, *motivational intent*, and *overall sentiment*. In addition to label prediction, the model generates an explanation E describing the interaction between visual elements, textual cues, and implied metaphors. The task therefore combines multi-label classification with explanation generation.

4.2 Prompt-Based Inference

We first evaluate pretrained vision-language models using structured prompting without parameter updates. For each meme, a template prompt specifies the prediction categories and enforces a structured output format.

Two prompting strategies are considered.

Zero-shot prompting. The model predicts the labels and generates an explanation directly from the input meme.

Chain-of-thought prompting. The model is encouraged to produce intermediate reasoning steps, including visual interpretation, textual understanding, and metaphor identification before generating the final prediction.
Refer to Appendix B for detailed prompt.

4.3 Role-Based LLM Council

We employ a three-stage, multi-model *LLM Council* to improve explanation quality through deliberative refinement. The council consists of two explainer models (Claude Opus 4.5 and Gemini-3 Flash), three critic models (Llama-4 Maverick, Qwen3-VL-8B-Instruct, and Gemma-3-12B-IT), and one synthesizer model (GPT-5.2). The explainer stage generates initial interpretations of the meme's literal meaning, figurative meaning, cultural context, and labels. The critic stage reviews these interpretations for grounding, cultural alignment, and possible interpretive gaps. The synthesizer then reconciles disagreements and produces the final structured prediction. Consensus is achieved through *sequential synthesis* rather than majority voting. All roles use fixed prompt templates, a shared predefined label schema, and constrained structured outputs. Images are resized to a maximum of 1024×1024 pixels and reused across all stages. Refer to Appendix B for detailed prompt.

4.4 Parameter-Efficient Fine-Tuning

To adapt models to culturally grounded supervision, we apply parameter-efficient fine-tuning using LoRA. Instead of updating all parameters, LoRA introduces trainable low-rank matrices into selected projection layers while keeping the pretrained weights fixed.

Given a weight matrix $W \in \mathbb{R}^{d\times k}$, LoRA updates the weight as

$$W' = W + \Delta W, \quad \Delta W = BA,$$

where $A \in \mathbb{R}^{r\times k}$ and $B \in \mathbb{R}^{d\times r}$ are low-rank matrices with $r \ll \min(d, k)$.

Training optimizes a joint objective combining multi-label classification and explanation generation losses.

4.5 Evaluation Metrics

Model performance is evaluated along two dimensions.

Classification performance. We compute standard multi-label metrics for humor, sarcasm, offensiveness, motivational intent, and overall sentiment.

Explanation quality. Explanation faithfulness is evaluated using BERTScore(Zhang et al., 2020)

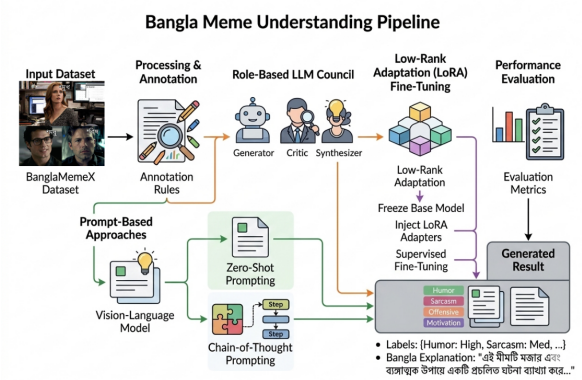


Figure 3: Proposed pipeline to benchmark BanglaMemeX dataset.

together with an LLM-as-a-judge evaluation inspired by LAVE(Mañas et al., 2024). The evaluation assesses textual grounding, cultural alignment, and metaphor interpretation.

4.6 Implementation Details

Experiments are conducted using the predefined train, validation, and test splits of BanglaMemeX. Images are resized according to the input resolution of each backbone model, and meme images are processed directly without external OCR. LoRA fine-tuning experiments are performed on an NVIDIA RTX 4090 GPU (24GB VRAM) using PyTorch with mixed-precision optimization. For prompt-based inference and the LLM Council experiments, models are accessed through OpenRouter-hosted inference endpoints. For LoRA fine-tuning, adapters are applied to attention projection layers with rank $r = 16$, scaling factor $\alpha = 32$, and dropout 0.05. Training runs for 5 epochs using AdamW with learning rate 2×10^{-4} and weight decay 0.01. Training uses batch size 1 with mixed-precision optimization. All experiments use fixed random seeds. Model selection is performed on the validation split, and the test split is used only for final evaluation.

5 Results Analysis

Table 4 reports classification and explanation performance across five meme interpretation tasks, while Table 5 summarizes textual and visual metaphor understanding measured by LAVE. Figure 4 provides a macro-level view of average classification accuracy (*avg*) across inference strategies.

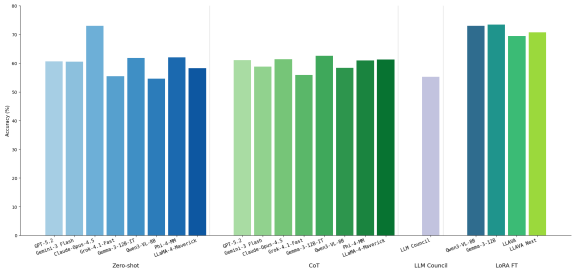


Figure 4: **Average classification accuracy across strategies on BanglaMemeX.** Bars report the *Avg* score (%) from Table 4 for each model under zero-shot prompting, chain-of-thought prompting, LLM Council reasoning, and LoRA fine-tuning. LoRA yields the highest average accuracy among fine-tuned open-source models, while the LLM Council prioritizes explanation quality over label accuracy.

Closed-source vs. open-source models Under zero-shot prompting, closed-source models form the strongest baseline overall (Figure 4). The highest average accuracy is achieved by Claude-Opus-4.5 (Avg=72.96). This advantage is concentrated in the more inference-intensive labels, particularly sarcasm and offensiveness (Sarc=68.50; Offe=66.10). By comparison, GPT-5.2 and Gemini-3 Flash remain competitive on labels like humor and motivational intent (Hum ≈ 80; Moti > 94), but exhibit weaker performance on sarcasm and overall sentiment. Open-source models exhibit a broader spread in zero-shot accuracy. Within this group, Phi-4-MM provides the strongest baseline (Avg=62.00), whereas Qwen3-VL-8B attains the lowest (Avg=54.56). Across architectures, motivational intent is consistently the easiest label, plausibly because of direct textual and visual patterns. Sarcasm, in contrast, remains consistently difficult reflecting the need to infer rhetorical reversal and culturally specific framing. Yet a different picture emerges when examining explanation quality. BERTScore (Zhang et al., 2020) remains high and relatively concentrated indicating fluent generations with substantial lexical and semantic overlap with references. But metaphor faithfulness measured by LAVE varies substantially. This divergence implies that explanations may appear coherent and reference-aligned but fail to capture the underlying metaphorical relation.

Do reasoning prompts help? Figure 4 suggests that chain-of-thought prompting produces an aggregate effect that is small and model-dependent. GPT-5.2 increases marginally in Avg from 60.59

Models	Classification						Explanation		
	Hum	Sarc	Offe	Moti	Over	Avg	BScore	LAVE	ExplAvg
Zero-Shot Prompting									
GPT-5.2	81.10	41.70	43.20	94.43	42.53	60.59	69.40	34.70	52.05
Gemini-3 Flash	79.63	35.25	56.95	95.07	35.85	60.55	70.00	33.61	51.81
Claude-Opus-4.5	79.90	68.50	66.10	94.50	55.80	**72.96**	69.19	52.98	**61.10**
Grok-4.1-Fast	77.68	38.91	43.21	89.62	27.59	55.40	68.16	7.47	37.82
Gemma-3-12B-IT	80.43	42.80	61.83	94.37	29.50	61.79	66.86	2.90	34.88
Qwen3-VL-8B	62.43	30.20	52.50	93.30	34.37	54.56	66.86	4.37	35.62
Phi-4-MM	80.27	38.10	58.20	94.40	39.03	62.00	67.62	3.63	35.63
LLaMA-4-Maverick	77.23	27.87	50.37	94.53	41.13	58.23	69.58	16.97	43.28
Chain-of-Thought Prompting									
GPT-5.2	81.73	45.33	52.47	90.60	34.90	61.01	69.26	40.17	54.72
Gemini-3 Flash	69.89	44.61	60.92	93.83	24.81	58.81	69.00	34.74	51.87
Claude-Opus-4.5	66.84	42.86	56.50	94.17	46.24	61.32	69.52	60.88	**65.20**
Grok-4.1-Fast	79.53	40.00	43.37	92.07	24.23	55.84	67.98	8.30	38.14
Gemma-3-12B-IT	77.07	45.20	60.90	94.00	35.80	**62.59**	66.88	4.10	35.49
Qwen3-VL-8B	75.67	45.00	53.87	94.47	23.00	58.40	67.58	7.33	37.46
Phi-4-MM	74.83	46.37	60.77	94.23	28.23	60.89	67.54	6.43	36.99
LLaMA-4-Maverick	79.63	46.87	61.27	94.53	24.07	61.27	69.53	20.53	45.03
LLM Council									
LLM Council	53.17	32.29	51.10	95.26	44.06	55.18	68.37	**68.82**	**68.60**
LoRA Fine-Tuning									
Qwen3-VL-8B	81.70	55.50	66.70	95.10	66.10	73.00	68.09	11.82	**39.96**
Gemma-3-12B	82.20	58.80	65.40	95.40	65.20	**73.40**	69.19	7.20	38.20
LLaVa-1.6-Mistral-7B	81.80	44.90	64.50	94.40	61.80	69.48	68.06	4.60	36.33
LLaVa-Next-8B	81.20	48.60	66.80	93.90	63.00	70.70	66.59	3.41	35.00

Table 4: Performance comparison on BanglaMemeX across zero-shot prompting, chain-of-thought prompting, LLM Council reasoning, and LoRA fine-tuning. Here ExplAvg is the average score of BScore and LAVE score

to 61.01, whereas Gemini-3 Flash declines from 60.55 to 58.81. Same story in Claude Opus 4.5. CoT appears most useful on the more inference-heavy labels like sarcasm. For instance, GPT-5.2 improves on sarcasm from 41.70 to 45.33 (+3.63), and LLaMA-4-Maverick exhibits a larger increase from 27.87 to 46.87. But the benefits are not stable across labels. Several models exhibit notable degradations on overall sentiment under CoT prompting; for example, GPT-5.2 decreases on Overall sentiment from 42.53 to 34.90, and Gemini-3 Flash drops from 35.85 to 24.81. Consequently, the net improvement in Avg remains modest across the evaluated suite.

Effect of LLM Council The LLM Council consistently delivers the strongest explanation quality among all evaluated approaches as shown in Table 4. These gains, however, are accompanied by a marked decline in label prediction accuracy. The multi-model deliberation in the LLM Council tends to prioritize broader, more narratively appealing interpretations, even when these diverge from the strict decision boundaries demanded by the classification rubric.

Tradeoffs of LoRA fine-tuning The most pronounced improvements in classification arise from LoRA-based adaptation. As summarized in Figure 4 and detailed in Table 4, Qwen3-VL-8B increases in Avg from 54.56 under zero-shot prompting to 73.00 after fine-tuning, and Gemma-3-12B attains Avg=73.40, marginally surpassing the strongest zero-shot baseline of Claude-Opus-4.5. The gains are concentrated on the more ambiguous labels—particularly sarcasm and overall sentiment. This suggests that task-specific supervision effectively calibrates decision boundaries and improves alignment with the dataset's label definitions. But metaphor faithfulness does not exhibit the same upward trajectory. Table 5 shows substantial drops in MetAvg for the fine-tuned models for example Qwen3-VL-8B: 22.38 to 11.57. The corresponding LAVE scores in Table 4 remain low post-adaptation (Qwen3-VL-8B LAVE=11.82; Gemma-3-12B LAVE=7.20). Under fine-tuning, models may learn label-predictive shortcuts that improve classification accuracy without strengthen-

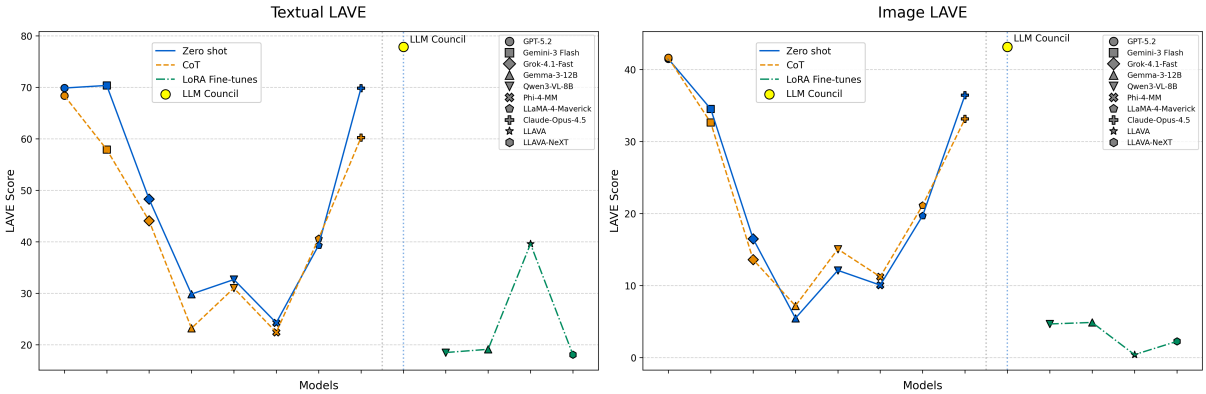


Figure 5: **Textual vs. visual metaphor faithfulness across strategies.** Textual and image LAVE scores for each model under zero-shot prompting, CoT prompting, LoRA fine-tuning, and the LLM Council. Visual metaphor understanding is consistently weaker than textual metaphor understanding, and LoRA fine-tuning often reduces both.

Models	Textual LAVE	Image LAVE	MetAvg
Zero-Shot Prompting			
GPT-5.2	69.86	41.46	**55.66**
Gemini-3 Flash	70.36	34.52	52.44
Claude-Opus-4.5	69.83	36.43	53.13
Grok-4.1-Fast	48.27	16.50	32.39
Gemma-3-12B-IT	29.83	5.47	17.65
Qwen3-VL-8B	32.65	12.11	22.38
Phi-4-MM	24.25	10.06	17.16
LLaMA-4-Maverick	39.30	19.69	29.50
Chain-of-Thought Prompting			
GPT-5.2	68.39	41.61	**55.00**
Gemini-3 Flash	57.91	32.62	45.27
Claude-Opus-4.5	60.23	33.13	46.68
Grok-4.1-Fast	44.07	13.61	28.84
Gemma-3-12B-IT	23.19	7.19	15.19
Qwen3-VL-8B	30.98	15.06	23.02
Phi-4-MM	22.40	11.22	16.81
LLaMA-4-Maverick	40.60	21.14	30.87
LLM Council			
LLM Council	**77.87**	**43.14**	**60.49**
LoRA Fine-Tuning			
Qwen3-VL-8B	18.47	4.67	11.57
Gemma-3-12B	19.10	4.88	11.99
LLAVA	39.58	0.41	**19.99**
LLAVA Next	18.06	2.25	10.16

Table 5: Metaphor understanding performance on BanglaMemeX measured using textual and image LAVE.

ing cross-modal metaphor grounding.

Why visual metaphors remain hard Table 5 and Figure 5 consistently reveal a performance gap between textual and visual metaphor comprehension across all strategies. For instance, in Zero-shot GPT-5.2 attains a Textual LAVE of 69.86 compared to an Image LAVE of 41.46, and Gemini-3 Flash reaches 70.36 versus 34.52. This discrepancy underscores a central challenge in culturally grounded meme understanding: textual metaphors usually contain explicit linguistic signals that models can align with known rhetorical structures, while visual metaphors demand interpreting symbolic objects, situations, and culturally specific cues to infer the intended meaning. The gap is especially severe for open-source models in the zero-shot configuration implying that weak multilingual visual grounding and difficulty reading stylized Bangla text can impede visual metaphor detection. Importantly, the LLM Council helps narrow this gap by boosting both textual and image LAVE (Textual=77.87, Image=43.14), suggesting that iterative critique and aggregation can restore missing cross-modal connections. By contrast, LoRA fine-tuning raises label accuracy but markedly degrades metaphor faithfulness often lowering both textual and image LAVE. Establishing the view that optimizing solely for classification metrics does not ensure faithful, metaphor-aware explanations.

6 Conclusion

The study introduced a culturally grounded Bangla multimodal metaphoric image benchmark that jointly evaluates multi-label meme interpretation and explanation faithfulness under metaphor-rich interactions. Across prompting, LLM Council reasoning, and LoRA adaptation, our results reveal a consistent decoupling between classification accuracy and metaphor-grounded explanations: strong label prediction does not guarantee faithful reasoning. While closed-source models provide strong zero-shot baselines and LoRA boosts open-source classification performance, explanation quality remains fragile for visual metaphors, highlighting the need for supervision and evaluation that directly target cross-modal metaphor understanding.

7 Limitations

This study has several limitations. BanglaMemeX is limited to Bangla memes collected from public online social media platforms and may not capture the full range of dialectal, regional, or community-specific humor and metaphor. Although the dataset includes explanation annotations, interpretation of memes remains inherently subjective despite multi-annotator agreement procedures. In addition, explanation evaluation relies partly on automated metrics and LLM-based judgment, which are informative but imperfect proxies for human assessment. Finally, the experimental study is limited to a selected set of models and one dataset, so the findings should not be taken as evidence of generalization to all multimodal or cross-cultural settings.

8 Ethical Considerations and potential risks

We put attention to content sensitivity, privacy, and responsible research use. During dataset construction, memes containing explicit sexual content were excluded. Because the data were collected from publicly accessible online memes, the benchmark may still contain socially sensitive, offensive, or polarizing expressions that reflect the nature of real world meme discourse. No personally identifiable information is included in the released dataset, and no personal information was collected from annotators beyond what was necessary to conduct the annotation process. To reduce privacy risk, the dataset excludes user-level metadata and is intended only as a research resource for studying culturally grounded multimodal understanding rather than for profiling individuals or communities. We further acknowledge that labels such as humor, sarcasm, offensiveness, sentiment, and metaphor-based explanation involve subjective judgment. Any public release of annotations, splits, and code will be carried out only where permitted, with appropriate attribution to original sources and with attention to platform and copyright constraints. We recommend that the benchmark be used for research and evaluation purposes only, and not as the sole basis for automated moderation or other high-stakes decisions without human oversight.

8.1 Use of AI Tools

We employed AI-assisted tools in a limited and controlled manner to support specific stages of the dataset construction pipeline. First, the Segment Anything Model (SAM) (Kirillov et al., 2023) was used to generate region-level segment proposals for each meme image. These segments served only as visual aids to assist annotators in identifying candidate regions corresponding to visual metaphors. All final selections and interpretations of visual metaphors were performed manually by human annotators. No generative AI systems were used to create or modify meme content. All textual annotations, including communicative labels, textual metaphors, visual metaphor descriptions, and final explanations, were written exclusively by human annotators. Additionally, AI models were not used to automatically assign labels or generate explanations at any stage of dataset construction. AI-based evaluation metrics, specifically BERTScore (Zhang et al., 2020), were used post hoc to measure semantic similarity between annotator-written explanations. These metrics were used solely for quality analysis and did not influence the annotation process itself. We emphasize that all annotations in BanglaMemeX reflect human judgment and culturally grounded interpretation. AI tools were used only as auxiliary support for visualization and evaluation, ensuring that the dataset remains a human-curated benchmark.

References

Newaz Ben Alam, Akm Moshiur Rahman Mazumder, Mir Sazzat Hossain, Mysha Samiha, Md Alvi Noor Hossain, Md Fahim, Amin Ahsan Ali, Ashraful Islam, M Ashraful Amin, and Akmmahbubur Rahman. 2025. Cmban: Cartoon-driven meme contextual classification dataset for bangla. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2178–2194, Mumbai, India. Association for Computational Linguistics.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736.

Susmoy Biswas, Mostafizur Rahman Zahid, Mst. Taposi Rabeya, Minhazul Abedin, and Md. Hasan Imam Bijoy. 2025a. Banglasarc3: A benchmark dataset for bangla sarcasm detection from social media to advance bangla nlp. *Data in Brief*.

Susmoy Biswas, Mostafizur Rahman Zahid, Mst. Taposi Rabeya, Minhazul Abedin, and Md. Hasan Imam Bijoy. 2025b. Banglasarc3: A benchmark dataset

for bangla sarcasm detection from social media to advance bangla nlp. *Data in Brief*.

Mithun Das and Animesh Mukherjee. 2023. Banglaabusememe: A dataset for bengali abusive meme classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15498–15512, Singapore. Association for Computational Linguistics.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022a. Memosen: A multimodal dataset for sentiment analysis of memes. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1542–1554, Marseille, France. European Language Resources Association.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022b. Mute: A multimodal dataset for detecting hateful memes. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 32–39, Online. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Douwe Kiela, Zeerak Waseem, Akiyoshi Handa, Lyle Ungar, and 1 others. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment anything. *Preprint*, arXiv:2304.02643.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, pages 12888–12900. PMLR.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. 2024. Improving automatic vqa evaluation using large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4170–4178. Also available as arXiv:2310.02567.

Md Ayon Mia and Md Fahim. 2025. Banhateme : Understanding hate in bangla memes thorough detection, categorization, and target profiling. In *Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025)*, pages 180–195, Mumbai, India. Association for Computational Linguistics.

Md Ayon Mia, Akm Moshiur Rahman Mazumder, Khadiza Sultana Sayma, Md Fahim, Md Tahmid Hasan Fuad, Muhammad Ibrahim Khan, and Akmmahbubur Rahman. 2025. Banmime : Misogyny detection with metaphor explanation on bangla memes. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17813–17839, Suzhou, China. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR.

F. H. Raquib and 1 others. 2025. An up-to-date and fine-grained bangla hate speech dataset. In *Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025)*. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM.

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*, pages 2887–2899. ACM.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

A Dataset Annotation Guidelines

This section presents the annotation guidelines developed for the **BanglaMemeX** dataset. The guidelines were designed to support consistent,

reproducible, and culturally grounded annotation of Bangla memes, which often rely on implicit metaphor, sarcasm, visual exaggeration, and socio-political context. All annotations were conducted by native Bangla speakers familiar with local cultural practices, public discourse, and social media conventions. Each meme was annotated independently across five categories: *Humor*, *Sarcasm*, *Offensiveness*, *Motivation*, and *Overall Sentiment*. Annotators were instructed to assign **exactly one label per category** based on the dominant communicative intent of the meme.

A.1 Annotator Recruitment and Qualification

Three annotators were recruited for the annotation task, all of whom were undergraduate students from the Department of Computer Science and Engineering at the University of Dhaka, Bangladesh. The annotators were aged between 21 and 24 years and had strong academic backgrounds in computational studies, with demonstrated familiarity with Bangla internet culture, social media discourse, and contemporary socio-political references commonly found in meme content. All three annotators were native Bangla speakers residing in Bangladesh. The selection criteria additionally required familiarity with common meme formats, culturally situated humor, and public discourse conventions in Bangla-language social media.

Compensation. Annotators were compensated at a rate of BDT 5 per meme instance annotated. This rate was determined in consultation with the annotators and was considered reasonable in the context of student research assistantship norms at the university.

Risk Disclosure. Prior to the annotation task, all annotators were informed that the dataset contained memes spanning a wide range of content types, including instances that may involve offensive language, hateful rhetoric, politically sensitive material, and culturally provocative imagery. Annotators were advised that they could skip or take breaks from any instance they found distressing and were encouraged to report any concerns to the supervising researchers at any point during the process.

Consent and Data Usage. All annotators were informed that their annotations would be used to construct a publicly released research dataset and would be reported in an academic publication. Informed consent was obtained from each annotator prior to the start of the annotation process, and annotators were made aware that no personally identifiable information would be collected or associated with their annotations in the released dataset.

Ethics Statement. The annotation task involved labeling publicly available internet memes and did not involve the collection of personal data, sensitive information, or interaction with vulnerable populations. The study was reviewed and determined to be exempt from full ethical review by the institutional authority, as it fell within the scope of minimal-risk research involving publicly available data and voluntary participation by informed adult participants.

Demographic Summary. All three annotators were Bangladeshi nationals, native speakers of Bangla, aged 21–24, and enrolled as undergraduate students in computer science at the time of annotation. The dataset does not contain any protected personal information as defined under GDPR or equivalent frameworks. No self-reported demographic attributes beyond native language proficiency and academic background were collected from the annotators.

Calibration. Prior to the main annotation phase, all annotators underwent a calibration session in which they independently labeled a shared subset of 50 memes. Disagreements from this pilot round were discussed collectively to align interpretation standards and resolve ambiguities in the annotation schema. Only after achieving satisfactory agreement on the calibration set did the annotators proceed to the full dataset.

A.2 Annotation Timeline and Session Protocol

The full annotation of 3,000 memes was completed over a period of three months. Annotation was carried out in structured daily sessions to ensure sustained quality and prevent cognitive degradation over time.

Session Structure. Each annotator worked in daily sessions lasting approximately 2–3 hours. Within each session, annotation was organized into focused blocks of 45–50 minutes of continuous labeling, followed by a mandatory break of 10–15 minutes. During break periods, annotators were in-

structed to disengage entirely from the annotation interface and the meme content. Annotators were encouraged to step away from their workstations, engage in unrelated activities, or rest during these intervals. No annotation was permitted during designated break windows.

Daily Limits. To prevent fatigue-induced degradation in annotation quality, a soft daily cap of approximately 40–50 memes per annotator per session was maintained. This rate was empirically determined during the calibration phase, where annotators reported that sustained annotation beyond this threshold led to diminished attentiveness and increased difficulty in distinguishing nuanced label boundaries, particularly for sarcasm and offensiveness categories.

Cognitive Well-Being and Content Exposure. A substantial proportion of the memes in BanglaMemeX contain hateful, misogynistic, communally provocative, or otherwise psychologically taxing content. Prolonged and continuous exposure to such material poses a recognized risk of cognitive fatigue, desensitization, and emotional distress. To mitigate these effects, the following well-being measures were enforced throughout the annotation period:

- **Mandatory Breaks:** As described above, annotators were required to take a 10–15 minute break after every 45–50 minutes of annotation. These breaks were non-negotiable and were enforced by the supervising researchers.
- **Content Flagging and Skipping:** Annotators were permitted to flag any individual meme as personally distressing and temporarily skip it. Flagged instances were revisited at a later session or reassigned to a different annotator if the original annotator preferred not to return to it.
- **Extended Cool-Down Periods:** On days when annotators reported encountering a high concentration of particularly offensive or hateful content, an additional extended break of 20–30 minutes was provided before resuming annotation.
- **Open Communication:** Annotators were encouraged to communicate any concerns, fatigue, or emotional discomfort to the supervising researchers at any point. No annotator was penalized for taking additional unscheduled breaks or for reducing their daily annotation volume on particularly difficult days.
- **Weekly Check-Ins:** The supervising researchers conducted brief weekly check-ins with each annotator to assess their well-being, discuss any recurring difficulties with specific content types, and adjust workload distribution if necessary.

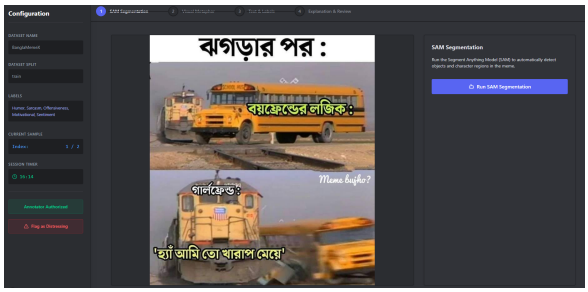
These measures were implemented to ensure that the annotation process remained sustainable over the full three-month duration, that annotator mental freshness was preserved throughout, and that annotation quality did not degrade as a result of cumulative exposure to harmful content.

A.3 SAM-Assisted Visual Metaphor Annotation Pipeline

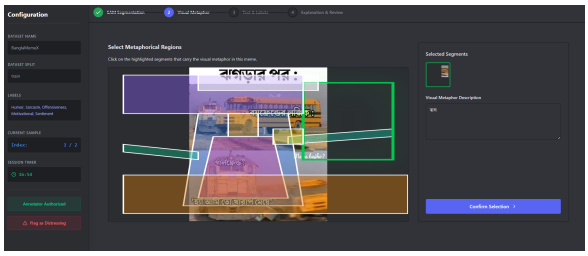
To support systematic and fine-grained identification of visual metaphors, we designed a custom annotation interface integrating the Segment Anything Model (SAM; (**?**)). The interface was motivated by the observation that visual metaphors in Bangla memes are typically carried by specific regions within a visually complex image, and that asking annotators to identify these regions without structural guidance introduces inconsistency in granularity and spatial precision across annotations.

The annotation interface is organized as a four-stage pipeline, presented to the annotator through a horizontal stepper indicating the current stage and completion status. A persistent left-hand configuration panel displays the dataset name, split, annotation label categories, current sample index, a running session timer, annotator authorization status, and a *Flag as Distressing* button that allows the annotator to skip any meme they find psychologically taxing. The four stages operate as follows and are illustrated in Figure 6:

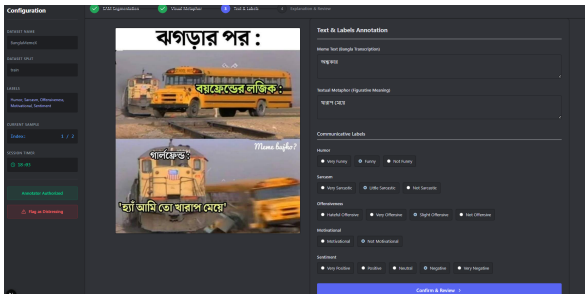
Stage 1: SAM Segmentation (Figure 6a). The original meme image is displayed in the central panel. The right-hand panel presents a description of the SAM segmentation process and a *Run SAM Segmentation* button. Upon clicking, the meme image is passed through SAM, which automatically detects and segments objects and character regions within the meme into semantically coherent visual regions.



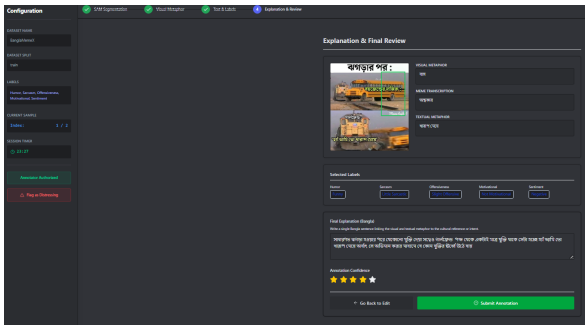
(a) Stage 1: SAM Segmentation



(b) Stage 2: Visual Metaphor Selection



(c) Stage 3: Text & Labels Annotation



(d) Stage 4: Explanation & Final Review

Figure 6: Interface of the web-based annotation tool developed for the BanglaMemeX dataset.

Stage 2: Visual Metaphor Selection (Figure 6b). After segmentation, the meme is re-displayed with all candidate segments highlighted as semi-transparent color-coded overlays with distinct borders. The annotator is instructed to click on the segment or segments that carry the visual metaphor. Selected segments are confirmed with a highlighted border and appear as thumbnails in the right-hand *Selected Segments* panel. Below the selected thumbnails, the annotator provides a short *Visual Metaphor Description* characterizing the symbolic role of the chosen region before clicking *Confirm Selection* to proceed.

Stage 3: Text and Labels Annotation (Figure 6c). The original meme is displayed in the left panel for reference. The right-hand panel presents the text and label annotation form. The annotator first manually transcribes the overlaid Bangla text in the *Meme Text (Bangla Transcription)* field, since external OCR tools were not employed. The annotator then provides the figurative meaning of the text in the *Textual Metaphor (Figurative Meaning)* field. Below these, the annotator assigns communicative labels via radio button selectors for each of the five categories: Humor (*Very Funny*, *Funny*, *Not Funny*), Sarcasm (*Very Sarcastic*, *Little Sarcastic*, *Not Sarcastic*), Offensiveness (*Hateful Offensive*, *Very Offensive*, *Slight Offensive*, *Not Offensive*), Motivational (*Motivational*, *Not Motivational*), and Sentiment (*Very Positive*, *Positive*, *Neutral*, *Negative*, *Very Negative*). The annotator then clicks *Confirm & Review* to proceed to the final stage.

Stage 4: Explanation and Final Review (Figure 6d). The final stage presents a consolidated review of all annotations. The top section displays the meme image with the selected visual metaphor segment highlighted, alongside summary fields showing the visual metaphor description, meme transcription, and textual metaphor. Below this, all assigned communicative labels are displayed as colored badges for quick verification. The annotator then writes a single Bangla sentence in the *Final Explanation (Bangla)* field, linking the visual and textual metaphor to the underlying cultural reference or intent. An *Annotation Confidence* rating on a 1–5 star scale is provided to indicate the annotator's certainty in the overall annotation quality. The annotator may click *Go Back to Edit* to revise any previous stage, or *Submit Annotation* to finalize and proceed to the next meme.

The session timer visible throughout the interface tracks cumulative annotation time per session and triggers a visual reminder after 45 minutes, prompting the annotator to take a mandatory break. The *Flag as Distressing* button remains accessible at every stage, allowing the annotator to skip any

meme at any point in the pipeline without penalty.

This SAM-assisted workflow standardizes the granularity at which visual metaphors are identified across annotators, reduces subjectivity in determining which part of a visually complex meme constitutes the metaphoric vehicle, and provides a spatially grounded reference that can be used for downstream evaluation of visual metaphor localization in multimodal models.

A.4 General Instructions

The annotation process followed the following core principles:

- **Multimodal Interpretation:** Annotators examined both visual and textual elements, considering their joint meaning rather than interpreting either modality in isolation.
- **Cultural Grounding:** Bangla-specific metaphors, idiomatic expressions, political references, and socially familiar exaggerations were prioritized over literal interpretation.
- **Intent-Oriented Labeling:** Labels were assigned based on the implied communicative intent of the meme creator, rather than the annotator's personal agreement, disagreement, or moral judgment.
- **Single-Label Assignment:** For each category, annotators selected exactly one label corresponding to the most dominant interpretation.
- **Consistency and Reproducibility:** Definitions were applied uniformly across samples to ensure annotation reliability and reduce subjective drift.
- **Ambiguity Resolution:** When a meme admitted multiple plausible readings, annotators selected the interpretation most strongly supported by common Bangla cultural understanding.

A.5 Category Definitions

Humor

This category captures the extent to which the meme is intended to evoke amusement, wit, or laughter.

Label	Class	κ
Humor	Very Funny	0.72
	Funny	0.76
	Not Funny	0.70
Sarcasm	Not Sarcastic	0.68
	Little Sarcastic	0.71
	Very Sarcastic	0.66
Offensive	Not Offensive	0.79
	Slight Offensive	0.74
	Very Offensive	0.73
	Hateful Offensive	0.81
Motivational	Not Motivational	0.69
	Motivational	0.72
Overall	Very Negative	0.63
	Negative	0.67
	Neutral	0.70
	Positive	0.64
	Very Positive	0.60

Table 6: Per-class Fleiss' κ agreement scores across three annotators.

- **Very Funny**
  Definition: The meme strongly evokes laughter through sharp exaggeration, absurd contrast, or an effective punchline.
  Indicators:
  - Strong comedic exaggeration
  - Clear absurdity or unexpected contrast
  - High entertainment value
- **Funny**
  Definition: The meme is mildly amusing or witty but does not produce a strong comedic effect.
  Indicators:
  - Light playfulness or irony
  - Familiar humorous comparison
  - Moderate amusement
- **Not Funny**
  Definition: The meme is serious, critical, informational, or otherwise not intended to be humorous.
  Indicators:
  - Serious or advocacy-oriented tone
  - Social or political messaging
  - No clear attempt at amusement

Figure 7(a) is labeled *Not Funny* because it depicts a serious socio-political metaphor where different student ideological groups are shown eventually entering the structure of a political student wing, emphasizing political assimilation rather than humor.

Figure 7(b) is labeled *Funny* because it humorously exaggerates the common excuse of poor Wi-Fi connectivity by metaphorically showing an astronaut leaving Earth to search for a better signal.

(a) Not Funny

(b) Funny

(c) Very Funny

Figure 7: Examples of humor annotations in BanglaMemeX across the three humor levels.

Figure 7(c) is labeled *Very Funny* because it uses an exaggerated visual metaphor of numerous figures descending from the sky to humorously represent mosquitoes attacking someone who tried to sleep without using a mosquito net.

Sarcasm

This category reflects the presence and strength of irony, ridicule, or indirect mockery.

- **Very Sarcastic**
  Definition: The meme contains strong, explicit sarcasm or overtly exaggerated mockery directed at a person, institution, or idea.
  Indicators:
  - Clear ridicule or scorn
  - Strong ironic reversal
  - Deliberate exaggeration for criticism
- **Little Sarcastic**
  Definition: The meme contains mild sarcasm or a subtle ironic undertone.
  Indicators:
  - Gentle teasing
  - Indirect irony
  - Mild subversion of literal meaning
- **Not Sarcastic**
  Definition: The meme communicates its message directly, without meaningful irony or sarcastic inversion.
  Indicators:
  - Literal expression
  - Straightforward criticism or observation
  - No ironic contrast

Figure 8(a) is labeled *Not Sarcastic* because it expresses a literal situation in which someone fasting keeps a packet of sweets nearby, portraying temptation directly without relying on irony or ridicule.

Figure 8(b) is labeled *Little Sarcastic* because the dramatic warrior imagery paired with patriotic slogans introduces a mild ironic exaggeration of nationalist enthusiasm, creating a playful but not strongly ridiculing tone.

Figure 8(c) is labeled *Very Sarcastic* because placing the label of a governmental advisory role related to women and children on a Hollywood celebrity image creates a strong ironic contrast that explicitly mocks the plausibility of such an appointment.

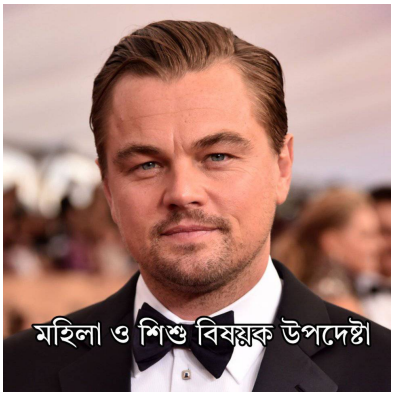
Offensiveness

This category assesses whether the meme contains insulting, degrading, harmful, or discriminatory content.

- **Hateful Offensive**
  Definition: The meme contains explicit hate, dehumanization, or discriminatory targeting based on identity, community, religion, nationality, ethnicity, or similar group membership.
  Indicators:
  - Identity-based hostility
  - Dehumanizing comparison
  - Group-directed hatred or exclusion
- **Very Offensive**
  Definition: The meme includes strong insults, degrading comparisons, or aggressive attacks toward individuals or groups without necessarily constituting hate speech.
  Indicators:

(a) Not Sarcastic

(b) Little Sarcastic



(c) Very Sarcastic

Figure 8: Examples of sarcasm annotations in BanglaMemeX across different levels of ironic intensity.

  - Direct humiliation
  - Harsh ridicule
  - Strong degrading metaphor

- **Slight Offensive**
  Definition: The meme contains mild insult, teasing, or edgy humor that may be disrespectful but is not strongly hostile.
  Indicators:
  - Light bullying
  - Suggestive mockery
  - Mildly degrading comparison

- **Not Offensive**
  Definition: The meme does not contain insulting, harmful, or degrading intent.
  Indicators:
  - Harmless observation
  - Neutral comparison
  - No hostile or demeaning target

Figure 9(a) is labeled *Hateful Offensive* because it targets a specific individual and frames the act of fasting alongside the temptation of sweets in a way that ridicules personal religious discipline through mocking insinuation.

Figure 9(b) is labeled *Very Offensive* because it depicts a threatening and violent scenario where a symbolic warrior figure confronts a nationalist slogan, using aggressive imagery and confrontation to demean the opposing viewpoint.

Figure 9(c) is labeled *Slight Offensive* because it employs dark humor by juxtaposing a historical ritual scene with a cooking metaphor, creating a sarcastic and mildly disrespectful comparison rather than direct hostility.

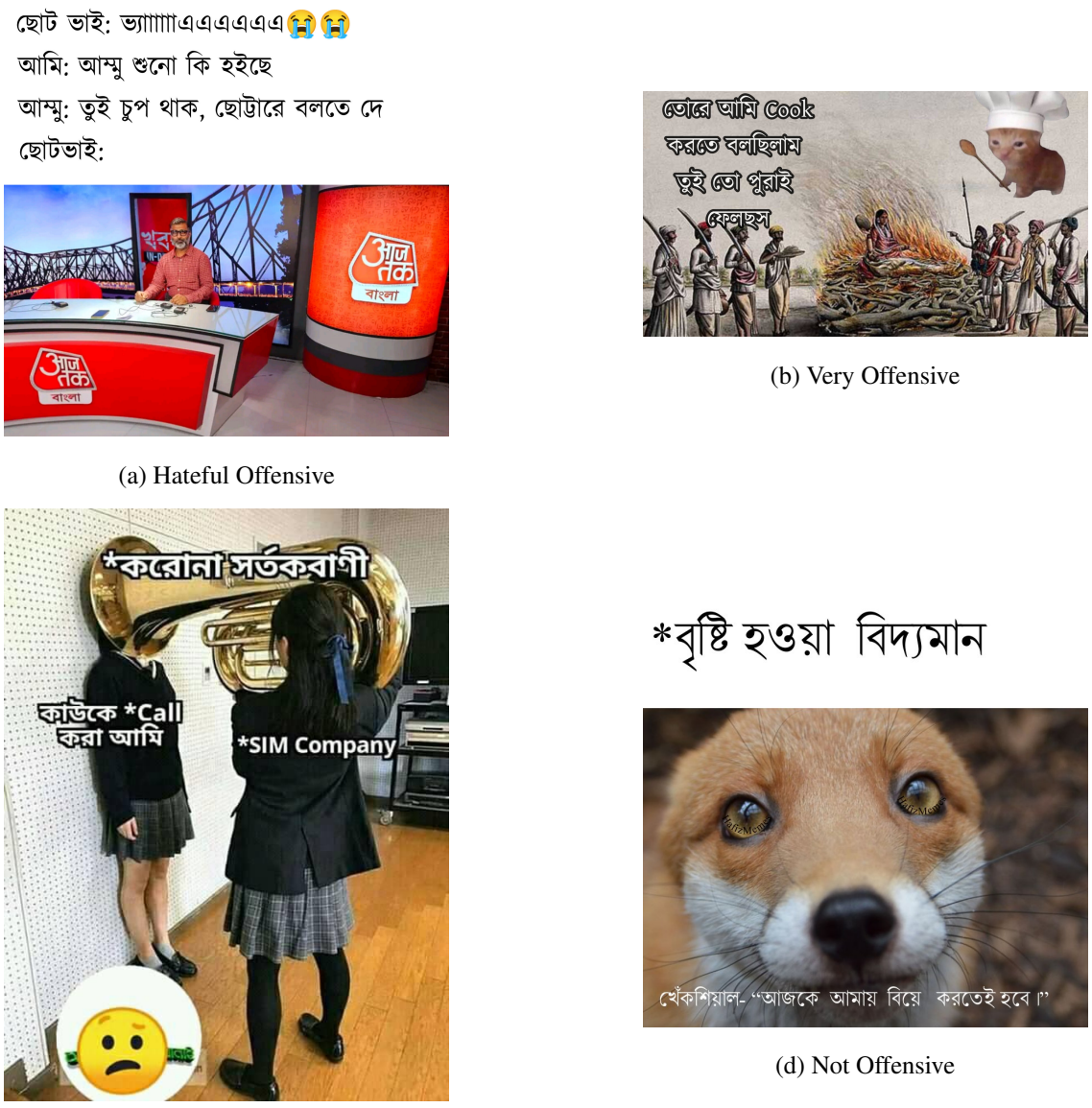
Figure 9(d) is labeled *Not Offensive* because it humorously illustrates a relatable communication problem during the pandemic using exaggerated visual comparison between a person making a call and a telecom company, without targeting or insulting any individual or group.

Motivational Intent

This category captures whether the meme encourages hope, effort, resilience, self-improvement, or positive action.

- **Motivational**
  Definition: The meme encourages constructive thinking, aspiration, resilience, or positive behavioral orientation.
  Indicators:
  - Encouragement toward improvement
  - Positive aspiration
  - Hopeful or uplifting framing

- **Not Motivational**
  Definition: The meme does not promote inspiration or constructive action and is instead neutral, critical, mocking, or merely descriptive.
  Indicators:
  - Critical or sarcastic message
  - No inspirational direction
  - Observational or discouraging tone



(a) Hateful Offensive

(b) Very Offensive

(c) Slight Offensive

(d) Not Offensive

Figure 9: Examples of offensiveness annotations in BanglaMemeX across four offensive intensity levels.

(a) Not Motivational

(b) Motivational

Figure 10: Examples of motivational intent annotations in BanglaMemeX.

Figure 10(a) is labeled *Not Motivational* because the meme uses humorous wordplay comparing a guest bird visiting the Jahangirnagar campus with the person themselves becoming a "bird," conveying playful self-reference rather than encouraging positive action or aspiration.

Figure 10(b) is labeled *Motivational* because it metaphorically depicts daily prayer as a protective force holding back life's hardships and emotional struggles, framing religious devotion as a source of strength and resilience.

Overall Sentiment

This category reflects the dominant emotional tone conveyed by the meme.

- **Very Positive**
  Definition: Strongly uplifting, reassuring, joyful, or emotionally supportive content.
- **Positive**
  Definition: Generally kind, hopeful, encouraging, or pleasant in tone.
- **Neutral**
  Definition: Emotionally balanced or observational content without strong positive or negative polarity.
- **Negative**
  Definition: Critical, disapproving, mocking, or emotionally discouraging content.
- **Very Negative**
  Definition: Strongly hostile, deeply critical, harshly condemning, or emotionally severe content.

Figure 11(a) is labeled *Very Positive* because it humorously portrays admiration and excitement about seeing guest birds at Jahangirnagar University, framing the speaker's transformation into a "bird" as an expression of joyful enthusiasm.

Figure 11(b) is labeled *Positive* because it encourages collective participation in prayer during iftar time, presenting religious practice as a constructive and socially supportive activity.

Figure 11(c) is labeled *Neutral* because the poetic metaphor of planting flowers in a beloved's chest evokes emotional imagery without clearly conveying optimism or criticism.

Figure 11(d) is labeled *Negative* because it expresses frustration about recurring problems by comparing boiling milk that repeatedly spills over with the effort of continuously trying to control a persistent situation.

Figure 11(e) is labeled *Very Negative* because it depicts a strong and hostile condemnation of a public figure by associating violent imagery with political commentary, conveying intense anger and moral outrage.

A.6 Inter-Annotator Agreement

To assess the consistency and reliability of the categorical label annotations, we computed Fleiss' Kappa (κ) across all three annotators for each of the five annotation categories. Fleiss' Kappa was chosen as the agreement metric because it generalizes Cohen's Kappa to settings involving more than two raters and accounts for agreement occurring by chance. The resulting Fleiss' Kappa scores for each annotation category are reported in Table 6 in the main paper.

The final label for each meme in every category was determined through majority voting among the three annotators. In cases where all three annotators assigned different labels, the instance was flagged for a subsequent discussion round, and a consensus label was agreed upon collectively.

A.7 Explanation Annotation and Evaluation

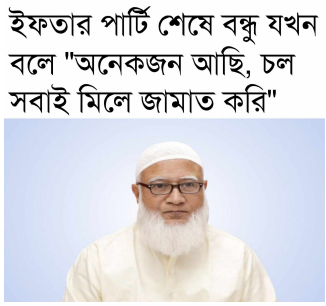
In addition to categorical labels, each meme was paired with a short human-written explanation in Bangla. The purpose of this annotation was to capture the underlying metaphorical meaning of the meme and provide a reference for evaluating explanation generation quality.

Annotators were instructed to follow these principles:

- **Single-Sentence Format:** Each explanation must be written as one complete Bangla sentence.
- **Metaphor-Focused Interpretation:** The explanation should identify the implicit metaphor, symbolic comparison, or intended meaning of the meme.
- **Contextual Grounding:** Where relevant, the explanation should incorporate cultural, social, political, or religious context necessary for interpretation.
- **Non-Literal Description:** Literal description of image content should be avoided unless

(a) Very Positive

(b) Positive



(c) Neutral

(d) Negative

(e) Very Negative

Figure 11: Examples of overall sentiment annotations in BanglaMemeX.

it directly supports the explanation of the metaphor.

Explanation Quality Evaluation

Unlike categorical labels, free-text explanations cannot be evaluated using standard inter-annotator agreement metrics such as Fleiss' Kappa. To assess the quality and acceptability of the annotated explanations, we adopted a confidence-based evaluation protocol. Each of the three annotators independently reviewed every explanation and assigned a confidence score on a scale of 1 to 5, where 1 indicated that the explanation was inadequate or failed to capture the meme's intended meaning, and 5 indicated that it was fully accurate, culturally grounded, and faithfully represented the metaphorical content.

The final acceptance decision for each explanation was determined through majority voting: an explanation was accepted if at least two of the three annotators assigned a confidence score of 3 or above. Explanations that did not meet this threshold were revised through a collaborative discussion among the annotators until a satisfactory version was agreed upon. This process ensured that all retained explanations met a minimum standard of interpretive quality while remaining grounded in shared cultural understanding.

These explanation annotations serve as ground-truth references for evaluating explanation faithfulness, metaphor understanding, and cultural alignment in multimodal models.

Overall, the annotation framework of BanglaMemeX was designed to capture not only surface-level semantic labels, but also the deeper metaphorical and culturally situated meanings necessary for robust Bangla meme understanding.

B Prompting

All the prompts are detailed in this section.

Figure 12: Example of human-written explanation annotation for a Bangla meme instance.

Zero shot prompt

You are an expert assistant who understands Bangla memes, their metaphors, and social context. Your job is to read the image (and text, if any), for each category, choose EXACTLY ONE option from the allowed list, then write a short explanation in Bangla (<= 50 words) describing the hidden metaphor and intent. Do NOT invent new labels. Do NOT translate the labels. Labels MUST be exactly one of the options given. The explanation MUST be in Bangla single sentence, but the labels MUST stay in English.

Figure 13: ZeroShot Prompt.

Chain-of-thought prompt (Part-2)

You are an expert assistant who understands Bangla memes, their metaphors, and social context. Your job is to read the image (and text, if any). Think step-by-step about the meme's meaning, context, and metaphor. For each category, choose EXACTLY ONE option from the allowed list. Then write a short explanation in Bangla (<= 50 words) describing the hidden metaphor and intent.Follow these reasoning steps.
STEP 1 - Analyze visual elements: What do you see in the image? What objects, people, or scenes are shown?
STEP 2 - Analyze textual elements: What does the text say? What is its literal meaning?
STEP 3 - Integrate visual and text: How do the image and text work together? What is the combined meaning?
STEP 4 - Identify the metaphor: What is being compared or symbolized? What hidden message is conveyed?
STEP 5 - Assess humor: Is it funny? Why or why not? Consider absurdity, wit, or cleverness.
STEP 6 - Assess sarcasm: Is there irony or mockery? How strong is it?
STEP 7 - Assess offensiveness: Could this offend anyone? Consider cultural, religious, or personal sensitivity.
STEP 8 - Assess motivation: Does this inspire or encourage positive action?
STEP 9 - Determine overall sentiment: Considering all factors, what is the overall emotional impact?

Figure 14: Chain-of-Thought Prompt.

LAVE Evaluation prompt

You are an expert evaluator of Bangla memes. Your task is to judge the QUALITY of the model's explanation by comparing it to the human reference. LAVE scoring focuses ONLY on the explanation correctness, coherence, and alignment with the reference. Classification labels (Humor, Sarcastic, etc.) are provided ONLY as context and should NOT determine the score. EVALUATION CRITERIA (for explanation):
Score = 1 (Good Explanation) if:
- The explanation captures the same core meaning as the reference
- Minor wording differences or phrasing variations are acceptable
- It is relevant, coherent, and consistent with the meme's meaning
- No hallucinations or contradictions

Score = 0 (Poor Explanation) if:
- The explanation misses the main meaning
- It is vague, incorrect, incomplete, or contradictory
- It introduces fabricated or irrelevant details
- The explanation does not align with the context of the meme

Labels DO NOT need to match exactly. They are NOT part of evaluation.
Now evaluate the model explanation:

Figure 15: LAVE score evaluation Prompt.

Explainer prompt

You are the EXPLAINER in an LLM Council analyzing Bangla memes and an expert who understands Bangla memes, their metaphors, and social context. YOUR role is to break down the metaphor in the meme - identify literal vs figurative meaning, cultural references, and hidden messages.
MEME TEXT (if available): {text_content}
{classification_guidelines}
Analyze the meme and provide:
1. LITERAL MEANING: What is shown/written directly
2. FIGURATIVE MEANING: The hidden metaphor or message
3. CULTURAL CONTEXT: Any Bangla/South Asian cultural references
Do NOT invent new labels. Labels MUST be exactly one of the options given. Be thorough but concise. Your analysis will be reviewed by a Critic.

Figure 16: System prompt for the Explainer agent in the LLM Council framework.

Critic prompt

You are the CRITIC in an LLM Council analyzing Bangla memes. You are an expert who understands Bangla memes, their metaphors, and social context. Your role is to review the Explainer's analysis and challenge it for accuracy, depth, and potential blind spots.
MEME TEXT (if available): {text_content}
{classification_guidelines}
EXPLAINER'S ANALYSIS:
{explainer_output}
Your task:
1. STRENGTHS: What did the Explainer get right?
2. WEAKNESSES: What was missed or incorrectly interpreted?
3. ALTERNATIVE INTERPRETATION: If you disagree, provide your view
4. REVISED CLASSIFICATION (if needed)
Do NOT invent new labels. Labels MUST be exactly one of the options given. Be constructive but rigorous.

Figure 17: System prompt for the Critic agent in the LLM Council framework.

Synthesizer prompt

You are the SYNTHESIZER in an LLM Council analyzing Bangla memes. You are an expert who understands Bangla memes, their metaphors, and social context. Your role is to compile the Explainer's and Critic's outputs into a final, cohesive analysis with definitive classifications.
MEME TEXT (if available): {text_content}
{classification_guidelines}
EXPLAINER'S ANALYSIS: {explainer_output}
CRITIC'S REVIEW: {critic_output}
YOUR TASK:
1) Consider all perspectives and resolve any disagreements with reasoned judgment.
2) For each category, choose EXACTLY ONE option from the allowed list.
3) Write a short explanation in Bangla (<= 50 words) describing the hidden metaphor and intent.
Do NOT invent new labels. Do NOT translate the labels. Labels MUST be exactly one of the options given. The explanation MUST be in Bangla single sentence, but the labels MUST stay in English.

Figure 18: System prompt for the Synthesizer agent in the LLM Council framework.