ZeroR@CHiPSAL 2026: Two-Stage Vision-Language Adaptation with Contrastive Learning for Nepali Meme Classification

Nitiz Khanal
Pulchowk Campus, Institute of Engineering, Tribhuvan University
Lalitpur, Nepal
khanalnitij20@gmail.com

Abstract

This paper presents our system for the CHiPSAL 2026 shared task on multimodal hate speech and sentiment detection in Nepali memes. We address both subtasks: binary hate speech classification and three-class sentiment analysis. Our approach adapts the Robust Adaptation of Hateful Meme Detection (RA-HMD) framework using Qwen3-VL-8B-Instruct, a state-of-the-art vision-language model with native Devanagari support. We employ a two-stage training pipeline: (1) LoRA fine-tuning with an MLP projection head for generative classification, and (2) contrastive backbone fine-tuning with supervised InfoNCE loss. We handle class imbalance through minority oversampling, image augmentation, and focal loss. At inference, we ensemble Stage 1 token probabilities with Stage 2 classifier scores using validation-tuned weights. Our end-to-end approach eliminates error propagation from separate OCR and translation pipelines by leveraging the model's native Devanagari understanding. Our system achieved **2nd place** on hate speech detection (F1: 0.797) and **4th place** on sentiment analysis (F1: 0.518). We provide detailed ablations, error analysis, and insights into adapting large vision-language models for low-resource South Asian languages.

Keywords: hate speech detection, sentiment analysis, vision-language models, Nepali, multimodal, contrastive learning, low-resource languages

1. Introduction

The proliferation of memes on social media platforms has created new challenges for content moderation systems. Unlike traditional text or image classification, memes require understanding the complex interplay between visual and textual elements, where meaning often emerges from their combination rather than either modality alone (Kiela et al., 2020). A seemingly innocuous image paired with specific text can convey hate, while aggressive text on a humorous image might be entirely benign. This multimodal reasoning challenge is further compounded for low-resource languages, where annotated datasets are scarce and pre-trained models have limited exposure to the language and its cultural context (Parihar et al., 2021).

Nepali, spoken by approximately 32 million people, presents unique challenges for automated content moderation. The language uses Devanagari script, which is embedded directly into meme images rather than appearing as separate text. Cultural references, humor styles, and implicit meanings rooted in Nepali society add layers of complexity that general-purpose models may not capture. Furthermore, the code-mixing of Nepali with English, common in social media, creates additional processing challenges.

The Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2026), co-located with LREC-COLING 2026, addresses these challenges through shared tasks focused on South Asian language processing (Sarveswaran et al., 2026). We participated in the multimodal hate and sentiment detection task (Thapa et al., 2026), which provides carefully curated Nepali meme datasets collected from social media platforms (Thapa et al., 2025b,c; Bhandari et al., 2023). The shared task comprises two subtasks:

- **Subtask A – Hate Speech Detection**: Binary classification of memes as hateful (content targeting groups based on ethnicity, religion, gender, or caste, or promoting violence/discrimination) or safe.
- **Subtask B – Sentiment Analysis**: Three-class classification of memes into negative, neutral, or positive sentiment categories.

Our approach leverages recent advances in vision-language models (VLMs), specifically adapting the Robust Adaptation of Large Multimodal Models (RA-HMD) framework (Mei et al., 2025) for low-resource Nepali content. Rather than constructing complex pipelines involving OCR extraction, machine translation, and separate text/image encoders, we leverage Qwen3-VL-8B-Instruct (Wang et al., 2024), a state-of-the-art VLM with native support for multiple scripts including Devanagari. This end-to-end approach eliminates error propagation that would occur in pipeline-based systems.

Our key contributions are:

1. A two-stage training framework combining generative fine-tuning with contrastive learning, adapted from RA-HMD for low-resource multimodal classification.

Proceedings of the Second workshop on Challenges in Processing South Asian Languages (CHiPSAL2026) @ LREC, pages 275–283

2. Task-specific configurations addressing the distinct challenges of binary hate speech detection versus three-class sentiment analysis.

Our system achieved **2nd place** on hate speech detection (F1: 0.797) and **4th place** on sentiment analysis (F1: 0.518), demonstrating the effectiveness of adapting modern VLMs for low-resource multimodal tasks.

2. Related Work

2.1. Multimodal Hate Speech Detection

The challenge of detecting hate speech in multimodal content gained significant attention with the Facebook Hateful Memes Challenge (Kiela et al., 2020), which demonstrated that simple fusion of unimodal representations is insufficient. The challenge showed that even state-of-the-art vision and language models struggled when hate was conveyed through subtle interactions between image and text.

Subsequent work has explored various fusion strategies, from early fusion approaches that concatenate features to more sophisticated cross-modal attention mechanisms. Pramanick et al. (2021) introduced methods for detecting harmful memes and identifying their targets, while Bhandari et al. (2023) extended multimodal hate speech analysis to crisis contexts with the CrisisHateMM dataset, providing annotation schemas for directed and undirected hate.

More recently, Mei et al. (2024) proposed retrieval-guided contrastive learning for hateful meme detection, using similar training samples to guide representation learning. This work was extended by Mei et al. (2025) with the RA-HMD framework, which combines LoRA fine-tuning with rationale-aware contrastive learning, achieving state-of-the-art results on multiple English hateful meme datasets including HatefulMemes, MAMI, and PrideMM. Our work adapts this framework for low-resource Nepali content.

2.2. Vision-Language Models

The emergence of large vision-language models has transformed multimodal understanding. Early approaches like CLIP (Radford et al., 2021) learned joint image-text representations through contrastive learning on web-scale data. More recent instruction-tuned models like LLaVA (Liu et al., 2024) and the Qwen-VL family (Wang et al., 2024) can follow complex multimodal instructions and generate detailed responses.

Qwen3-VL-8B-Instruct, which we use as our backbone, represents a significant advancement in multimodal understanding. The model supports dynamic image resolution, multiple languages including those using non-Latin scripts, and demonstrates strong performance across diverse vision-language benchmarks. Critically for our application, it has native support for Devanagari script, enabling end-to-end processing of Nepali text embedded in images without requiring separate OCR.

2.3. Parameter-Efficient Fine-Tuning

Fine-tuning large models on downstream tasks is computationally expensive and can lead to catastrophic forgetting. Low-Rank Adaptation (LoRA) (Hu et al., 2022) addresses this by freezing pre-trained weights and injecting trainable rank decomposition matrices into transformer layers. This approach reduces trainable parameters by orders of magnitude while maintaining competitive performance.

For vision-language models, LoRA has proven particularly effective, allowing adaptation to new tasks and domains without the computational cost of full fine-tuning. We apply LoRA to all linear projections in Qwen3-VL, enabling efficient adaptation to Nepali meme classification.

2.4. Contrastive Learning for Classification

Supervised contrastive learning (Khosla et al., 2020) extends the self-supervised contrastive paradigm by leveraging label information. The objective pulls together representations of samples from the same class while pushing apart representations from different classes. The InfoNCE loss (van den Oord et al., 2018) provides a principled framework for this objective.

For content moderation tasks, contrastive learning helps learn discriminative representations that capture subtle differences between classes. Mei et al. (2024) and Mei et al. (2025) demonstrated its effectiveness for hateful meme detection, and we adapt this approach for Nepali content.

2.5. Low-Resource Language Processing

Processing low-resource languages presents unique challenges due to limited training data, fewer pre-trained resources, and distinct linguistic characteristics. Large language models have shown promise for such languages through cross-lingual transfer (Thapa et al., 2025a), but multimodal tasks in low-resource settings remain underexplored.

The CHiPSAL workshop series (Sarveswaran et al., 2026) aims to address these gaps for South Asian languages. Prior work on Nepali NLP has focused primarily on text classification and named entity recognition, with multimodal analysis receiv-

ing limited attention. Our work contributes to this emerging area by demonstrating effective adaptation of state-of-the-art VLMs for Nepali multimodal content.

3. Dataset and Task Description

3.1. Dataset Overview

The shared task provides Nepali meme datasets collected from various social media platforms (Thapa et al., 2025b,c). The memes contain Devanagari text embedded in images, often with code-mixing between Nepali and English. The annotation follows established protocols for multimodal hate speech (Bhandari et al., 2023).

3.2. Subtask A: Hate Speech Detection

The hate speech dataset contains approximately 1,068 memes with binary labels:

- **Hate (1)**: Content that attacks, demeans, or incites hatred against individuals or groups based on protected characteristics including ethnicity, religion, gender, caste, nationality, or disability. This also includes content promoting violence or discrimination.
- **Safe (0)**: Content that does not contain hate speech, including neutral information, positive messages, and humor that does not target protected groups.

The dataset exhibits class imbalance, with safe memes being more frequent than hateful ones. This imbalance reflects the natural distribution of content on social media but poses challenges for classification.

3.3. Subtask B: Sentiment Analysis

The sentiment dataset contains memes labeled with three categories:

- **Negative (0)**: Content expressing criticism, mockery, sadness, anger, frustration, disappointment, or pessimism.
- **Neutral (1)**: Factual information, observations, or content without clear emotional valence.
- **Positive (2)**: Content expressing joy, humor, encouragement, celebration, love, hope, or optimism.

The neutral class dominates the distribution, creating a significant class imbalance challenge. Additionally, the boundary between categories can be subjective, particularly for humorous content that may express negativity in a playful manner.

3.4. Evaluation Metric

Both subtasks use macro F1-score as the primary evaluation metric:

$$\text{Macro F1} = \frac{1}{C}\sum_{c=1}^{C}\text{F1}_c \tag{1}$$

where C is the number of classes and F1_c is the F1-score for class c. This metric gives equal weight to each class regardless of frequency, making it particularly appropriate for imbalanced datasets.

3.5. Data Preparation

We create an 85/15 stratified train/validation split (random seed 42) to tune hyperparameters and ensemble weights. Stratification ensures that class proportions are preserved in both splits. All images are converted to RGB format for consistency. We do not use any external data or pre-training beyond the provided training set and the base Qwen3-VL model.

4. Methodology

Our system follows a two-stage training pipeline adapted from the RA-HMD framework (Mei et al., 2025). Figure 1 illustrates the overall architecture.

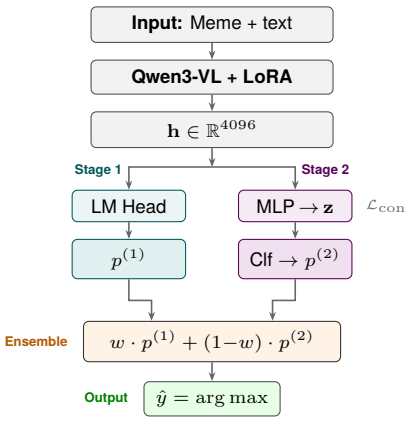

Figure 1: Architecture overview. Stage 1 uses the LM head for token probabilities. Stage 2 adds an MLP projection with contrastive loss ($\mathcal{L}_{\text{con}}$) and a classifier. Outputs are ensembled and the final prediction is made via argmax.

4.1. Base Model Selection

We select Qwen3-VL-8B-Instruct as our backbone for several reasons:

- **Native Devanagari support**: The model can read and understand text in Devanagari script directly from images, eliminating the need for separate OCR.

- **Strong multilingual capabilities**: Pre-training on diverse multilingual data enables transfer to low-resource languages like Nepali.
- **Dynamic resolution**: The model handles images of varying sizes (min: 256×28^2, max: 512×28^2 pixels) without fixed resizing.
- **Instruction following**: The model can follow complex task instructions, enabling prompt-based classification.

4.2. Stage 1: Generative Fine-Tuning with LoRA

In Stage 1, we fine-tune the model to generate class labels directly from meme images. This generative approach leverages the model's pre-trained language understanding capabilities.

4.2.1. Low-Rank Adaptation (LoRA)

Rather than fine-tuning all parameters, we apply LoRA (Hu et al., 2022) to inject trainable low-rank matrices into the model. For a pre-trained weight matrix $W_0 \in \mathbb{R}^{d \times k}$, LoRA adds:

$$W = W_0 + \Delta W = W_0 + BA \tag{2}$$

where $B \in \mathbb{R}^{d \times r}$, $A \in \mathbb{R}^{r \times k}$, and $r \ll \min(d, k)$ is the rank. The scaling factor α/r modulates the contribution of the adaptation. Table 1 shows task-specific LoRA configurations.

Parameter	Hate Speech	Sentiment
Rank (r)	64	32
Alpha (α)	128	64
Scaling (α/r)	2.0	2.0
Dropout	0.05	0.1
Trainable params	$\sim$0.5%	$\sim$0.25%

Table 1: LoRA configurations. Sentiment uses smaller rank and higher dropout to prevent overfitting on the three-class problem.

4.2.2. Task-Specific Prompts

We design prompts that establish the model's role and provide clear classification criteria.

Hate Speech Detection Prompts:

System prompt: *"You are an expert at detecting hate speech in Nepali memes. Analyze both the visual content and any Nepali/Devanagari text in the image. Consider cultural context, targeted groups (ethnicity, religion, gender, caste), and implicit meanings."*

User prompt: *"Is this Nepali meme hateful?"*

Categories:

- *Hate: Content that targets or attacks groups based on ethnicity, religion, gender, caste, or promotes violence/discrimination*
- *Safe: Non-hateful content (neutral, positive, jokes without targeting groups)*

Answer with one word only: Hate or Safe.
Answer:

Sentiment Analysis Prompts:

System prompt: *"You are an expert at analyzing sentiment in Nepali memes and social media content. You understand Nepali culture, humor, sarcasm, and emotional expression. Analyze both the visual elements (images, facial expressions, symbols) and Devanagari/English text. Pay attention to cultural references, meme templates, and contextual meaning."*

User prompt: *"Analyze the overall sentiment of this Nepali meme."*

Categories:

- *Negative: Expresses criticism, mockery, sadness, anger, frustration, disappointment, or pessimism*
- *Neutral: Factual information, observations, or content without clear emotional valence*
- *Positive: Expresses joy, humor, encouragement, celebration, love, hope, or optimism*

Consider: (1) Visual content and facial expressions, (2) Text meaning (literal and contextual), (3) Cultural context and references, (4) Overall emotional tone.

Respond with ONLY one word: Negative, Neutral, or Positive.

Answer:

4.2.3. MLP Projection Head

Alongside LoRA, we train an MLP projection head that maps the model's last hidden state to a lower-dimensional embedding space:

$$\mathbf{h}_1 = \text{GELU}(W_1\mathbf{h} + b_1) \tag{3}$$

$$\mathbf{h}_2 = \text{LayerNorm}(\mathbf{h}_1) \tag{4}$$

$$\mathbf{z} = W_2\mathbf{h}_2 + b_2 \tag{5}$$

where $\mathbf{h} \in \mathbb{R}^{4096}$ is the last hidden state, $W_1 \in \mathbb{R}^{2048 \times 4096}$, $W_2 \in \mathbb{R}^{512 \times 2048}$, and $\mathbf{z} \in \mathbb{R}^{512}$ is the projected embedding.

This projection head serves two purposes: (1) dimensionality reduction for efficient contrastive learning in Stage 2, and (2) learning task-specific representations beyond what the language model head captures.

4.2.4. Training Objective

The Stage 1 training objective is the standard language modeling loss, but computed only over the target tokens. Given a prompt x and target label y (e.g., "Hate" or "Safe"), we minimize:

$$\mathcal{L}_{\text{LM}} = -\log P(y|x, \text{image}) \tag{6}$$

We mask the prompt tokens (set labels to -100) so that gradients flow only through the answer tokens. This focuses learning on the classification decision.

4.2.5. Class Imbalance Handling

Both datasets exhibit significant class imbalance. We employ three complementary strategies:

Minority Class Oversampling: We duplicate minority class samples to achieve balanced class frequencies. For hate speech, we target a 1:1 ratio (hate:safe). For sentiment, we target 1:1:1 (negative:neutral:positive). The majority class is not downsampled to preserve all available information.

Image Augmentation: Oversampled (duplicated) images undergo random transformations to increase diversity and prevent the model from memorizing specific samples. With 80% probability, we apply:

- Horizontal flip: 50% probability
- Rotation: uniform in $[-15°, +15°]$
- Brightness adjustment: factor in $[0.8, 1.2]$
- Contrast adjustment: factor in $[0.8, 1.2]$
- Saturation adjustment: factor in $[0.8, 1.2]$

Each transformation is applied independently with 50% probability, creating diverse augmented versions.

Focal Loss (Sentiment Only): For the three-class sentiment task, we additionally use focal loss (Lin et al., 2017):

$$\mathcal{L}_{\text{focal}} = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \tag{7}$$

with focusing parameter $\gamma = 2.0$. The modulating factor $(1 - p_t)^{\gamma}$ reduces the loss contribution from easy examples, focusing training on hard cases. Class weights α_t are set to inverse class frequencies.

4.2.6. Training Configuration

The sentiment task uses a higher learning rate and larger effective batch size. We found this necessary for the more complex three-class decision boundary.

Parameter	Hate Speech	Sentiment
Learning rate	5×10^{-5}	1×10^{-4}
Batch size	2	2
Gradient accum.	8	12
Effective batch	16	24
Epochs	5	5
Optimizer	AdamW (β_1=0.9, β_2=0.999)	
Weight decay	0.01	
Epsilon	10^{-8}	
LR scheduler	Cosine, 10% warmup	
Grad. clipping	1.0	
Precision	bfloat16 mixed	

Table 2: Stage 1 training configurations.

4.3. Stage 2: Contrastive Backbone Fine-Tuning

Stage 2 continues training the LoRA layers and MLP with a joint classification and contrastive objective. The goal is to learn embeddings where same-class samples cluster together while different-class samples are pushed apart.

4.3.1. Linear Classifier

We add a linear classifier on the projected embeddings:

$$\mathbf{y} = W_c \mathbf{z} + b_c \tag{8}$$

where $W_c \in \mathbb{R}^{C \times 512}$, $b_c \in \mathbb{R}^C$, and C is the number of classes (2 for hate speech, 3 for sentiment).

4.3.2. Supervised Contrastive Loss

We use a supervised variant of InfoNCE loss (van den Oord et al., 2018; Khosla et al., 2020). For each anchor embedding $\mathbf{z}_i$ with label y_i, we sample k^+ positive embeddings (same label) and k^- negative embeddings (different labels) from the training set.

The contrastive loss is:

$$\mathcal{L}_{\text{con}} = -\frac{1}{k^+} \sum_{j=1}^{k^+} \log \frac{\exp(s_{ij}^+ / \tau)}{D_i} \tag{9}$$

where $s_{ij}^+ = \text{sim}(\mathbf{z}_i, \mathbf{z}_j^+)$ is the cosine similarity between anchor i and positive j, and the denominator is:

$$D_i = \sum_{p=1}^{k^+} \exp(s_{ip}^+ / \tau) + \sum_{n=1}^{k^-} \exp(s_{in}^- / \tau) \tag{10}$$

where $\text{sim}(\mathbf{a}, \mathbf{b}) = \mathbf{a}^\top \mathbf{b} / (\|\mathbf{a}\| \|\mathbf{b}\|)$ is cosine similarity and $\tau = 0.07$ is the temperature parameter. All embeddings are L2-normalized before computing similarity.

The temperature τ controls the concentration of the distribution. Lower temperatures make the

model more sensitive to hard negatives but can lead to training instability. We use $\tau = 0.07$ following prior work (Khosla et al., 2020).

4.3.3. Joint Training Objective

The Stage 2 loss combines classification and contrastive terms:

$$\mathcal{L} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}} \tag{11}$$

For hate speech: $\lambda_{\text{cls}} = 1.0$, $\lambda_{\text{con}} = 0.3$. The classification loss uses label smoothing of 0.05 to prevent overconfident predictions.

For sentiment: $\lambda_{\text{cls}} = 1.0$, $\lambda_{\text{con}} = 0.4$. The classification loss uses inverse-frequency class weights to handle imbalance.

4.3.4. Stage 2 Training Configuration

Table 3 shows Stage 2 hyperparameters. A key difference from Stage 1 is the much lower backbone learning rate and stronger regularization, especially for sentiment.

Parameter	Hate Speech	Sentiment
Backbone LR	2×10^{-5}	5×10^{-6}
Classifier LR	1×10^{-4}	2×10^{-5}
Batch size	2	2
Gradient accum.	8	12
Epochs	5	3
Weight decay	0.01	0.1
Early stopping	–	2
Positives (k^+)	3	4
Negatives (k^-)	5	6
Temperature (τ)	0.07	0.07
λ_{con}	0.3	0.4
Label smoothing	0.05	–
Class weights	–	Inverse freq.

Table 3: Stage 2 training configurations. Sentiment uses aggressive regularization to prevent overfitting.

The sentiment task required substantially different settings:

- 4× lower backbone learning rate (5×10^{-6} vs 2×10^{-5})
- 10× higher weight decay (0.1 vs 0.01)
- Fewer epochs (3 vs 5) with early stopping

These aggressive regularization choices were necessary because the three-class sentiment task has a smaller effective training set per class and showed signs of overfitting early in our experiments.

4.4. Inference and Ensemble

At inference time, we combine predictions from both stages to leverage their complementary strengths.

4.4.1. Stage 1 Predictions

We extract token-level probabilities from the language model head. For hate speech:

$$p^{(1)}_{\text{hate}} = \frac{\exp(l_{\text{Hate}})}{\exp(l_{\text{Hate}}) + \exp(l_{\text{Safe}})} \tag{12}$$

where l_{Hate} and l_{Safe} are logits for the respective tokens at the first generated position.

For sentiment, we similarly compute:

$$p^{(1)}_c = \frac{\exp(l_c)}{\sum_{c'} \exp(l_{c'})} \tag{13}$$

for each class $c \in \{\text{Negative}, \text{Neutral}, \text{Positive}\}$.

4.4.2. Stage 2 Predictions

We apply softmax to the linear classifier outputs:

$$p^{(2)} = \text{softmax}(W_c\mathbf{z} + b_c) \tag{14}$$

4.4.3. Ensemble Strategy

The final prediction is a weighted combination:

$$p_{\text{ensemble}} = w \cdot p^{(1)} + (1 - w) \cdot p^{(2)} \tag{15}$$

We tune the ensemble weight w and decision threshold jointly via grid search on the validation set, optimizing for macro F1. For hate speech:

- Ensemble weight w: search in $[0.1, 0.9]$ with step 0.02
- Decision threshold: search in $[0.25, 0.75]$ with step 0.005

For sentiment (multi-class), we use argmax on the ensembled probabilities and only tune the ensemble weight.

5. Experimental Setup

5.1. Implementation Details

We implement our system using:

- PyTorch 2.4.0 with CUDA
- HuggingFace Transformers $\geq$ 4.46.0 (Wolf et al., 2020)
- PEFT $\geq$ 0.13.0 for LoRA (Mangrulkar et al., 2024)
- Accelerate $\geq$ 1.0.0 for distributed training

Training runs on NVIDIA H100 GPUs via Modal Labs cloud infrastructure. Each training run takes approximately 2–4 hours per stage.

5.2. Reproducibility

We use fixed random seeds (42) for data splitting, model initialization, and training. All hyperparameters are specified in Tables 1–3. Our code will be made available upon publication.

6. Results and Discussion

6.1. Official Results

Subtask	Macro F1	Rank
A: Hate Speech Detection	0.7970	2nd
B: Sentiment Analysis	0.5177	4th

Table 4: Official leaderboard rankings on CHiPSAL 2026.

Table 4 shows our official results. We achieved strong performance on hate speech detection with F1 of 0.797, securing 2nd place. For sentiment analysis, we achieved F1 of 0.518, placing 4th. The binary hate speech task proved more tractable, likely due to the clearer class boundary compared to the three-way sentiment distinction.

6.2. Ablation Study

Configuration	Test F1
Subtask A: Hate Speech Detection	
Stage 1 (LoRA baseline)	0.74
+ Oversampling	0.76
+ Image augmentation	0.77
+ Stage 2 contrastive	0.79
+ Threshold tuning	**0.80**
Subtask B: Sentiment Analysis	
Stage 1 (LoRA baseline)	0.49
+ Focal loss	0.50
+ Oversampling & augmentation	0.51
+ Stage 2 contrastive	**0.52**

Table 5: Ablation study. Each row adds one component cumulatively. F1 scores are reported on the held-out test set, except threshold tuning which was optimized on the validation split and then applied to the test set.

Table 5 presents detailed ablation results. Key findings:

LoRA scope matters. Applying LoRA to all attention and feed-forward layers outperformed applying it only to attention layers (+3 points on hate speech).

Class imbalance handling is crucial. The combination of oversampling and augmentation contributes 2–3 points. Focal loss adds another point for the three-class sentiment task by addressing neutral class bias.

Contrastive learning provides consistent gains. Stage 2 adds 2 points on hate speech and contributes to sentiment, demonstrating that the contrastive objective learns more discriminative representations.

Threshold tuning improves hate speech detection. Optimizing the classification threshold on the validation split and applying it to the test set adds 1 point, reaching the final score of 0.80.

6.3. Qualitative Examples

Successful Cases: The model correctly identifies explicit hate speech with visual symbols (e.g., derogatory depictions) combined with text. For sentiment, clear emotional expressions in both image and text (e.g., smiling faces with celebratory text) are reliably classified.

Failure Cases: A meme showing a historical figure with a quote that is hateful in Nepali cultural context but neutral in isolation was misclassified as safe. The model lacks sufficient cultural knowledge to interpret such references.

7. Conclusion

We presented a two-stage vision-language system for Nepali meme classification at CHiPSAL 2026. By combining LoRA fine-tuning with contrastive learning on Qwen3-VL-8B-Instruct, we achieved 2nd place on hate speech detection (F1: 0.797) and 4th place on sentiment analysis (F1: 0.518).

Our results highlight several key findings. Modern VLMs with native Devanagari support enable end-to-end processing of South Asian language memes without OCR pipelines, eliminating error propagation from intermediate steps. Two-stage training (generative then contrastive) outperforms either approach alone by 3–5 F1 points, demonstrating the complementarity of language modeling and metric learning objectives. Task-specific hyperparameter tuning proved essential, as sentiment required much stronger regularization than hate speech. Finally, careful handling of class imbalance is crucial for macro F1 optimization, contributing up to 7 points across our ablation.

Promising future directions include retrieval-augmented approaches to incorporate cultural knowledge, synthetic data augmentation to expand limited training sets, multi-task learning jointly across hate speech and sentiment, and explainability analysis to better understand model decisions for culturally sensitive content.

8. Limitations

- **Data scarcity:** With approximately 1,000 samples per task, overfitting remains a concern. Our validation results may not fully generalize to the

test distribution. Due to the high computational cost of each training run, we did not evaluate variance across multiple random seeds; results may vary with different initializations.

- **Cultural context:** Nepali cultural references, historical context, and humor styles may not be well-represented in Qwen3-VL's pretraining data, limiting performance on culturally-specific content.
- **Prompt sensitivity:** We did not exhaustively search prompt variations, which could further improve performance.
- **Code-mixing:** While Qwen3-VL handles Nepali-English code-mixing to some extent, highly mixed content may still pose challenges.
- **Baseline comparisons:** We did not evaluate Qwen3-VL-8B-Instruct in a zero-shot (no fine-tuning) setting or compare against an OCR + text-only pipeline, which would better quantify the contribution of our end-to-end multimodal approach. We leave these comparisons for future work.

9. Ethical Considerations

Hate speech detection systems must be deployed carefully to avoid over-censorship or bias. Our model may have false positives on legitimate political speech or cultural content unfamiliar to the base model. We recommend human review for content moderation decisions.

10. Acknowledgements

We thank the CHiPSAL 2026 organizers for creating this valuable benchmark and shared task infrastructure. We acknowledge Modal Labs for computational resources.

11. Bibliographical References

Aashish Bhandari, Siddhant B Shah, Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. Crisishatemm: Multimodal analysis of directed and undirected hate speech in text-embedded images from russia-ukraine conflict. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1994–2003.

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*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673.

Douwe Kiela, Hamed Firooz, Aravind Mober, Vedanuj Goswami, Amanpreet Jambhekar, Patrick Hwang, Arthur Kiela, Holger Schwenk, et al. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2024. PEFT: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

Jingyu Mei, Jingyuan Chen, Weihua Lin, Bill Byrne, and Marcus Tomalin. 2024. Improving hateful meme detection through retrieval-guided contrastive learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5333–5347.

Jingyu Mei, Jingyuan Chen, Guang Yang, Weihua Lin, and Bill Byrne. 2025. Robust adaptation of large multimodal models for retrieval augmented hateful meme detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 23817–23839.

Anil Singh Parihar, Surendrabikram Thapa, and Sushruti Mishra. 2021. Hate speech detection using natural language processing: Applications and challenges. In *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 1302–1308. IEEE.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. MOMENTA: A multimodal framework for detecting harmful memes

and their targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455.

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 *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Kengatharaiyer Sarveswaran, Surendrabikram Thapa, Ashwini Vaidya, Tafseer Ahmed Khan, and Bal Krishna Bal. 2026. Findings of the second workshop on challenges in processing south asian languages (chipsal 2026). In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*.

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Surabhi Adhikari, Hariram Veeramani, Mehwish Nasim, and Usman Naseem. 2025a. Large language models (llm) in computational social science: prospects, current state, and challenges. *Social Network Analysis and Mining*, 15(1):1–30.

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Kritesh Rauniyar, Laxmi Thapa, Surabhi Adhikari, Kristina T. Johnson, Kengatharaiyer Sarveswaran, Bal Krishna Bal, and Usman Naseem. 2026. Multimodal hate and sentiment understanding in low-resource text-embedded images for online safety and digital well-being. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*.

Surendrabikram Thapa, Hariram Veeramani, Liang Hu, Qi Zhang, Wei Wang, and Usman Naseem. 2025b. A multimodal prompt-based framework for analyzing code-mixed and low-resource memes. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 1913–1923.

Surendrabikram Thapa, Hariram Veeramani, Imran Razzak, Roy Ka-Wei Lee, and Usman Naseem. 2025c. Cross platform multimodal retrieval augmented distillation for code-switched content understanding. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2042–2051.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.