Multi-Modal-Minds@CHiPSAL 2026: A Comparative Study of Textual, Visual and Multimodal Architecture for Nepali Meme Moderation

Sandesh Shrestha[1], **Bikram K.C.**[1], **Akshyat Shah**[2], **Ashish Acharya**[3], **Rabin Thapa**[1]
[1] IIMS College, Kathmandu, Nepal
{sandeshrestha115, bikramkc9810}@gmail.com, rabin@iimscollege.edu.np
[2] Delhi Technological University, Delhi, India
akshyatshah@gmail.com
[3] Kathmandu University, Dhulikhel, Nepal
ashishacharya@ku.edu.np

Abstract

Memes have become ubiquitous on social media platforms blending text and imagery to express complex and culturally nuanced messages. While a high degree of automation in meme moderation has been achieved for high-resource languages, low-resource languages, such as Nepali, still remain largely neglected. In this paper, we describe our system submission to the CHiPSAL 2026 Shared Task on Multi-modal Hate and Sentiment Understanding in Low-Resource Nepali Memes, which features two main sub-tasks: (1) Detection of Hate Speech as binary classification and (2) Sentiment Analysis as multi-class classification in Nepali memes. We perform a comprehensive analysis of the following models: uni-modal textual models (mBERT, XLM-RoBERTa, MuRIL), uni-modal visual models (ResNet, ConvNeXt, ViT), nine different early-fusion multimodal models, and the vision-language foundation model, SigLIP. Among all models, the ViT model achieved the best macro F1-score (0.6278) for the hate speech detection task, while SigLIP achieved the best score (0.5481) for the sentiment analysis task. We hypothesize that the under-performance of fusion models may be attributed to OCR noise and inadequate low-resource textual representations that act as a bottleneck when paired with more advanced visual encoders. These results highlight the unique challenges of multimodal meme comprehension in low-resource contexts and underscores the requirement for culturally grounded, noise-robust approaches to content moderation in Nepali.

Keywords: Nepali memes, Hate Speech Detection, Sentiment Analysis, Multi-modal Learning, Low-Resource NLP

1. Introduction

There are numerous social networks that have significantly changed the way varying sentiments, humor, and opinions are disseminated. With memes becoming one of the most popular formats for information exchange in contemporary society, the dynamics of social discourse have also become more multifaceted. Unlike traditional approaches that use text for expression, memes combine textual and visual media to convey nuanced meaning, often relying on subtle cultural contexts (Kiela et al., 2020a). Although many memes are humorous or benign, a significant subset tends to skew towards hate speech or invoke sentiment that may influence social attitudes. Automatic identification of harmful or sentiment-laden memes is a growing area of research within natural language processing (NLP) and computer vision.

Detecting hate speech and classifying sentiment in memes is particularly tedious and challenging due to multi-modal interactions as neither text nor image alone captures the full meaning(Parihar et al., 2021). Most of the existing work on ML and language processing is specifically limited to high-resource languages such as English, while low-resource languages such as Nepali remain underserved and under-explored (Thapa et al., 2025a). The lack of sufficient data and pre-trained models to support robust multi-modal understanding in under-resourced languages is a pressing need. Thus, research that organically focuses on such languages is essential for equitable and worldwide content moderation.

The Shared Task (Thapa et al., 2026) on Multimodal Hate and Sentiment Understanding in Low-Resource Memes at CHiPSAL 2026 (Sarveswaran et al., 2026) provides a benchmark for addressing this problem specifically in Nepali. The task uses a dataset curated from (Thapa et al., 2025a) and contains Nepali-only memes annotated for hate speech and sentiment, and challenges participants to develop capable systems for binary hate detection and three-class sentiment classification (negative, neutral, positive).

2. Related Works

Although substantial work has been done on hate speech and sentiment analysis in English, non-English languages particularly low resource ones have remained largely underrepresented in literature. We therefore focus on extensions into the Nepali language, as limited work has been done in

Proceedings of the Second workshop on Challenges in Processing South Asian Languages (CHiPSAL2026) @ LREC, pages 258–266

hate speech detection and sentiment analysis.

In their work, (Kiela et al., 2020a) introduce the Hateful Memes Challenge, a large-scale benchmark of multimodal memes. The paper shows that text-only models marginally outperform vision-only models. They observe that the more advanced the fusion, the better the model performs. The paper suggests that early fusion models (MMBT, ViLBERT and Visual BERT) broadly outperform middle (Concat) and late fusion approaches with ViLBERT CC giving the highest accuracy of 66.10 among all the methods. (Bhandari et al., 2023) introduce CrisisHateMM, a multimodal dataset of over 4,700 text-embedded images collected from the Russia-Ukraine conflict and annotated for directed and undirected hate speech using a CLIP-based multimodal fusion architecture. Their best-performing model achieves F1-scores of 0.786, 0.609, and 0.615 for hate speech, its direction and targets respectively, demonstrating that combining visual and textual modalities significantly outperforms text-only and image-only baselines.

(Chakravarthi et al., 2020, 2021) address the lack of resources for Dravidian languages by constructing an annotated corpus of 15,744 sentences and organizing shared tasks on sentiment analysis and offensive language detection for code-mixed social media text. They released datasets comprising 43,919 Tamil, 20,010 Malayalam, and 7,772 Kannada comments collected from YouTube. The highest-performing system used an ensemble of transformer models selected via a genetic algorithm, achieving weighted F1-scores of 0.78 for Tamil, 0.97 for Malayalam, and 0.75 for Kannada through task-adaptive pretraining of mBERT and XLM-RoBERTa (Sitaula et al., 2021).

(Sitaula et al., 2021) presented one of the first large-scale benchmarks for Nepali social media sentiment analysis, focusing on 33,247 COVID-19-related tweets. Their experiments with deep learning models, including CNNs, yielded a best overall classification accuracy of 68.7

(Thapa et al., 2023) introduced NEHATE, a manually annotated dataset of 13,505 Nepali tweets for hate speech and target identification in local election discourse. Using multilingual BERT variants, NepBERTa achieved the best performance with macro F1-scores of 0.68 for hate speech detection and 0.60 for target identification, establishing NEHATE as one of the first large-scale hate speech resources for Nepali. (Thapa et al., 2025a) propose a multi-modal prompt-based framework evaluated on low resource Nepali and code-mixed Nepali meme datasets. By using their proposed MemeNePAL model, they achieved an F1-score of 0.5301 and 0.6331 across sentiment analysis and hate detection respectively, higher compared to other text-only, vision-only and multi-modal models. The paper also introduces NeMeme, the first Nepali meme dataset annotated for hate speech and sentiment analysis. (Thapa et al., 2025b) present a cross-platform retrieval-augmented distillation approach for 4,211 code-switched memes in the Nepali-English language classified for sentiment and hate speech. They propose MM-RAD model for code switched context achieving an F1-score of 43.24 in sentiment analysis and 64.96 in hate speech detection.

Building on the dataset published in (Thapa et al., 2025a), the shared task (Thapa et al., 2026) at CHiPSAL@LREC 2026 (Sarveswaran et al., 2026) extends the challenge to the community. They invite methods for multi-modal hate and sentiment understanding on low-resource Nepali memes serving as the basis for this paper.

3. Dataset and Task

This work is part of the Shared Task on Multimodal Hate and Sentiment Understanding in Low-Resource Memes (Thapa et al., 2026) at CHiPSAL 2026 (Sarveswaran et al., 2026). The shared task consists of two different subtasks: Subtask A aims to detect hate speech via binary classification into hate and non-hate while Subtask B is associated with the three way classification of sentiment into positive, negative and neutral classes. The shared task is hosted on Codabench and a dataset for this task has already been published in ICWSM 2025 (Thapa et al., 2025a) as a part of NeMeme. The dataset contains memes collected from social media platforms, and textual content embedded within meme images is extracted using EasyOCR with Nepali language support. However, EasyOCR's performance on informal, colloquial Devanagari script as commonly found in internet memes is known to be inconsistent, often introducing transcription noise that can degrade downstream textual representations. Table 1 summarizes the distribution between all splits for the shared task.

The relatively small training set of approximately 1,000 samples per subtask poses a particular challenge for multimodal fusion models, which introduce significantly more parameters than their unimodal counterparts. Learning effective cross-modal alignment under such data constraints increases the risk of overfitting, which may partially explain the comparatively weaker performance of explicit fusion architectures observed in our experiments.

3.1. Subtask A: Hate Speech Detection

Subtask A requires binary classification of Nepali-only memes into hate and non-hate categories. The training set comprises 1,068 memes, of which

Subtask	Label	Train	Val	Test
A (Hate Detection)	Hate	348	35	–
	Non-Hate	720	98	–
	Total	1,068	133	134
B (Sentiment Analysis)	Negative	341	39	–
	Neutral	473	65	–
	Positive	247	29	–
	Total	1,061	133	133
Total		**2,129**	**266**	**267**

Table 1: Distribution across train, validation, and test splits for Subtask A (Hate Detection) and Subtask B (Sentiment Analysis) of the CHiPSAL 2026 shared task (Thapa et al., 2026). Test set class-level labels were withheld by the shared task organizers; only the total instance count is available.

720 are Non-Hate(labeled as 0) and 348 are Hate(labeled as 1). The validation set consists of 133 memes with 98 Non-Hate and 35 Hate instances. The test set, on the other hand, contains 134 unlabeled memes for the final evaluation. Figure 1 gives an idea of the instances in the Hate Detection subtask.

(a) Hate

(b) Non-Hate

Figure 1: Examples of Nepali memes from the shared task (Thapa et al., 2026) for Subtask A (Hate Speech Detection).

3.2. Subtask B: Sentiment Analysis

Subtask B requires a multi-class sentiment classification of Nepali-only memes into Negative(labeled as 0), Neutral(labeled as 1), and Positive(labeled as 2) categories. The training set contains 1,061 memes with 341 negative, 473 neutral and 247 positive text embedded images. The validation set consists of 133 memes with 39 negative, 65 neutral, and 29 positive instances. The test set contains 133 unlabeled memes for the final evaluation. Figure 2 gives an idea of the instances in the Sentiment Analysis subtask.

(a) Negative

(b) Neutral

(c) Positive

Figure 2: Examples of Nepali memes from the shared task (Thapa et al., 2026) for Subtask B (Sentiment Analysis).

4. Methodology

We developed a comprehensive experimental framework to process the raw dataset and establish robust unimodal baselines. Using these baselines, we evaluated two multimodal paradigms: an exhaustive grid of nine early-fusion architectures and a pre-aligned vision-language foundation model.

4.1. Textual & Visual Prepossessing

The EasyOCR framework, configured with Devanagari and English language support, was utilized to extract text from the meme images. To evaluate the models under realistic, "in-the-wild" moderation conditions, we applied EasyOCR directly to the raw images without explicit image-level text preprocessing (such as binarization, deskewing, or contrast adjustment). While this preserves the original structure of the memes, the highly variable nature of meme typography (e.g., erratic stroke widths, dynamic colors, and chaotic backgrounds) introduces an inherent baseline of OCR noise. The raw strings resulting from this extraction were then passed directly to our subword tokenizers to accommodate unknown vocabulary items and local slang.

We built a stochastic enhancement pipeline for our visual modality through the application of the *albumentations* library (Buslaev et al., 2020), which enabled our encoders to acquire strong semantic characteristics that protected it from learning specific patterns of internet memes. The images were resized and scaled to 224×224 pixels. The model was trained using the following data augmentation pipeline:

- **Web Degradation:** We simulate the characteristic artifact of the platform using two methods, which include JPEG compression at quality

levels between 30 and 70 and aggressive spatial down-scaling using scale factors between 0.5 and 0.8, with a probability of 40%.

- **Spatial Regularization:** We use coarse dropout (cutout) to up to four rectangular regions with a 30% probability, preventing the model from over-relying on isolated visual cues.

- **Color Variation:** We dynamically apply random hue and saturation adjustments with the probability of 30% which prevents the model from over-relying on isolated visual cues.

Finally, all augmented image tensors are normalized using standard ImageNet statistics before being passed to the visual encoders.

4.2. Experimental Setup

Hyperparameter	Value
Optimizer	AdamW ($\epsilon = 1 \times 10^{-8}$)
β_1, β_2	0.9, 0.999
Batch Size	16
Dropout (Head)	0.3
Epochs (Max)	15
Early Stopping	Patience = 5
Hardware	NVIDIA Tesla P100
Random Seed	42

Table 2: General training configurations applied across all experiments.

A consistent training pipeline was employed across both the Hate Speech Detection and Sentiment Analysis subtasks The CHIPSAL 2026 Nepali meme dataset (Thapa et al., 2026) served as the primary benchmark for all model evaluations

The preprocessed inputs, as detailed in the previous section, were utilized to train the various architectures Specifically, the textual modality relies on OCR-extracted Nepali text, while the visual modality leverages stochastically augmented meme images This standardized setup ensured a fair and unbiased evaluation across all unimodal and early-fusion multimodal systems

As summarized in Table 2, key hyperparameters included a batch size of 16, the AdamW optimizer, and a maximum of 15 training epochs with early stopping (patience = 5) Parameter-efficient fine-tuning was achieved using LoRA ($r = 16$, $\alpha = 32$) for both text and image encoders A complete list of model-specific learning rates, is provided in Appendix 9 for full reproducibility

4.2.1. Unimodal and Multimodal Models Selection

Unimodal Textual Models: We evaluated multiple text-only transformer-based models for Nepali meme analysis. As justified above, we utilized mBERT-base-uncased (Devlin et al., 2019), XLM-RoBERTa-base (Conneau et al., 2020), and MuRIL (Khanuja et al., 2021). From these models, we extracted independent textual embeddings $\mathbf{E}_{\text{text}} \in \mathbb{R}^{d_t}$ to serve as our baseline textual framework.

Uni-modal Visual Models: We extracted visual embeddings $\mathbf{E}_{\text{image}} \in \mathbb{R}^{d_v}$ using a diverse set of advanced visual architectures: ResNet (He et al., 2016), ConvNeXt (Liu et al., 2022), and Vision Transformer (ViT) (Dosovitskiy et al., 2021).

Multi-modal Models: To capture cross-modal interactions, we experimented with early-fusion architectures by pairing each visual encoder with each textual encoder. Specifically, we combined the set of visual models $\{\text{ResNet}, \text{ConvNeXt}, \text{ViT}\}$ with the set of textual models $\{\text{mBERT}, \text{XLM-R}, \text{MuRIL}\}$, resulting in a total of nine multimodal configurations.

Following standard early-fusion (feature-level) paradigms for multi-modal meme classification (Baltrušaitis et al., 2018; Kiela et al., 2020b), textual and visual representations are extracted independently before being combined for classification. Formally, given an image-text pair (x_v, x_t), let f_v and f_t denote the visual and textual encoders, respectively. We first extract the independent dense feature vectors $\mathbf{h}_v = f_v(x_v)$ and $\mathbf{h}_t = f_t(x_t)$. These representations are then concatenated to form a fused multimodal vector:

$$\mathbf{h}_{fused} = [\mathbf{h}_v \oplus \mathbf{h}_t]$$

where $\oplus$ denotes the concatenation operation. Finally, this joint representation is passed through a linear classification head to obtain the predicted probabilities $\hat{y}$:

$$\hat{y} = \text{Softmax}(\mathbf{W}\mathbf{h}_{fused} + \mathbf{b})$$

where $\mathbf{W}$ and $\mathbf{b}$ represent the learnable weights and bias of the classifier. This constitutes an early-fusion paradigm, as the representations from both modalities are merged prior to the final classification layer, in contrast to decision-level late fusion where independent model predictions are combined.

Vision-Language Model: In addition to encoder-based fusion approaches, we also evaluated SigLIP (Zhai et al., 2023), a pretrained vision-language model that jointly learns aligned representations between images and text.

4.2.2. Architectural Justification:

The selection of our diverse model architectures is deeply grounded in the linguistic and visual realities of Nepali digital culture, as highlighted in recent

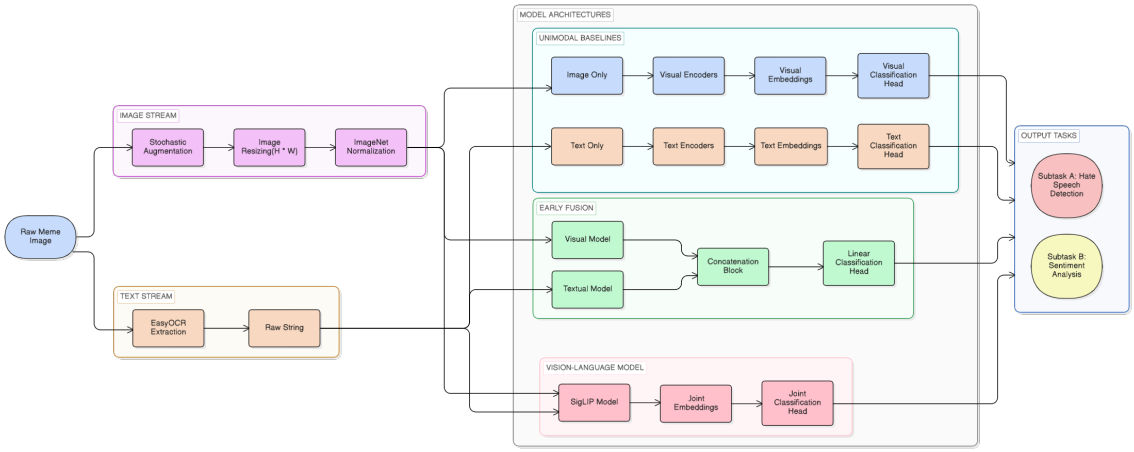


Figure 3: Unified system architecture illustrating the experimental framework. The raw meme input is split into independent preprocessing streams: the Image Stream utilizes stochastic augmentation to prevent pattern memorization, while the Text Stream extracts raw Devanagari strings via EasyOCR. The framework evaluates three distinct modeling paradigms—Unimodal Baselines, explicit Early Fusion networks, and a pre-aligned Vision-Language Model (SigLIP)—across the final output tasks of Hate Speech Detection and Sentiment Analysis.

multimodal literature (Kiela et al., 2020b; Bhandari et al., 2023). For the textual modality, we deliberately prioritized multilingual models (mBERT, XLM-ROBERTa) and specifically MuRIL, rather than relying solely on monolingual Nepali encoders. Nepali memes frequently exhibit complex code-switching not only with English but heavily with Hindi, driven by the pervasive influence of Bollywood and the Indian entertainment industry on regional internet culture (Thapa et al., 2025a). MuRIL's specific pre-training on Indic languages makes it uniquely equipped to handle this nuanced, cross-lingual meme vocabulary. For the visual modality, we selected ResNet to serve as a proven, robust baseline, ConvNeXt to evaluate modern CNN optimizations, and the Vision Transformer (ViT) to capture the global context of visual semiotics (e.g., emojis and specific meme templates). Finally, our inclusion of the SigLIP foundation model was driven by the proven efficacy of pre-aligned cross-modal representations in noisy, low-resource moderation tasks.

5. Results & Discussion

Table 3 presents a comparative performance analysis of the evaluated architectures across the Hate Speech Detection and Sentiment Classification subtasks.

5.1. Hate Speech Detection:

Transformer-based visual architectures proved highly effective for the hate speech detection task. While the prevailing narrative suggests that multimodal fusion is inherently superior for meme moderation, our standalone Vision Transformer (ViT) achieved the highest overall Macro F1-score of 0.6278, driven by the highest recall (0.6278) among all evaluated models.

Equally notable is the performance of the vision-language foundation model, SigLIP. While its F1-score of 0.5999 slightly trailed the pure ViT, SigLIP achieved the highest overall accuracy (0.6940) and a strong precision of 0.6424. This demonstrates that SigLIP's pre-trained cross-modal alignment makes it highly reliable for general classification across both classes, even if the pure visual signal of the ViT was slightly more effective at retrieving specific hateful instances.

Further examination of the metrics revealed distinct behavioral differences across the models. The text-only baseline, XLM-RoBERTa, achieved a high precision of 0.8383 but an extremely low recall of 0.5114. This indicates that while XLM-RoBERTa rarely misclassified regular memes as hateful, it overlooked a considerable amount of genuinely hateful content. Explicit multi-modal architectures, such as MuRIL-ConvNeXt (F1-score of 0.6072), performed competitively but did not surpass the performance of the pure visual signal represented by the ViT.

5.2. Sentiment Analysis:

A similar reliance on pre-trained visual alignment for sentiment classification was observed in Subtask B. The SigLIP model achieved the best performance, sweeping all evaluation metrics with the highest F1-score of 0.5481, a high precision of 0.5067, a recall of 0.6091 and an outstanding accuracy of

	Models	Hate Detection				Sentiment Analysis			
		Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1	Acc.
Text	mBERT	0.5322	0.5318	0.5320	0.5896	0.4254	0.4208	0.3831	0.3985
	XLM-RoBERTa	**0.8383**	0.5114	0.4258	0.6791	0.1579	0.3333	0.2143	0.4737
	MuRIL	0.3358	0.5000	0.4018	0.6716	0.4254	0.4208	0.3831	0.3985
Visual	ResNet	0.6178	0.5811	0.5804	0.6791	0.4955	0.4753	0.4793	0.5188
	ConvNeXt	0.5872	0.5763	0.5782	0.6493	0.3905	0.3985	0.3897	0.4511
	ViT	0.6278	**0.6278**	**0.6278**	0.6716	0.4140	0.4104	0.4112	0.4436
Explicit Fusion	mBERT + ResNet	0.5890	0.5944	0.5904	0.6269	0.3051	0.3049	0.2840	0.2857
	mBERT + ConvNeXt	0.5795	0.5902	0.5682	0.5821	0.4083	0.4123	0.4076	0.4211
	mBERT + ViT	0.5916	0.6005	0.5915	0.6194	0.4828	0.4984	0.4826	0.4887
	MuRIL + ResNet	0.5601	0.5601	0.5601	0.6119	0.2634	0.2579	0.2402	0.2406
	MuRIL + ConvNeXt	0.6101	0.6235	0.6072	0.6269	0.4320	0.4362	0.4276	0.4436
	MuRIL + ViT	0.5774	0.5833	0.5780	0.6119	0.4713	0.4657	0.4677	0.4887
	XLM-RoBERTa + ResNet	0.6014	0.6119	0.6012	0.6269	0.1322	0.3083	0.1682	0.2180
	XLM-RoBERTa + ConvNeXt	0.5589	0.5710	0.5598	0.5675	0.3698	0.3698	0.3685	0.3985
	XLM-RoBERTa + ViT	0.5352	0.5384	0.5335	0.5672	0.4561	0.4562	0.4491	0.4662
V-L	**SigLIP**	0.6424	0.5980	0.5999	**0.6940**	**0.5067**	**0.6091**	**0.5481**	**0.7143**

Table 3: Performance comparison of unimodal and multimodal architectures for sentiment and hate analysis. Bold values denote the best-performing model for each metric within a given subtask.

0.7143.

On the other hand, unimodal text models underperformed on this subtle task. For instance, XLM-RoBERTa struggled significantly, achieving a precision of only 0.1579. Although explicit fusion methods such as mBERT-ViT achieved a reasonable F1 score of 0.4826 and an accuracy of 0.4887, it was significantly lower than the SigLIP baseline. Since SigLIP is already trained on a high-quality, cross-learning image-text alignment, its performance implies that the use of a foundation model with acquired cross-modal knowledge is much more effective than directly combining low-resource Nepali text encoders with standard vision models.

5.3. Discussion of Findings

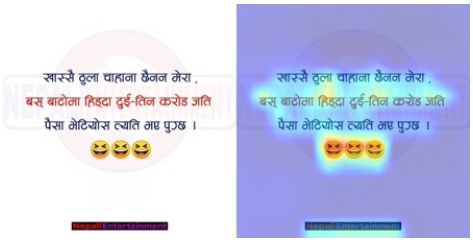
The primary finding across both subtasks is that unimodal visual architectures and pre-aligned vision-language foundation models unexpectedly outperformed explicit early-fusion integration networks. While prevailing assumptions in meme moderation suggest that multimodal fusion is inherently superior, our empirical results demonstrate that the visual modality carries the primary semantic burden in this specific low-resource, high-noise environment.

We postulate that this performance gap is directly attributable to the "distractor" effect caused by degraded textual signals. In high-noise environments characterized by complex code-switching, casual slang, and intentional visual degradation (such as "deep-fried" aesthetics), the OCR extraction pipeline frequently produces noisy or incoherent text inputs. When these degraded textual representations are concatenated with strong visual features during early fusion, the textual modality acts as an active distractor rather than a complementary signal. The linear classification head struggles to isolate the predictive visual signal from the high-entropy textual noise, creating a significant performance bottleneck.

Consequently, standalone visual models like the Vision Transformer (ViT) perform better precisely because they are insulated from this textual noise. By treating typographical overlays and emojis strictly as visual semiotics, the ViT avoids being confused by OCR transcription errors. Similarly, foundation models like SigLIP, which are pre-trained on massive datasets to align images and text holistically, demonstrate a superior ability to filter out this modality-specific noise compared to fusion networks trained from scratch on our limited dataset. Ultimately, these findings underscore that forcing multimodal fusion with unreliable text inputs is actively detrimental, and robust moderation in low-resource languages must prioritize noise-resilient or entirely vision-dominant approaches.

6. Interpretability and Modality Reliance

We conducted a two-fold interpretability analysis: (1) mapping spatial attention on the unimodal Vision Transformer (ViT) using Gradient-weighted Class Activation Mapping (Grad-CAM) to visualize modality-specific focus areas, and (2) performing



(a) Correct Classification (True: Hate, Pred: Hate)

(b) Incorrect Classification (True: Hate, Pred: Non-Hate)

Figure 4: Grad-CAM heatmaps from the penultimate layer of the ViT model. Subfigure (a) correctly identifies hateful intent by focusing on visual semiotics (emojis), while (b) shows a failure to capture context, misidentifying a hateful meme as Non-Hate

counterfactual modality ablation using the vision-language foundation model (SigLIP).

6.1. Visual Attention and Semiotics (ViT)

To visualize how the text-agnostic ViT handles multimodal inputs, we generated Grad-CAM heatmaps from the penultimate layer of the model. Because ViT lacks a natural language tokenizer, it processes typographical overlays strictly as visual tokens.

As shown in Figure 4 (a), a hateful meme was successfully classified. The activation map shows that the model is largely invariant to the structural Devanagari text; instead, it centers its attention on the laughing emojis (specifically, the *face with tears of joy* emoji). Emojis play a significant role and can be regarded as strong visual semiotics that carry significant emotive weight in meme culture. In most cases, emojis express mockery, sarcasm, or malevolent intent. Standard low-resource NLP pipelines often do not recognize the semantic content of such Unicode characters. Nevertheless, these are leveraged effectively by the ViT as highly discriminative visual characteristics, hence its superior F1-score in hate speech classification.

On the other hand, Figure 4 (b) shows a critical failure mode (True: 1, Predicted: 0). The Grad-CAM heatmap shows that the model focuses disproportionately on a visually secondary object (a monkey). The model fails to capture the contextual hatefulness created by the interaction of the text and image because it lacks the ability to process the accompanying text. This proves that while visual semiotics are powerful, this reliance creates significant blindspots regarding context-specific memes where the image alone does not convey the full intent.

Modality Reliance	Hate Speech	Sentiment
Image Bias	34 (25.6%)	12 (9.0%)
Text Bias	6 (4.5%)	17 (12.8%)
True Multimodal	93 (69.9%)	104 (78.2%)
Total	**133**	**133**

Table 4: Counterfactual modality ablation results for the SigLIP model, detailing unimodal bias versus true cross-modal synergy across both subtasks.

6.2. Counterfactual Modality Ablation (SigLIP)

We employed a counterfactual ablation framework to evaluate modality bias and inter-modality dynamics in the SigLIP architecture. By applying both text and image ablation to each validation sample, we isolated the modality driving the baseline multimodal prediction. An instance was defined as having an image or text bias if only one of the unimodal predictions matched the multimodal baseline.

The ablation results (Table 4) provide empirical evidence for our hypothesis of visual dominance in low-resource settings. In Subtask A (Hate Speech Detection), SigLIP exhibited a strong image bias (34 instances), significantly outpacing the minimal text bias (6 instances). This suggests that the model relies extensively on visual signals for hate detection, effectively bypassing the noisy or code-mixed Nepali textual stream.

Conversely, in Subtask B (Sentiment Classification), the modality reliance was considerably more balanced. The ablation identified 12 image-biased and 17 text-biased cases.However, the vast majority (104 instances) represented true multimodal synergy.

7. Conclusion

This paper has compared different unimodal, early-fusion, and foundation models of the CHiPSAL 2026 Shared Task on Nepali meme moderation. Contrary to the intuitive expectation that explicit text-vision fusion is necessary, we find that single visual architectures and pre-aligned foundation models carry the major semantic burden in this low-resource, high-noise environment. In particular, the Vision Transformer (ViT) had the largest F1-score (0.6278) on binary Hate Speech Detection

with exploiting the visual semiotics, and the SigLIP foundation model excelled at the multi-class Sentiment Analysis subtask (0.5481 F1-score) with the strong cross-modal synergy. These findings finally show that the performance of noisy OCR-extracted text with powerful visual features through standard late fusion is actually worse, which emphasizes the importance of noise-resilient text encoders or more intensive utilization of vision-language models with inherent alignment on low-resource languages.

8. Limitations

Despite establishing encouraging baselines, our study presents several limitations. The most prominent bottleneck stems from OCR transcription errors and the casual, code-mixed nature of Nepali internet slang, which significantly degraded the quality of the textual embeddings utilized by our early-fusion models. We did not employ explicit text-denoising or image-restoration techniques prior to text extraction. This decision was twofold: first, there is a current lack of robust, off-the-shelf denoising or spell-correction pipelines optimized for code-mixed, low-resource Nepali internet data. Second, within meme culture, visual noise (e.g., low resolution, grain, compression artifacts, or "deep-fried" effects) is frequently considered part of the intentional aesthetic. Aggressively removing these artifacts to improve OCR accuracy risks destroying subtle semiotic cues that heavily influence the meme's overall sentiment or hateful intent.

Furthermore, our interpretability analysis indicates that visual models, such as ViT, are over-reliant on specific visual markers like emojis. Consequently, these models often fail to identify malicious intent in highly contextual memes where the subtle synergy between text and image is the primary driver of the sentiment.

Finally, our experiments were constrained by a relatively small training dataset. This data scarcity likely contributed directly to the underperformance of our explicit early-fusion architectures compared to unimodal baselines and pre-aligned foundation models, the fusion networks simply lacked sufficient examples to effectively learn complex cross-modal alignments from scratch. This constrained our models' capacity to detect nuanced inter-modal semantic dissonance and generalize across the rapidly evolving landscape of Nepali digital culture. To bypass the fragile text-extraction phase entirely and mitigate OCR-induced noise without destroying the underlying visual aesthetic, future research should prioritize the use of true OCR-free Vision-Language Foundation Models (such as Donut or PaliGemma). These specialized, end-to-end architectures represent the most promising primary next step for multimodal comprehension in low-resource, high-noise environments.

9. References

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

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.

Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. 2020. Albumentations: Fast and flexible image augmentations. *Information*, 11(2).

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John P. McCrae. 2020. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210, Marseille, France. European Language Resources association.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Anand Kumar M, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan R L, John Philip McCrae, and Elizabeth Sherly. 2021. Findings of the shared task on offensive language identification in Tamil, Malayalam, and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 133–145, Kyiv. Association for Computational Linguistics.

Alexis Conneau et al. 2020. Unsupervised cross-lingual representation learning at scale. *ACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL*.

Alexey Dosovitskiy et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*.

Kaiming He et al. 2016. Deep residual learning for image recognition. *CVPR*.

Simran Khanuja et al. 2021. Muril: Multilingual representations for indian languages. *EMNLP*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020a. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624. Curran Associates, Inc.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020b. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624.

Zhuang Liu et al. 2022. A convnet for the 2020s. *CVPR*.

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.

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)*.

Chiranjibi Sitaula, Anish Basnet, Ashish Mainali, and Tej Bahadur Shahi. 2021. Deep learning-based methods for sentiment analysis on nepali covid-19-related tweets. *Computational Intelligence and Neuroscience*, 2021(1):2158184.

Surendrabikram Thapa, Kritesh Rauniyar, Shuvam Shiwakoti, Sweta Poudel, Usman Naseem, and Mehwish Nasim. 2023. Nehate: Large-scale annotated data shedding light on hate speech in nepali local election discourse. *Frontiers in Artificial Intelligence and Applications*, 372:2346–2353.

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. 2025a. 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. 2025b. Cross platform multimodal retrieval augmented distillation for code-switched content understanding. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2042–2051.

Xiaohua Zhai et al. 2023. Sigmoid loss for language image pre-training. *ICCV*.

Appendix A. Detailed Experimental Setup

Model Type	Learning Rate / LoRA
Fusion	1×10^{-5} or LoRA ($r = 16$, $\alpha = 32$)
Image-Only	1×10^{-5} or LoRA ($r = 16$, $\alpha = 32$)
Text-Only	LoRA Fine-tuning ($r = 16$, $\alpha = 32$)
Classifier	1×10^{-4} (All models)

Table 5: Model-specific learning rates and LoRA parameters.