When Meaning Travels: A Granular Lens on Hybrid-MoE's Role in Idiomatic Understanding for Language Models

Sarmistha Das[1*], **Vaibhav Vishal**[1*], **Shreyas Guha**[1*], **Amaan Ali**[1*],
Kitsuchart Pasupa[2], **Sriparna Saha**[1]
[1]Department of Computer Science and Engineering, Indian Institute of Technology Patna, India
[2]School of Information Technology, King Mongkut's Institute of Technology Ladkrabang, Thailand
{sarmistha1515, vvaibhav728, shreyas.slg, ali2003.amaan}@gmail.com
kitsuchart@it.kmitl.ac.th, sriparna@iitp.ac.in

Abstract

In the contemporary epoch of multilingual education, learning idioms provides a fascinating gateway towards creativity, cultural values, historical context, and diverse perspectives inherent to various linguistic traditions. This paper showcases the navigation of retaining figurative and cultural semantics in low-resource Southeast Asian languages such as Hindi, Bengali, and Thai, where culturally rich idioms pose significant obstacles for computational modeling and cross-linguistic transfer due to their deep metaphorical complexity. To tackle such complexity, we present *Varnika* (वर्णिका), a reconstructed multimodal idiom corpus comprising 3,533 multilingual idioms, enriched with seven idiomatic tones aligned with both textual and visual representations. Additionally, to infer informative idiomatic understanding, we introduce a Hybrid Mixture-of-Experts (HybridMoE) framework that embeds multiple idiomatic expert opinions while mitigating expert sparsity by integrating outputs from both selected and unselected experts through controlled hybridization, further augmented with Idiomatic Property Signals via masked multimodal embeddings. To analyze the performance across multiple dimensions, we propose the *IDIO-TONE* and Idiomatic Validation Score, a three-stage evaluation pipeline measuring (i) literal translation fidelity, (ii) visual-semantic alignment, and (iii) idiomatic meaning retention. Empirical evaluations highlight that HybridMoE achieves 5–6% performance gains across advanced vision language models, demonstrating improved representation of figurative language and culturally embedded meaning in multilingual multimodal settings[1].

1 Introduction

Idioms constitute a pivotal conduit for expressive language, encapsulating the figurative essence of human experience, culture, history, and creativity. Cross-linguistic idiom learning plays a critical role in enhancing language proficiency, deepening cultural literacy, and strengthening both semantic awareness and affective engagement with language (Liontas, 2017). Despite rapid advances in large language models and their growing integration into educational and social ecosystems (Kasneci et al., 2023; Nayeem and Rafiei, 2024; Rong et al., 2024; Maji et al., 2025a,b), accurately decoding culture-specific lexical patterns and capturing figurative metaphors grounded in linguistic convention remains a persistent challenge (Liu et al., 2025). Languages across South and Southeast Asia, including Hindi, Bengali, and Thai, exhibit shared phonological traits and deeply rooted cultural motifs, often giving rise to semantically parallel idiomatic expressions. For example, the Hindi idiom एक पत्थर से दो शिकार (*ek patthar se do shikār*, "two outcomes from a single action") finds close counterparts in Bengali এক পাথর দুই পাখি মারা (*ek pāthor dui pākhi mārā*) and Thai ยิงปืนนัดเดียวได้นกสองตัว (*ying peun nát diao dâi nók sŏng tua*). Nevertheless, idioms fundamentally resist compositional interpretation; their meanings are intrinsically tied to cultural grounding, shared world knowledge, and conventionalized contextual usage (Sporleder and Li, 2009; Fornaciari et al., 2024). Consequently, figurative language understanding has emerged as a central research theme in contemporary Natural Language Processing (NLP), with substantial progress in related tasks such as metaphor identification (Dagan et al., 2005; Gao et al., 2018; Chakrabarty et al., 2021), simile detection (Niculae and Danescu-Niculescu-Mizil, 2014; Mpouli, 2017; Zeng et al., 2020), pun recognition (Poliak et al., 2018), and idiom retrieval (Tan et al., 2016; Lee et al., 2016). Despite these advances, a unified and comprehensive evaluation frame-

[1]Resources are available at https://github.com/sarmistha-D/Hybrid_MOE.
[*] These authors contributed equally.

Findings of the Association for Computational Linguistics: ACL 2026, pages 33229–33245

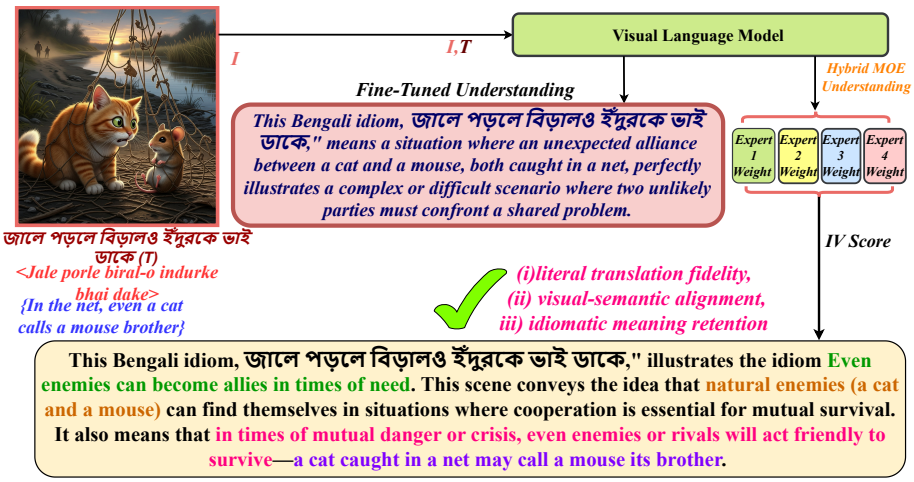

Figure 1: Visual Representation of Idiomatic Understanding via Hybrid Mixture-of-Experts.

work for idiomatic interpretation remains largely absent. FLUTE (Chakrabarty et al., 2022) introduced a large-scale figurative language inference benchmark spanning sarcasm, simile, metaphor, and idiom understanding, while its multimodal extension, V-FLUTE (Saakyan et al., 2024), further demonstrated the critical role of visual–linguistic grounding in robust idiom comprehension.

However, both fall short in addressing the multilingual and cultural complexities of low-resource languages such as Thai, Bengali, and Hindi, resulting in a persistent gap in figurative understanding across languages (Haagsma et al., 2020). While scaling enhances model performance, dense architectures remain computationally expensive. Mixture-of-Experts (MoE) frameworks (Shazeer et al., 2017) offer a solution by activating only a small subset of expert modules for each input (Artetxe et al., 2021; Du et al., 2022), making it possible to scale efficiently. MoE routes tokens through a small subset of experts, reducing computational load while preserving performance. In pragmatically rich tasks such as idiom explanation, relying solely on top-k expert selection may constrain expressiveness. Leveraging HyperNetworks (Zhao et al., 2023) enables cross-expert knowledge sharing, enhancing task-specific representation and semantic transfer as depicted in Figure 1. However, preserving multimodal information in idioms, where the literal and intended meanings are semantically divergent, remains a challenging task.

Despite the growing interest in idiomatic understanding, cross-lingual research that jointly incorporates visual representations and idiomatic tones remains limited, particularly for culturally rich languages such as Thai, Bengali, and Hindi (Table 1). This gap is critical, as idioms encapsulate nuanced intercultural meaning and social cognition, yet no existing dataset provides visually grounded idioms suitable for both educational and cross-modal learning. To address this, we introduce *Varnika* (वर्णिका), a multimodal idiom corpus with seven idiomatic tones. Building on this dataset, we propose a Hybrid Mixture-of-Experts (HybridMoE) framework that integrates expert-specific modules with a globally shared hypernetwork to capture diverse idiomatic properties while enabling positive expert transfer. To further enhance contextual fidelity, HybridMoE is augmented with an Idiomatic Property Signal (IPS) that selectively conditions cross-modal embeddings, and an Idiomatic Validation (IV) score that systematically evaluates literal fidelity, visual–semantic alignment, and idiomatic meaning retention. Additionally, we propose *IDIO-TONE*, a label-based quality assurance and performance metric for emotional and stylistic consistency of idioms.

The research objectives of the current work are as follows:

(i) Evaluate how Vision-Language Models (VLMs) internalize culturally grounded idiomatic knowledge, and analyze the impact of HybridMoE with IPS integration on enhancing fine-grained cultural inference and interpretative accuracy.

(ii) Assess the cross-lingual and cross-architectural generalization capabilities of both the proposed dataset and models,

Table 1: Comparison of the proposed *Varnika* dataset with leading existing resources. Abbreviations: E (English); ML (Multilingual: Thai, Hindi, and Bengali). Multimodality indicates the inclusion of paired text–image data.

Corpus Name	Count	Language	Explanations	Multimodality	Tone	Idiomatic Part
SemEval-2013 (Korkontzelos et al., 2013)	4,350	E	×	×	×	×
FLUTE (Chakrabarty et al., 2022)	8,962	E	✓	×	×	✓(497)
V-FLUTE (Saakyan et al., 2024)	6,027	E	✓	✓	×	✓(370)
MAGPIE (Haagsma et al., 2020)	56,622	E	✓	✓	×	✓(1,756)
IRFL (Yosef et al., 2023)	6,027	E	×	✓	×	✓(628)
Varnika (वर्णिका)(Proposed)	3,533	ML	✓	✓	✓	3,533

evaluating their ability to adapt to diverse linguistic settings and unseen idiomatic constructs.

The primary contributions of this work are as follows:

(i) The proposal of HybridMoE, a multimodal hypernetwork-based framework equipped with IPS designed to enhance reasoning over idiomatic constructs, alongside the formulation of the novel Idiomatic Validation (IV) and *IDIO-TONE* metric to quantify idiomatic understanding.

(ii) The design and execution of two core tasks: (a) multimodal idiomatic interpretation leveraging state-of-the-art VLMs, and (b) evaluation of the HybridMoE mechanism for performance enhancement, aimed at benchmarking the effectiveness of the proposed dataset and model design.

(iii) We reconstruct the existing multimodal idiom dataset *Mediom* (Das et al., 2026) by augmenting each instance with fine-grained idiomatic tonal annotations, and present the resulting resource as *Varnika* (वर्णिका). This enhanced dataset targets low-resource languages such as Hindi, Bengali, and Thai and incorporates seven pragmatic idiomatic tones: *Humor*, *Ridicule*, *Affection*, *Aspiration*, *Fear*, *Sorrow*, and *Deception*.

2 Formulation of *Varnika* (वर्णिका) dataset.

2.1 Dataset Collection

We reconstruct *Mediom* (Das et al., 2026), a multilingual, multimodal idiom dataset comprising 3,533 expressions across Hindi, Bengali, and Thai. The dataset is curated from culturally grounded and linguistically diverse sources, including online repositories, literary compilations, and archival linguistic resources. Specifically, Hindi idioms are sourced from *The Simple Help*[2], Bengali idioms from *Bangla Probad*[3], and Thai idioms from a comprehensive collection (Udomporn, 2014), ensuring broad cultural coverage and authenticity across the three languages.

While *Mediom* provides broad multilingual and multimodal coverage, it does not explicitly capture idiomatic tonality, a key pragmatic dimension that often governs how idioms are interpreted, expressed, and visually grounded. This aspect is particularly important because idiomatic meaning extends beyond literal semantics to encode affective, social, and culturally situated intent. To address this gap, we reconstruct the dataset by augmenting each entry with a taxonomy of seven pragmatic tone labels: Humor, Ridicule, Affection, Aspiration, Fear, Sorrow, and Deception. This refinement yields a richer and more nuanced resource for studying idiomatic understanding at the intersection of language, culture, and visual representation.

Mediom corpus spans a wide range of idiomatic constructions, including fixed expressions (e.g., น้ำท่วมปาก, *nam thuam pak*, unable to speak out), which often convey fear, hesitation, or emotional suppression; semi-fixed idioms (e.g., आसमान से गिरे, खजूर में अटके, *āsmān se gire, khajūr mein aṭke*, out of the frying pan into the fire), which encode sorrow, frustration, or helplessness; verb-object constructions (e.g., ঘি না পেয়ে নাক কাটা, *ghi nā peyē nāk kāṭā*, to overreact over a loss), which frequently evoke ridicule or social mockery; adjective-noun phrases (e.g., ฝันหวาน, *fan wān*, sweet dreams), which reflect affection, tenderness, and aspiration; prepositional idioms (e.g., কানে কানে বলা, *kāne*

[2] https://thesimplehelp.com/hindi-idioms-with-meanings-and-sentences

[3] https://archive.org/details/in.ernet.dli.2015.455639/page/n557/mode/2up

kāne balā, to whisper), which may express deception, secrecy, or intimate affection depending on context; and binomial formations (e.g., धीरे धीरे, *dhīre dhīre*, step by step), which convey gradual progress, patience, and aspiration. More broadly, several curated idioms instantiate humor through playful exaggeration, ridicule through sarcastic social commentary, fear through cautionary imagery, sorrow through expressions of loss and resignation, and deception through motifs of concealment, misdirection, and duplicity.

Motivated by the need to capture these pragmatically salient dimensions, we reconstruct *Mediom* (Das et al., 2026) into a more expressive and semantically enriched resource, which we name *Varnika* (वर्णिका). Derived from the notion of varnan (description or depiction), *Varnika* is designed to foreground not only the figurative semantics of idioms but also their tonal, cultural, and affective portrayals, thereby yielding a more comprehensive multimodal representation of idiomatic meaning.

Syntactically flexible idioms were normalized by standardizing verb inflections and pronouns, whereas rigid forms were retained in their original surface realization to preserve cultural authenticity. A representative view of *Varnika* is presented in Figure 2.

2.2 Data Quality Assurance

To ensure annotation quality and reliability, each instance underwent a two-stage validation process. First, annotations were peer-reviewed by at least two additional fluent annotators to minimize subjective bias and enhance consistency. To further assess cross-modal pragmatic alignment, we quantified the overlap between tone labels assigned from textual interpretations and those derived from corresponding visual representations, yielding an overlap of 54.75%, indicating strong alignment between idiomatic tonal intent and visual grounding (Please refer to Appendix A.2.1). Subsequently, a panel of cultural and linguistic experts performed a comprehensive final validation, evaluating each instance against the predefined tonal alignment criteria and assigning scores based on the degree of compliance. The overall annotation process achieved a substantial inter-annotator agreement of 0.65 (Cohen's kappa (Gwet, 2014)), demonstrating a high level of consistency and reliability in the resulting dataset.

3 Methodology

Given an idiom instance represented as a multimodal pair (x_t, x_i), where x_t denotes the textual idiom expression and x_i corresponds to its associated visual depiction, this study proposes a unified framework for idiomatic interpretation through a HybridMoE architecture coupled with an IV (Idiomatic Validation) mechanism. The dataset comprises two aligned sets: textual explanation pairs $\mathcal{D}_t = \{(x_{t_j}, y_j)\}_{j=1}^{N}$ and image explanation pairs $\mathcal{D}_i = \{(x_{i_j}, y_j)\}_{j=1}^{N}$, where N denotes the total number of idioms and y_j represents the idiomatic meaning.

The proposed framework operates in two sequential stages, as illustrated in Figure 3. The first stage focuses on idiom interpretation, employing HybridMoE to model and integrate semantic features from both textual and visual cues. The second stage introduces the IV (Idiomatic Validation), a tri-level evaluation mechanism designed to assess idiomatic understanding.

3.1 Formulation of HybridMoE

Given an idiomatic expression T and its corresponding image I, the proposed framework begins by encoding I using a ViT-patch16 based vision encoder (Dosovitskiy et al., 2020) to obtain visual embeddings E_i, and encoding T with a BERT-based text encoder (Devlin et al., 2019) to produce textual embeddings E_t. The resulting hidden representations are denoted as $H_i \in \mathbb{R}^{B\times N\times d}$ and $H_t \in \mathbb{R}^{B\times L\times d}$, where B is the batch size, N the number of image patches, L the tokenized text length, and d the hidden dimensionality. From these, the CLS tokens are extracted via attention mechanisms as E_i and E_t each in $\mathbb{R}^{B\times d}$. To encode idiomatic specificity, the CLS tokens are enriched via IPS, capturing idiom-aware visual and textual cues (see Figure 3), replacing the original embeddings. The IPS-enhanced embeddings are independently propagated through $M = 3$ parallel Hybrid-MoE modules. Each module consists of $K = 4$ expert networks with a shared architecture $\mathbb{R}^d \to \mathbb{R}^{1024} \to \mathbb{R}^d$, and a trainable gating function defined as $g(E) = \text{Softmax}(EW_g + b_g) \in \mathbb{R}^{B\times 4}$, which dynamically computes expert weights $\alpha_1, \ldots, \alpha_4$. Within each block, the expert outputs are aggregated as $\mathbf{E} = [K_i(E)]_{i=1}^{4} \in \mathbb{R}^{B\times 4\times d}$, and further refined via a mean-pull strategy to yield an enhanced representation as $\text{Mean}(\hat{y}_{i,t})$ and subsequently passed to a fusion layer for final integra-

Figure 2: Sample instances of our proposed *Varnika* (वर्णिका) dataset.

tion.

In the fusion block, we employ Einstein summation (EinSum) to compute a weighted aggregation of the expert outputs. Specifically, each expert's adjusted output is multiplied by its corresponding gating weight and summed across the K experts to produce a single fused vector $\hat{y} \in \mathbb{R}^{B\times d}$ per Hybrid-MoE module. Upon obtaining the outputs from all M parallel modules, we perform an element-wise averaging to derive a unified representation. This aggregated representation is then combined $\{\hat{E}_t, \hat{E}_i\}$, effectively integrating modality-specific idiomatic signals. The fused embedding $\hat{E}_{\text{fused}}$ is subsequently passed to the decoder of the VLM. Each decoder layer performs cross-modal attention using the standard scaled dot-product formulation, defined as $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$, where $Q = H_{\text{dec}}W_Q$, $K = H_f W_K$, $V = H_f W_V$, and $d_k = \frac{d}{H}$, with H denoting the number of attention heads. Finally, the decoder generates a target sequence $\hat{y} \in \mathbb{Z}^{B\times L'}$, resulting in a descriptive output D that encapsulates the idiomatic interpretation, effectively leveraging multimodal alignment and expert-guided reasoning.

3.2 Novel Evaluation Metrics

To systematically evaluate a model's multimodal understanding of idiomatic expressions, we introduce two metrics.

3.2.1 Idiomatic Validation (IV) Score

The IV Score is a structured three-stage framework that assesses (i) literal translation fidelity, (ii) visual-semantic alignment, and (iii) idiomatic meaning retention. Given an idiomatic description d, a pretrained VLM, specifically Qwen2.5-VL-7B, is used as a zero-shot evaluator to generate a meaning pair $\tau = \langle \xi, \mu \rangle$, where ξ de-

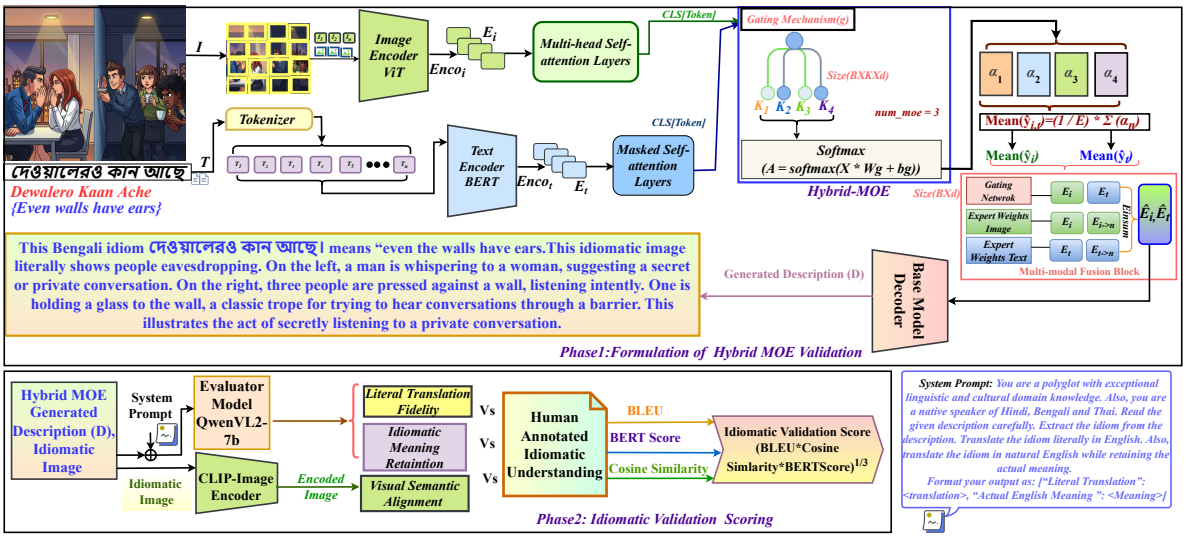

Figure 3: Architectural Viewpoint of Proposed Multimodal HybridMoE Model.

notes the literal translation and μ captures its inferred meaning. The pair is sampled via $\mathcal{T} \sim \pi_{\theta'}(\tau \mid h_1, d)$, where h_1 is a system prompt guiding Qwen-VL's behavior. The fidelity score λ is computed as the BLEU score between ξ and the reference literal translation $\texttt{trans}(I)$, i.e., $\lambda(\xi, \texttt{trans}(I)) = \text{BLEU}(\xi, \texttt{trans}(I))$. Given the idiomatic image ν, visual-semantic alignment δ with the idiom I is computed as the cosine similarity between CLIP encodings of image $\phi_v(\nu)$ and idiom $\phi_t(I)$, normalized to the $[0, 1]$ range: $\delta(\nu, I) = \frac{\cos(\phi_v(\nu), \phi_t(I)) + 1}{2}$. Idiomatic meaning retention ρ is quantified using BERTScore between the generated meaning μ and the ground truth: $\rho(\mu, \texttt{groundtruth}) = \text{BERTScore}(\mu, \texttt{groundtruth})$. The final IV Score is computed as the geometric mean of the three components to penalize underperformance in any one modality: $\texttt{IV\_Score}(d) = (\lambda \cdot \delta \cdot \rho)^{1/3}$. During each validation stage, Qwen2.5-VL-7B performs zero-shot inference and selects the most appropriate candidate via internal voting mechanisms, ensuring contextual fidelity and robustness in evaluation.

3.2.2 *IDIO-TONE* Score

We use the evaluation dataset for generating *IDIO-TONE* scores for our models (refer to Figure 7 in Appendix A.2). The dataset is defined as $\mathcal{D} = \{(t_i, v_i, L_i)\}_{i=1}^{N}$, where for the i-th instance, t_i represents the idiom, v_i represents the visual input (image) and $L_i \subseteq \mathcal{C}$ is the set of ground-truth idiomatic tone labels associated with that image. The predefined label space is denoted as $\mathcal{C}$ = {Humor, Ridicule, Affection, Aspiration, Fear, Sorrow, Deception}. Given the idiom t_i and the visual input v_i, our trained VLM, denoted by f_θ, generates a tuple consisting of the predicted idiom translation $\hat{t}_i$ and its corresponding explanation $\hat{e}_i$, i.e., $(\hat{t}_i, \hat{e}_i) = f_\theta(t_i, v_i)$. To map the generated output back to the discrete label space $\mathcal{C}$, we employ a distilled DeepSeek-R1 (Guo et al., 2025) model as an evaluator, denoted by g_ϕ, which processes the generated explanation to extract the predicted set of idiomatic labels $\hat{L}_i$, given by $\hat{L}_i = g_\phi(\hat{e}_i)$, where $\hat{L}_i \subseteq \mathcal{C}$.

To quantify the alignment between the predicted idiomatic tone labels and the gold standard, we compute a set-based F_1 score for each instance. For the i-th instance, the F_1 score is defined as the harmonic mean of precision and recall over the label sets. To handle the edge case where both the ground-truth and predicted sets are empty, the formulation is defined as:

$$F_1^{(i)} = \begin{cases} 1, & \text{if } |L_i| = 0 \ \& \ |\hat{L}_i| = 0, \\ \frac{2|L_i \cap \hat{L}_i|}{|L_i| + |\hat{L}_i|}, & \text{otherwise.} \end{cases} \quad (1)$$

Finally, the overall *IDIO-TONE* Score, $S_{\text{IDIO-TONE}}$, for the model f_θ over the dataset $\mathcal{D}$ is calculated as the sample-averaged F_1-score:

$$S_{\text{IDIO-TONE}} = \frac{1}{N} \sum_{i=1}^{N} F_1^{(i)}. \quad (2)$$

4 Experiments & Resultant Discussion

This section outlines the experimental setup, baseline configurations, and a head-to-head evaluation

of VLMs. We supply a qualitative error analysis that pinpoints interpretive shortcomings and highlights unresolved research challenges. Our study addresses two focal research questions (RQs):

- **RQ1**: To what extent does HybridMoE, in concert with the IPS module, improve idiom comprehension and contextual reasoning while preserving or boosting overall metrics?
- **RQ2**: What are the societal implications and cross-lingual generalizability of the proposed dataset?

The models were fine-tuned using a learning rate of 2×10^{-4} on 3 epochs on an NVIDIA A100 GPU with 80GB VRAM. Training utilized a batch size of 4 with gradient accumulation (8 steps) to manage GPU memory efficiently. We adopted the fused AdamW optimizer to enhance generative diversity. A constant learning rate scheduler was applied for stable convergence and mixed-precision training was enabled for computational efficiency. We fine-tuned the model using PEFT with LoRA and supervised fine-tuning (SFT) to enable efficient adaptation during training and evaluation. The dataset was split into 70% for training, 20% for validation, and 10% for testing purposes. The dataset comprises 1,277 Hindi, 1,751 Thai, and 505 Bengali idiomatic samples. A comparative evaluation was performed across a range of prominent VLMs, including Blip2-7B (Li et al., 2023), Qwen2.5-VL-7B-Instruct (Qwen2.5VL, 2025), SmolVLM-Instruct (Marafioti et al., 2025), Paligemma2-10B (Steiner et al., 2024), Llava-1.5-7B (Liu et al., 2024), Gemma3-12b-pt (Gemma, 2025). Model outputs were benchmarked using a comprehensive suite of evaluation metrics designed to capture lexical overlap, semantic alignment, and distributional similarity. Specifically, we employed ROUGE (R1, R2, R-L, and R-Lsum) and BLEU (B1–B3 and overall score BS) to assess lexical fidelity. Semantic and structural properties were evaluated via BERTScore (BTS). Finally, we measured linguistic complexity and clarity using the IV (Idiomatic Validation) Score and the *IDIO-TONE* Score.

4.1 Results and Discussion

This section synthesizes the findings in response to the stated RQs, supported by qualitative insights and error analyses across model generations.

4.1.1 Response to RQ1–Impact of HybridMoE

The ablation results in Table 2 reveal a marked elevation in idiomatic reasoning performance when transitioning from conventional fine-tuning to the proposed HybridMoE framework. Standard VLMs exhibit moderate gains in syntactic fidelity and surface-level lexical overlap; however, their ability to generalize over complex, culturally embedded figurative constructs remains constrained, particularly in low-resource multilingual contexts. By contrast, HybridMoE introduces a dynamic expert routing topology that enables token-level modulation across modality- and task-specialized subspaces, yielding significantly sharper representations for idiomatic abstraction and cultural disambiguation. Qwen2.5-VL-7B-Instruct + HybridMoE exhibits significant performance gains. ROUGE-L improves from 0.36 to 0.41, BLEU-3 from 16.51 to 21.48, and BERTScore from 0.93 to 0.95, indicating enhanced semantic precision and idiomatic fluency. Qwen2.5-VL-7B+HybridMoE has the highest IV (Idiomatic Validation) score of 0.82. Qualitative analysis further reveals improved figurative anchoring within culturally grounded narratives. Similar gains in SmolVLM-Instruct and Gemma3-12b-pt confirm that HybridMoE's modular design offers architecture-agnostic improvements by enabling cross-modal grounding, idiom-specific alignment, and reducing overfitting through shared expert knowledge. However, the *IDIO-TONE* results reveal a clear limitation in modeling idiomatic tone, with most models scoring around $\sim$0.30-0.35. Even the best-performing models, such as *Gemma3-12b-pt* (0.55) and *Qwen2.5-VL-7B-Instruct* (0.47), achieve only moderate improvements. This highlights a persistent gap in capturing pragmatic and culturally grounded nuances, indicating the need for more targeted learning strategies. All reported results are statistically significant at $p < 0.05$ (Welch, 1947).

Notably, the efficacy of the HybridMoE architecture, instantiated with Qwen2.5-VL-7B-Instruct, is further substantiated through evaluation on the multimodal V-FLUTE (Saakyan et al., 2024) corpus. As shown in Table 3, HybridMoE consistently yields significant performance gains over standard supervised fine-tuning (SFT).

Table 2: Performance variance between fine-tuned VLMs and HybridMoE-based models. Bold values denote the best scores; ↑ indicates metrics where higher is better, and ↓ denotes metrics where lower values are optimal.

Experimental Setting	Model Names	Metrics										
		R-1	R-2	R-L	R-LSum	B-1	B-2	B-3	BS	BTS	IVS	IDIO-TONE
VLM+Finetune	Blip2-7B	0.26	0.07	0.17	0.18	18.64	08.62	06.28	09.17	0.88	0.52	0.32
	Paligemma2-10B	0.39	0.15	0.27	0.29	32.62	15.73	08.95	13.73	0.88	0.53	0.32
	Llava-1.5-7B	0.35	0.15	0.26	0.26	27.46	12.74	08.20	11.40	0.90	0.57	0.31
	SmolVLM-Instruct	0.43	0.18	0.33	0.33	47.48	21.82	14.06	17.70	0.92	0.59	0.35
	Gemma3-12b-pt	0.4	0.15	0.29	0.29	32.86	13.36	07.90	11.64	0.90	0.74	**0.54**↑
	Qwen2.5-VL-7B-Instruct	**0.48**↑	**0.22**↑	**0.36**↑	**0.37**↑	**51.06**↑	**24.81**↑	**16.51**↑	**20.68**↑	**0.93**↑	**0.75**↑	0.46
VLM+HybridMoE	Blip2-7B	0.3	0.13	0.23	0.28	23.76	12.93	13.29	15.64	0.89	0.55	0.33
	Paligemma2-10B	0.45	0.2	0.34	0.35	35.97	17.25	14.2	15.94	0.89	0.59	0.32
	Llava-1.5-7B	0.4	0.17	0.32	0.34	32.52	16.64	12.95	14.25	0.91	0.61	0.33
	SmolVLM-Instruct	0.48	0.21	0.36	0.36	50.46	24.16	16.59	20.92	0.92	0.66	0.35
	Gemma3-12b-pt	0.46	0.19	0.35	0.35	38.81	17.82	11.93	14.58	0.92	0.79	**0.55**↑
	Qwen2.5-VL-7B-Instruct	**0.54**↑	**0.32**↑	**0.41**↑	**0.43**↑	**56.32**↑	**28.92**↑	**21.48**↑	**24.19**↑	**0.95**↑	**0.82**↑	0.47

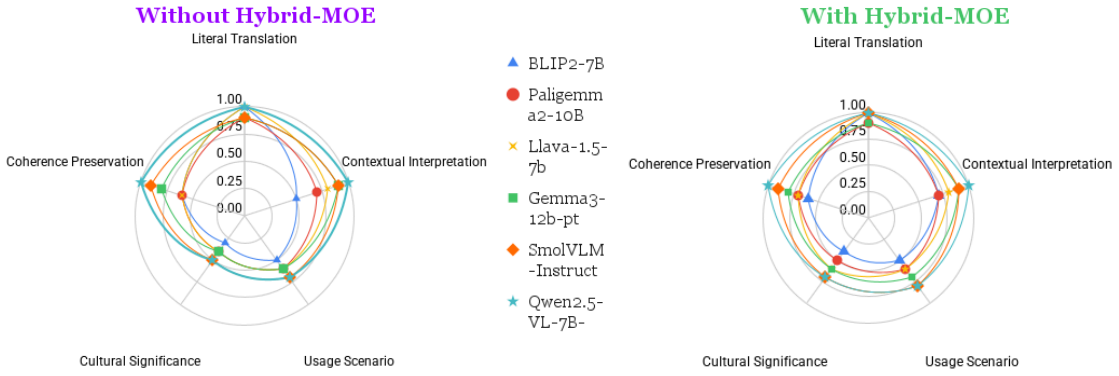

Figure 4: Expert Evaluation of Idiom Understanding Performance of VLM models with Idiomatic Property Retention (IPR) Criteria.

4.1.2 Response to RQ2–Generalizability and Societal Relevance

The *Varnika* corpus serves as a comprehensive multimodal benchmark for idiomatic understanding across Hindi, Thai, and Bengali. As evidenced in Table 2, even VLMs not explicitly designed for multilingual settings demonstrate strong adaptability to *Varnika*, achieving effective idiomatic interpretation under both fine-tuning and HybridMoE configurations. Furthermore, as shown in Table 3, language-specific and relatively low-resource models such as *Ganga* (Hindi) (Group and IITGN, 2025) and *Typhoon* (Thai) (Pipatanakul et al., 2023) also exhibit the ability to infer idiomatic meanings when trained on *Varnika*. However, *TitulLaMA* (Bengali) (Nahin et al., 2025) underperforms, likely due to its pretraining bias toward Bengali combined with the relatively lower representation of Bengali idioms in the dataset. Notably, despite their smaller scale, these language-specific models achieve *IDIO-TONE* scores comparable to larger LLMs and VLMs (Table 2), suggesting that idiomatic tone understanding is not solely dependent on model size but also on alignment with linguistic and cultural context. These findings demonstrate the strong generalizability of *Varnika* across diverse model architectures and linguistic settings. Beyond technical performance, *Varnika* holds substantial educational and societal value.

4.2 Analytical Discussion

4.2.1 Human Evaluation

We conducted a human evaluation involving three native-speaking cultural experts with domain expertise in Hindi, Bengali, and Thai, who assessed 709 instances using Information Persistence Ratings (IPR) (Das et al., 2026) across five dimensions: literal accuracy, contextual fit, usage naturalness, cultural depth, and overall coherence. As depicted in Figure 4, models enhanced with HybridMoE consistently outperformed their finetuned counterparts. Qwen2.5-VL-7B-Instruct + HybridMoE showed the most significant gains, particularly in contextual interpretation (+0.1), cultural significance (+0.2), and usage scenario (+0.1). These improvements

Table 3: Additional Experiments on language-specific LLMs and also SFT on Flute and V-Flute dataset.

Experimental Setting	Model Names	Metrics										
		R-1	R-2	R-L	R-LSum	B-1	B-2	B-3	BS	BTS	IVS	IDIO-TONE
SFT on *Varnika*	Ganga-2-1B	0.314	0.127	0.268	0.269	33.56	13.13	7.83	12.84	0.83	0.32	0.31
	Titulm-Llama-3.2-1B	0.025	0.01	0.018	0.02	2.13	0.16	0.031	0.15	0.8	0.25	0.28
	Typhoon-7B-Instruct	0.293	0.18	0.248	0.249	19.32	12.32	9.53	12.11	0.88	0.51	0.46
SFT on Flute	Qwen2.5-VL-7B-Instruct	0.39	0.21	0.33	0.33	27.43	16.77	11.92	18.41	0.84	0.58	-
SFT on V-flute	Qwen2.5-VL-7B-Instruct	0.34	0.19	0.22	0.26	21.14	11.82	9.02	15.59	0.75	0.44	-
HybridMoE on V-flute	Qwen2.5-VL-7B-Instruct	0.43	0.21	0.35	0.35	46.59	18.13	9.80	14.34	0.89	0.59	-

(a) Qualitative Analysis of the Hindi idiom सोने पे सुहागा (*Sone pe Suhāgā*; literal meaning: gold with a touch of *suhāgā*, a traditional polishing substance; metaphorical meaning: an added bonus or improvement to something already good)

(b) Error Analysis of the Hindi idiom दफा होना (*Dafā honā*; literal meaning: to disappear or go away; metaphorical meaning: dismissal or rejection, implying exclusion)

Figure 5: Comparative Qualitative and Error Analyses of Idiom Understanding in Vision Language Models

stem from HybridMoE's dynamic expert routing and task-specific reasoning, enabling better disentanglement of semantic, cultural, and contextual representations. In contrast, conventional fine-tuning lacks such specialization, leading to shallow generalizations. Consistent patterns observed in SmolVLM-Instruct and LLaVA-1.5-7B further substantiate HybridMoE as a resilient framework for cross-modal idiomatic reasoning in low-resource, culturally nuanced settings.

4.2.2 Qualitative Analysis

Figure 5a highlights that models integrated with HybridMoE exhibit significantly improved idiomatic understanding. While finetuned models tend to rely on object-centric cues (e.g., gold, gems) and fail to capture the relational enhancement (good → better), HybridMoE-augmented models perform stronger semantic alignment and figurative abstraction. Interestingly, Qwen2.5-VL-7B-Instruct + HybridMoE, SmolVLM-Instruct + HybridMoE, and Gemma3-12b-pt + HybridMoE accurately capture both the literal origin and idiomatic meaning, often producing coherent visual metaphors of transformation or enhancement. These improvements stem from HybridMoE's dynamic expert routing, which activates culturally informed reasoning pathways for precise and context-aware idiom interpretation.

4.2.3 Error Analysis

Figure 5b illustrates a failure case where current VLMs struggle to grasp the idiomatic meaning of दफा होना. These models tend to rely on shallow lexical or visual cues, produce interpretations that miss the cultural and contextual depth of the idiom. For example, SmolVLM-Instruct and LLaVA-1.5-7B generated responses such as "to be on the right path" or "to go through a cycle," which are far from the idiom's actual meaning of disappearance or being dismissed. These mismatches highlight the shortcomings of standard fine-tuning when it comes to handling figurative, culturally grounded language. In contrast, Qwen2.5-VL-7B-Instruct enhanced with HybridMoE effectively captures both the literal and figurative aspects, producing a meaningful visual metaphor, such as pushing away shadowy figures that accurately reflect the idiom's true intent.

5 Conclusion

This paper introduces a HybridMoE architecture that enhances idiomatic understanding with IPS in VLMs through task-adaptive expert routing. Alongside this framework, we present *Varnika* (वर्णिका), a reconstructed extension of the multimodal idiom dataset *Mediom* (Das et al., 2026), enriched with fine-grained annotations across seven idiomatic tones such as *Humor*, *Ridicule*, *Affection*, *Aspiration*, *Fear*, *Sorrow*, and *Deception*. Empirical results show improved lexical abstraction, figurative disambiguation, and cultural grounding in models, such as Qwen2.5-VL-7B, validated by the proposed IV (Idiomatic Validation) and *IDIO-TONE* scores. *Varnika* advances inclusive, culturally aware AI while enabling applications in creative modeling and education for social good.

6 Limitations

Despite the improvements introduced by the HybridMoE framework and the *Varnika* dataset, several limitations remain:

- **Dependence on Visual Grounding:** The idiomatic reasoning capabilities of current VLMs remain sensitive to the availability of meaningful visual cues. For highly abstract or culturally implicit idioms lacking clear visual metaphors, models often default to surface-level or literal interpretations. This exposes a limitation of HybridMoE, which, despite enhanced feature routing, does not fully bridge deeper figurative reasoning gaps.

- **Subjectivity in IDIO-TONE Annotations:** The proposed seven-tone taxonomy captures broad pragmatic dimensions; however, idiomatic tone interpretation is inherently subjective. Fine-grained emotional nuances and hierarchical relationships between tones are not explicitly modeled, which may limit the expressiveness and consistency of the *IDIO-TONE* score.

- **Limitations of IV (Idiomatic Validation) Score:** The Idiomatic Validation (IV) score, while effective in combining literal fidelity, semantic alignment, and multimodal grounding, is an aggregated metric. It may obscure component-level weaknesses, thereby reducing the interpretability of model-specific strengths and failure modes.

- **Static Evaluation Setting:** The dataset and evaluation framework are static, whereas idiomatic meaning is dynamic and context-dependent. This limits the assessment of model performance in real-world conversational or evolving discourse scenarios.

While HybridMoE enhances idiomatic abstraction through expert specialization, it does not fully bridge the gap between cultural reasoning and perceptual grounding. The results indicate that idiomatic understanding remains challenging for VLMs without explicit cultural or visual cues, highlighting the need for integrating external knowledge and advanced reasoning mechanisms beyond static image-text alignment.

7 Ethical Considerations

While *Varnika* enhances the *Mediom* dataset through the integration of fine-grained idiomatic tonal annotations, several ethical considerations arise from the process of tone-centric labeling:

1. **Subjectivity in Tonal Interpretation:** Idiomatic tone assignment inherently involves subjective judgment, as tones such as *Humor*, *Ridicule*, or *Affection* may be perceived differently across individuals and cultural contexts. Despite expert validation and annotation guidelines, residual interpretative variability may influence label consistency.

2. **Cultural Sensitivity and Misrepresentation:** Idioms are deeply embedded in cultural, historical, and social contexts. Assigning discrete tonal categories risks oversimplifying or misrepresenting nuanced cultural meanings, particularly for expressions that carry layered or context-dependent interpretations across Hindi, Bengali, and Thai.

3. **Multi-Label Ambiguity and Overlap:** Many idioms naturally express multiple overlapping tones (e.g., *Ridicule* intertwined with *Humor* or *Sorrow*). While the multi-label framework captures this complexity, it may still fail to fully represent subtle tonal gradations or hierarchical relationships between tones.

4. **Bias in Tonal Distribution:** The frequency and distribution of selected *IDIO-TONE* categories may reflect underlying biases in data sources or annotation practices, potentially leading to over-representation of certain emotional or pragmatic tones while under-representing others.

5. **Static Representation of Dynamic Semantics:** Idiomatic tones are context-sensitive and may evolve across discourse, media, and cultural settings. The current dataset provides a static snapshot of tonal interpretations, which may not fully capture the dynamic and evolving nature of idiomatic usage in real-world communication.

8 Future Work

In future work, we aim to extend *Varnika* by expanding its coverage to a broader range of Asian

languages, enabling richer cross-lingual and cultural analysis of idiomatic understanding. We also plan to explore reinforcement learning (RL)-based frameworks to model idiomatic reasoning as a dynamic, feedback-driven process within multimodal settings. Additionally, we seek to enhance idiomatic interpretation by incorporating more refined semantic and pragmatic modeling, including context-aware reasoning and structured tone representations.

9 Acknowledgement

All the authors sincerely thank the dataset annotators Tannu, Jheel, and Paavnee for their valuable contributions in curating and validating the idiomatic tone annotations. Their efforts were instrumental in ensuring the quality and reliability of the *Varnika* dataset.

References

Mikel Artetxe, Shruti Bhosale, Naman Goyal, Todor Mihaylov, Myle Ott, Sam Shleifer, Xi Victoria Lin, Jingfei Du, Srinivasan Iyer, Ramakanth Pasunuru, and 1 others. 2021. Efficient large scale language modeling with mixtures of experts. *arXiv preprint arXiv:2112.10684*.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. Flute: Figurative language understanding through textual explanations. *arXiv preprint arXiv:2205.12404*.

Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan, and Nanyun Peng. 2021. Mermaid: Metaphor generation with symbolism and discriminative decoding. *arXiv preprint arXiv:2103.06779*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

Sarmistha Das, Shreyas Guha, Suvrayan Bandyopadhyay, Salisa Phosit, Kitsuchart Pasupa, and Sriparna Saha. 2026. When meaning isn't literal: Exploring idiomatic meaning across languages and modalities. *Preprint*, arXiv:2604.10787.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, and 1 others. 2022. Glam: Efficient scaling of language models with mixture-of-experts. In *International conference on machine learning*, pages 5547–5569. PMLR.

Francesca De Luca Fornaciari, Begoña Altuna, Itziar Gonzalez-Dios, and Maite Melero. 2024. A hard nut to crack: Idiom detection with conversational large language models. *arXiv preprint arXiv:2405.10579*.

Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. 2018. Neural metaphor detection in context. *arXiv preprint arXiv:1808.09653*.

Team Gemma. 2025. Gemma 3. Technical report, Google.

Lingo Research Group and India IITGN. 2025. LingoIITGN/Ganga-2-1B.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Kilem L Gwet. 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. Magpie: A large corpus of potentially idiomatic expressions. In *12th Language Resources and Evaluation Conference: LREC 2020*, pages 279–287. European Language Resources Association (ELRA).

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, and 1 others. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Ioannis Korkontzelos, Torsten Zesch, Fabio Massimo Zanzotto, and Chris Biemann. 2013. Semeval-2013 task 5: Evaluating phrasal semantics. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 39–47.

Hanbit Lee, Yeonchan Ahn, Haejun Lee, Seungdo Ha, and Sang-goo Lee. 2016. Quote recommendation in dialogue using deep neural network. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 957–960.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.

John I Liontas. 2017. Why teach idioms? a challenge to the profession. *Iranian Journal of Language Teaching Research*, 5(3):5–25.

Chen Cecilia Liu, Anna Korhonen, and Iryna Gurevych. 2025. Cultural learning-based culture adaptation of language models. *arXiv preprint arXiv:2504.02953*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Arijit Maji, Raghvendra Kumar, Akash Ghosh, Anushka Anushka, and Sriparna Saha. 2025a. Sanskriti: A comprehensive benchmark for evaluating language models' knowledge of indian culture. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4434–4451.

Arijit Maji, Raghvendra Kumar, Akash Ghosh, Nemil Shah, Abhilekh Borah, Vanshika Shah, Nishant Mishra, Sriparna Saha, and 1 others. 2025b. Drishtikon: A multimodal multilingual benchmark for testing language models' understanding on indian culture. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1289–1313.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, and 1 others. 2025. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*.

Suzanne Mpouli. 2017. Annotating similes in literary texts. In *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-13)*.

Shahriar Kabir Nahin, Rabindra Nath Nandi, Sagor Sarker, Quazi Sarwar Muhtaseem, Md Kowsher, Apu Chandraw Shill, Md Ibrahim, Mehadi Hasan Menon, Tareq Al Muntasir, and Firoj Alam. 2025. Titullms: A family of bangla llms with comprehensive benchmarking. *Preprint*, arXiv:2502.11187.

Mir Tafseer Nayeem and Davood Rafiei. 2024. Kidlm: Advancing language models for children–early insights and future directions. *arXiv preprint arXiv:2410.03884*.

Vlad Niculae and Cristian Danescu-Niculescu-Mizil. 2014. Brighter than gold: Figurative language in user generated comparisons. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 2008–2018.

Kunat Pipatanakul, Phatrasek Jirabovonvisut, Potsawee Manakul, Sittipong Sripaisarnmongkol, Ruangsak Patomwong, Pathomporn Chokchainant, and Kasima Tharnpipitchai. 2023. Typhoon: Thai large language models. *arXiv preprint arXiv:2312.13951*.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018. Collecting diverse natural language inference problems for sentence representation evaluation. *arXiv preprint arXiv:1804.08207*.

Team Qwen2.5VL. 2025. Qwen2.5-vl.

Junwei Rong, Kostas Terzidis, and Junfeng Ding. 2024. Kids ai design thinking education for creativity development. *Archives of Design Research*, 37(3):119–133.

Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. 2024. V-flute: Visual figurative language understanding with textual explanations. *CoRR*.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Caroline Sporleder and Linlin Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 754–762.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, Siyang Qin, Reeve Ingle, Emanuele Bugliarello, Sahar Kazemzadeh, Thomas Mesnard, Ibrahim Alabdulmohsin, Lucas Beyer, and Xiaohua Zhai. 2024. Paligemma 2: A family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*.

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2016. A neural network approach to quote recommendation in writings. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 65–74.

Ekarat Udomporn. 2014. *The Book on 5000 Thai idioms: from the past right on up to now!* P.S. Pattana Publishing.

Bernard L Welch. 1947. The generalization of 'student's'problem when several different population varlances are involved. *Biometrika*, 34(1-2):28–35.

Ron Yosef, Yonatan Bitton, and Dafna Shahaf. 2023. Irfl: Image recognition of figurative language. *Preprint*, arXiv:2303.15445.

Jiali Zeng, Linfeng Song, Jinsong Su, Jun Xie, Wei Song, and Jiebo Luo. 2020. Neural simile recognition with cyclic multitask learning and local attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9515–9522.

Hao Zhao, Jie Fu, and Zhaofeng He. 2023. Prototype-based hyperadapter for sample-efficient multi-task tuning. *arXiv preprint arXiv:2310.11670*.

A Appendix

Idioms constitute a rich form of figurative expression, wherein meaning extends beyond literal composition to encode culturally grounded emotions, social intent, and contextual nuance. Their interpretation is inherently tied not only to semantics but also to underlying affective and pragmatic signals, making tone an essential dimension of idiomatic understanding. For instance, the Hindi idiom ऊँट के मुँह में जीरा (*oont ke muh mein jeera*, "a cumin seed in a camel's mouth") conveys insufficiency and is often associated with tones of *Ridicule* and subtle *Sorrow*, reflecting dissatisfaction in disproportionate situations. Similarly, the Bengali idiom নাচতে না জানলে উঠোন বাঁকা (*nachte na janle uthoon banka*, "blaming the courtyard for not knowing how to dance") embodies *Ridicule* and *Deception*, exposing the human tendency to deflect personal shortcomings.

In contrast, idioms may also encode positive or aspirational sentiments. The Thai idiom ช้า ๆ ได้พร้าเล่มงาม (*cha cha dai phra lem ngam*, "slowly, one obtains a beautiful blade") reflects patience and perseverance, aligning with tones of *Aspiration* and calm *Affection* toward disciplined effort. Likewise, expressions such as *breaking the ice* often carry *Humor* and social *Affection*, facilitating interpersonal connection. Conversely, idioms like *cry over spilled milk* encapsulate *Sorrow*, while *a wolf in sheep's clothing* signals *Deception* and latent *Fear*, highlighting cautionary undertones embedded in everyday language.

These examples illustrate that idioms inherently operate across a spectrum of emotional and pragmatic tones, including {*Humor*, *Ridicule*, *Affection*, *Aspiration*, *Fear*, *Sorrow*, *Deception*} which shape their intended meaning and usage. Therefore, capturing idiomatic meaning without accounting for such tonal dimensions results in incomplete or misleading interpretations. Despite this, existing research remains predominantly English-centric (Haagsma et al., 2020), with limited focus on multilingual and multimodal settings where emotional, cultural, and visual grounding jointly influence idiomatic understanding.

A.1 Justification of the selected languages

The selection of Hindi, Bengali, and Thai in this work is motivated by a unique combination of typological diversity and deep-rooted cultural interconnectedness across South and Southeast Asia. While Hindi and Bengali belong to the Indo-Aryan family and Thai is a tonal, analytic language from the Kra-Dai family, these languages are historically linked through centuries of cultural exchange shaped by Sanskrit and Pali traditions. This shared heritage is reflected not only in vocabulary but also in religious narratives, philosophical doctrines, and storytelling traditions. For instance, epics such as the *Ramayana* are preserved in India and reinterpreted in Thailand as the *Ramakien*, demonstrating parallel mythological structures and moral frameworks that influence idiomatic expressions and figurative language.

Such cultural alignment manifests in idioms that encode similar emotional and pragmatic intents despite linguistic variation. Expressions conveying moral causality, fate, or human behavior often draw upon shared belief systems such as karma, duty, and social conduct. These idioms are not merely linguistic constructs but carriers of culturally embedded sentiments ranging from cautionary *Fear* and reflective *Sorrow* to social *Ridicule*, aspirational values, and interpersonal *Affection*.

This cross-cultural consistency directly motivates the adoption of the seven idiomatic tone categories {*Humor*, *Ridicule*, *Affection*, *Aspiration*, *Fear*, *Sorrow*, *Deception*} as they capture the most recurrent and semantically stable pragmatic dimensions observed across these languages. Alternative tonal categories were explored during dataset construction; however, they either exhibited significant overlap or lacked consistent representation across linguistic and cultural contexts. In contrast, the selected seven tones provide a balanced and expressive framework for modeling idiomatic meaning, aligning closely with the emotional and narrative functions embedded in these traditions.

From a modeling perspective, incorporating Thai introduces a controlled typological contrast, enabling the evaluation of whether models can generalize idiomatic reasoning beyond closely related language families. Success in transferring idiomatic understanding across these languages indicates abstraction beyond lexical patterns toward deeper cultural and semantic comprehension. Therefore, the inclusion of Hindi, Bengali, and

Thai not only enhances linguistic diversity but also establishes a culturally grounded and theoretically meaningful testbed for studying multimodal idiomatic reasoning.

A.2 Dataset

Building upon the existing *Mediom* dataset, which already captures rich syntactic diversity across idioms in Hindi, Bengali, and Thai, we extend it by incorporating fine-grained pragmatic tonality annotations. The dataset includes a wide range of idiomatic constructions, such as fixed expressions (e.g., น้ำท่วมปาก, *unable to speak out*), semi-fixed constructions (e.g., নিজের পায়ে কুড়াল মারা, *to harm oneself*), verb–object structures (e.g., नाक रगड़ना, *to plead intensely*), adjective–noun phrases (e.g., মিষ্টি স্বপ্ন, *sweet dreams*), prepositional idioms (e.g., เข้าหูซ้ายทะลุหูขวา, *in one ear and out the other*), and binomial expressions (e.g., उल्टा सीधा, *nonsensical behavior*).

While *Mediom* effectively captures structural and semantic richness, it does not explicitly model the underlying emotional and pragmatic tones that govern idiomatic usage. To address this gap, we augment each instance with a set of idiomatic tone labels, grounded in a predefined taxonomy. During the initial design phase, we explored a broader set of ten tonal categories, including additional dimensions such as *Surprise*, *Anger*, and *Neutral*. However, upon systematic analysis and expert-driven validation, these categories were found to be either semantically overlapping with existing tones or inconsistently represented across languages and modalities.

Consequently, we refine the taxonomy to seven core tones {*Humor*, *Ridicule*, *Affection*, *Aspiration*, *Fear*, *Sorrow*, *Deception*}, as depicted in Figure 6 and demonstrate higher cross-lingual consistency, clearer semantic boundaries, and stronger alignment with both textual and visual representations. This augmentation enables a more comprehensive representation of idiomatic meaning by integrating structural, semantic, and affective dimensions within a unified multimodal framework.

A.2.1 Data Quality Assurance

To ensure high-quality annotations and robust validation of cross-modal tonal alignment, we engaged a team of domain experts comprising one doctoral researcher and two literature professors, all of whom are native speakers of Thai, Hindi, and Bengali. Their combined linguistic proficiency and academic expertise enabled the precise mapping of culturally nuanced idiomatic expressions to their corresponding pragmatic tone labels.

A.2.2 Annotation Procedure

The annotation process was designed to systematically capture the pragmatic and emotional tones of idioms in the *Mediom* dataset. Each annotator was provided with the complete multimodal instance, including the idiom text, its literal translation, figurative meaning, idiomatic interpretation, and the corresponding generated image. The goal was to assign one or more labels from a predefined idiomatic tone taxonomy: $\mathcal{C}$ = {Humor, Ridicule, Affection, Aspiration, Fear, Sorrow, Deception}

To establish a reliable foundation, expert annotators first independently labeled a subset of 100 samples. Each instance was carefully examined using its full multimodal context, and multiple tone labels were assigned when necessary. This process resulted in a high-quality reference set of 300 gold-standard annotations, which was subsequently used to guide the broader annotation effort.

To ensure annotation quality and consistency, we developed a structured evaluation framework focusing on five key aspects:

(i) Accurate understanding of tone from multimodal inputs.

(ii) Consistency of tone across text and image.

(iii) Alignment with real-world usage and intended meaning.

(iv) Preservation of cultural and pragmatic nuances.

(v) Coherence among multiple assigned tones.

This framework served as a guiding principle for both annotation and validation.

Following this, a group of trained undergraduate annotators conducted the main annotation process. They were first introduced to a curated set of reference examples and detailed annotation guidelines, enabling them to understand how to interpret and assign idiomatic tones effectively. During this phase, annotators received continuous feedback and clarification from the expert panel to resolve ambiguities and maintain consistency.

Finally, all annotated instances underwent a rigorous validation stage. Expert annotators reviewed and verified each annotation to ensure semantic correctness, cultural fidelity, and adherence

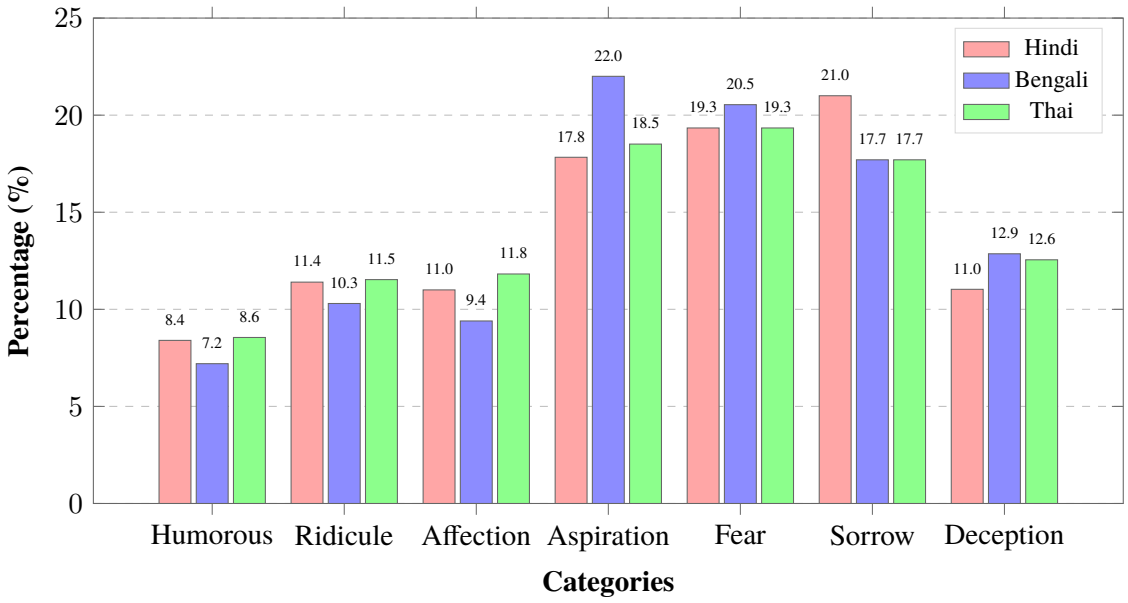

Figure 6: Distribution of Idiomatic Tone categories across Hindi (h), Bengali (b), and Thai (t).

to the defined *IDIO-TONE* taxonomy. This two-stage pipeline annotation followed by expert validation ensured high-quality, reliable, and culturally grounded labels across the dataset. Annotators were compensated at a rate of $0.5 per sample.

(i) **Peer Review:** Each annotated instance was independently reviewed by at least two additional fluent annotators to minimize subjective bias and ensure consistency in tone assignment. Furthermore, to evaluate cross-modal pragmatic alignment, we quantified the overlap between tone labels assigned based on textual interpretations and those inferred from corresponding visual representations as described in Figure 7. This analysis yielded an overlap of **54.75%**, indicating a strong correlation between idiomatic tonal intent and its visual grounding.

(ii) **Expert Validation:** A panel of cultural and linguistic experts conducted a comprehensive final review of the annotated samples. Each instance was evaluated against the predefined assessment criteria, with scores assigned based on the extent to which the five aspects of tonal alignment were preserved (e.g., a score of 5 indicates full compliance across all criteria).

The final annotation process achieved a substantial inter-annotator agreement score of 0.65 (Cohen's kappa (Gwet, 2014)), reflecting a high degree of consistency and reliability in the tone annotation framework.

A.3 Results & Analysis

To understand the precise contribution of the IPS in our HybridMoE framework, we conduct a rigorous ablation comparing two configurations: (i) HybridMoE-augmented VLMs without IPS, and (ii) full HybridMoE with IPS injection in Table 4. The comparative results indicate that HybridMoE-augmented VLMs, even without IPS injection, provide a strong baseline improvement across models by enhancing multimodal fusion and representation learning. In this setting, models such as *Qwen2.5-VL-7B-Instruct* and *Gemma3-12B-pt* achieve competitive performance on lexical and sequence-level metrics, with ROUGE-LSum values around 0.41–0.43 and BLEU-3 scores exceeding 21, reflecting improved coherence and n-gram alignment. However, despite these gains, the absence of IPS limits the model's ability to capture deeper semantic and idiomatic nuances. This is evident from comparatively moderate BERTScore values ($\approx$ 0.89–0.95) and lower Idiomatic Validation Scores and *IDIO-TONE* metrics, suggesting that the improvements are largely driven by surface-level alignment rather than true figurative understanding.

In contrast, the full HybridMoE framework with IPS injection yields consistent and more comprehensive performance improvements across all evaluation dimensions. The inclusion of IPS acts as

Validation Centric Idiomatic Tonality Label Generation from Textual Idioms

System Role: You are a linguistic expert and expert at multi-cultural idiomatic understanding.
Instruction: You will be provided with an idiom, along with its English translation and meaning. It is your responsibility to figure out which of the following labels best fits the given idiom. Here are the labels: Humorous, Ridicule, Affection, Aspiration, Fear, Sorrow, Deception Here is the output format: label0: reasoning for assigning label0, label1: reasoning for assigning label1, …
Return nothing, except the required dictionary.

Validation Centric Idiomatic Tonality Label Generation from Images

System Role: You are an expert image analyzer and cross-cultural idiomatic expert.
Instruction: You will be provided with an image which represents an idiom, along with its English translation and meaning. It is your responsibility to figure out which of the following labels best fits the given idiom. Here are the labels: Humorous, Ridicule, Affection, Aspiration, Fear, Sorrow, Deception. Here is the output format: label0: reasoning for assigning label0, label1: reasoning for assigning label1, …
Return nothing, except the required dictionary.
Input: {Output of Stage 1}

Figure 7: Idiomatic Tonality label validations by VLM DeepSeek-R1

Table 4: The variance in performance metrics between HybridMoE-enhanced and HybridMoE-without IPS is presented below. Here, R-1, R-2, R-L, and R-LSum denote ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-LSum scores, respectively. B-1, B-2, B-3, and BS denote BLEU-1, BLEU-2, BLEU-3, and the BLEU score. BTS, IVS, and *IDIO-TONE* correspond to BERTScore, Idiomatic Validation Score, and the proposed idiomatic tone score. Bold values indicate the best performance; ↑ denotes that higher values are better.

Experimental Setting	Model Names	Metrics										
		R-1	R-2	R-L	R-LSum	B-1	B-2	B-3	BS	BTS	IVS	IDIO-TONE
HybridMoE w/o IPS	Blip2-7B	0.27	0.10	0.19	0.23	19.91	9.62	8.02	11.73	0.88	0.53	0.33
	Paligemma2-10B	0.42	0.17	0.30	0.30	33.20	16.64	11.82	14.80	0.89	0.56	0.32
	Llava-1.5-7B	0.37	0.15	0.27	0.29	29.53	14.91	9.36	12.94	0.90	0.58	0.33
	SmolVLM-Instruct	0.44	0.18	0.35	0.35	48.92	22.75	15.37	19.00	0.92	0.63	0.36
	Gemma3-12b-pt	0.42	0.17	0.31	0.32	35.04	15.43	8.99	12.69	0.91	0.76	**0.54**↑
	Qwen2.5-VL-7B-Instruct	**0.49**↑	**0.23**↑	**0.39**↑	**0.40**↑	**53.22**↑	**25.90**↑	**18.36**↑	**22.11**↑	**0.94**↑	**0.77**↑	0.46
HybridMoE	Blip2-7B	0.30	0.13	0.23	0.28	23.76	12.93	13.29	15.64	0.89	0.55	0.33
	Paligemma2-10B	0.45	0.20	0.34	0.35	35.97	17.25	14.20	15.94	0.89	0.59	0.32
	Llava-1.5-7B	0.40	0.17	0.32	0.34	32.52	16.64	12.95	14.25	0.91	0.61	0.33
	SmolVLM-Instruct	0.48	0.21	0.36	0.36	50.46	24.16	16.59	20.92	0.92	0.66	0.35
	Gemma3-12b-pt	0.46	0.19	0.35	0.35	38.81	17.82	11.93	14.58	0.92	0.79	**0.55**↑
	Qwen2.5-VL-7B-Instruct	**0.54**↑	**0.32**↑	**0.41**↑	**0.43**↑	**56.32**↑	**28.92**↑	**21.48**↑	**24.19**↑	**0.95**↑	**0.82**↑	0.47

a task-specific semantic signal that enhances contextual grounding and idiomatic reasoning, leading to stronger alignment in both textual and multimodal representations. This is reflected in higher BERTScore values (reaching ≈ 0.94), improved IVS (Idiomatic Validation Score) (≈ 0.73–0.77), and notable gains in *IDIO-TONE*, indicating better modeling of pragmatic and affective aspects of idiomatic expressions. Additionally, performance across ROUGE and BLEU metrics remains stable or improves marginally, demonstrating that semantic enrichment does not come at the cost of lexical fidelity. Overall, while HybridMoE without IPS primarily strengthens structural fusion and surface-level reasoning, the integration of IPS enables a more balanced and semantically aware system, effectively bridging the gap between lexical accuracy and deeper idiomatic understanding.

This demonstrates that although MoE structures can introduce expert diversity, without IPS, the routing mechanism lacks semantic specificity. Experts are activated based on undifferentiated patterns, leading to diluted idiomatic representation and weaker cross-modal alignment. By contrast, the HybridMoE with IPS explicitly encodes idiomatic abstraction signals into the CLS representations during encoding, enabling task-aware expert routing that prioritizes idiom-grounded con-

Table 5: Qualitative Analysis of IDIO-TONE Performance in HybridMoE-Configured Vision-Language Models

Idiom: মানুষের অভ্যাসই দেবতা।
Pronounciation:(manusher obhyash-i debota)
Literal Translation: A person's habits are like their God.
Idiomatic Understanding: This bengali idiom " মানুষের অভ্যাসই দেবতা " emphasizes that habit shapes human behavior and character, often becoming so powerful that it governs one's actions like a higher authority. It reflects the idea that repeated practices strongly influence decision-making and identity, for better or worse.
Ground Truth: ['Aspiration']

Qwen2.5-VL-7B-Instruct: ['Aspiration']
SmolVLM-Instruct: ['Affection', 'Aspiration']
Gemma3-12b-pt : ['Aspiration', 'Fear']
Paligemma2-10B: ['Aspiration']
Llava-1.5-7B: ['Aspiration']
Blip2-7B: ['Aspiration']

***Idiom*:** ปิดควันไฟไม่มิด
Pronounciation: (pit khwan fai mai mit)
Literal Translation: You cannot completely hide smoke from a fire.
Idiomatic Understanding: This thai idiom " ปิดควันไฟไม่มิด " conveys that the truth cannot be fully concealed, no matter how hard one tries. Just as smoke inevitably reveals the presence of fire, hidden actions, lies, or wrongdoing will eventually become visible. It is often used in contexts involving secrets, deception, or wrongdoing, emphasizing the inevitability of exposure.
Ground Truth- ['Fear', 'Deception']

Qwen2.5-VL-7B-Instruct: ['Deception', 'Fear']
SmolVLM-Instruct: ['Affection', 'Aspiration']
Gemma3-12b-pt: ['Fear']
Paligemma2-10B: ['Fear', 'Sorrow']
Llava-1.5-7B: ['Fear']
Blip2-7B: ['Fear']

text, figurative meaning, and symbolic grounding. This semantic conditioning significantly improves linguistic coherence. Furthermore, idiom-sensitive evaluation via IVS reveals that IPS-enhanced models better capture cultural nuance and metaphorical intent, which are often lost in finetuning or naïve MoE setups. In essence, this ablation confirms that without IPS, models struggle to disambiguate literal vs. figurative content, leading to representational collapse; whereas IPS-enabled HybridMoE transforms multimodal fusion into an idiom-aware, semantically enriched, and performance-consistent mechanism across diverse VLM architectures.

Qualitative Analysis on *IDIO-TONE* Metric. As illustrated in Table 5, we conduct a qualitative analysis using the *IDIO-TONE* metric to examine model behavior beyond aggregate scores. The results indicate that VLMs generally align well with the dominant idiomatic tone, while exhibiting variability in capturing secondary and nuanced tonal signals. For the thai idiom ปิดควันไฟไม่มิด (truth cannot be concealed), the ground truth reflects a combination of *Aspiration* and *Deception*. Most models successfully identify *Fear*, demonstrating strong sensitivity to explicit cues, although they often underrepresent the complementary *Deception* aspect, suggesting challenges in modeling implicit moral or reflective undertones.

Similarly, for the bengali idiom মানুষের অভ্যাসই দেবতা , where the ground truth emphasizes *Aspiration*, models consistently capture the primary tone, while occasionally introducing additional labels such as *Fear* or *Affection*. This reflects a tendency toward broader associative reasoning in abstract or philosophical contexts.

Moreover, Qwen2.5-VL-7B-Instruct consistently predicts the correct *IDIO-TONE* labels for both the Bengali and Thai idioms, further demonstrating its superior capability in capturing nuanced idiomatic semantics and pragmatic tone compared to other models. Overall, these observations highlight that while models are effective at identifying dominant tonal categories, capturing multi-label coherence and subtle pragmatic nuances remains an evolving capability. Conclusively, *IDIO-TONE* serves as a crucial and sensitive evaluation metric, enabling a more fine-grained assessment of idiomatic understanding beyond surface-level semantic alignment.

A.4 LLM usage

We used large language models (LLMs) to assist with code development and minor editing of the final manuscript.