Bridging the Visual Gap: Resource-Efficient VLM Adaptation for Meitei Mayek using Synthetic Multi-Modal Mixtures

Anonymous ACL submission

Abstract

Vision-Language Models (VLMs) have demonstrated remarkable capabilities in text recognition and visual reasoning, yet these advancements remain largely confined to high-resource languages, leaving indigenous scripts like Meitei Mayek severely marginalized. We introduce a resource-efficient synthetic bootstrapping methodology for adapting a pre-trained VLM to an entirely unseen, zero-resource orthography. Our pipeline synthesizes a massive 5-piece multi-modal mixture combining custom-rendered Meitei OCR data, dynamically translated VQA datasets, and instruction-tuning pairs, validated by native speakers through a Human-in-the-Loop (HiTL) protocol. Applying QLoRA to PaliGemma-3B, our fine-tuned adapter establishes initial token-level script recognition, raising Average Normalized Levenshtein Similarity (ANLS) by +13.0 absolute points (8.62% to 21.62%) and improving VQA Exact Match to 16.00%. A paired McNemar test ($p = 0.013$) confirms the improvement over base PaliGemma is statistically significant. Crucially, our ablation reveals the *Generative Collapse*, a tokenizer byte-fallback bottleneck that encourages statistical memorization over genuine reasoning, demonstrating that vocabulary expansion is strongly motivated for robust zero-resource adaptation. We release the adapter weights, datasets, and rendering pipeline as a replicable blueprint and failure-mode analysis for extending VLMs to marginalized scripts.[1]

1 Introduction

Vision-Language Models (VLMs) (Radford et al., 2021) like PaliGemma (Beyer et al., 2024) and Qwen2.5-VL (Bai et al., 2025) excel in text recognition and visual reasoning (Antol et al., 2015; Singh et al., 2019). Driven by massive image-text datasets, they nonetheless exclude the vast majority of the world's 7,000 languages (Vaswani et al., 2017), creating a "digital cliff" for under-resourced indigenous communities.

This disparity is severe for indigenous orthographies like Meitei Mayek (Manipuri), spoken in Northeast India. Suppressed for two centuries and only added to Unicode in 2009, its digital footprint is nascent. The resulting absence of natural image-text pairs creates a severe cold-start problem for multi-modal AI.

Consequently, pre-trained VLMs exhibit near-total script blindness. Standard vision encoders extract degraded, out-of-distribution embeddings for unseen orthographies, depriving the language head of typographical features. Zero-shot evaluations confirm this: Qwen2.5-VL-3B achieves 0.0% Exact Match, and base PaliGemma-3B yields just 8.62% Average Normalized Levenshtein Similarity (ANLS). Manually annotating data to train a visual encoder from scratch is resource-prohibitive.

To overcome this, we propose a resource-efficient synthetic bootstrapping methodology requiring no massive human annotation. We synthesize a 5-piece multi-modal mixture integrating custom-rendered Meitei OCR with English benchmarks (VQAv2, Alpaca) and dynamically translated Meitei counterparts. Using text-to-image rendering and automated translation, we inject script comprehension into the visual encoder and language head. We apply Quantized Low-Rank Adaptation (QLoRA) (Hu et al., 2022; Dettmers et al., 2023) to PaliGemma-3B, circumventing full fine-tuning costs. Our contributions are:

- A resource-efficient data synthesis pipeline for bootstrapping visual-linguistic alignment for zero-resource scripts, requiring only a font file and a translation API, validated by an Inter-Annotator Agreement (IAA) study.
- The identification and empirical proof of the

[1] https://anonymous.4open.science/r/meitei_mayek_VLM-5EF8/

Generative Collapse, a fundamental tokenizer bottleneck that cripples zero-resource adaptation, accompanied by a Vocabulary Extension Proof-of-Concept that mitigates it.

- The first open-source VLM adapter fine-tuned for the Meitei Mayek script, establishing a foundational baseline (ANLS +13.0%, VQA EM 16.00%, McNemar p = 0.013 vs. base PaliGemma) and releasing our instruction-tuning corpus as a replicable blueprint for Tibeto-Burman multi-modal AI.

2 Related Work

VLMs in Low-Resource Contexts. Our research sits at the intersection of multimodal adaptation for low-resource languages, Manipuri (Meitei) computational processing, and parameter-efficient fine-tuning via synthetic data. While VLMs excel in high-resource settings, adapting them to non-Latin scripts remains challenging. Although visual signals can bridge textual data scarcity, tokenization bottlenecks and alignment difficulties limit zero-shot transfer (Minaee et al., 2025). Evaluation suites like IndicVisionBench (Faraz et al., 2026), BharatBench (Khan et al., 2025), and DRISHTIKON (Maji et al., 2025) reveal stark VLM deficits on Indic scripts. While models like Chitrarth and Maya address major Indian languages, indigenous scripts like Meitei Mayek remain unsupported.

Computational Processing for Manipuri. Manipuri NLP has historically focused on Machine Translation to overcome data scarcity. Recent efforts use byte-level modeling to bypass Out-Of-Vocabulary bottlenecks (Lin et al., 2024) and apply PEFT to models like NLLB-200 and IndicTrans2 for WMT shared tasks (Acharya et al., 2025). Moving beyond text-only models, Meetei et al. (2024) pioneered a Multimodal Machine Translation framework for Manipuri, demonstrating that auxiliary visual cues boost translation fluency. Our work extends this trajectory to native VLM comprehension, explicitly training a multimodal architecture to reason over the Meitei Mayek script.

Synthetic Data and PEFT. The absence of massive multi-modal corpora for Manipuri necessitates aggressive bootstrapping. Drawing from text-based synthetic data generation (Schick and Schütze, 2021; Lancheros et al., 2025), we utilize dynamic cross-lingual translation to synthesize VQA data. To adapt large architectures efficiently, Low-Rank Adaptation (Hu et al., 2022) is now standard, with recent work successfully integrating LoRA into vision-language frameworks for few-shot adaptation (Ghiasvand et al., 2025). We build on this paradigm, using QLoRA to inject synthesized Meitei Mayek comprehension into a VLM, providing a scalable blueprint for zero-resource orthographies.

3 Proposed Work

This section outlines our framework for adapting a pre-trained VLM to the extremely low-resource Meitei (Manipuri) language. The methodology is structured around three core pillars: (1) synthesis and curation of a multi-modal instruction-tuning dataset with HiTL validation; (2) a high-capacity PEFT strategy that maps the novel script to an existing visual latent space without inducing catastrophic forgetting; and (3) a rigorous multi-metric evaluation framework. Figure 1 provides an overview.

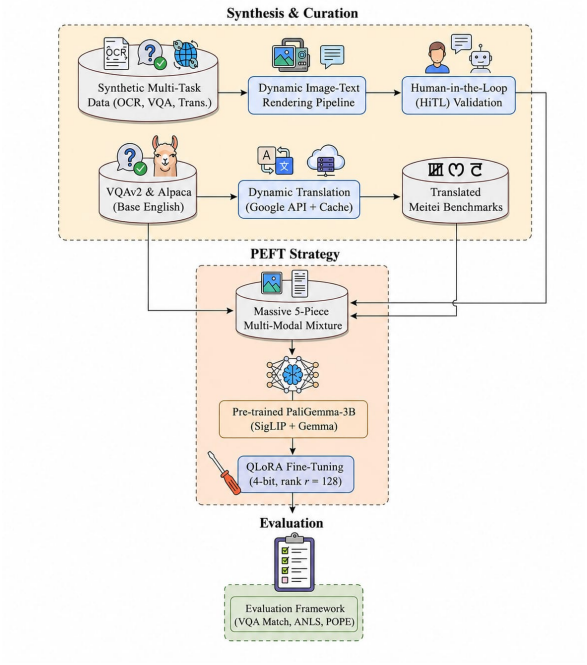

Figure 1: Visual-Linguistic bootstrapping methodology for the Meitei Mayek script

3.1 Data Collection, Synthesis, and Curation

To overcome the absence of native image-text pairs and the architectural blindness of Latin-centric VLMs, we engineered a heterogeneous, human-validated dataset comprising synthesized tasks and translated English benchmarks.

Synthetic Multi-Task Dataset. We constructed a structured synthetic corpus (`batch_deduplicated.json`) spanning six distinct task types, forcing the model to learn a unified representation of the language rather than over-fitting to a single objective. The task taxonomy includes: *OCR* (transcribing Meitei text from rendered images), *Visual Question Answering* (answering contextual questions from image cues in Meitei), *Visual Translation* (interpreting Meitei text within an image and outputting its English semantic equivalent), *Text Translation* in both directions (Meitei $\leftrightarrow$ English), and *Instruction Following* (modifying or summarizing Meitei statements into interrogative or condensed forms). The full taxonomy with representative examples is provided in Appendix A.

Dynamic Image-Text Rendering. Because our VLM requires visual inputs to ground OCR and visual translation tasks, we developed a deterministic text-to-image rendering pipeline using PIL with Google's Noto fonts (`NotoSansMeeteiMayek-Regular`). Text strings are rendered onto 224×224 pixel canvases with a custom word-wrapping algorithm that continuously tracks x/y spatial coordinates and switches seamlessly between Meitei Unicode characters (> 127) and Latin characters within the same visual bounding box, ensuring text does not truncate abruptly. For text-only tasks (where no image is required), the pipeline generates a blank black canvas (RGB: 0, 0, 0). This is designed to encourage the vision encoder to output near-null embeddings for such inputs, nudging the language modeling head to rely on text embeddings when no visual data is present, a design choice intended to reduce visual hallucinations.

Human-in-the-Loop (HiTL) Validation. Relying strictly on synthetic generation and automated translation introduces a high risk of propagating morphological errors, unnatural phrasing, and code-switching artifacts. To mitigate this, we recruited two native Meitei speakers via community outreach who independently reviewed all outputs for grammatical correctness, cultural relevance, semantic equivalence in translations, and OCR fidelity against rendered images. Disagreements were resolved by consensus, with erroneous samples discarded rather than corrected to avoid annotator bias (Mosqueira-Rey et al., 2023). Importantly, this destructive filtering approach renders a pre/post HiTL ablation methodologically invalid: training a baseline on the unvalidated dataset would expose the visual encoder to corrupted script renderings, actively misleading spatial reasoning rather than providing a fair generative baseline. This human-validated approach therefore transformed a potentially noisy corpus into a linguistically coherent foundational resource. Inter-annotator agreement for this validation phase was assessed via Cohen's Kappa on a 100-sample audit (see §4).

Cross-Lingual Benchmark Translation. To infuse general visual reasoning and logical instruction-following capabilities into the newly mapped Meitei semantic space, we sampled and translated established English benchmarks into Meitei (`mni-Mtei`) via the Google Translator API with intelligent local caching: 5,000 VQAv2 samples (Goyal et al., 2017) and 2,000 text-only Alpaca logic samples (filtered for output lengths ≤ 100 words) (Taori et al., 2023). In the event of an API failure, the system gracefully fell back to the original English text. This fallback mechanism inadvertently functioned as a natural regularization strategy, exposing the model to code-switched bilingual data and reinforcing English retention.

3.2 Formulation of the Massive Multi-Modal Mixture

Fine-tuning a pre-trained model exclusively on a new, low-resource language historically triggers catastrophic forgetting, wherein established spatial reasoning and English comprehension rapidly degrade. To counteract this phenomenon, we formulated a balanced 5-component composite dataset via PyTorch's `ConcatDataset`, detailed in Table 1. The deliberate 50/50 English–Meitei split across VQA and Alpaca components is intentional: it is designed to mitigate catastrophic forgetting of pre-trained English representations while simultaneously acquiring Meitei script understanding. By interleaving English and Meitei prompts dynamically, the model's attention mechanism is forced to map newly acquired Meitei tokens into the same latent conceptual space as its pre-trained English representations, the core mechanism enabling cross-lingual transfer without starting from scratch.

3.3 Model Architecture and Input Formulation

We adopted PaliGemma-3B as our foundational architecture, which integrates a SigLIP (Zhai et al.,

Table 1: Training mixture composition. The deliberate 50/50 English–Meitei split is designed to mitigate catastrophic forgetting while enabling script acquisition.

Component	Samples
Custom HiTL-Validated Meitei Multi-Task	All validated
Translated VQAv2 (Meitei)	2,500
Base VQAv2 (English)	2,500
Translated Alpaca (Meitei)	1,000
Base Alpaca (English)	1,000
Train / Validation Split	90% / 10%

2023) vision encoder with a Gemma (Team et al., 2024) causal language model, bridged by a learned projector. Images were processed and resized to a fixed resolution of 224×224 pixels. Input formatting and masking details are provided in Appendix B.

3.4 PEFT Setup (QLoRA)

Adapting a 3B parameter multi-modal model to a completely unseen script requires significantly higher parameter capacity than standard downstream fine-tuning. We employed QLoRA (Dettmers et al., 2023) to quantize the base PaliGemma weights to 4-bit NormalFloat (NF4) with double quantization. LoRA adapters were configured with an exceptionally high rank $r = 128$, scaling alpha $\alpha = 256$, and dropout 0.05, targeting both the attention projection matrices (q_proj, k_proj, v_proj, o_proj) and the feed-forward network projections ($gate_proj$, up_proj, $down_proj$) in both the visual encoder and language modeling head. The QLoRA configuration, combined with PyTorch's expandable segments memory management, ensures full functionality on an NVIDIA T4 (16GB), confirming accessibility on widely available cloud compute.

To justify this massive parameter capacity requirement, we conducted a rank sensitivity study on a preliminary data split. As visualized in Figure 2, scaling the adapter rank from $r = 16$ to $r = 64$ yielded a direct, monotonic decrease in final training loss ($r = 16 \rightarrow 5.383$, $r = 32 \rightarrow 4.805$, $r = 64 \rightarrow 4.434$). Based on this downward trajectory demonstrating the model's hunger for parameters when mapping a novel orthography, we deployed an aggressively parameterized $r = 128$ for the final massive mixture to help mitigate catastrophic forgetting. While computational constraints precluded a full validation sweep across all possible ranks, this $r = 128$ configuration ultimately yielded the downstream benchmark improvements reported in §4.

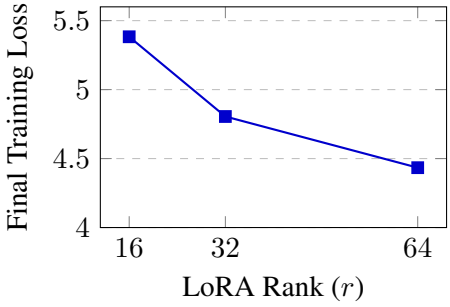

Figure 2: LoRA rank sensitivity ablation. Scaling the adapter rank demonstrates a monotonic decrease in training loss on a preliminary split. Based on this trajectory, the maximal r=128 was selected for the final full-capacity model.

3.5 Evaluation Framework and Metrics

Evaluating cross-lingual visual models requires metrics that holistically assess linguistic fluency, visual grounding, and hallucination resistance. Benchmark datasets were strictly isolated from the training distribution, sampled from TextVQA by LMMS-Lab and Flickr30k (Plummer et al., 2015). The benchmark comprises three task types (VQA/ANLS, POPE, and Captioning), with 100 queries per task type (50 English, 50 dynamically translated Meitei), yielding 300 total evaluation queries. Bootstrap confidence intervals (1,000 iterations) in §4 are computed per task type: VQA/ANLS CIs use the 100 VQA-task queries; POPE CIs use the 100 POPE-task queries. We report four complementary metrics:

- **VQA Exact Match**: Strict string equivalence between prediction and ground truth. Zero partial credit.
- **ANLS** (Biten et al., 2019): Average Normalized Levenshtein Similarity. Measures character-level edit distance normalized to $[0, 1]$, superior to exact match for morphologically rich, agglutinative scripts.
- **POPE** (Li et al., 2023): Polling-based Object Probing Evaluation. Binary Yes/No visual probes measuring hallucination propensity. We report Accuracy, Precision, Recall, and F1.

- **BLEU & ROUGE-L** (Papineni et al., 2002; Lin, 2004): n-gram overlap metrics. Reported for completeness but noted as structurally incompatible for non-Latin, instruction-tuned VLMs (see §4).

4 Results and Analysis

4.1 Zero-Shot Baseline Script Blindness

Before analyzing our fine-tuned model, we establish the baseline performance of current state-of-the-art VLMs on Meitei. Table 2 reveals a stark dichotomy between general visual comprehension and localized script acquisition. Qwen2.5-VL-3B achieves impressive POPE scores (86.70% Acc, 92.10% F1) on general English-centric object recognition benchmarks. However, it scores an absolute 0.0% on both VQA Exact Match and ANLS for Meitei, failing to generate a single character-by-character guess close enough to receive partial credit. Chitrarth-7.5B, despite being tailored for Indian languages, similarly flatlines at 2.50% EM and 3.24% ANLS. Crucially, Chitrarth-7.5B was designed for major Indic scripts; Meitei Mayek falls outside its documented training distribution, making its 2.50% EM an upper bound on zero-shot transfer rather than a tuned baseline. These results are consistent with a tokenizer-level failure hypothesis, though disentangling encoder from decoder failure would require probing intermediate representations beyond the scope of this work. Base PaliGemma-3B achieves 8.00% EM and 8.62% ANLS only by outputting random English artifacts that coincidentally aligned with target strings.

4.2 Script Comprehension (VQA & ANLS)

The most critical indicators of linguistic adaptation are the VQA Exact Match and ANLS metrics. These are computed exclusively on the 100 VQA-type queries from the evaluation set, with captioning and POPE probes evaluated separately to avoid metric contamination. A paired McNemar test on the 100 matched VQA predictions (fine-tuned vs. base PaliGemma) yields $\chi^2 = 6.13$, $p = 0.013$, confirming the improvement is statistically significant at the 5% level. Discordant pairs: $b = 0$ (baseline correct, fine-tuned wrong), $c = 8$ (fine-tuned correct, baseline wrong) — a one-directional improvement with no regressions. VQA Exact Match improved from 8.00% to 16.00% (95% bootstrap CI: [9.00%, 23.00%]). ANLS rose from 8.62% to 21.62% (Fine-Tuned CI: [13.84%, 29.01%] vs. Baseline CI: [3.62%, 14.62%]). We note that the evaluation queries are API-translated from English benchmarks rather than organically collected Meitei visual queries; the CIs accordingly estimate performance stability on this translated distribution and should not be interpreted as estimates of generalization to naturally occurring Meitei data.

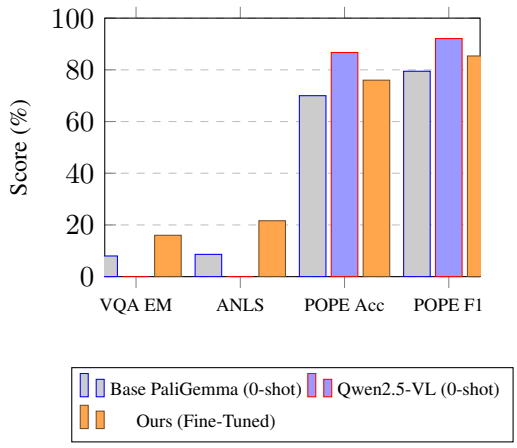

Figure 3: Metric comparison. Our adapter outperforms all baselines on script comprehension (VQA EM, ANLS) and maintains competitive visual grounding (POPE) relative to base PaliGemma.

The divergence between EM and ANLS is particularly informative for Meitei. As an agglutinative language, Meitei root words are heavily modified by affixes to convey tense, case, and gender; exact matching therefore yields false negatives even for semantically correct answers. The ANLS jump demonstrates the model has stopped outputting random English tokens and has partially learned the structural orthography of Meitei Mayek, consistently predicting correct root morphological characters even when complex suffixes are mispredicted (see Appendix D for qualitative examples of this agglutinative scoring dynamic).

Inter-Annotator Agreement (IAA) Study. To validate automatic metrics, we recruited two native Meitei volunteers to independently evaluate 100 randomly sampled model predictions across three dimensions. We computed Inter-Annotator Agreement (IAA) using Cohen's Kappa (κ). For **Translation Quality**, annotators achieved 90.0% absolute agreement (Quadratic Weighted $\kappa = 0.783$, indicating substantial alignment on semantic fidelity). For **Prediction Correctness**, they reached near-perfect alignment with 98.0% agreement ($\kappa = 0.790$). For **Prediction Fluency**, agreement was moderate at 72.0% ($\kappa = 0.555$), with variance

Table 2: Comparative Evaluation (%). POPE improvements stated relative to PaliGemma-3B. †Base PaliGemma scores on VQA EM/ANLS are derived entirely from random English artifacts, not genuine Meitei predictions.

Metric	Qwen2.5-VL 3B (0-shot)	Base Pali 3B (Artifacts)†	Chitrarth 7.5B (0-shot)	Ours (Fine-Tuned)
VQA Exact Match	0.00	8.00	2.50	**16.00** (↑+8.0)
ANLS	0.00	8.62	3.24	**21.62** (↑+13.0)
POPE Accuracy	86.70	70.00	89.00	**76.00** (↑+6.0 vs Pali)
POPE F1-Score	92.10	79.45	86.42	**85.36** (↑+5.9 vs Pali)
BLEU	0.00	8.47	1.80	0.15 (↓-8.3)
ROUGE-L	0.00	5.02	10.20	3.50 (↓-1.5)

Table 3: Inter-Annotator Agreement (IAA) Study on 100 randomly sampled predictions from two native Meitei evaluators.

Dimension	Agreement	Unweighted κ	Quadratic Weighted κ
Translation Quality	90.0%	0.651	0.783
Prediction Fluency	72.0%	0.555	0.439
Prediction Correctness	98.0%	0.790	0.790

stemming from differing thresholds: one evaluator penalized task-mismatched outputs (e.g., generating a fluent "yes" for a captioning task) while the other rated solely on orthographic validity. These results are summarized in Table 3.

4.3 Hallucination Resistance (POPE Analysis)

To measure hallucination, we utilized Yes/No POPE probes. Since zero-shot models like Qwen2.5-VL score 0% on Meitei VQA, their strong POPE F1 on English benchmarks is not a valid Meitei baseline. The relevant comparison is base PaliGemma vs. our adapter. Our 100-sample subset is affirmative-biased (88 yes, 12 no), meaning POPE serves primarily as an anti-hallucination probe relative to the base model.

Table 4 dissects the hallucination dynamics. The base PaliGemma achieved 100% Precision but a severely crippled Recall of 65.90%: it was highly conservative, defaulting to null or negative responses when confronted with Meitei queries. Our fine-tuned adapter raises Recall by +13.64% (to 79.54%), boosting F1 from 79.45% to 85.36%. To verify this improvement is not a degenerate "always say yes" loop, we examined the full confusion matrix over the 100-sample POPE probe set (visualized in Figure 4): TP=70, FP=6, FN=18, TN=6. These values reproduce the reported metrics exactly: Accuracy=$(70+6)/100=76\%$,

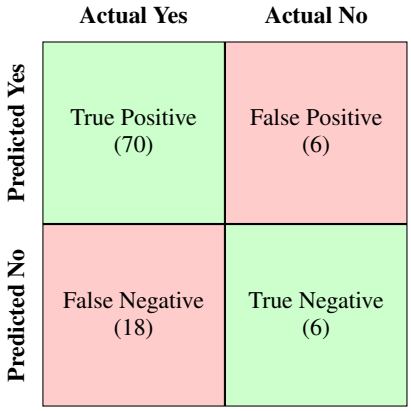

Figure 4: POPE Confusion Matrix (n=100 probing subset). Values reproduce Table 4 exactly: Acc=76%, Prec=92.1%, Recall=79.5%, F1=85.4%. The TP:FP ratio of 70:6 confirms the adapter grounds affirmative answers in visual evidence.

Precision=$70/76=92.1\%$, Recall=$70/88=79.5\%$, F1=$140/164=85.4\%$. The substantial True Positive count relative to False Positives confirms the model is grounding its affirmative answers in visual evidence. The precision trade-off is expected; a model that abstains less will inevitably make more errors, and the net F1 gain confirms the trade-off is favorable.

4.4 The Generative Collapse: A Tokenizer Bottleneck

While reasoning and grounding metrics improved substantially, the generative BLEU score dropped precipitously from 8.47% to 0.15%. This is a compound effect of two distinct phenomena.

First, BLEU and ROUGE-L are legacy metrics engineered for high-resource Latin-based languages with clear whitespace delimitation; standard English-centric tokenizers fail to properly

Table 4: POPE Analysis relative to base PaliGemma-3B. The Recall gain indicates reduced abstention on Meitei queries.

Metric	Base PaliGemma	Fine-Tuned	Change
Accuracy	70.00%	76.00%	↑+6.00%
Precision	100.00%	92.10%	↓-7.90%
Recall	65.90%	79.54%	↑**+13.64%**
F1-Score	79.45%	85.36%	↑+5.91%

segment continuous Meitei script, making BLEU structurally incompatible. Furthermore, instruction tuning drives the model toward concise, factual answers rather than verbose captions, a factually correct Meitei noun phrase incurs a heavy BLEU Brevity Penalty against verbose Flickr30k references.

Second, deeper analysis revealed an architectural collapse induced by tokenizer byte-fallback. Because PaliGemma lacks a native Meitei vocabulary, the LoRA adapter was forced to memorize raw UTF-8 byte sequences rather than learning compact, token-level orthographic representations. The model successfully memorized short, high-frequency VQA macro-labels (e.g., ꯑꯣꯟ for "No", ꯀꯦꯟꯆꯥꯄ for "Ketchup") via their compact byte-barcodes, explaining the VQA EM and ANLS gains. However, generating novel, multi-word image captions exceeded this memorization capacity: the model either crashed into byte-loops or resorted to Template Echoing (regurgitating the instruction prompt verbatim).

To probe whether this is a tokenizer bottleneck rather than a visual encoder limitation, we conducted a *Vocabulary Extension Proof-of-Concept*. By manually adding just 50 Meitei script tokens to the vocabulary, unfreezing the `lm_head` and `embed_tokens`, and running a highly constrained fine-tuning step, training loss dropped from 11.58 to 7.085. We note that unfreezing `lm_head` and `embed_tokens` independently changes the optimization landscape; a control condition (unfreezing these layers without adding new tokens) would be required to fully isolate the vocabulary extension's contribution. Nevertheless, the magnitude of the loss drop relative to the constrained training budget is consistent with the byte-fallback bottleneck hypothesis, establishing full vocabulary expansion as the most promising direction for eliminating the Generative Collapse.

4.5 Ablation Study: Deconstructing the Multi-Modal Mixture

To isolate the specific variables driving script acquisition, we evaluated checkpoints at three distinct phases of the data mixture pipeline (Table 5):

1. `only_custom`: Trained exclusively on our HiTL-validated synthetic Meitei multi-task dataset.

2. `custom_vqa`: Custom data combined with the translated and base VQAv2 visual reasoning mixture.

3. **Final (Massive)**: The full high-capacity model incorporating the complete visual and logical data mixture.

The ablation results offer meaningful insights into the mechanics of cross-lingual fine-tuning. In stages `only_custom` and `custom_vqa`, VQA Exact Match and ANLS remain at 0.00%, and POPE Accuracy hovers near random chance (22.00%–24.00%). Paradoxically, during `custom_vqa`, BLEU skyrockets to an anomalous 76.44%. Without the deductive logic introduced by the Alpaca dataset, the VLM appears to have memorized the syntactic structure of descriptive image captions, generating fluent sentences that may be decoupled from factual visual grounding (*Semantic Parroting*). This is directly evidenced by a systematic *Template Echoing* pattern in the stage-2 prediction logs: for every Meitei captioning prompt (e.g., "Describe this image"), the model echoes the verbatim prompt back as its output prefix, regardless of image content — for instance, predicting "Describe this image.\nThe couple is" for a crowd scene and "Describe this image.\nThe are jumping" for a ballet image. This behavior ceased completely in the Final (Massive) adapter, confirming the Alpaca component's role in breaking the template-memorization loop, though the causal mechanism cannot be fully isolated from the confounding effects of increased training volume and dataset diversity.

Upon introduction of the full massive mixture (`Final (Massive)`), the metric landscape shifted substantially. The addition of the Alpaca component appears to have been the critical catalyst, though we cannot fully isolate its contribution from the confounding effects of increased training volume and dataset diversity. VQA jumped to 16.00%, ANLS to 21.62%, and POPE Accuracy to 76.00%,

Table 5: Ablation across training mixture stages. The BLEU collapse at the final stage reflects the Generative Collapse, while VQA/ANLS/POPE gains confirm genuine script-level adaptation.

Stage	VQA EM	ANLS	POPE Acc	POPE F1	BLEU	ROUGE-L
`only_custom`	0.00	0.00	22.00	29.09	27.41	4.64
`custom_vqa`	0.00	0.00	24.00	34.48	**76.44**	**8.07**
`Final (Massive)`	**16.00**	**21.62**	**76.00**	**85.36**	0.15	3.50

while BLEU collapsed due to the Generative Collapse. Our sequential ablation confirms that both the foundational OCR/VQA stages and the Alpaca reasoning subset contribute critically to the final capability, though an exhaustive combinatorial study was computationally prohibitive. This finding establishes ANLS and POPE as the appropriate metrics for low-resource, non-Latin VLM evaluation, where BLEU's whitespace tokenization assumptions break down entirely.

5 Conclusion

We confronted severe multi-modal data scarcity for Meitei Mayek by introducing a resource-efficient synthetic bootstrapping methodology. We demonstrated that state-of-the-art VLMs suffer from complete zero-shot script blindness. By engineering a HiTL-validated 5-piece multi-modal mixture and applying QLoRA to PaliGemma-3B, we established a concrete baseline: ANLS rose +13.0 points and VQA Exact Match improved significantly (McNemar $p = 0.013$). Our ablation identifies the Generative Collapse, motivating vocabulary expansion as the critical next step. Our pipeline—dynamic font rendering, translation caching, and multi-modal interleaving—forms a scalable blueprint for adapting VLMs to marginalized scripts worldwide.

Limitations

Despite the gains reported, several limitations merit acknowledgment. First, VQA Exact Match of 16.00% remains low in absolute terms; this reflects partial, foundational script adaptation rather than robust multi-modal mastery. Second, our translation pipeline heavily relies on the Google Translate API for Meitei (`mni-Mtei`), introducing the risk of inheriting translation artifacts, unnatural phrasing, and a loss of culturally specific idiomatic nuances. Because our evaluation set consists of translated English benchmarks rather than organically collected Meitei visual data, it is difficult to fully disentangle genuine script-level reasoning from adaptation to these translation regularities. Third, our image rendering depends exclusively on the Noto Meetei Mayek font family, meaning font-style diversity, critical for real-world OCR generalization on handwritten signs or creative typography, is entirely absent. Consequently, while this work provides a replicable synthetic blueprint, future efforts must integrate naturally occurring Meitei visual data to bridge the synthetic-to-real domain gap.

Future Work

Several natural extensions emerge. The bootstrapping methodology is script-agnostic; applying it to other low-resource Indic scripts such as Chakma, Mizo, or Gondi requires only a font file and a translation API. The Generative Collapse directly motivates vocabulary-extension: adding a Meitei-specific token vocabulary to PaliGemma's tokenizer before fine-tuning would enable genuine long-sequence generation rather than byte-sequence memorization. Scaling to PaliGemma-10B or 7B-class models and integrating real-world scanned Meitei documents, as they become digitized, would further bridge the synthetic-to-real domain gap. We also intend to evaluate cross-lingual transfer to closely related Tibeto-Burman scripts such as Tangsa and Thadou.

Ethical Considerations

This work relies heavily on synthetic data generation and automated translation APIs, which introduces inherent risks of propagating bias, unnatural phrasing, and cultural erasure. While we mitigated these risks via Human-in-the-Loop validation, the underlying models and APIs were trained on distributions that historically under-represent indigenous perspectives. The resulting dataset and adapter weights are intended exclusively for academic research and foundational capability development; they should not be deployed in production or high-

stakes settings without extensive, community-led evaluation and native alignment.

Acknowledgements

We acknowledge the use of Gemini as an AI assistant during this project. Specifically, it was used for structural editing and stylistic refinement of the manuscript, as well as for code generation and debugging during the development of our synthetic data pipeline and training infrastructure.

References

Priyobroto Acharya, Haranath Mondal, Dipanjan Saha, Dipankar Das, and Sivaji Bandyopadhyay. 2025. Junlp: Improving low-resource indic translation system with efficient lora-based adaptation. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1201–1209.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, and 1 others. 2024. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Ali Faraz, Akash, Shaharukh Khan, Raja Kolla, Akshat Patidar, Suranjan Goswami, Abhinav Ravi, Chandra Khatri, and Shubham Agarwal. 2026. Indicvisionbench: Benchmarking cultural and multilingual understanding in VLMs. In *The Fourteenth International Conference on Learning Representations*.

Sajjad Ghiasvand, Haniyeh Ehsani Oskouie, Mahnoosh Alizadeh, and Ramtin Pedarsani. 2025. Few-shot adversarial low-rank fine-tuning of vision-language models. *arXiv preprint arXiv:2505.15130*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Shaharukh Khan, Ayush Tarun, Abhinav Ravi, Ali Faraz, Akshat Patidar, Praveen Kumar Pokala, Anagha Bhangare, Raja Kolla, Chandra Khatri, and Shubham Agarwal. 2025. Chitrarth: Bridging vision and language for a billion people. *Preprint*, arXiv:2502.15392.

Brayan Stiven Lancheros, Gloria Corpas Pastor, and Ruslan Mitkov. 2025. Data augmentation and transfer learning for cross-lingual named entity recognition in the biomedical domain. *Language Resources and Evaluation*, 59(2):665–684.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Sen Lin, Turdi Tohti, Yi Liang, and Askar Hamdulla. 2024. From bytes to words: Enhancing multilingual nmt through explicit byte integration. In *International Conference on Intelligent Multilingual Information Processing*, pages 262–276. Springer.

Arijit Maji, Raghvendra Kumar, Akash Ghosh, Anushka, Nemil Shah, Abhilekh Borah, Vanshika Shah, Nishant Mishra, and Sriparna Saha. 2025. 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, Suzhou, China. Association for Computational Linguistics.

Loitongbam Sanayai Meetei, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2024. Exploiting multiple correlated modalities can enhance low-resource machine translation quality. *Multimedia Tools and Applications*, 83(5):13137–13157.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2025. Large language models: A survey. *Preprint*, arXiv:2402.06196.

Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. 2023. Human-in-the-loop machine

learning: a state of the art. *Artificial Intelligence Review*, 56(4):3005–3054.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Timo Schick and Hinrich Schütze. 2021. Few-shot text generation with natural language instructions. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 390–402.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.

A Synthetic Task Examples

The six task categories in our synthetic Meitei multi-task corpus, with representative input-output pairs, are described in Table 6. The taxonomy was designed to force the model to learn a single unified Meitei representation rather than task-specific heuristics, each task type requires a different interaction between the visual encoder and the language head, preventing any single pathway from dominating.

B Training Configuration Details

Table 7 details the full set of QLoRA fine-tuning hyperparameters used in our experiments. The exceptionally high LoRA rank ($r = 128$) was a deliberate design choice: adapting to an entirely new orthography requires substantially more adapter capacity than standard downstream task fine-tuning, where the model already has a strong linguistic prior. Regarding input formatting, all input prompts follow the uniform template `"<image>answer {prompt}"`. During tokenization, a custom data collator dynamically pads sequences to the maximum length within each batch. Crucially, target label sequences are formatted such that prompt and image tokens are masked using PyTorch's ignore index (`-100`), ensuring that Cross-Entropy loss is computed exclusively over the autoregressive generation of the target answer.

C Evaluation Data Sources

Test sets were drawn from TextVQA by LMMS-Lab and Flickr30k (Plummer et al., 2015), strictly isolated from the training distribution. The combined evaluation subset comprised 300 queries (150 English, 150 dynamically translated Meitei). Zero-shot baselines (Qwen2.5-VL-3B-Instruct and Chitrarth-7.5B) were evaluated on identical data splits for fair comparison. Translation into Meitei was performed using the Google Translate API (`mni-Mtei`) with the same caching infrastructure as the training pipeline, ensuring consistency between training-time and evaluation-time translation quality.

D Morphological Scoring Dynamics

To illustrate why ANLS is a structurally superior metric to Exact Match for agglutinative scripts like Meitei Mayek, consider the following structural examples of root-morphology matching (presented

Table 6: Synthetic task taxonomy with examples.

Task	Input	Output
`ocr`	Image of rendered Meitei text	Verbatim Meitei Unicode transcription
`vqa`	Image + Meitei question	Meitei factual answer
`visual_translation`	Image of Meitei text	English semantic equivalent
`text_translation_m2e`	Meitei text query	English translation
`text_translation_e2m`	English text query	Meitei (Meetei Mayek) translation
`instruct`	Declarative Meitei sentence	Interrogative / summarized form

Table 7: QLoRA fine-tuning hyperparameters.

Hyperparameter	Value
Base model	PaliGemma-3B
Quantization	4-bit NF4 (`float16` compute)
LoRA rank r	128
LoRA α	256
LoRA dropout	0.05
Target modules	`q/k/v/o_proj`, `gate/up/down_proj`
Learning rate	2×10^{-4}
Epochs	4
Per-device batch size	2
Gradient accumulation	8 steps (effective batch: 16)
Mixed precision	BF16 / FP16
Hardware	NVIDIA RTX A6000 (48GB)

here in Latin transliteration for typographic clarity). In Meitei, root words are heavily modified by suffixes for tense, case, and plurality.

- **Example 1 (Tense Variation):** Target is *chatli* (going), Prediction is *chatkhi* (went). Both share the root *chat* (go). Exact Match yields 0.00, whereas ANLS yields 0.50, correctly crediting the model for recognizing the core action.
- **Example 2 (Plurality vs Case):** Target is *mioishing* (people), Prediction is *mioina* (by a person). Both share the root *mioi* (person). Exact Match yields 0.00, whereas ANLS yields 0.60, correctly capturing the foundational semantic grounding despite the suffix mismatch.

These variations are exceedingly common in VQA tasks where the model correctly identifies the primary object or action but defaults to a baseline suffix rather than the specific suffix demanded by the question's grammatical context.