Culturally Grounded Image Captioning in Indigenous Languages with Vision-Language Models: Cascaded and Single-Stage Approaches

Mirelle Bueno
Motorola Mobility
mirellec@motorola.com

Sushil Garg
Motorola Mobility
sushilgarg@motorola.com

Abstract

Culturally grounded image captioning for under-resourced Indigenous languages is challenging due to severe data scarcity and the need to describe culturally specific visual content. This paper describes our submission to the AmericasNLP 2026 shared task, where we evaluate two architectural paradigms for caption generation across Bribri, Guaraní, Yucatec Maya, Wixárika, and Orizaba Nahuatl. First, we implement a cascaded system that combines a large vision-language model with a machine translation pipeline, showing that culturally contextualized, persona-based prompting improves over the official baseline in most comparable settings. Second, we develop a direct, end-to-end Single-stage approach by adapting PaliGemma 2 using LoRA fine-tuning, continued pre-training, and multilingual joint training. Our single-stage experiments show that, despite severe domain mismatch and reliance on synthetic training data, multilingual training and continued pre-training improve automatic chrF++ relative to single-language LoRA fine-tuning in some settings. Overall, cascaded pipelines remain the strongest among the evaluated approaches under current data constraints, while single-stage models remain a promising but currently data-limited path toward direct Indigenous-language image captioning.

1 Introduction

Recent progress in vision-language modeling has led to strong performance on image captioning benchmarks, especially in high-resource languages and domains with large-scale annotated data. However, culturally grounded image captioning for Indigenous languages remains substantially more difficult.

In this paper, we describe our system for the AmericasNLP 2026 shared task (Bui et al., 2026) on culturally grounded image captioning for under-resourced Indigenous languages. The task requires participants to generate natural-language image captions for five target languages: Bribri, Guaraní, Yucatec Maya (hereafter Maya), Wixárika, and Orizaba Nahuatl (hereafter Nahuatl). We approach the task through two paradigms. The first is a cascaded architecture in which a general-purpose vision-language model generates a Spanish caption, which is then translated into the target Indigenous language using a machine translation system derived from the AmericasNLP 2023 shared task submission (Gow-Smith and Sánchez Villegas, 2023) designed to translate from Spanish into eleven Indigenous languages. The second is a single-stage architecture that adapts a vision-language model to generate captions directly in the target language, avoiding an explicit intermediate translation step. Furthermore, to the best of our knowledge, this is among the first efforts to adapt a single-stage image-captioning system specifically for Indigenous-language caption generation in this shared-task setting.

Our experiments demonstrate that the cascaded persona-based system is the strongest among the evaluated approaches overall across the development and test sets. Specifically, on the test set, it outperforms the baseline by 1.46 chrF++ for Wixárika, 0.92 for Bribri, and 4.54 for Nahuatl, while underperforming the baseline for Guaraní. These results indicate that persona-based prompting is beneficial in most settings but remains sensitive to language-specific translation quality. Furthermore, we observe that employing a powerful vision-language model for caption generation does not inherently guarantee superior final quality, as performance remains heavily contingent upon the quality of the translation model.

For the Single-stage approach, we explored diverse training methodologies, including continued pre-training, LoRA supervised fine-tuning (Hu et al., 2021), and multilingual training. Our findings show that multilingual training improves over

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 248–256

single-language LoRA fine-tuning in our development experiments. This suggests that multilingual training may provide useful transfer for image captioning, although our experiments do not isolate transfer from the effects of increased training data volume. A similar phenomenon has been widely documented in multilingual neural machine translation (Aharoni et al., 2019; Iyer et al., 2024). For Wixárika, continued pre-training provided substantial gains over alternative techniques, narrowing the gap with the official baseline to 2.99 chrF++ points on the test set. These insights may support the development of more effective vision-language models (VLMs) for Indigenous languages, reducing dependence on translation steps or intermediate modules.

Consequently, the primary contributions of this paper are twofold: first, we describe the systems submitted to the AmericasNLP 2026 shared task and evaluate the impact of prompting on culturally contextualized caption generation; second, we present an initial empirical comparison of several VLM adaptation strategies aimed at integrating extremely low-resource languages previously unseen by VLMs. The datasets and code are available at: https://github.com/MirelleB/InclusionVLM

2 Task and Data Description

The shared task challenges participants to develop multimodal systems that generate culturally grounded natural-language descriptions of images in under-resourced Indigenous languages. Formally, given an input image I, the objective is to generate a caption $S = \{w_1, w_2, ..., w_n\}$ in a target Indigenous language L. In contrast to conventional benchmarks such as MS-COCO (Lin et al., 2014), which prioritize generic object detection, this task emphasizes cultural granularity. Models must move beyond superficial labeling (e.g., "a colorful craft"), where visually appropriate, to identify culturally specific entities, such as Wixárika ceremonial artifacts, requiring a deeper integration of visual features and community-specific semantic knowledge.

For the evaluation phase, the shared task adopted a two-stage ranking methodology. Initially, all submitted systems were ranked automatically using the chrF++ metric. Subsequently, the top five systems advanced to a human evaluation stage, where they were assessed against a fixed set of standardized criteria.

2.1 Data Description

The corpus curated for this shared task is designed to reflect the sociocultural heritage and daily practices of the participating Indigenous communities. In contrast to generic image-captioning benchmarks, this dataset prioritizes domain-specific semantic density and cultural authenticity. The primary tracks encompass Bribri, Guaraní, Maya, Wixárika and Nahuatl. For each language, the development set consists of approximately 50 high-quality, manually annotated image-caption pairs. A blinded test set is utilized for the final benchmarking process. To ensure the integrity of the evaluation and prevent overfitting to the test distribution, ground-truth references are withheld until the conclusion of the competitive cycle. The annotations exhibit a high degree of descriptive specificity that poses a challenge to standard multimodal architectures. While a conventional vision-language model might generate a generic description such as "a person sitting," the shared task dataset annotations frequently employ a specialized lexicon to denote specific traditional activities. Consequently, successful systems must move beyond broad visual categorization toward a fine-grained understanding situated within the community's cultural context.

2.2 Evaluation Metric

We evaluated all developed systems using the mean sentence-level chrF++ score (Popović, 2017) via sacreBLEU (Post, 2018) implementation across the development subset, as recommended by the shared task organizers.

3 System Description

Two distinct architectural approaches were implemented and evaluated:

(i) A Cascaded framework, which utilizes a frozen Vision-Language Model (VLM) to generate culturally situated captions, followed by a frozen machine translation (MT) model to translate the output into the target language; a similar configuration also served as the baseline of the shared task.

(ii) A Single-stage approach, involving the development of a unified, end-to-end model capable of simultaneously interpreting the image and generating culturally situated captions directly in the target language.

Although the cascaded system yielded results competitive with the baseline, we hypothesize that

the single-stage architecture offers potential long-term advantages for under-resourced Indigenous-language settings. By generating captions natively, these models avoid an explicit translation step at inference time and may reduce the semantic distortion and information loss introduced by intermediate translation pipelines.

3.1 Cascaded approach

The cascaded architecture utilizes a VLM to generate an initial caption, which is subsequently processed by an MT model for translation. This framework is frequently adopted in low-resource linguistic scenarios, particularly where foundation models lack exposure to the target languages during pretraining (Geigle et al., 2025). A key architectural advantage is its modularity, allowing for the independent substitution of either the VLM or the MT component. However, the use of machine translation can lead to a loss of cultural nuance (Venuti, 2008; Szymańska, 2017).

In our cascaded pipeline, which is similar to the official shared task baseline, Gemini 2.5 Flash (Gemini Team et al., 2025) generated the Spanish image captions. These captions were subsequently translated into the target Indigenous languages utilizing the Sheffield translation system developed for the AmericasNLP 2023 Shared Task (Gow-Smith and Sánchez Villegas, 2023) for Bribri, Guaraní, Wixárika, and Nahuatl; for Maya, we utilized the separately trained Spanish-to-Maya MT model described in Section 3.7.

The Spanish captioning phase was executed with hyperparameters set to a temperature of 0.01 and top-p of 0.01. We conducted an ablation study of the following prompting strategies. The prompts were instantiated separately for each target community by replacing the language- and culture-specific descriptors; Appendix A shows the Wixárika instantiation used in the prompt ablation.

i) Baseline: The official shared task prompt, which includes comprehensive instructions regarding the cultural context of the target languages.

ii) Persona-based: A significantly more concise prompt focused on cultural perspective. This approach uses an explicit cultural-role framing to encourage concise captions that include culturally relevant details. Because such prompting may also encourage over-specific interpretation, we interpret improvements primarily in terms of automatic chrF++ rather than validated cultural fidelity.

iii) Question-Answer Chain (QA Chain): Drawing inspiration from prior work (Ibrahim et al., 2025; Bai et al., 2025), this three-stage approach begins with the LLM generating five questions based on image elements. In the second stage, the model answers these questions by leveraging its internal knowledge of the target culture. Finally, the model synthesizes these inputs to produce a contextualized caption.

While the QA Chain strategy offers a more exhaustive extraction of cultural detail, it is susceptible to error propagation; any inaccuracies in the intermediate reasoning phase are carried through to the final caption and may remain undetected by standard automatic metrics.

These strategies were initially evaluated on the Wixárika development set. To further investigate potential performance gains, we also assessed the impact of Gemini 2.5 Pro, constraining this analysis strictly to the Persona-based approach.

As demonstrated in Table 1, the Persona-based approach outperformed the baseline by a margin of 1.5 chrF++ points. Interestingly, the results also revealed a 0.53-point drop in the score when employing Gemini 2.5 Pro. These results suggest that using the larger Gemini 2.5 Pro model does not necessarily translate into higher final chrF++ scores, likely because downstream MT quality can dominate the final output. Consequently, we adopted the Gemini 2.5 Flash in conjunction with the Persona-based approach for the final evaluation across all target languages.

Method	**chrF++**
Baseline	17.77
Persona + Gemini 2.5 Flash	19.25
Persona + Gemini 2.5 Pro	18.72
QA Chain	17.07

Table 1: chrF++ comparison between prompt strategies on the Wixárika development set.

Table 2 presents the evaluation results on the development and test sets for the persona-driven prompting strategy across the target languages. These findings demonstrate that the Persona-based prompting improves over the baseline on the development set for all comparable languages and on the test set for Bribri, Wixárika, and Nahuatl. However, it underperforms the baseline on the Guaraní test set, and no Maya baseline is available for direct comparison. We hypothesize that persona-based prompting improves chrF++ in most

comparable settings, suggesting that culturally oriented prompts may help, though automatic scores alone do not establish cultural fidelity.

Language	Approach	Dev	Test
Bribri	Baseline	7.57	7.01
	Persona-based	8.49	7.93
Wixárika	Baseline	17.77	16.91
	Persona-based	19.25	18.37
Maya	Baseline	–	–
	Persona-based	9.00	16.96
Guaraní	Baseline	20.82	20.13
	Persona-based	21.75	16.48
Nahuatl	Baseline	11.53	9.52
	Persona-based	15.68	14.06

Table 2: chrF++ scores for the cascaded baseline and persona-based prompting approach across target languages. Dashes indicate scores that were unavailable.

3.2 Single-stage approach

Our primary objective is to train a single-stage VLM capable of generating culturally situated captions in the target languages without relying on intermediate modules. While many high-resource image-captioning settings are now effectively handled by proprietary and open-source models, it presents a formidable challenge for extremely low-resource languages due to the acute scarcity of both monolingual corpora and task-specific datasets.

To address this challenge, we evaluated three distinct training strategies:

(i) LoRA fine-tuning: Training directly on image-caption pairs in the target languages.

(ii) Continued Pre-training: Adapting the VLM's language component to the target language through unsupervised learning, followed by fine-tuning on image-caption pairs.

(iii) Multilingual Joint Fine-tuning: Developing a single unified model trained simultaneously on image-caption pairs across Wixárika, Bribri, Guaraní, and Maya.

Each strategy serves a specific investigative purpose: approach (i) assesses the model's linguistic abilities in languages not seen during pre-training; approach (ii) quantifies the impact of continued pre-training on the language backbone; and approach (iii) investigates whether cross-lingual transfer is effective for image captioning.

All experiments utilized the PaliGemma 2 (3B) pre-trained model (Steiner et al., 2024). This architecture was selected for its competitive performance in image captioning relative to significantly larger models and for its integration within the Gemma ecosystem, which streamlines the optimization of the language module. For strategy (ii), our experiments were focused exclusively on Wixárika.

The following sections provide a description of the training and evaluation datasets, alongside the technical specifications of our training protocols.

3.3 Data preparation

While constructing massive image-captioning datasets for high-resource languages remains a significant undertaking, doing so for extremely low-resource languages presents a formidable challenge. In response to this data scarcity, existing literature frequently leverages machine translation models to synthesize captions across multiple target languages. Adopting a similar methodology, we utilized the winning system from the AmericasNLP 2023 Shared Task to translate captions from the Polaris dataset (Wada et al., 2024). Polaris was selected due to its status as a large-scale, human-annotated corpus specifically designed for evaluating image captioning systems, encompassing approximately 131,000 human judgments. To accommodate training time constraints and ensure data quality, we filtered the dataset to include only examples with a human-annotated score exceeding 0.75. This threshold reflects high confidence in the captions' quality, accounting for critical factors such as fluency, relevance, and descriptive granularity. The filtering yielded a training dataset of 13,562 samples. It is important to note that Polaris captions differ fundamentally from the shared task dataset in that they lack specific cultural context. Our primary objective in utilizing Polaris is to facilitate cross-modal alignment, enabling the model to map visual elements to their corresponding linguistic representations in the target language. However, a potential limitation of this approach is that the use of synthetic (translated) data may introduce cascading noise during the training phase.

3.4 Monolingual data

For the continued pre-training phase, Wixárika source data were aggregated from previous AmericasNLP shared tasks; specifically, 93,247 tokens were sourced from AmericasNLI (Ebrahimi et al., 2022) and 189,362 tokens from AmericasNLP 2023 (Ebrahimi et al., 2023). These corpora provide human-curated or human-translated text rather

than model-generated synthetic captions. In total, the monolingual corpus utilized in this study comprised 282,609 tokens for Wixárika; it is important to note that this corpus is extremely small relative to the data typically used for language-model adaptation, highlighting the severe resource constraints under which the model was trained and the importance of exploring external datasets.

3.5 LoRA Fine-tuning

The training was conducted over ten epochs using the Low-Rank Adaptation (LoRA) approach with a rank (r) of 8. We used LoRA for parameter-efficient adaptation while limiting updates to the base model parameters; specifically, we aimed to preserve the model's pre-established cross-modal capabilities while simultaneously extending its linguistic proficiency to previously unseen target languages. The optimization protocol utilized a learning rate of 1×10^{-5} and a batch size of 4. For the training objective, we implemented a concise instruction format: each input was prepended with the mandatory PaliGemma <image> token, followed by the Spanish prompt *Generar pie de foto*.

3.6 Continued Pre-training and VLM Alignment

The continued pre-training framework was executed in two distinct stages. The initial phase focused on domain adaptation of the language backbone—specifically the Gemma-2 2B model. This stage aimed to integrate knowledge of a previously unseen language into the model using the causal language modeling pretext task of next-token prediction. To prevent the degradation of existing pre-trained knowledge, we employed LoRA with a rank (r) of 16. Furthermore, we incorporated a replay buffer consisting of approximately 4% English monolingual data to serve as a regularization mechanism, ensuring the retention of the model's primary language capabilities. This phase was conducted with a batch size of 4 and a learning rate of 2×10^{-4} over three epochs, totaling 1,071 optimization steps. Following the adaptation of the language model, we proceeded to the cross-modal alignment phase. During this stage, the Vision Tower and Language Model were kept frozen, with the training objective concentrated solely on optimizing the multimodal projector. This alignment phase utilized the same instructional prefix, learning rate, and hyperparameter configuration as described in the previous LoRA fine-tuning section.

3.7 Multilingual Joint Fine-tuning

This approach extends the methodology described in the previous fine-tuning section by incorporating multiple Indigenous languages into a single training objective. The primary motivation behind this experiment is to investigate the potential for cross-lingual transfer and determine whether synergistic effects between different low-resource languages can enhance performance in the target task.

The training corpus comprised data from Bribri, Maya, Guaraní, and Wixárika. In accordance with our established fine-tuning protocol, we translated the curated Polaris dataset into each of these target languages. For Bribri, Guaraní, and Wixárika, we employed the winning translation system from the AmericasNLP 2023 Shared Task. However, because Maya was unsupported by this Sheffield-developed system, we trained an independent Spanish-to-Maya machine translation model to facilitate both cascaded inference and synthetic data generation.

To manage the multi-target nature of this stage, a multilingual prompting strategy was adopted. Each input was formatted with the <image> prefix, followed by the instruction: *'Generar pie de foto en [language]'*, where the placeholder was dynamically substituted with the corresponding target language of the training sample.

4 Results

Table 3 reports the performance of the proposed methodologies. Single-stage experiments were conducted for Bribri, Guaraní, Maya, and Wixárika; Nahuatl was evaluated only in the cascaded setting. For clarity, the Low-Rank Adaptation fine-tuning is denoted as LoRA SFT, the Continued Pre-training and VLM Alignment as CPT + Alignment, and the Multilingual Joint Fine-tuning as Multilingual SFT. The baseline values cited correspond to the official scores reported by the shared task organizers.

Our results indicate that, despite the significant domain shift between our training set and the shared task dataset—specifically regarding visual stylistic diversity and the required descriptive complexity—the trained models demonstrated promising but limited performance in both Bribri and Wixárika languages. On the development set, the best single-stage models trailed the baseline by 2.32 chrF++ for Bribri and 2.64 chrF++ for Wixárika. On the test set, the corresponding gaps were 4.47 for Bribri and 2.99 for Wixárika. Per-

formance was less competitive for Guaraní, where the best single-stage model remained 10.51 chrF++ below the development baseline and 12.52 chrF++ below the test baseline. For Maya, no official baseline was available, so the single-stage results should be interpreted without direct baseline comparison.

Furthermore, the Multilingual Joint Fine-tuning approach improves over standard LoRA fine-tuning on the development set, suggesting that multilingual training and resource sharing can improve direct caption generation under low-resource conditions.

These findings suggest that the models acquired some cross-modal mapping capabilities even under suboptimal data conditions. We posit that by refining the training corpora to include more culturally grounded samples and specifically incentivizing the model to prioritize Indigenous sociocultural nuances, the performance of these single-stage systems may narrow the gap to current baselines.

Language	Approach	Dev	Test
Wixárika	Baseline	17.77	16.91
	LoRA SFT	11.79	–
	CPT + Alignment	15.13	13.92
	Multilingual SFT	12.72	10.51
Bribri	Baseline	7.57	7.01
	LoRA SFT	2.71	–
	Multilingual SFT	5.25	2.54
Maya	Baseline	–	–
	LoRA SFT	7.69	–
	Multilingual SFT	9.00	9.13
Guaraní	Baseline	20.82	20.13
	LoRA SFT	7.65	–
	Multilingual SFT	10.31	7.61

Table 3: chrF++ scores for Single-stage approaches across target languages on the development and test sets. Nahuatl was not evaluated in the single-stage setting due to resource and development timeline constraints. Dashes indicate scores that were unavailable or configurations that were not evaluated.

5 Related Work

Recent vision-language models (VLMs) remain predominantly English-centric, struggling to capture culturally embedded concepts in low-resource settings. While prior works extend architectures to multiple languages (e.g., Maya, a multilingual VLM (Alam et al., 2025), M-MiniGPT4 (Han et al., 2026)), they primarily target broad multilingual coverage rather than culturally grounded Indigenous-language captioning. Furthermore, these efforts often rely heavily on translated data. Although translation provides a scalable foundation, models adapted with native multimodal pairs (e.g., Chinese CLIP (Yang et al., 2022), DanQing (Shen et al., 2026)) demonstrate that culturally situated data is crucial for deep semantic fidelity.

In the absence of native end-to-end models, cascaded pipelines combining high-resource VLMs with machine translation offer a practical alternative (Geigle et al., 2025). However, these pipelines risk compounding errors by generating generic descriptions (Lin et al., 2014) that strip away community-specific nuances. Unlike recent retrieval-augmented cultural captioning methods (Ibrahim et al., 2025), we operate under severe resource constraints. We therefore contrast cascaded prompting with single-stage PaliGemma adaptation strategies (e.g., LoRA, continued pretraining) to assess the viability of end-to-end culturally grounded captioning for Indigenous languages.

6 Conclusion

This paper presents the development of multimodal systems for generating culturally grounded, natural-language descriptions of images in under-resourced Indigenous languages within the AmericasNLP shared-task framework. We propose and evaluate two distinct methodologies: a cascaded pipeline comprising VLM-generated captions followed by a machine translation (MT) system, and a direct, single-stage VLM. In the cascaded setup, persona-based prompting improved over the official baseline in most comparable settings, suggesting that explicitly encouraging a culturally grounded perspective can improve automatic captioning scores.

Experiments with the single-stage VLM indicate that direct caption generation can be trained, but it remains substantially less competitive than the Cascaded approach under the present data constraints. On the development set, the best single-stage models remained 2.64 chrF++ below the Wixárika baseline and 2.32 chrF++ below the Bribri baseline, indicating partial but incomplete progress toward direct caption generation due to severe data scarcity and cross-domain mismatch. Single-stage VLMs exhibit potential, provided they are trained on extensive, culturally grounded datasets that reward community-specific semantic fidelity over generic

visual descriptions.

Overall, our findings highlight a practical trade-off: while cascaded systems currently yield stronger short-term performance than the evaluated single-stage systems, single-stage systems offer a direct approach to Indigenous-language modeling, albeit requiring more robust data curation and language adaptation. Our final shared-task submission ranked 7th overall, placing in the top five for three of five evaluated languages: Maya, Nahuatl, and Wixárika.

7 Limitations

This study is subject to several limitations. First, our findings are constrained by the shared task's restricted set of target languages and small development sets, which limits both statistical power and generalizability to other Indigenous languages. Second, reliance on synthetically translated training data from the Polaris dataset inevitably introduces translation artifacts and omits crucial community-specific cultural grounding, potentially restricting models to surface-level cross-modal alignment. Third, our evaluation relies predominantly on automatic chrF++ scores; while effective for measuring character-level similarity, this metric cannot adequately capture cultural fidelity, factual grounding, or terminological accuracy. Fourth, our prompting experiments explore a limited set of inference strategies, omitting advanced paradigms such as chain-of-thought or retrieval-augmented generation. Fifth, our persona-based and QA-chain prompts rely on explicit cultural-role framing. While this framing improved chrF++ in several settings, it may also encourage over-specific cultural interpretations or hallucinated cultural associations, and we did not conduct community or speaker validation of the generated captions. Finally, our architectural comparisons are confined to the cascaded Gemini pipeline and single-stage PaliGemma 2 models; consequently, these results may not generalize to broader vision-language architectures.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. *Preprint*, arXiv:1903.00089.

Nahid Alam, Karthik Reddy Kanjula, Surya Guthikonda, Timothy Chung, Bala Krishna S. Vegesna, Abhipsha Das, Anthony Susevski, Ryan Sze-Yin Chan, S. M. Iftekhar Uddin, Shayekh Bin Islam, Roshan Santhosh, Snegha A, Drishti Sharma, Chen Liu, Isha Chaturvedi, Genta Indra Winata, Ashvanth.S, Snehanshu Mukherjee, and Alham Fikri Aji. 2025. Behind Maya: Building a multilingual vision language model. *Preprint*, arXiv:2505.08910. Accepted at VLMs4ALL CVPR 2025 Workshop.

Longju Bai, Angana Borah, Oana Ignat, and Rada Mihalcea. 2025. The power of many: Multi-agent multimodal models for cultural image captioning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2970–2993, Albuquerque, New Mexico. Association for Computational Linguistics.

Minh Duc Bui, David Guzmán, Abteen Ebrahimi, Franklin Morales, Marvin Agüero-Torales, Raquel Insfrán, Cecilia González, Ramón Araujo, Luca Cernuzzi, Carlos Raul Noh Chi, Carlos Eduardo Tec Cahun, Sindi Estrella Poot Cohuo, Daniel Ricardo Benítez Chi, Santos Natanael Palomo Arévalo, Jessica Elizabeth Canul Canche, Deysi Aracely Poot Poot, Wendy Marleny Dzib Dzib, Eduardo José Ake Pool, Reynaldo Alexander Couoh Martin, Silvia Fernandez Sabido, Luis Samuel Santiago Melchor, Sotero Silverio, Robert Pugh, Raúl Vázquez, John E. Ortega, Arturo Oncevay, Rubén Manrique, Luis Chiruzzo, Rolando Coto-Solano, Elisabeth Mager, Shruti Rijhwani, David Ifeoluwa Adelani, Manuel Mager, and Katharina von der Wense. 2026. Findings of the AmericasNLP 2026 shared task on cultural image captioning for Indigenous languages. In *Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, San Diego, California. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Shruti Rijhwani, Enora Rice, Arturo Oncevay, Claudia Baltazar, María Cortés, Cynthia Montaño, John E. Ortega, Rolando Coto-Solano, Hilaria Cruz, Alexis Palmer, and Katharina Kann. 2023. Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 206–219, Toronto, Canada. Association for Computational Linguistics.

Gregor Geigle, Florian Schneider, Carolin Holtermann, Chris Biemann, Radu Timofte, Anne Lauscher, and Goran Glavaš. 2025. Centurio: On drivers of multilingual ability of large vision-language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2831–2881, Vienna, Austria. Association for Computational Linguistics.

Gemini Team et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Edward Gow-Smith and Danae Sánchez Villegas. 2023. Sheffield's submission to the AmericasNLP shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 192–199, Toronto, Canada. Association for Computational Linguistics.

Seung Hun Eddie Han, Youssef Mohamed, and Mohamed Elhoseiny. 2026. M-MiniGPT4: Multilingual VLLM alignment via translated data. In *Proceedings of the 7th Workshop on African Natural Language Processing (AfricaNLP 2026)*, pages 11–16, Rabat, Morocco. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

George Ibrahim, Rita Ramos, and Yova Kementchedjhieva. 2025. CONCAP: Seeing beyond English with concepts retrieval-augmented captioning. *Preprint*, arXiv:2507.20411. Published as a conference paper at COLM 2025.

Vivek Iyer, Bhavitvya Malik, Pavel Stepachev, Pinzhen Chen, Barry Haddow, and Alexandra Birch. 2024. Quality or quantity? on data scale and diversity in adapting large language models for low-resource translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1393–1409, Miami, Florida, USA. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Hengyu Shen, Tiancheng Gu, Bin Qin, Lan Wu, Yuling Wu, Shuo Tan, Zelong Sun, Jun Wang, Nan Wu, Xiang An, Weidong Cai, Ziyong Feng, and Kaicheng Yang. 2026. DanQing: An up-to-date large-scale Chinese vision-language pre-training dataset. *Preprint*, arXiv:2601.10305.

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. *Preprint*, arXiv:2412.03555.

Izabela Szymańska. 2017. The treatment of geographical dialect in literary translation from the perspective of relevance theory. *Research in Language*, 15:61–77.

Lawrence Venuti. 2008. *The Translator's Invisibility: A History of Translation*, 2nd edition. Routledge, London; New York.

Yuiga Wada, Kanta Kaneda, Daichi Saito, and Komei Sugiura. 2024. Polos: Multimodal metric learning from human feedback for image captioning. *Preprint*, arXiv:2402.18091.

An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2022. Chinese CLIP: Contrastive vision-language pretraining in Chinese. *Preprint*, arXiv:2211.01335.

A Cascaded Approach: Prompts

The following prompts are reported verbatim for reproducibility. They reflect the prompts used during the official submission and were not revised after evaluation; therefore, the results should be interpreted as measuring the behavior of these specific prompts rather than an endorsement of their cultural assumptions.

(A) PERSONA-BASED PROMPT

Eres wixárika (huichol), conoces y practicas la cultura.
Basándote en tu cultura, usa todo tu conocimiento para generar un pie de foto conciso, respetuoso y culturalmente preciso.
Ejemplo corto (PREFERIDO):
"Cuadro de estambre wixárika con patrones que representan visiones chamánicas. Los colores brillantes y diseños simbólicos son característicos del arte ceremonial contemporáneo."
Genera pies de foto concisos siguiendo este formato.

(B) QUESTION-ANSWER CHAIN
[Question generator prompt]

Considerando la imagen, crea 5 preguntas únicas relacionadas con la cultura.
Es importante mencionar que las imágenes pertenecen a la cultura wixárika (huichol), una comunidad indígena de la Sierra Madre Occidental, México.
Formato: Pregunta 1: , Pregunta 2: ,...

[Prompt that answers the questions]
Eres un experto en la cultura Wixárika (Huichol), una comunidad indígena de la Sierra Madre Occidental en México y debes responder correctamente a las preguntas; utiliza la imagen para contextualizar tu respuesta. Formato: pregunta 1: respuesta 1 / pregunta 2: respuesta 2

[Captioning prompt]
Teniendo en cuenta la imagen y las preguntas y respuestas relacionadas con la cultura Wixárika (Huichol), una comunidad indígena de la Sierra Madre Occidental en México, utiliza esta información para generar un pie de foto conciso de la imagen en una o dos frases.

Figure 1: Prompts used to evaluate the cascaded approach: (A) persona-based prompting and (B) chained question-answer prompting.