# USP at AmericasNLP 2026 Shared Task: Culturally-Aware Image Captioning for Indigenous Languages via Vision-Language Models and Fine-Tuned Neural Machine Translation

**Rafael M. Fernandes**
University of São Paulo (USP)
São Paulo, Brazil
rafael.macario@usp.br

## Abstract

We describe the USP system for the AmericasNLP 2026 Shared Task on Culturally Relevant Image Captioning for Indigenous Languages, covering Guaraní (grn), Maya Yucateco (yua), Nahuatl (nah), Wixárika (hch), and Bribri (bzd). We propose a two-stage cascade: Qwen3-VL-8B-Instruct (Bai et al., 2025) generates Spanish captions via language-specific cultural prompts; language-specific fine-tuned NLLB-200-distilled-600M (NLLB Team et al., 2022) models then translate them into each target language. We train on AmericasNLP 2023 data (Ebrahimi et al., 2023) augmented with public parallel corpora. Our system achieves competitive results, including **3rd place in Guaraní human evaluation** (2.41/5.0) and 5th in Bribri (1.09/5.0) among 8 teams. We also report that NLLB-200-distilled-600M silently lacks vocabulary entries for Bribri and Maya Yucateco, producing English output without error.

## 1 Introduction

Culturally relevant image captioning for Indigenous languages requires understanding not just what is depicted, but what it *means* within a specific community (Yun and Kim, 2024; Gao et al., 2025). Standard captioning benchmarks are predominantly Western-centric (Lin et al., 2014), and even state-of-the-art VLMs exhibit substantial cultural blind spots for non-Western communities (Burda-Lassen et al., 2025; Romero et al., 2024; Lupascu et al., 2025). A picture of a ceramic vessel might be described as "a clay pot" when the culturally meaningful description is *guampa para tereré*—a vessel central to Guaraní identity.

The AmericasNLP 2026 Shared Task on Culturally Relevant Image Captioning (Bui et al., 2026) provides images from five Indigenous communities: Guaraní (Paraguay, Brazil), Maya Yucateco (Mexico), Nahuatl (Mexico), Wixárika (Mexico), and Bribri (Costa Rica). These languages have been central to AmericasNLP MT tasks since 2021 (Ebrahimi et al., 2023, 2024; de Gibert et al., 2025) and educational materials creation (Lupicki et al., 2025; Vasselli et al., 2025), providing established corpora on which our system builds.

We present a two-stage cascade (Figure 1): a VLM generates culturally-prompted Spanish captions, which are then translated per language by fine-tuned NLLB-200. Using Spanish as a pivot (Utiyama and Isahara, 2007) is natural for our setting: it is the dominant contact language for all five communities, all major parallel corpora are Spanish-indexed, and recent low-resource captioning systems adopt the same strategy (Oduwole et al., 2026; Jain et al., 2021).

Our contributions are:

1. A two-stage cascade covering all five shared task languages, with language-specific cultural prompting requiring no VLM fine-tuning.
2. Documentation of a previously unreported failure mode: NLLB-200 silently generates English for Bribri and Maya Yucateco due to missing vocabulary tokens.
3. Empirical evidence that domain mismatch between general-domain MT corpora and image captions can outweigh gains from language-specific fine-tuning.

## 2 Related Work

**Cultural captioning.** Yun and Kim (2024) propose CIC, a VQA+LLM framework that elicits cultural elements (traditional clothing, ritual objects) to enrich captions. Karamolegkou et al. (2024) show that cultural prompting improves *human-judged* quality even when ChrF++ declines—suggesting automatic metrics undervalue culturally enriched outputs. Buettner et al. (2025) use multimodal recaptioning with native-speaker examples to correct English-centric perceptual bias. Gao et al. (2025) construct MELLA, a dataset of cultur-



*Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 264–271

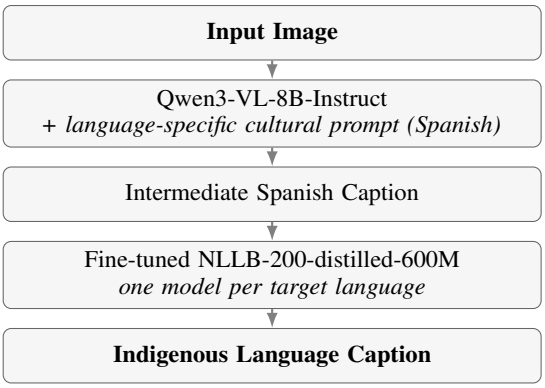


Figure 1: Two-stage cascade. Stage 1 generates a culturally-prompted Spanish caption; Stage 2 translates it into the target Indigenous language.

ally grounded "thick descriptions" for low-resource MLLMs.

**Cascade MT for low-resource captioning.** Jain et al. (2021) show that combining image-text with bitext training (MURAL) compensates for sparse caption data in low-resource languages. Oduwole et al. (2026) apply a nearly identical NLLB-based pivot cascade to African languages, finding that while the architecture is effective, domain mismatch between MT training data and image captions is the primary limitation—a finding we corroborate.

**Cultural bias in VLMs.** VLMs trained on Western-centric data fail substantially on culturally specific visual content (Burda-Lassen et al., 2025; Cao et al., 2024; Lupascu et al., 2025). The CVQA benchmark (Romero et al., 2024) quantifies this gap across 30 countries and 31 languages. Standard metrics like ChrF++ do not capture cultural correctness (Yun and Kim, 2024; Burda-Lassen et al., 2025); fine-tuned semantic embeddings (LaBSE) correlate better with human judgments for Guaraní and Bribri (Krasner et al., 2025).

**AmericasNLP community.** NLLB-200 fine-tuning has been the dominant approach in recent AmericasNLP MT tasks for our target languages (DeGenaro and Lupicki, 2024; García Gilabert et al., 2024; de Gibert et al., 2025). LLM prompting with community-specific context is competitive for Guaraní, Maya, Bribri, and Nahuatl (Lupicki et al., 2025; Vasselli et al., 2025). Bribri morphology presents particular challenges for sequence-to-sequence models (Anderson et al., 2025).

| Language | Key cultural categories in prompt |
|---|---|
| Guaraní | *Tereré*, *ñandutí*, *tatakua*, traditional foods (chipa, sopa paraguaya), mythology (Pombero, Jasy Jatere), Jesuit missions |
| Maya | Henequén, *huipil*, *cenotes*, Hanal Pixán, Cochinita Pibil, Kukulkán, Chaac, archaeological sites |
| Wixárika | Peyote (*hikuri*), *nierika* yarn painting, beadwork (*chaquira*), *mara'akame*, Wirikuta pilgrimage |
| Nahuatl | Milpa, Quetzalcóatl, Day of the Dead, mole, Voladores de Papantla, copal, temazcal |
| Bribri | Cacao, *Sibö*, conical housing, *sukia* healer, chicha, Talamanca territory |

Table 1: Cultural vocabulary categories targeted by each language-specific prompt. Full prompts in Appendix A.

## 3 System Description

### 3.1 Stage 1: Culturally-Prompted Visual Captioning

We use **Qwen3-VL-8B-Instruct** (Bai et al., 2025) to generate a Spanish description for each image (resized to 384×384 px). We design a *language-specific cultural prompt* for each of the five languages, following the approach of Yun and Kim (2024). Each prompt is written in Spanish and instructs the model to generate a concise (2–4 sentence) culturally-aware caption while foregrounding community-specific objects, practices, and symbols.

Table 1 shows the cultural vocabulary categories included in each prompt. The complete prompts are provided in Appendix A.

For Guaraní, we iterated over two prompt versions on the development set: an initial version (V1) with categorical vocabulary lists, and a refined version (V2) that additionally includes three few-shot examples of culturally correct captions. V2 was used for the final test submission; all other languages used a single prompt version with no few-shot examples.

Inference parameters: `do_sample=False`, `max_new_tokens=128`, `repetition_penalty=1.3`. All inference runs on NVIDIA T4 GPUs (Kaggle environment). The VLM is unmodified; no fine-tuning is performed.

| Language | ANLP23 | Total |
|---|---|---|
| Guaraní[†] | ∼14k | ∼46k |
| Nahuatl[‡] | ∼16k | ∼36k |
| Wixárika[§] | ∼9k | ∼16k |
| Bribri[¶] | ∼7.5k | ∼9.2k |
| Maya[‖] | — | ∼11.8k |

Table 2: Training data. ANLP23 = AmericasNLP 2023 (Ebrahimi et al., 2023). Extra corpora: [†]Jojajovai (Chiruzzo et al., 2022); [‡]Axolotl (Gutierrez-Vasques et al., 2016); [§]Pywirrarika (Mager et al., 2018); [¶]Feldman et al. (Feldman and Coto-Solano, 2020); [‖]Iikim Translator (Rangel and Kobayashi, 2024) (no ANLP23 for Maya).

### 3.2 Stage 2: Language-Specific Neural Machine Translation

We fine-tune **NLLB-200-distilled-600M** (NLLB Team et al., 2022) separately for each language, following the dominant approach in recent AmericasNLP MT tasks (DeGenaro and Lupicki, 2024; García Gilabert et al., 2024). For each language we evaluate both the fine-tuned and the base NLLB model on the development set and select the best performer.

**Training data.** We start from the **AmericasNLP 2023 parallel corpus** (Ebrahimi et al., 2023) and augment with additional public resources (Table 2). All data is Unicode NFC-normalized and whitespace-cleaned.

**Vocabulary issues: Bribri and Maya.** We discover that `bzd_Latn` (Bribri) and `yua_Latn` (Maya Yucateco) are absent from NLLB-200's vocabulary. Inspection of the pretrained tokenizer configuration confirms that neither identifier appears in the special tokens map or the sentence piece vocabulary. Querying either token returns id 3 (`<s>`), so the base model silently generates English. We add `yua_Latn` as a new special token (id 256204) for Maya, followed by `model.resize_token_embeddings()` to extend the embedding matrix; the new token's embedding is randomly initialized and trained from scratch. For Bribri, `bzd_Latn` was similarly added and the embedding matrix resized before fine-tuning. This is consistent with vocabulary gaps documented by Lupascu et al. (2025) for other under-represented languages in large multilingual models.

**Training configuration.** For Wixárika, Nahuatl, Bribri, and Guaraní: 3 epochs; batch 1 + 32 gradient accumulation steps (effective batch 32); lr $2\times10^{-5}$; 200 warmup steps; weight decay 0.01; Adafactor (Shazeer and Stern, 2018); gradient checkpointing; `fp16=False`; max sequence length 64 tokens. Hardware: NVIDIA A100 (Google Colab Pro).

Maya used different settings due to platform constraints (Kaggle T4): 5 epochs with early stopping (patience 2); batch 4 + 4 gradient accumulation steps (effective batch 16); same lr and optimizer; `fp16=True`. Training data comprised the Iikim Translator corpus augmented with the 50 AmericasNLP 2026 development captions repeated 20 times to expose the model to the image caption domain; task rules explicitly permitted using the development set for training.

| Lang. | Dev ChrF++ | | Sub. | Test | Human (test) | |
|---|---|---|---|---|---|---|
| | Base | FT | | ChrF++ | Rating | Rank |
| Guaraní | 19.49 | 17.57 | FT | 19.73 | 2.41 | 3rd/5 |
| Wixárika | 2.48 | 12.50 | FT | 13.68 | — | — |
| Maya | — | 9.03 | FT | 10.83 | — | — |
| Nahuatl | 3.29 | 5.23 | FT | 9.49 | — | — |
| Bribri | — | 2.65 | FT | 10.95 | 1.09 | 5th/5 |

Table 3: Dev and test ChrF++ for the submitted (FT) system, and test human evaluation. "—" = not evaluated or did not qualify for human evaluation (Bui et al., 2026). Human ratings on a 1–5 scale; 8 teams participated.

**Inference.** Fine-tuned models: beam search ($k$=4), `max_length=256`. For Bribri, `repetition_penalty=2.5` and `no_repeat_ngram_size=3` suppress degenerate repetition loops (Holtzman et al., 2020) caused by the combination of small training data, domain mismatch, and Bribri's complex morphology (Anderson et al., 2025).

## 4 Results

Table 3 reports development ChrF++ (Popović, 2017) for both base and fine-tuned NLLB models, and test-set human evaluation scores. The shared task ranked systems first by ChrF++ on the test set; the top 5 per language advanced to human annotation (1–5 scale).

For Wixárika, fine-tuning yields a large improvement (+10.02 ChrF++) over the base model, likely because the base NLLB has almost no Wixárika coverage. For Bribri, only the fine-tuned model is viable due to the vocabulary issue described above.

**Qualitative analysis.** Table 4 shows real system inputs and outputs from our test submissions, il-

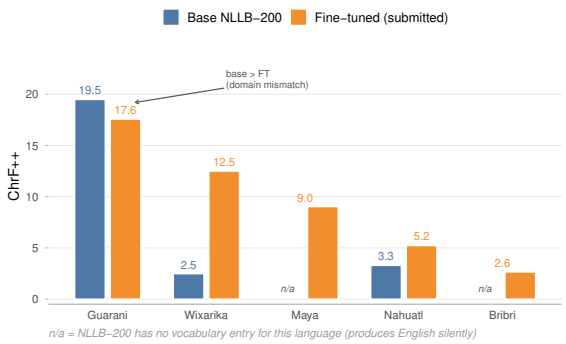


Figure 2: Dev ChrF++ for base vs. fine-tuned NLLB-200. Fine-tuning substantially improves Wixárika (+10.02) but slightly hurts Guaraní (domain mismatch). Bribri and Maya have no base score due to the vocabulary gap.

lustrating the pipeline in practice. Examples are unedited.

## 5 Analysis

**Cultural prompting and the metric gap.** Our 3rd-place Guaraní result was achieved without fine-tuning the VLM. For the other four languages, we used zero-shot cultural prompts; for Guaraní, we refined the Stage 1 prompt from a zero-shot version (used on dev) to a few-shot version with three culturally grounded caption examples (used on test), motivated by the observation that NMT fine-tuning slightly hurt dev ChrF++. The strong human evaluation result supports that culturally grounded prompting improved perceived caption quality. Notably, the Guaraní test ChrF++ (19.73) slightly exceeds the dev fine-tuned score (17.57), consistent with the few-shot V2 prompt providing richer Stage 1 captions than the zero-shot V1 used on dev. Karamolegkou et al. (2024) explain this pattern: cultural prompting improves human-judged quality even when ChrF++ declines, because culturally enriched descriptions include community-specific terms absent from generic references. Our dev ChrF++ scores likely underestimate cultural quality (Yun and Kim, 2024; Burda-Lassen et al., 2025); metrics capturing semantic and cultural fidelity (Krasner et al., 2025) would be more appropriate.

**Fine-tuning does not always help.** For Guaraní, the base NLLB model (19.49) outperforms the fine-tuned version (17.57) on dev. This suggests domain mismatch between our training data (general parallel text) and image captions (short, present-tense descriptions) can outweigh language-specific fine-tuning signal. Oduwole et al. (2026) and de Gibert et al. (2025) report similar findings. For Wixárika, where the base model has almost no coverage (2.48 ChrF++), fine-tuning improves substantially (+10.02).

| Stage 1 — Spanish (VLM) | Stage 2 — target language |
|---|---|
| *Guaraní, id grn_052 (tereré image)* | |
| Tereré, bebida tradicional guaraní preparada con yerba mate, agua fría y una bombilla llamada "guampa". Se aprecian las calabazas nativas usadas como cántaras. | Tereré, mba'yrumýiva guaraníme ojejapóva yvyra'ípe, y ro'y ha peteî bombilla hérava guampa. Ko ta'angápe ojehecha umi ka'aguy guasúva ojeporúva cántaras. |
| *Guaraní, id grn_053 (ñandutí image)* | |
| Encaje ñandutí, tejido manual con hilos multicolores en forma circular, característico de Itauguá, Paraguay. Símbolo de la herencia textil guaraní. | Ñemboja'o ñandutí, ñongatupyrýva umi mba'apohára ha'éva Itauguá, Paraguái. Ha'e símbolo herencia textil guarani. |
| *Bribri, id bzd_051 (cacao/Sibö image)* | |
| Este es el cacao, una planta sagrada que honra a Sibö, Dios creador bribri. Sus semillas son parte de las ofrendas rituales y simbolizan vida y fertilidad. | s e'r kákoa, tsékea … Sibòa. Se' wöyök tö íyikaalaa ie'pa rö i wèna- awa- ataa. [Bribri; see submission bzd_051] |
| *Nahuatl, id nlv_052 (failure: repetition collapse)* | |
| Una mujer nahua descansa sobre los escalones frente a una puerta antigua, vistiendo su huipil tejido con patrones que reflejan la riqueza artesanal de sus raíces. Su postura tranquila evoca las prácticas cotidianas del campo veracricense, donde aún se cultiva milpa y se preparan tamales. | *[Repetition collapse: 119 of 200 Nahuatl outputs consist entirely of repeated Unicode variation- selector characters. No repetition penalty was applied for Nahuatl.]* |

Table 4: Examples from our test submissions (unedited). Stage 1 correctly elicits culturally specific terms (*tereré*, *guampa*, *ñandutí*, *Sibö*) without VLM fine-tuning. The Nahuatl row shows a real failure: 119 of 200 Nahuatl test outputs collapsed to emoji repetition loops in Stage 2, as no repetition penalty was applied for that language.

**NLLB vocabulary gaps are a silent failure.** The absent `bzd_Latn` and `yua_Latn` tokens produce fluent English output with no warning. We recommend explicit verification before applying NLLB to any new language:

```
tokenizer.convert_tokens_to_ids(
```

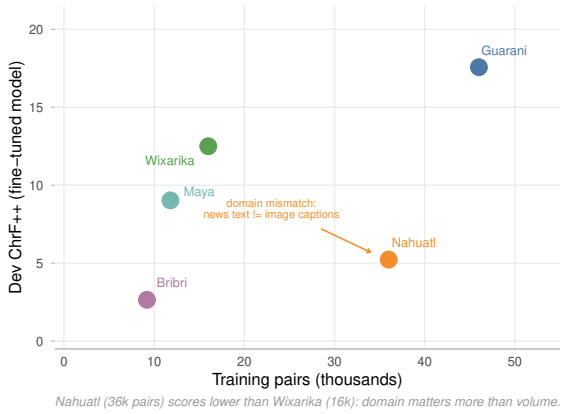


Figure 3: Training data size vs. dev ChrF++ (fine-tuned model). Nahuatl (36k pairs) underperforms Wixárika (16k pairs) due to domain mismatch between the Axolotl corpus (news text) and image captions.

```
"bzd_Latn") # -> 3 (<s>): NOT supported
```

Lupascu et al. (2025) document similar uneven coverage across large multilingual models.

**Repetition collapse.** Bribri's degenerate decoding (Holtzman et al., 2020) likely reflects the interaction of scarce training data (9.2k pairs), high domain mismatch, and morphological complexity (Anderson et al., 2025). A repetition penalty of 2.5 resolves the symptom; future work should address the cause via back-translation (Sennrich et al., 2016) or community-sourced captioning data (Gao et al., 2025).

## 6 Conclusion

We presented the USP system for the AmericasNLP 2026 image captioning shared task: a two-stage cascade combining culturally-prompted VLM captioning with per-language fine-tuned NLLB-200 NMT. Building on the AmericasNLP community's parallel corpora (Ebrahimi et al., 2023, 2024; de Gibert et al., 2025) and cultural captioning literature (Yun and Kim, 2024; Buettner et al., 2025; Gao et al., 2025), our system achieves 3rd place in Guaraní human evaluation. We contribute two findings: (1) NLLB-200 silently lacks Bribri and Maya Yucateco vocabulary; and (2) cultural prompting may improve human-judged quality in ways ChrF++ does not reflect (Karamolegkou et al., 2024). Taken together, these findings suggest that large multilingual models fail Indigenous captioning not only due to data scarcity, but through hidden infrastructure mismatches: tokenizer coverage gaps, domain misalignment, and cultural grounding deficiencies. We release our prompts and training scripts.[1]

## Ethics Statement

This work involves languages spoken by Indigenous communities whose cultural knowledge informed our prompting strategy. The cultural prompts were designed based on publicly available literature and existing NLP resources, without direct community consultation. We acknowledge that prompt-based cultural grounding risks misrepresenting or essentializing community practices, particularly if cultural elements are applied to images where they are not present. We do not claim that our system produces culturally authoritative captions, and we encourage community-led evaluation and correction of any outputs deployed in practice.

## Limitations

Cultural prompts were authored without native-speaker consultation, limiting their cultural validity (Oduwole et al., 2026; Gao et al., 2025). NMT models trained on general-domain text face domain mismatch on image captions; submitted systems may not reflect the optimal configuration for all languages. Bribri and Maya models use non-standard vocabulary configurations. Human evaluation sample sizes vary by language. For Guaraní, the development and test sets used different prompt versions (V1 zero-shot vs. V2 few-shot), which introduces a confound: the observed human evaluation gain cannot be attributed solely to cultural prompting, as prompt format also changed.

## Acknowledgments

The author thanks the AmericasNLP 2026 organizing committee. Compute via Google Colab Pro and Kaggle.

## References

Carter Anderson, Mien Nguyen, and Rolando Coto-Solano. 2025. Unsupervised, semi-supervised and LLM-based morphological segmentation for Bribri. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 63–76, Albuquerque, New Mexico. Association for Computational Linguistics.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei

[1] https://github.com/rmaacario/americasnlp2026-usp

Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 5 others. 2025. Qwen3-VL technical report. *Preprint*, arXiv:2511.21631.

Kyle Buettner, Jacob T. Emmerson, and Adriana Kovashka. 2025. A multimodal recaptioning framework to account for perceptual diversity across languages in vision-language modeling. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1989–2006, Mumbai, India. The Asian Federation of Natural Language Processing and The 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, and 15 others. 2026. Findings of the AmericasNLP 2026 shared task on image captioning in Indigenous languages. In *Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, San Diego, California. Association for Computational Linguistics.

Oliver Burda-Lassen, Aman Chadha, Sanjay Goswami, and Vikas Jain. 2025. How culturally aware are vision-language models? In *2025 IEEE 6th International Conference on Image Processing, Applications and Systems (IPAS)*, pages 1–6.

Yong Cao, Wenlong Li, Jiaang Li, Yifei Yuan, and Daniel Hershcovich. 2024. Exploring visual culture awareness in GPT-4V: A comprehensive probing. *arXiv preprint arXiv:2402.06015*.

Luis Chiruzzo, Santiago Alemán, Santiago Góngora, Aldo Alvarez, Lili Ferrari, and Yliana Rodríguez. 2022. Jojajovai: A parallel Guaraní-Spanish corpus for MT benchmarking. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, pages 2098–2107, Marseille, France. European Language Resources Association.

Ona de Gibert, Robert Pugh, Ali Marashian, Raul Vazquez, Abteen Ebrahimi, Pavel Denisov, Enora Rice, Edward Gow-Smith, Juan Prieto, Melissa Robles, Rubén Manrique, Oscar Moreno, Angel Lino, Rolando Coto-Solano, Aldo Alvarez, Marvin Agüero-Torales, John E. Ortega, Luis Chiruzzo, Arturo Oncevay, and 3 others. 2025. Findings of the AmericasNLP 2025 shared tasks on machine translation, creation of educational material, and translation metrics for indigenous languages of the americas. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 134–152, Albuquerque, New Mexico. Association for Computational Linguistics.

Dan DeGenaro and Tom Lupicki. 2024. Experiments in Mamba sequence modeling and NLLB-200 fine-tuning for low resource multilingual machine translation. In *Proceedings of the 4th Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 188–194, Mexico City, Mexico. Association for Computational Linguistics.

Abteen Ebrahimi, Ona de Gibert, Luis Chiruzzo, Javier Garcia Gilabert, Manuel Mager, Arturo Oncevay, Robert Pugh, Shruti Rijhwani, and Katharina von der Wense. 2023. Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages. In *Proceedings of the Third Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 1–17, Toronto, Canada. Association for Computational Linguistics.

Abteen Ebrahimi, Ona de Gibert, Raul Vazquez, Rolando Coto-Solano, Pavel Denisov, Robert Pugh, Manuel Mager, Arturo Oncevay, Luis Chiruzzo, Katharina von der Wense, and Shruti Rijhwani. 2024. Findings of the AmericasNLP 2024 shared task on machine translation into indigenous languages. In *Proceedings of the 4th Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 117–130, Mexico City, Mexico. Association for Computational Linguistics.

Isaac Feldman and Rolando Coto-Solano. 2020. A pipeline for spoken language documentation of Bribri. In *Proceedings of the WILDRE-5 Workshop at LREC*, Marseille, France. European Language Resources Association.

Yanbo Gao, Jianfei Fei, Nuo Chen, Ruonan Chen, Guofeng Yan, Yuanmeng Lan, and Bojun Shi. 2025. MELLA: Bridging linguistic capability and cultural groundedness for low-resource language MLLMs. *arXiv preprint arXiv:2508.05502*.

Javier García Gilabert, Aleix Sant, Carlos Escolano, Francesca De Luca Fornaciari, Audrey Mash, and Maite Melero. 2024. BSC submission to the AmericasNLP 2024 shared task. In *Proceedings of the 4th Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 143–149, Mexico City, Mexico. Association for Computational Linguistics.

Ximena Gutierrez-Vasques, Gerardo Sierra, and Isaac Pompa. 2016. Axolotl: A large corpus of Spanish-Nahuatl parallel text. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, Portorož, Slovenia. European Language Resources Association.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. 2021. MURAL: Multimodal, multitask

representations across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3449–3463, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Antonia Karamolegkou, Phillip Rust, Ruixiang Cui, Yong Cao, Anders Søgaard, and Daniel Hershcovich. 2024. Vision-language models under cultural and inclusive considerations. In *Proceedings of the 1st Human-Centered Large Language Modeling Workshop (HuCLLM)*, pages 53–66, Bangkok, Thailand. Association for Computational Linguistics.

Nathaniel Krasner, Justin Vasselli, Belu Ticona, Antonios Anastasopoulos, and Chi-Kiu Lo. 2025. Machine translation metrics for indigenous languages using fine-tuned semantic embeddings. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, Albuquerque, New Mexico. 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 13th European Conference on Computer Vision (ECCV)*, pages 740–755. Springer.

Marian Lupascu, Ana-Cristina Rogoz, Mihai Sorin Stupariu, and Radu Tudor Ionescu. 2025. Large multimodal models for low-resource languages: A survey. *arXiv preprint arXiv:2502.05568*. Accepted in Information Fusion.

Tom Lupicki, Lavanya Shankar, Kaavya Chaparala, and David Yarowsky. 2025. JHU's submission to the AmericasNLP 2025 shared task on the creation of educational materials for indigenous languages. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 105–111, Albuquerque, New Mexico. Association for Computational Linguistics.

Manuel Mager, Diónico Carrillo, and Ivan Meza. 2018. Probabilistic finite-state morphological segmenter for Wixarika (Huichol) language. *Journal of Intelligent & Fuzzy Systems*, 34(5):3081–3087.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Mardiyyah Oduwole, Prince Mireku, Fatimo Adebanjo, Oluwatosin Olajide, Mahi Aminu Aliyu, and Jekaterina Novikova. 2026. AfriCaption: Establishing a new paradigm for image captioning in African languages. In *Proceedings of the 7th Workshop on African Natural Language Processing (AfricaNLP 2026)*, pages 44–55, Rabat, Morocco. Association for Computational Linguistics.

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.

Julio Rangel and Norio Kobayashi. 2024. Advancing NMT for indigenous languages: A case study on Yucatec Mayan and Chol. In *Proceedings of the 4th Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 155–164, Mexico City, Mexico. Association for Computational Linguistics.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, and 57 others. 2024. CVQA: Culturally-diverse multilingual visual question answering benchmark. *Preprint*, arXiv:2406.05967.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4596–4604. PMLR.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Proceedings of NAACL-HLT 2007*, pages 484–491, Rochester, New York. Association for Computational Linguistics.

Justin Vasselli, Haruki Sakajo, Arturo Martínez Peguero, Frederikus Hudi, and Taro Watanabe. 2025. Leveraging dictionaries and grammar rules for the creation of educational materials for indigenous languages. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 112–118, Albuquerque, New Mexico. Association for Computational Linguistics.

Youngsik Yun and Jihie Kim. 2024. CIC: A framework for culturally-aware image captioning. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1634–1642.

## A Cultural Prompts

Below are the full Spanish-language cultural prompts used in Stage 1 for each language. Each prompt instructs the model to generate a 2–4 sentence Spanish caption while foregrounding culturally relevant elements.

**Guaraní (V1 — zero-shot, used on dev).** *"Eres un sistema de subtitulado de imágenes diseñado para describir imágenes con relevancia cultural para el pueblo Guaraní de Paraguay. Tu tarea: Generar subtítulos concisos, respetuosos y culturalmente precisos (2–4 oraciones máximo). Contexto cultural a reconocer: tereré (bebida fría de yerba mate), chipa, sopa paraguaya, mbejú, ñandutí, ao po'i, jacarandá (tajy), mburucuyá, opy (casa ceremonial), tatakua, kambuchi, guampa, misiones jesuíticas, Jasy Jatere, Pombero, Kurupi, Luison."*

**Guaraní (V2 — few-shot, used on test).** *"Eres un sistema de subtitulado de imágenes para el pueblo Guaraní de Paraguay. Genera un subtítulo en ESPAÑOL, conciso y culturalmente preciso (2–4 oraciones).*

*EJEMPLOS DE SUBTÍTULOS CORRECTOS:*

*Imagen de encaje artesanal colorido: "Ñandutí, encaje artesanal tradicional de Itauguá, Paraguay. Se teje a mano con hilos de colores formando patrones circulares como telas de araña."*

*Imagen de sopa con bolitas amarillas: "Vori vori, sopa tradicional paraguaya elaborada con bolitas de harina de maíz y queso. Es un plato típico de la gastronomía guaraní, especialmente consumido en invierno."*

*Imagen de vasija de barro con tela: "Kambuchi, vasija de barro tradicional guaraní, utilizada para transportar y conservar agua fresca."*

*CONTEXTO CULTURAL: tereré, chipa, sopa paraguaya, ñandutí, tatakua, kambuchi, guampa, Jasy Jatere, Pombero, Kurupi, Luison, misiones jesuíticas."*

**Maya Yucateco.** *"Eres un sistema de subtitulado de imágenes para el pueblo Maya de México (Yucatán). Contexto cultural: henequén, huipil, milpa, cenotes, Chichen Itzá / Uxmal / Tulum, Hanal Pixán, jarana, pib/mucbipollo, sopa de Lima, cochinita pibil, Xtabay, Alux, Chaac, Kukulkán."*

**Wixárika.** *"Eres un sistema de subtitulado de imágenes para el pueblo Wixárika (Huichol) de México. Contexto cultural: peyote (hikuri), Wirikuta, nierika (tabletas rituales), cuadros de estambre, arte con chaquira, ojo de Dios (tsikiri), mara'akame, kuchuri, tatewarí (Dios del Fuego), Tatei Haramara, ceremonias Mitote, nawá/tejuino. Prefiere "Wixárika" sobre "Huichol"."*

**Nahuatl.** *"Eres un sistema de subtitulado de imágenes para el pueblo Nahua de México (Orizaba, Veracruz). Contexto cultural: mole, tamales, tlayudas, chinampas, Teotihuacán / Tenochtitlán, Quetzalcóatl, Xochitl, Día de Muertos (cempasúchil), huipil, copal, temazcal, milpa, Voladores de Papantla."*

**Bribri.** *"Eres un sistema de subtitulado de imágenes para el pueblo Bribri de Costa Rica. Contexto cultural: Talamanca, cacao (planta sagrada), Sibö (dios creador), clanes matrilineales, sukia (chamán), casa cónica circular, chicha de maíz/pejibaye, pejibaye, cestería, usure (ceremonia de muerte), Kéköldi, Cordillera de Talamanca."*