6fanle Submission to the AmericasNLP 2026 Shared Task on Wixarika Image Captioning

Ji Wang and Hanqi Yang
Uppsala University

Abstract

This paper describes the 6fanle system for the Wixarika track of the AmericasNLP 2026 Shared Task on Cultural Image Captioning for Indigenous Languages. We report the data, preprocessing, model components, development experiments, and final results. Our system uses Spanish as a pivot language: Qwen3-VL generates Spanish caption candidates, CLIP retrieval supplies visually related examples, the official Sheffield-compatible machine translation model translates candidates into Wixarika, and a character n-gram language model reranks the translated outputs. The selected configuration achieved 19.1468 chrF++ in local 5-fold validation. In official Wixarika automatic evaluation, our v0 submission obtained 19.1569 chrF++ and 0.02145 CIDEr. In the final overall ranking, the team placed third.

1 Introduction

The AmericasNLP 2026 Cultural Image Captioning shared task asks systems to produce captions for images in Indigenous languages of the Americas (Bui et al., 2026). We participated in the Wixarika track, where only 70 official image-caption examples were available before test inference: 20 pilot examples and 50 development examples. The test set contains images and metadata, and the required output is a JSONL file with a `predicted_caption` field.

Because the available paired image-caption data are very small, we did not train an end-to-end image-to-Wixarika model. Instead, we used a modular pipeline that separates visual description from low-resource text generation. A vision-language model produces Spanish descriptions, and the official machine translation component converts them to Wixarika. Retrieval and Wixarika-side reranking are used to adapt the outputs to the limited in-domain data. We release the code and configuration files needed to reproduce the pipeline at `https://github.com/ousyu66-pixel/americasnlp2026-wixarika-captioning`.

2 Related Work

Our system follows the general pattern of previous AmericasNLP system descriptions, which document the resources used, preprocessing decisions, model variants, validation protocols, and final submissions. The official baseline and the Sheffield-compatible translation component motivate our use of Spanish-Wixarika MT as the low-resource generation stage (Gow-Smith and Haddow, 2023).

We also considered ideas from low-resource multimodal data construction and weakly supervised visual data generation (Xie et al., 2024; Xiao et al., 2025). However, the final system does not synthesize new images and does not train a new multimodal model. Instead, we use a lightweight version of this idea: additional Spanish descriptions are rewritten into short visual captions and used only as retrieval support.

3 Data

Table 1 summarizes the data sources used by the submitted system and by a rejected ablation involving lexical filtering. We distinguish final-system resources from resources used only in rejected experiments to avoid overstating what entered the submitted run.

3.1 Preprocessing and Post-processing

The preprocessing stage normalizes input JSONL records, resolves image paths, and prepares three kinds of retrieval examples: pilot examples with Spanish captions, development examples back-translated from Wixarika into Spanish, and augmented examples rewritten from longer Spanish descriptions into short caption-style Spanish. Held-out validation examples are removed from the retrieval bank and from the Wixarika language-model

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 243–247

Source	Count / size	Use in this work
Official Wixarika pilot images	20	Used for validation, retrieval support, and Wixarika language-model training. Pilot Spanish captions are used when available.
Official Wixarika development images	50	Used for validation and Wixarika language-model training. Spanish retrieval text is produced by Wixarika-to-Spanish backtranslation.
Official Wixarika test images	201	Used only for final inference. No test labels were used.
Wixarika Research Center augmented images	65	Spanish descriptions were rewritten into short caption style and used as auxiliary retrieval support (Wixarika Research Center, 2026).
AmericasNLP 2023 Spanish-Wixarika parallel text	9,960 pairs	Used as Wixarika-side text for the character language model.
Native Languages Huichol word list	small word list	Used only in the rejected filtering experiment, not in the final submitted configuration (Native Languages of the Americas, 2026).

Table 1: Data sources and their use.

training text during cross-validation.

The final post-processing stage is audit-based candidate cleanup. The pipeline stores all Spanish and Wixarika candidates in `predictions.audit.jsonl`. If the selected Wixarika output shows clear degeneration, such as repeated plus-marked fragments or repeated tokens, the cleanup script chooses a better already generated candidate for the same image. It does not manually write captions and does not use test references. The cleanup step was automatic and candidate-constrained. It never created a new caption from scratch: for each affected test image, it selected one of the Wixarika candidates that had already been generated by the same pipeline for that image. No test references or manually written test captions were used. This makes the cleanup a deterministic post-selection step rather than manual test-set annotation. On the final test run, this changed 37 captions and reduced the heuristic bad-output count from 20 to 7 out of 201 predictions.

4 Methods

The final pipeline is: image → CLIP retrieval → Qwen3-VL Spanish caption candidates → official Spanish-to-Wixarika MT → Wixarika character-LM reranking → audit cleanup → JSONL submission.

4.1 Retrieval

We embed images with `openai/clip-vit-large-patch14-336`, a CLIP model (Radford et al., 2021), and retrieve the top $k = 4$ visually similar examples. The retrieval code uses `CLIPModel.get_image_features` and normalizes image vectors before nearest-neighbor search. Retrieved Spanish captions are inserted into the Qwen3-VL prompt as style and content support.

4.2 Spanish Caption Generation

We use Qwen/Qwen3-VL-8B-Instruct (Qwen Team, 2025) to generate Spanish caption candidates. Spanish is used as a pivot because the VLM is more reliable in Spanish than in Wixarika. The selected configuration requests eight candidates per image with `max_new_tokens=72`, temperature 0.85, `top_p=0.95`, and sampling enabled. The prompt asks for concise descriptions based only on visible content, with attention to people, clothing, textiles, art objects, ritual objects, music, labor, architecture, and actions.

4.3 Machine Translation

The official Sheffield-compatible translation model is used in both directions. Wixarika-to-Spanish translation provides retrieval text for development examples, while Spanish-to-Wixarika translation converts generated Spanish candidates into target-language candidates. The MT beam size is 5. To avoid memory and device-mixing issues on a single A100 40GB instance, Qwen3-VL and CLIP run on GPU, while the official MT subprocess can run on CPU when needed. We observed that the official model produced more useful outputs in the Spanish-to-Wixarika direction than in the Wixarika-to-Spanish direction. Therefore, Wixarika-to-Spanish translation was used only as an auxiliary step for building retrieval text, while final caption generation always used the Spanish-to-Wixarika direction.

System	Protocol	chrF++	Decision
Organizer baseline	Baseline report	∼17.7	Reference
Selected pipeline	Local 5-fold CV	19.1468	Final
Output filter + word list	Local 5-fold CV	18.9689	Rejected
Restrictive prompt	Local 5-fold CV	16.9403	Rejected
Decoding sweep	Local 2-fold screen	18.4174	Rejected

Table 2: Development validation and ablation results.

4.4 Reranking

Each image receives multiple Wixarika candidates. A character 5-gram language model scores target-language fluency, and the final score combines fluency, a length preference, candidate order, and a penalty for very short outputs. For final test inference, the language model is trained on all available non-test Wixarika text.

5 Experiments

We used deterministic 5-fold cross-validation over the 70 official pilot and development examples. Each fold contained 14 validation examples. Held-out examples were excluded from both the retrieval bank and the character language model. We used chrF++ for local validation, matching the automatic evaluation emphasis of the shared task.

5.1 Development Results

Table 2 reports the selected configuration and the main rejected variants. We mention the organizer baseline only as a reference point; the main comparison is among our own validated variants.

The decoding sweep was an early 2-fold screening run used to discard an unpromising setting under limited compute; it is reported for transparency but is not directly comparable to the 5-fold validation runs.

The selected pipeline achieved fold scores of 18.1857, 19.7800, 19.4415, 19.5702, and 18.7564. The filtering patch improved two folds but lowered the overall mean, as shown in Table 3. The selected pipeline was also more stable across folds, with a sample standard deviation of 0.6601 compared with 0.9309 for the filtering patch. The restrictive prompt reduced chrF++ substantially, suggesting that over-constraining the Spanish caption removed useful in-domain wording.

Fold	Selected	Filter	Delta
1	18.1857	19.1539	+0.9682
2	19.7800	19.0545	-0.7255
3	19.4415	20.1932	+0.7517
4	19.5702	18.8574	-0.7128
5	18.7564	17.5856	-1.1708
Mean	19.1468	18.9689	-0.1779

Table 3: Fold-level comparison between the selected pipeline and filtering patch.

5.2 Test Results

The final inference run produced 201 predictions for 201 Wixarika test images. In the official automatic Wixarika evaluation, our submitted v0 system obtained 19.1569 chrF++. In the final human evaluation, the team ranked third overall with a mean rating of 2.48 over 201 test images. This suggests that, despite competitive automatic chrF++, fluency, naturalness, and image-specific adequacy remained important limitations.

5.3 Error Analysis

We analyzed the final 201-line submission file and the saved audit logs. The submission itself was complete: all 201 test images received a non-empty prediction, all IDs were unique, and all required JSONL fields were present. The remaining errors were therefore generation-quality errors.

Table 4 summarizes the main automatic error categories observed in the final output. The most frequent issue was residual Spanish or untranslated lexical material. We found 35 predictions containing likely Spanish words or named expressions, such as words for streets, buildings, materials, murals, and other concrete objects. This suggests that the Spanish-to-Wixarika MT component often preserved source-language nouns when the relevant Wixarika lexical item was unavailable or uncertain.

A second issue was repetition and translation degeneration. Although the final cleanup stage replaced 37 suspicious predictions by selecting alternative candidates from the audit file, 17 final predictions still showed repeated words, repeated character sequences, or abnormal forms. These cases likely reduced human-perceived fluency. We also found three groups of exactly repeated captions, affecting five additional rows, which indicates that some outputs were too generic and insufficiently image-specific.

For example, `hch_111` preserved Spanish words such as *camioneta*, *calle*, *aceras*, and *baldosas*.

Error type	Count
Complete predictions	201 / 201
Empty predictions	0
Likely Spanish residue	35
Repetition / degeneration suspects	17
Over-long predictions	10
Exact duplicate caption groups	3
Extra rows affected by exact duplicates	5
Cleanup replacements before submission	37

Table 4: Automatic error analysis of the final Wixarika test submission.

Similarly, hch_234 retained words such as *mural*, *girasoles*, and *bordados*. Repetition also remained in some outputs, such as hch_246, where *katixex-exeiyat+* contains an abnormal repeated character sequence.

The audit logs also showed that some Spanish caption candidates were visually underspecified. When the Spanish description only mentioned generic people, rural scenes, or objects, the translated Wixarika caption could be fluent but not sufficiently tied to the image. This helps explain why the system was competitive under character-level chrF++ but weaker in image-specific adequacy and fluency. Overall, the main bottleneck was the interaction between generic visual descriptions, untranslated Spanish lexical items, and MT degeneration in the final Wixarika output.

6 Conclusions

We presented a translation-centered system for Wixarika image captioning. The selected system combines CLIP retrieval, Qwen3-VL Spanish caption generation, official Spanish-Wixarika translation, and Wixarika character-LM reranking. It achieved competitive automatic results and produced a complete 201-image submission, but the final ranking shows that Wixarika fluency and human-perceived naturalness remain the main areas for improvement.

Limitations

The system relies on Spanish as a pivot language, so visual details lost during Spanish captioning cannot be recovered by translation. The Wixarika language model is a surface-level character model and cannot guarantee semantic correctness. The cleanup step removes obvious degeneration but may still select semantically imperfect candidates.

Ethics Statement

This system is an assistive research prototype, not an authoritative Wixarika captioning tool. Automatic captions may be inaccurate or culturally inappropriate. We did not manually create test captions, and any real deployment should involve Wixarika speakers and community stakeholders.

References

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 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.

Edward Gow-Smith and Barry Haddow. 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*.

Native Languages of the Americas. 2026. Vocabulary in Native American Languages: Huichol Words. https://www.native-languages.org/huichol_words.htm. Accessed May 2026.

Qwen Team. 2025. Qwen3-VL-8B-Instruct. https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct. Accessed May 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*.

Wixarika Research Center. 2026. Wixarika Research Center Online Archive. https://www.wixarika.org/. Accessed May 2026.

Bushi Xiao, Qian Shen, and Daisy Zhe Wang. 2025. From text to multi-modal: Advancing low-resource-language translation through synthetic data generation and cross-modal alignments. In *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages*, pages 24–35.

Yangchen Xie, Xinyuan Chen, Hongjian Zhan, Palaiahnakote Shivakumara, Bing Yin, Cong Liu, and Yue Lu. 2024. Weakly supervised scene text generation for low-resource languages. *Expert Systems with Applications*, 237:121622.