From Machine Translation to Image Captioning: Training Vision-Language Models for Indigenous Languages of the Americas

Luis Lara[1], **Param Raval**[1]
[1]Mila – Quebec AI Institute
luis.lara@mila.quebec

Abstract

We describe our system for the AmericasNLP 2026 Shared Task on Cultural Image Captioning for Indigenous Languages of the Americas. Our post-training pipeline starts from Aya Vision 32B: the vision-language model is first fine-tuned on machine translation data from prior AmericasNLP shared tasks and then further fine-tuned on the cultural Image Captioning data. This approach uses translation as an intermediate training task, while the final system produces captions directly in the requested Indigenous language rather than translating a Spanish caption afterward. Our experiments show that machine translation fine-tuning is an important initialization step. The resulting fine-tuned vision-language model also shows translation capabilities for the languages considered in this work. In addition, our zero-shot GPT-5.5 submission ranks first in the Maya language track under the official human-evaluation stage. [1]

1 Introduction

The AmericasNLP 2026 shared task studies cultural Image Captioning for Indigenous languages of the Americas (Bui et al., 2026). Given an image and a requested language, a system must produce a caption that is both visually grounded and appropriate for the linguistic and cultural setting. This is challenging because many of these languages have little or no representation in the data used to pretrain current large language models (LLMs) and vision-language models (VLMs). It is also challenging because the shared task provides only a small set of image-caption pairs, so caption-only fine-tuning has relatively few examples to learn from.

At the same time, the AmericasNLP community has built machine translation resources for several Indigenous languages of the Americas across previous shared tasks (Mager et al., 2021; Ebrahimi et al., 2022, 2023, 2024; De Gibert et al., 2025). These resources do not teach a model to ground language in images, but they do provide substantially more examples of how to generate text in the relevant languages. This suggests a transfer strategy: before asking a VLM to caption images, first fine-tune it on bilingual translation data so it learns to produce text in the relevant Indigenous languages.

We implement this strategy with a multi-stage supervised fine-tuning pipeline. We use Aya Vision 32B, a VLM, as the backbone (Dash et al., 2025). We first perform supervised fine-tuning for machine translation, SFT (MT), on Spanish–Indigenous bilingual text. We then further fine-tune the same model on the cultural Image Captioning (IC) data, SFT (IC), so that the final system generates captions directly in the requested Indigenous language instead of first producing a Spanish caption and translating it afterward. All stages use Low-Rank Adaptation (LoRA) (Hu et al., 2022). We also test reinforcement learning with verifiable rewards for machine translation, RLVR (MT), after SFT (MT), but find that it provides smaller and less consistent gains than the supervised transfer step.

Our experiments support the main motivation for the pipeline: caption fine-tuning works better when it starts from the translation-tuned model. In addition, the same fine-tuned VLM can produce useful translation candidates for these languages under reference-based oracle candidate selection. We also evaluate a zero-shot direct Image Captioning submission with GPT-5.5 (OpenAI, 2026) as a frontier-model comparison, which ranks first in the Maya track under the official human-evaluation stage. Figure 1 summarizes our two submission routes: the fine-tuned Aya Vision pipeline and the separate GPT-5.5 zero-shot direct Image Captioning submission.

[1]Project code is available at https://github.com/ludolara/americasnlp2026

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 224–235

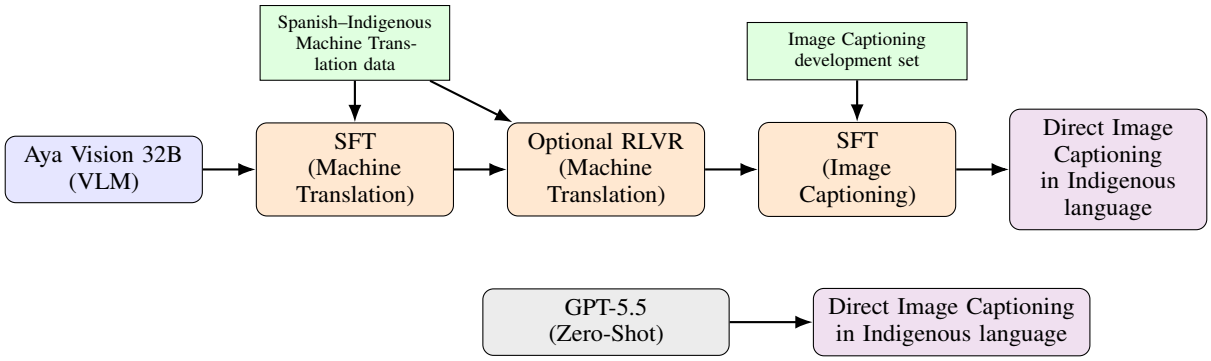

Figure 1: Overview of our Image Captioning systems. The main pipeline adapts Aya Vision 32B through supervised fine-tuning on Spanish–Indigenous machine translation data, optional RLVR on the same task, and supervised fine-tuning on the Image Captioning development set. The lower branch shows our separate GPT-5.5 zero-shot submission. Both systems generate captions directly in the target Indigenous language.

Our contributions are as follows:

1. We present a multi-stage post-training pipeline for cultural Image Captioning in Indigenous languages of the Americas, using machine translation fine-tuning as an intermediate training stage before image-captioning fine-tuning.

2. We show that the same translation-tuned VLM can produce useful translation candidates for the overlapping Indigenous languages under reference-based oracle candidate selection.

3. We include a separate zero-shot GPT-5.5 direct Image Captioning submission, which reaches the strongest human-evaluation result for the Maya track under the official shared-task evaluation.

2 Related Work

AmericasNLP shared tasks. Across the 2021, 2022, 2023, 2024, and 2025 editions, the AmericasNLP shared tasks (STs) evolve from an open machine translation benchmark into a broader suite of tasks for Indigenous languages of the Americas (Mager et al., 2021; Ebrahimi et al., 2022, 2023, 2024; De Gibert et al., 2025). The 2021 ST established the benchmark with two tracks, ten languages, official baselines, and large improvements over baseline for many languages (Mager et al., 2021). The 2022 competition expanded the scope to speech by adding automatic speech recognition and speech-to-text translation alongside text-based machine translation, covering Bribri, Guaraní, Kotiria, Wa'ikhana, and Quechua (Ebrahimi et al., 2022). The 2023 edition returned the focus to machine translation but expanded the benchmark to eleven language pairs, added a new Chatino–Spanish evaluation set from the legal domain, and complemented automatic ranking with human evaluation (Ebrahimi et al., 2023). The 2024 findings emphasize a harder competitive setting: organizers released strong Sheffield and Helsinki baselines from 2023 and a repository of prior shared task training data, yet improvements over the best baseline were observed for only a subset of languages (Ebrahimi et al., 2024). The 2025 STs broadened the scope beyond machine translation to include educational-material creation and translation metrics, showing a shift from pure text-to-text translation toward a wider range of community-relevant language technologies (De Gibert et al., 2025).

This trajectory is directly relevant to our setting. First, the findings papers repeatedly show that progress depends on assembling, reusing, and extending scarce bilingual and speech resources (Mager et al., 2021; Ebrahimi et al., 2022, 2023, 2024). Second, they show that stronger baselines and human evaluation matter, because automatic gains alone do not fully characterize quality for low-resource Indigenous-language generation (Ebrahimi et al., 2023, 2024; De Gibert et al., 2025). The AmericasNLP 2026 shared task on cultural image captioning extends this line of work from text-to-text translation to visually grounded generation.

LLMs for Indigenous-language translation. Results from the AmericasNLP 2025 ST system papers do not provide strong evidence that off-the-shelf LLM-based methods are strong direct generators in the Spanish-to-Indigenous direction. Yahan and Islam (2025) report that NLLB-200

Table 1: Image Captioning test chrF++ results using the automatic evaluation metric. Maya was included only in v3, which corresponds to the GPT-5.5 zero-shot submission and was not produced by our fine-tuned image-captioning method. Versions v0 and v2 use the same training recipe, but correspond to different checkpoints.

Version	Submission	Overall	Bribri	Guaraní	Nahuatl	Wixárika	Maya
v0	SFT (MT) + SFT (IC)	17.398	**11.728**	19.634	19.421	18.811	–
v2	SFT (MT) + SFT (IC)	**17.599**	11.309	19.417	**20.655**	**19.013**	–
v1	SFT (MT) + RLVR (MT) + SFT (IC)	17.341	10.886	**19.772**	19.853	18.853	–
v3	GPT-5.5 (Zero-Shot)	12.812	4.555	12.804	19.725	10.984	**15.993**

outperforms LLaMA 3.1 and XGLM in Track 1, while Hus et al. (2025) use an LLM as a post-correction component and report stronger gains in the Indigenous-to-Spanish direction. Together with the official findings, these results motivate our setting: rather than treating a large generative model as a strong zero-shot translator, we study whether targeted supervised fine-tuning can make it a better generator for low-resource Indigenous languages (De Gibert et al., 2025) without intermediate Spanish translations.

Intermediate-task fine-tuning from translation to Image Captioning. Transfer learning in NLP commonly uses knowledge from one task, domain, or dataset to improve adaptation to another setting (Ruder et al., 2019). A more specific version of this idea is intermediate-task fine-tuning, where a pretrained model is first fine-tuned on an additional supervised task before being adapted to the final target task (Phang et al., 2018). We use this framework as methodological motivation rather than as direct evidence for our specific task combination. In our setting, machine translation serves as the intermediate task: it exposes the model to substantially more Spanish–Indigenous text and target-language generation examples before the model is fine-tuned on the much smaller Image Captioning data. This stage does not teach visual grounding. Instead, it is intended to improve the model's ability to generate text in the target Indigenous languages, while the subsequent Image Captioning fine-tuning stage teaches the model to connect visual inputs with concise target-language descriptions.

3 Shared Task and Data

3.1 Shared Task and Baseline

The shared task requires generating one caption for each input image in the requested Indigenous language. We treat this as direct image-to-Indigenous-text generation: the model is trained to produce the final caption itself, rather than first generating a Spanish caption and translating it afterward. Compared with generic image captioning, this task combines visual grounding with low-resource language generation. Further details on the shared task evaluation protocol are provided in the official findings paper (Bui et al., 2026).

The baseline for the ST follows the *generate-then-translate* pipeline where a caption is generated in Spanish using the Qwen3-VL-8B-Instruct VLM and then translated into the target language using a NLLB-200 model trained for that language. The latter stage uses the approach proposed by the winner of the AmericasNLP 2023 ST on MT (Gow-Smith and Sánchez Villegas, 2023).

3.2 Image Captioning Data

The development set consists of 250 labeled images, with 50 examples for each of Bribri, Guaraní, Yucatec Maya, Nahuatl, and Wixárika. We refer to Yucatec Maya as Maya in the result tables for compactness. We also use Nahuatl as a compact label for the shared-task Nahuatl track; where relevant, Appendix C preserves the Orizaba Nahuatl label used by the example IDs. Teams were allowed to train with the development set. The test set used to rank the submissions of the ST contains 267 Bribri images, 101 Guaraní images, 212 Yucatec Maya images, 200 Nahuatl images, and 201 Wixárika images.

3.3 Machine Translation Data

We restrict the translation data to the four overlapping languages: Bribri (`bzd`), Guaraní (`grn`), Nahuatl (`nah`), and Wixárika (`hch`). For Nahuatl, we use the available data from previous AmericasNLP shared tasks. We note that this data is not specifically labeled as *Orizaba Nahuatl*. Each training example contains Spanish text, Indigenous-language text, a language name, and a language code. The dataset is prepared by labeling these pairs bidirectionally so the model sees both Spanish-to-Indigenous and Indigenous-to-Spanish translation prompts.

The primary resources we use include Axolotl

Table 2: Per-language chrF++ on the internal development subset for Image Captioning systems.

System	Overall	Bribri	Guaraní	Nahuatl	Wixárika
None	6.053	2.077	7.688	6.836	6.687
SFT (IC)	14.589	11.570	15.714	15.444	14.848
SFT (MT) + SFT (IC)	19.313	11.457	22.143	**17.149**	**22.602**
SFT (MT) + RLVR (MT) + SFT (IC)	**19.805**	**11.508**	**23.767**	16.982	22.496

for Spanish–Nahuatl (Gutierrez-Vasques et al., 2016), a Wixárika resource derived from work on morphological segmentation (Mager et al., 2018), a Guaraní–Spanish parallel corpus (Chiruzzo et al., 2020), and a Bribri back-translation resource (Feldman and Coto-Solano, 2020). We also use additional data from the AmericasNLP 2025 ST1 language pairs, following the shared-task setup described by De Gibert et al. (2025); these additions are linked to Helsinki-NLP's earlier shared-task work (De Gibert et al., 2023).

4 System Overview

Our system is based on Aya Vision 32B (Dash et al., 2025), an open-weight multilingual multimodal model designed for image understanding, image captioning, visual question answering, text generation, and translation across 23 languages. Its open weights make it suitable for efficient fine-tuning on vision-language alignment tasks.

We use the same base model for translation and image captioning. The complete pipeline has three possible stages: SFT (MT), supervised fine-tuning for machine translation; optional RLVR (MT), reinforcement learning with verifiable rewards for machine translation; and SFT (IC), supervised fine-tuning for image captioning. The strongest and simplest configuration in our development experiments is SFT (MT) followed by SFT (IC).

4.1 Supervised Fine-Tuning (Machine Translation)

The first stage uses supervised fine-tuning on bilingual text with Aya Vision 32B. Each example is formatted as an instruction-style chat prompt, with the target sentence used as the assistant response. Training is bidirectional, so the model learns both Spanish-to-Indigenous and Indigenous-to-Spanish generation. The machine translation prompt template is listed in Appendix B.

4.2 Supervised Fine-Tuning (Image Captioning)

The second stage fine-tunes the model using the image captioning development set. Each sample provides an image and a target language, and the model learns to produce a caption directly in that language. Our approach skips intermediate translation and generates the final caption directly in the requested Indigenous language.

The image captioning prompt is intentionally simple. We avoid long explanations and multi-step instructions to minimize prompt engineering. The Spanish prompt asks for a single culturally appropriate caption in the target language and instructs the model not to include explanations. The image captioning prompt template is listed in Appendix B.

4.3 Reinforcement Learning with Verifiable Rewards (Machine Translation)

We also evaluate RLVR (MT) after SFT (MT). In this stage, we use the Group Relative Policy Optimization (GRPO) algorithm (Shao et al., 2024) for optimization. For the automated candidate-selection procedure used in these experiments, we use sentence-level chrF++ to select high-scoring translations from multiple sampled candidates. In our development experiments, this stage gives a small overall improvement but the gains are not consistent across languages. We therefore treat RLVR (MT) as an exploratory ablation rather than as the central contribution of the system.

4.4 Zero-Shot Direct Image Captioning with GPT-5.5

In addition to the fine-tuned Aya Vision systems, we submitted a zero-shot direct Image Captioning run with GPT-5.5 (OpenAI, 2026). For each example, we provided the image and requested target language, and asked the model to generate the final caption directly in that language. This run does not use our machine-translation fine-tuning stage, RLVR, or image captioning fine-tuning, and it does not follow a generate-then-translate pipeline. We include it as an exploratory frontier-VLM com-

Table 3: Official human-evaluation results for our submissions. Mean Rating is the official average human-evaluation score reported by the organizers, based on 1–5 human ratings where higher is better. When multiple annotators rated the same example, their scores were averaged per example before computing the final mean. N Ratings and N Images reproduce the counts reported in the official results. Points follow the shared-task overall ranking rule: 5 points for first place, 4 for second, 3 for third, 2 for fourth, and 1 for fifth.

Language	Version	Submission	Rank	N Ratings	N Images	Mean Rating
Bribri	v0	SFT (MT) + SFT (IC)	4	320	267	1.994
Guaraní	v1	SFT (MT) + RLVR (MT) + SFT (IC)	5	228	101	1.764
Maya	v3	GPT-5.5 (Zero-Shot)	**1**	212	212	**3.203**
Nahuatl	v2	SFT (MT) + SFT (IC)	3	200	200	1.560
Wixárika	v2	SFT (MT) + SFT (IC)	5	201	201	2.210
Overall human-evaluation ranking			4		Total points: 12	

parison, especially for Yucatec Maya, which is included in the shared task but not covered by our translation-initialized Aya Vision submissions. The exact prompt and inference details are listed in Appendix B.2.

5 Experimental Setup

5.1 Image Captioning Evaluation

We evaluate Image Captioning with chrF++ (Popović, 2017). Following the shared-task setting, we train SFT (IC) with the development set. To select the best candidate system before submission, we hold out a 20-example subset from the development data for internal model selection. This subset contains 5 examples per evaluated language: Bribri, Guaraní, Nahuatl, and Wixárika, corresponding to 10% of the 50 development examples available for each language. Because this model-selection subset is small, scores on it should be interpreted cautiously. We list the selected example IDs in Appendix C.

5.2 Machine Translation Evaluation

For translation, we evaluate the SFT (MT) and SFT (MT) + RLVR (MT) models on the AmericasNLP machine translation test split originally introduced in the 2021 shared task. Based on the later findings papers and shared task data releases, we believe this is the same test set that continued to be used in subsequent AmericasNLP machine translation evaluations for the overlapping languages (Mager et al., 2021; Ebrahimi et al., 2024; De Gibert et al., 2025). We report chrF++ for Spanish-to-Indigenous and Indigenous-to-Spanish directions.

All translation results use best-of-100 candidate selection under sentence-level chrF++ against the reference. We report chrF++ on the shared test set, but because candidate selection uses the reference, this setting should be interpreted as a reference-based oracle candidate-quality analysis rather than as single-output interactive translation.

5.3 Baselines and Comparisons

For image captioning, our main internal comparison is a step-by-step ablation of the training pipeline. We begin with the simplest setting, where Aya Vision is trained only on the image-captioning data. We then add translation fine-tuning before image-captioning training, so the model enters the captioning stage with stronger target-language generation ability. Finally, we test whether inserting the reward-based translation stage between translation fine-tuning and captioning provides any additional benefit.

For translation, we compare against reported AmericasNLP 2025 systems where the language and test-set overlap is available. Because our results use reference-based best-of-100 candidate selection, this comparison should be read as contextual rather than as a strict interactive MT leaderboard comparison. In particular, we are interested in whether SFT (MT) can move a large generative model from weak off-the-shelf behavior toward useful candidate generation for these languages.

6 Results

6.1 Image Captioning Results

Table 1 reports the main shared task results on the test set. Under the official automatic metric, v2 is the strongest fine-tuned Aya Vision submission overall, with 17.599 chrF++ averaged over Bribri, Guaraní, Nahuatl, and Wixárika. The RLVR-based v1 submission is slightly lower overall, which is consistent with our interpretation that RLVR (MT) is exploratory and does not provide consistent gains in this setting.

Table 2 reports chrF++ on the 20-example development subset. On this subset, the two systems

are close overall: SFT (MT) + SFT (IC) reaches 19.313 chrF++, while SFT (MT) + RLVR (MT) + SFT (IC) reaches 19.805. The per-language pattern is mixed: RLVR (MT) improves Bribri and Guaraní slightly, while Nahuatl and Wixárika are slightly lower.

Table 3 reports our official human-evaluation results. Our fine-tuned Aya Vision submissions reach the human-evaluation stage for all four languages covered by our translation-initialized pipeline: Mila ranks fourth for Bribri, fifth for Guaraní, third for Nahuatl, and fifth for Wixárika. Our separate zero-shot GPT-5.5 submission ranks first for Maya, which is the only language for which v3 was used in the official human-evaluation stage. Under the shared-task point system, Mila ranks fourth overall with 12 points. This result suggests that a frontier VLM can be a strong direct caption generator for Maya in this shared-task setting, even without our task-specific fine-tuning pipeline. However, this result should be interpreted alongside the automatic chrF++ results, where the GPT-5.5 Maya submission does not rank first. Since GPT-5.5 is a closed model, we cannot determine whether its pretraining or instruction tuning included Maya data.

6.2 Machine Translation Results

Table 4 compares our SFT (MT) model with selected AmericasNLP 2025 systems for overlapping languages. Because our SFT rows use reference-based best-of-100 candidate selection, they should be interpreted as reference-based oracle-selected candidate-quality scores rather than as single-output interactive translation results. In the Spanish-to-Indigenous direction, SFT (MT) remains below the reported multilingual baseline for most overlapping languages. However, it reaches the same broad range as several submitted systems, especially for Bribri, Wixárika, and Nahuatl. In the Indigenous-to-Spanish direction, the reference-based oracle-selected SFT candidates obtain strong chrF++ scores across all four languages. These results suggest that targeted supervised fine-tuning can make a large generative model produce useful candidate translations for these languages, while also showing that fairer comparisons require direct single-candidate generation rather than oracle selection. Appendix A reports qualitative translation examples for both translation directions (Tables 7 and 8) and sentence-level chrF++ score distributions (Figure 2).

Table 4: Comparison with selected AmericasNLP 2025 translation systems using chrF++ for overlapping languages. Our SFT rows use best-of-100 reference-based candidate selection and should be interpreted as reference-based oracle-selected candidate-quality scores, not as single-output interactive translation results. Bold values indicate the best performance, while underlined values indicate the second-best performance.

System	BZD	GRN	HCH	NAH
SPA→XXX				
Baseline	**25.52**	**35.68**	**28.26**	22.42
GMU	22.51	<u>29.95</u>	26.14	20.33
Syntax Squad	22.77	16.21	26.77	12.64
SFT (MT)	24.22	29.39	27.69	<u>26.39</u>
SFT (MT) + RLVR (MT)	<u>24.31</u>	29.46	<u>27.83</u>	**26.52**
XXX→SPA				
Baseline	30.14	35.91	26.33	26.36
GMU	27.86	33.84	24.37	25.58
Syntax Squad	26.22	24.70	22.02	13.88
SFT (MT)	**33.45**	<u>38.17</u>	<u>30.14</u>	**31.95**
SFT (MT) + RLVR (MT)	<u>33.42</u>	**38.29**	**30.18**	<u>31.73</u>

Table 5: Image Captioning pipeline ablation on the 20-example internal development subset. Rows incrementally add SFT (IC), SFT (MT), and RLVR (MT). Δ is measured against the previous row.

System	chrF++	Δ
No SFT	6.053	–
SFT (IC)	14.589	**+8.536**
SFT (MT) + SFT (IC)	19.313	+4.724
SFT (MT) + RLVR (MT) + SFT (IC)	**19.805**	+0.492

6.3 Pipeline Ablation

Table 5 makes the main pipeline ablation explicit. Starting from no supervised fine-tuning, SFT (IC) gives the largest improvement in chrF++ on the internal development subset, reaching 14.589. Adding SFT (MT) before SFT (IC) further improves the score to 19.313, and adding RLVR (MT) gives a smaller additional gain to 19.805.

This ablation clarifies the role of each component. The largest single contribution comes from SFT (IC), which teaches the model the image captioning task. SFT (MT) remains useful as a language-generation initialization for the Indigenous languages, while RLVR (MT) has the smallest incremental effect in this setup.

7 Conclusion

We presented a multi-stage supervised fine-tuning approach to cultural Image Captioning for Indige-

nous languages of the Americas. Instead of relying only on scarce image captioning supervision, we first run SFT (MT) with Spanish–Indigenous translation data and then run SFT (IC). Development experiments show that SFT (MT) provides a useful initialization in our internal model-selection setting: SFT (IC) from the base model is much weaker than SFT (IC) from the SFT (MT) model. We also show that the same fine-tuned VLM can produce useful translation candidates for the overlapping languages under reference-based oracle candidate selection.

Under the official human-evaluation stage, our submissions rank in the top five for each submitted language, and our zero-shot GPT-5.5 run ranks first in the Maya language track under the same protocol. More broadly, our results suggest that low-resource cultural image captioning in this setting is not only a multimodal grounding problem but also a target-language generation problem. Translation data can therefore serve as a practical bridge toward multimodal generation in low-resource settings: SFT (MT) improves sample efficiency for image captioning because it teaches the model to produce stable text in the target Indigenous language, while SFT (IC) teaches the model to connect that language to the image. When image-caption supervision is scarce but bilingual text exists, translation fine-tuning can be an effective initialization.

Limitations

Our image captioning evaluation is limited by the small size of the internal development subset. Twenty examples are useful for internal model selection, but they cannot fully characterize performance across images, cultures, and linguistic forms. The shared task description allowed participants to use the development set for training, and we follow that setting.

Our translation results use best-of-100 candidate selection with reference-based chrF++ scoring. Although this uses the machine translation test set and chrF++, it is not the same as interactive translation: in real use, the system would not have the reference translation available when choosing an output. These scores should therefore be read as reference-based oracle-selected candidate-quality results.

The system is also limited by automatic evaluation. chrF++ is useful for low-resource translation and image captioning, but it cannot determine whether a caption is culturally appropriate or whether a translation is acceptable to speakers. Future work should include evaluation by language experts and community members.

Finally, the model may produce incorrect, offensive, or culturally inappropriate outputs. We do not recommend deployment without careful human review, community consultation, and safety procedures appropriate for each language context.

Ethical Considerations

Work on Indigenous languages requires care because language data is connected to communities, identity, and cultural knowledge. A model that generates text in an Indigenous language should not be presented as a replacement for speakers, translators, teachers, or community institutions. Our goal is to study supervised fine-tuning and reinforcement learning methods and provide a technical system for shared task evaluation, not to claim authority over language use.

The cultural image captioning setting also raises questions about what visual content should be described and how. Captions may encode cultural assumptions, and automatic systems can easily produce descriptions that are linguistically plausible but culturally wrong. Any real use of such a model should involve community-centered evaluation and control over data, outputs, and deployment contexts.

Acknowledgments

We thank the AmericasNLP organizers for creating the shared task and for supporting evaluation on Indigenous languages of the Americas.

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.

Luis Chiruzzo, Pedro Amarilla, Adolfo Ríos, and Gustavo Giménez Lugo. 2020. Development of a Guarani-Spanish parallel corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2629–2633.

Saurabh Dash, Yiyang Nan, John Dang, Arash Ahmadian, Shivalika Singh, Madeline Smith, Bharat Venkitesh, Vlad Shmyhlo, Viraat Aryabumi, Walter Beller-Morales, Jeremy Pekmez, Jason Ozuzu, Pierre Richemond, Acyr Locatelli, Nick Frosst, Phil Blunsom, Aidan Gomez, Ivan Zhang, Marzieh Fadaee, and 6 others. 2025. Aya vision: Advancing the frontier of multilingual multimodality. *Preprint*, arXiv:2505.08751.

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.

Ona De Gibert, Raúl Vázquez, Mikko Aulamo, Yves Scherrer, Sami Virpioja, and Jörg Tiedemann. 2023. Four approaches to low-resource multilingual NMT: The Helsinki submission to the AmericasNLP 2023 shared task. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 177–191, 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 Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 236–246, Mexico City, Mexico. 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.

Abteen Ebrahimi, Manuel Mager, Adam Wiemerslage, Pavel Denisov, Arturo Oncevay, Danni Liu, Sai Koneru, Enes Yavuz Ugan, Zhaolin Li, Jan Niehues, Monica Romero, Ivan G. Torre, Tanel Alumäe, Jiaming Kong, Sergey Polezhaev, Yury Belousov, Wei-Rui Chen, Peter Sullivan, Ife Adebara, and 15 others. 2022. Findings of the second AmericasNLP competition on speech-to-text translation. In *Proceedings of the NeurIPS 2022 Competitions Track*, volume 220 of *Proceedings of Machine Learning Research*, pages 217–232. PMLR.

Isaac Feldman and Rolando Coto-Solano. 2020. Neural machine translation models with back-translation for the extremely low-resource indigenous language Bribri. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3965–3976.

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.

Ximena Gutierrez-Vasques, Gerardo Sierra, and Isaac Hernandez Pompa. 2016. Axolotl: a web accessible parallel corpus for Spanish-Nahuatl. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4210–4214, Portorož, Slovenia. European Language Resources Association (ELRA).

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Jonathan Hus, Nathaniel Krasner, and Antonios Anastasopoulos. 2025. Machine translation using grammar materials for LLM post-correction. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 92–99, Albuquerque, New Mexico. Association for Computational Linguistics.

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

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2021. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the*

Americas, pages 202–217, Online. Association for Computational Linguistics.

OpenAI. 2026. Introducing GPT-5.5. https://openai.com/index/introducing-gpt-5-5/. Accessed: 2026-05-19.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2018. Sentence encoders on STILTs: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

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.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Mahshar Yahan and Mohammad Islam. 2025. Leveraging large language models for Spanish-indigenous language machine translation at AmericasNLP 2025. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 126–133, Albuquerque, New Mexico. Association for Computational Linguistics.

A Machine Translation Examples

Tables 7 and 8 show qualitative examples for both translation directions. Within each direction and language block, rows are ordered from a high-scoring example to an approximately average example and then a low-scoring example according to sentence-level chrF++. Figure 2 shows the corresponding sentence-level chrF++ score distributions for both translation directions and the four overlapping languages.

B Training Details and Prompts

All local experiments use Aya Vision 32B as the base model. We keep the base model frozen and train only LoRA parameters, using LoRA rank 32 and LoRA alpha 64. Both SFT (MT) and SFT (IC) use a learning rate of 2×10^{-5}, a linear scheduler, and a warmup ratio of 0.03. SFT (MT) is configured for 60 epochs with batch size 32 and gradient accumulation 1, but early stopping terminates training after 6.80 epochs. SFT (IC) is configured for 10 epochs with batch size 1 and gradient accumulation 1, with the selected checkpoints obtained after 9 epochs, initialized either from Aya Vision 32B or from the SFT (MT) checkpoint depending on the experiment.

B.1 Fine-Tuning Prompt Templates

For SFT (MT), each bilingual training example uses the following instruction-style prompt, and the target sentence is used as the assistant response:

```
Traduce del <source language> al <target language>.
```

For SFT (IC), each image-captioning example uses the following Spanish prompt:

```
Escribe un solo pie de foto en {language} para esta imagen. Debe ser una descripción culturalmente adecuada de la imagen. Responde solo con el pie de foto en {language}, sin explicaciones.
```

B.2 GPT-5.5 Zero-Shot Prompt and Technical Details

For the GPT-5.5 submission (v3), we used direct zero-shot Image Captioning through the OpenAI Responses API. Each input consisted of one image and a target language. We did not provide in-context examples, reference captions, intermediate Spanish captions, or translation outputs. The instruction string was:

```
You create image captions for an AmericasNLP submission. The target language is specified in the user prompt. Answer with exactly one caption and no surrounding text.
```

The user prompt template was:

```
Look at the image and write one concise, culturally appropriate caption directly in {language} (ISO {iso_lang}). Do not write in Spanish or English. Do not include labels, explanations, markdown, quotation marks, or translations. Return only the final caption in {language}.
```

We used model `gpt-5.5`, reasoning effort `medium`, image detail `high`, and maximum output tokens 4096. We generated one caption per image. No supervised fine-tuning, RLVR, or multi-candidate reranking was used for this run.

C Internal Development Subset IDs

The internal Image Captioning development subset contains 20 examples, selected as 10% of the development set with 5 examples per language and seed 42. The IDs are listed in Table 6 for reproducibility.

Table 6: Internal development subset example IDs.

Language	IDs
Bribri	bzd_040, bzd_007, bzd_001, bzd_047, bzd_017
Guaraní	grn_016, grn_015, grn_009, grn_048, grn_007
Orizaba Nahuatl	nlv_045, nlv_049, nlv_036, nlv_006, nlv_039
Wixárika	hch_048, hch_023, hch_022, hch_026, hch_034

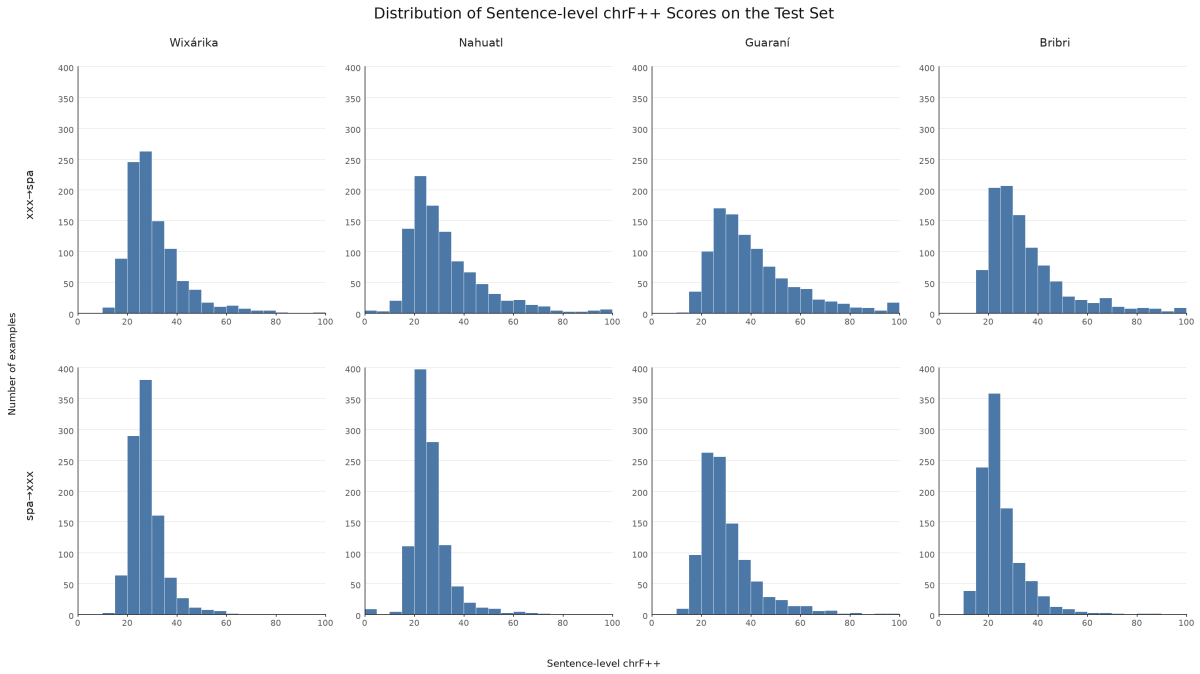

Figure 2: Distribution of sentence-level chrF++ scores on the machine translation test set for both translation directions and the four overlapping languages. These distributions complement the aggregate scores in Table 4 by showing the spread of reference-based oracle-selected candidate quality across examples.

Table 7: Qualitative translation examples from the SFT (MT) + RLVR (MT) model for the XXX→SPA direction. Within each language block, rows are ordered from a high-scoring example to an approximately average example and then a low-scoring example according to sentence-level chrF++.

chrF++	Language	Source	Reference	Model translation
100.00	Bribri	Sa' ù bak tkër kãmîx káx íxẽ tã.	Nuestra casa estaba lejos de este lugar.	Nuestra casa estaba lejos de este lugar.
35.26	Bribri	Îxñẽ káxm ye' dö, e' wa tã káxm se' i sãũ.	Hoy no he ido, así que no lo he visto.	Pues por ahora yo todavía no lo he visto, después algún momento lo veré.
17.88	Bribri	Ie' bak kèq klòqnẽ tã irir sãwîx ie.	Ella era increíblemente vulnerable.	Ella no había venido desde ese entonces.
100.00	Guaraní	Kennedy oñe'ê umi pilotokuéra ndive.	Kennedy habló con los pilotos.	Kennedy habló con los pilotos.
40.79	Guaraní	Ndaikuaaái ha'épa oipota umimarandu ñesê rehegua aravo.	No sabía si él quería información sobre los horarios de viaje.	No está claro si quería la información sobre el plazo de presentación.
18.84	Guaraní	Nameméi añemonde porâ.	No me pongo ropa de vestir a menudo.	No me dejes en la oscuridad.
100.00	Wixárika	waik+, waik+ Locust Hill.	No, no Locust Hill.	No, no Locust Hill.
30.66	Wixárika	m+k+ peuku yeix+a, p+reuyepi kitenie nepa eriwa nemana kunierix+ metá nepei xei kem+raka eriekai, tsiere nepeixei mey+ h+awekai.	Él vino, abrió la puerta y recuerdo mirar atrás y ver la expresión de su rostro, y pude ver que estaba decepcionado.	El joven se puso de pie, lo pegó en la puerta de su habitación, y fue a ver que estaba allí, y fue a ver que estaba en un sueño.
17.54	Wixárika	aix+ nuereu erie ena aimieme kwanetsi nake.	Me siento bastante seguro aquí y por eso me gusta.	En las orillas estaba una serpiente inmensa y negra.
100.00	Nahuatl	Fuerza Aerea quicohqui 18 aviones	La Fuerza Aérea compró 18 aviones.	La Fuerza Aérea compró 18 aviones.
32.37	Nahuatl	Coneme quiiqueh ce achi campán	Los niños bebieron un poco de champán.	Ellos quieren tocar un poco de campana.
12.40	Nahuatl	Ya nelpano eltoya yolquentzi	Ella era increíblemente vulnerable.	Pero ya viene mi madre, no los molestes.

Table 8: Qualitative translation examples from the SFT (MT) + RLVR (MT) model for the SPA→XXX direction. Within each language block, rows are ordered from a high-scoring example to an approximately average example and then a low-scoring example according to sentence-level chrF++.

chrF++	Language	Source	Reference	Model translation
86.70	Bribri	Viví en Texas.	Ye' sèxnẽ Texas.	Ye' sèxnẽ Texas árô.
25.00	Bribri	Siempre disfrutaba escuchando las historias de mi abuela.	Wake' ye' wîxke káxpake e' tsèke yö bua'ë.	Ye' ãmìx tso' e' tö ye' wã míxköla i chèke e' tsáx.
11.81	Bribri	El diseño era horrible	I dióqköl bak sulû ichóqk.	I mîxk sa' dör kóqyoq.
100.00	Guaraní	Vivían en Atlanta.	Oikova'ekue Atlanta-pe.	Oikova'ekue Atlanta-pe.
30.27	Guaraní	Me dijo que era urgente.	He'i chéve pya'eva'erâha.	He'i chupe ha'e péicha ha'erãjeýma.
14.05	Guaraní	Está bien si cometimos algunos errores.	Oî porâ, rojavy mimi.	Oĩmi jepémo, opavavévo oiko mba'e oporomongusúvo ko tendápe.
61.15	Wixárika	Ella dijo que su madre tiene el mismo caminar.	m+k+ petay+ kename mamaya yaxeik+a reyeikakai.	m+k+ petay+ kename mama payeikakai kename y+k+
27.78	Wixárika	Tu tiempo sigue aumentando cada año.	atukari p+t+ +rir+me witari manye yeika.	kuinie timaye yeitait+ mepeku mieme m+k+k+ ne hamikiekame muwa tukari wahiawariya.
15.85	Wixárika	No sé cuánto tiempo durará.	neukwara maté kem+ reutere,	ni nekai ukaratsi eriwaniya kemariyariyari,
61.06	Nahuatl	Augusta no es una gran ciudad.	Augusta amo hueyiatepetl.	Augusta niman amo hueyi altepetl.
26.13	Nahuatl	Pon un anuncio de refresco.	xihtlali tlamachiltilli cececatl.	Tlamaquetza un tomachiliztli ic in zan tlatlamacoz.
16.07	Nahuatl	Ella era increíblemente vulnerable.	Ya nelpano eltoya yolquentzi	Quipiya in itlacenpaj tlalquetzaloyan.