Retrieval-Augmented Long-Context Translation for Cultural Image Captioning: Gators submission for AmericasNLP 2026 shared task

Aashish Dhawan
University of Florida
aashish.dhawan@ufl.edu

Christopher Driggers-Ellis
University of Florida
driggersellis.cw@ufl.edu

Dzmitry Kasinets
University of Florida
dkasinets@ufl.edu

Christan Grant
University of Florida
christan@ufl.edu

Daisy Wang
University of Florida
daisyw@cise.ufl.edu

Abstract

We present the University of Florida Gators submission to the AmericasNLP 2026 shared task on cultural image captioning for Indigenous languages. Our two-stage pipeline generates a Spanish intermediate caption with Qwen2.5-VL, then produces the target-language caption using retrieval-augmented many-shot prompting with Gemini 2.5 Flash. We achieve 164.1%, 131.7%, and 122.6% improvements over the shared task baseline for Bribri, Guaraní, and Orizaba Nahuatl captioning, respectively, in our dev set evaluation and maintain >150% improvements for the Bribri and Orizaba Nahuatl languages in the test set evaluation. We find retrieval is highly language-dependent, beneficial only for large, in-domain corpora, and that synthetic data augmentation accounts for around 28 chrF++ of the dev set Guaraní performance gain. Our submission is the overall winner of the shared task, placing second out of five finalist submissions in human evaluations of target-language captions. Code and prompts are available on GitHub.[1]

1 Introduction

The AmericasNLP 2026 (Bui et al., 2026) consists of generating culturally grounded captions for images in Indigenous languages of the Americas. We identify three major challenges. First, the target languages are low-resource. Second, the captions are culturally specific rather than generic visual descriptions. Third, the task requires not only lexical transfer, but also stylistic control. Successful systems must produce short, natural captions that match the reference register expected by the organizers.

[1] https://github.com/dhawan98/AmericasNLP2026-Gators-Submission

While the cultural image captioning task is new to AmericasNLP 2026, we find that current vision-language models (VLMs) are not able to directly caption images in the target languages. This is unsurprising given the limited training data available for the target languages. We therefore formulate the problem as a compound of image captioning in a high-resource language followed by machine translation from that language into one of the shared task targets, which mirrors the shared task's baseline approach.

We first attempt to build on the work of Dhawan et al. (2026) by providing intermediate Spanish (Es) VLM captions to an mBART-based machine translation (MT) model. Dhawan et al. (2026) show that synthetic parallel data and language-specific preprocessing improve low-resource Indigenous MT, including Es-Guaraní (Grn) translation. In the 2026 shared task, however, a standard neural machine translation pipeline proves insufficient. When we apply the existing Es-Grn translation model to Spanish intermediate captions produced from the dev images, the resulting captions are often fluent enough at the sentence level but do not match the target caption register. In other words, the performance bottleneck shifts from generic translation quality to domain adaptation and culturally grounded caption style.

To address this mismatch, we move from sequence-to-sequence translation toward retrieval-augmented in-context translation with Large Language Models (LLMs). The core idea is simple. Instead of relying on a single model fine-tuned on mixed-domain low-resource language corpora, we retrieve Es-low-resource (LoRes) examples that are similar to the current caption and provide them as in-context exemplars. This design lets the decoder

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 212–223

adapt its lexical choices and stylistic register at inference time. We evaluate multiple OpenAI and Gemini models, vary the number of retrieved training examples and development exemplars, and test prompt variants and retrieval heuristics.

Our experiments lead to three main findings. First, among OpenAI models, many-shot direct translation outperforms both our mBART baseline and LLM post-correction, but gains are modest and highly sensitive to prompt composition. Second, Gemini 2.5 Flash (Comanici et al., 2025) is dramatically stronger in Es-LoRes in-context translation than the GPT-family models. Third, development references are useful as in-context exemplars for matching the expected caption register, but they must be interpreted carefully because same-pool development prompting can inflate development-set scores.

Our primary contributions are as follows. We present a retrieval-augmented LLM caption translation pipeline for low-resource cultural image captioning. We document an extensive set of negative and positive ablations, including model substitutions, prompt revisions, and development-exemplar hyperparameter searches. Finally, we highlight evaluation design as a central methodological issue for dev-conditioned in-context learning in low-resource captioning.

2 Background

Machine translation for Indigenous languages of the Americas has advanced largely through the AmericasNLP shared tasks, which have established evaluation benchmarks and made parallel corpora available for many low-resource languages (Mager et al., 2021; Ebrahimi et al., 2023, 2024). Strong systems in these shared tasks have typically relied on multilingual pretrained models such as mBART, M2M-100, or NLLB-200, often combined with synthetic data generation, multilingual transfer, and additional corpus collection (Gow-Smith and Sánchez Villegas, 2023; Tonja et al., 2023; Costa-Jussà et al., 2022). A recurring pattern in this literature is that performance improvements come not only from larger models, but also from better alignment between model capacity, augmentation strategy, and the target domain.

Character-based evaluation metrics are also especially relevant in this area. chrF and chrF++ are widely used for morphologically rich languages because they are more robust than BLEU to inflectional variation and spelling differences (Popović, 2015, 2017). The AmericasNLP 2026 organizers likewise use chrF++ as the first-stage ranking metric for the shared task.

Finally, our work relates to the broader use of in-context learning for low-resource generation. Instead of relying exclusively on fixed model parameters after fine-tuning, retrieval-augmented prompting can adapt the decoder to the current example at inference time. In our setting, this is especially attractive because the target outputs are short and stylistically constrained: parallel demonstrations can serve not only as semantic guides, but also as direct evidence of the desired caption register.

3 Dataset and Methodology

3.1 Baseline System

We compare our retrieval-augmented LLM pipeline against the Qwen3VL-8B (Qwen Team, 2025) Captioning and Sheffield 2023 (Gow-Smith and Sánchez Villegas, 2023) MT method, which is the stated baseline for the current shared task on image captioning. We apply this baseline to our submission results for dev set captioning in each of the applicable target languages and report percent improvement over the baseline scores.

For MT in Guaraní captioning, we also compare against the mBART-based Es-Grn model from Dhawan et al. (2026), trained on curated and synthetic parallel data. This serves as the strongest conventional translation baseline in our pipeline and allows us to assess whether retrieval-augmented in-context generation improves over standard sequence-to-sequence translation. We evaluate several retrieval-augmented GPT-4-family (OpenAI et al., 2024) model variants, summarized in Table 2: GPT-4o-mini, GPT-4.1-mini, GPT-4.1, along with the competing Gemini 2.5 Flash.

Across these models, we vary the number of retrieved training examples r, development exemplars d, and prompting strategy.

3.2 Task Data

The official development set $\mathcal{D}$ (dev set) contains 50 examples, and the organizers release the data in JSONL format paired with images. The pilot set includes Spanish captions for reference, but the organizers explicitly note that these are pilot-only and will not be present in development or test.

Our final submission pipeline is two-stage. Stage 1 produces a Spanish caption from the image using

Qwen 2.5B (Bai et al., 2025b). Stage 2 utilizes Gemini 2.5 Flash (Comanici et al., 2025) for Es-LoRes translation.

3.3 Datasets by Language

The shared task covers five target languages: Guaraní, Yucatec Maya, Orizaba Nahuatl, Bribri, and Wixárika. As shown in Table 1, the available retrieval data $\mathcal{R}$ differs substantially across languages, which motivates treating the retrieval size r and the number of development exemplars d as inference-time hyperparameters that vary across target language submissions.

Guaraní We use the largest retrieval bank in our setup. The retrieval Es-Grn bank of 53,183 pairs contains AmericasNLP 2023 (Ebrahimi et al., 2023) training data augmented with synthetic examples from the MultiScript30k project (Driggers-Ellis et al., 2025). The Guaraní retrieval dataset is relatively well aligned with the captioning task, as it contains culturally specific terms, proper nouns, and short descriptive examples useful for visual caption generation.

Yucatec Maya We do not have a comparable parallel training corpus for retrieval. The development set $\mathcal{D}$ is the only retrieval source, so we rely on dev exemplars and the pretrained knowledge of the LLM and we fix $r = 0$ in all experiments.

Other Target Languages We use the available Es-LoRes parallel data as retrieval banks. Their retrieval banks vary in size and domain match: Orizaba Nahuatl has 16,145 pairs, Bribri has 7,508 pairs, and Wixárika has 8,966 pairs. The Wixárika retrieval data is notably less caption-like, since much of the available corpus is narrative or literary rather than visual-description oriented. These differences motivate the language-specific hyperparameter choices reported in Table 1.

3.4 Retrieval Bank and Prompt Construction

For each target language, we construct a language-specific Es-LoRes retrieval bank $\mathcal{R}$ from the available parallel data described above. The Spanish side of each retrieval bank is indexed with BM25 (Robertson and Zaragoza, 2009), a TF-IDF-style lexical retrieval method that ranks candidate examples by query-term overlap while accounting for term importance and document length normalization. At inference time, the Spanish caption generated in Stage 1 is used as a query q, and the top-r retrieved pairs are selected as $\mathcal{R}_r(q) \subset \mathcal{R}$.

In addition to retrieved training pairs, some configurations include development exemplars. Let $\mathcal{D}$ denote the shared-task development set for a target language, and let $\mathcal{D}_d \subset \mathcal{D}$ denote the d development examples included in the prompt. For a query caption q, the full prompt context is defined as

$$P(q) = \mathcal{D}_d \cup \mathcal{R}_r(q),$$

where $\mathcal{R}_r(q)$ provides retrieval-based semantic and lexical grounding, and $\mathcal{D}_d$ provides examples of the caption style expected by the task. The model then generates the target-language caption conditioned on $P(q)$.

We sweep r and d where applicable. The values selected for submissions are in Table 1, and detailed grid-search results appear in our appendix.

3.5 Prompting Strategy

The MT prompt structure is deliberately simple. The system prompt instructs the model to translate from Spanish into the target language, match the example style, stay concise, preserve culturally specific nouns when appropriate, and produce exactly one line. The user prompt contains two evidence blocks, development exemplars $\mathcal{D}_d$ and retrieved Spanish–target-language pairs $\mathcal{R}_r(q)$, followed by the current Spanish caption.

We use the same general prompt structure across target languages, adding language-specific modifications only when required. After several prompt-engineering attempts, we found that more aggressive instructions, such as suppressing generic lead-ins or forcing noun-first phrasing, reduced performance. The final prompts therefore keep the system instruction minimal and rely on in-context examples for lexical and stylistic guidance. We summarize the final and ablation prompt files in Table 5 of the appendix.

3.6 University of Florida Gators Submission

Our submission system features a retrieval-augmented long-context translator embedded in a two-stage image captioning pipeline. Figure 1 illustrates the full system, which is organized into five steps. Stage 1 corresponds to Step 1 in the diagram. A vision-language model generates one Spanish caption q for each target image. We use either Qwen2.5-VL-72B-Instruct in 4-bit precision or Qwen3-VL-8B (Bai et al., 2025a) for this stage.

The prompt is culturally aware; follows a noun-first style; and encourages concise descriptions of visible entities, objects, clothing, actions, and scene context. We treat this stage as fixed and do not optimize it extensively in this paper.

Stage 2 corresponds to Steps 2–5 in Figure 1. It transforms the generated Spanish caption q into the final target-language caption using retrieval-augmented many-shot MT. The q is used as a BM25 query over the Spanish side of the available Es-LoRes retrieval bank $\mathcal{R}$. In Step 3, the retrieved subset $\mathcal{R}_r(q)$ and development subset $\mathcal{D}_d$ are assembled into a many-shot prompt with an instruction to translate from Spanish to the target language while matching the example style. Each prompt contains approximately 3K–5K tokens. The r nearest Es-LoRes training pairs provide semantic and lexical grounding, while the d development pairs provide direct evidence of the target caption register. In Step 4 of Figure 1, Gemini 2.5 Flash performs the final target-language generation with temperature set to 0.0, thinking disabled, and a maximum output length of 120 tokens. We then strip prefixes in Step 5 and normalize whitespace to produce the final JSONL submission. The system is evaluated using chrF++, followed by human judgment for the top-ranked submissions.

4 Results

Table 1 gives our final submission's performance for each of the target languages in the shared task while Table 2 summarizes the main progression of Es-Grn dev set experiments from an mBART translation model to in-context MT with Gemini 2.5 Flash, which is competent in all target languages. Several trends emerge immediately. First, direct many-shot translation is better than post-correction, confirming that it is more effective to generate the caption in one step than to repair the mBART output after the fact. Second, among the OpenAI models we tested, GPT-4o-mini remains the strongest, but the gains over the mBART baseline are modest. Third, Gemini 2.5 Flash yields much larger improvements, even before adding any development exemplars.

In our final dev set results, the Guaraní target achieves the highest absolute chrF++ score among the five target languages by at least 10 chrF++ and more than doubles the Bribri and Wixárika target performances. For Guaraní we achieve a remarkable 131.7% improvement over the baseline

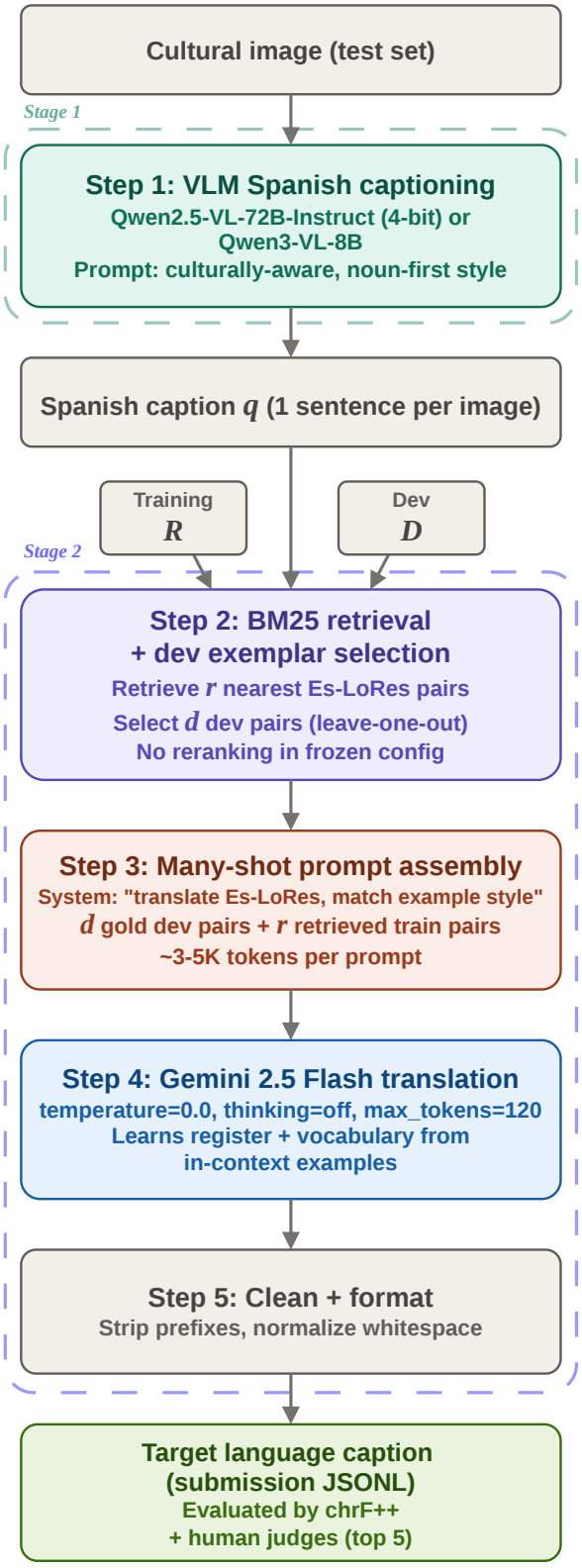

Figure 1: Overview of the proposed two-stage image captioning pipeline. Here, $\mathcal{R}$ denotes the full Es-LoRes retrieval bank, $\mathcal{R}_r(q) \subset \mathcal{R}$ denotes the r nearest Es-LoRes training pairs retrieved for Spanish caption q, $\mathcal{D}$ denotes the full development set, and $\mathcal{D}_d \subset \mathcal{D}$ denotes the d gold dev pairs used in the many-shot prompt.

Configuration					test		dev	
Language	r	d	Test Examples	Retrieval Pairs (R)	Test chrF++	% Imp.	Dev chrF++	% Imp.
Bribri	80	20	267	7,508	17.90	155.2%	19.99	**164.1%**
Guaraní	80	49	101	53,183*	23.10	14.72%	**48.24**	131.7%
Orizaba Nahuatl	40	20	200	16,145	**25.42**	**166.9%**	25.67	122.6%
Wixárika	40	20	201	8,966	17.58	3.952%	18.99	6.866%
Yucatec Maya	–	49	212	0†	21.11	–	26.29	–

Table 1: Submission chrF++ results across languages for the proposed system. Here, r is the number of retrieved training pairs included in each prompt, d is the number of development exemplars included in each prompt, and R denotes the total size of the available retrieval bank for each language. Baseline results are those provided for the shared task (Gow-Smith and Sánchez Villegas, 2023). **Bold** and underlined entries in the right-side columns indicate the best and second-best results, respectively. * Includes parallel Es-Grn MultiScript30k (Driggers-Ellis et al., 2025) synthetic exemplars. † No external retrieval bank is used; only development exemplars are included.

System	r	d	Notes	Setting type	chrF++
mBART Baseline	–	–	Inherited MT Baseline	Baseline	22.40
GPT-4o-mini Post-Correction	–	8	Revise mBART Draft	Post-Editing	19.81
GPT-4o-mini	24	4	Direct Many-Shot Translation	Direct Prompting	23.41
GPT-4o-mini	28	4	Best OpenAI Setting	Direct Prompting	23.60
GPT-4.1-mini	28	4	Model Swap	Direct Prompting	23.23
GPT-4.1	28	4	Model Swap	Direct Prompting	23.49
Gemini 2.5 Flash	28	0	No dev Exemplars	Train-Only Prompt	32.14
Gemini 2.5 Flash	28	4	Same-Pool dev Prompting	Dev-Assisted	39.25
Gemini 2.5 Flash	28	20	Same-Pool dev Prompting	Dev-Assisted	42.90
Gemini 2.5 Flash	28	49	Same-Pool dev Prompting	Dev-Assisted	42.19
Gemini 2.5 Flash	80	49	Same-Pool dev Prompting	Dev-Assisted	**48.24**

Table 2: Dev chrF++ Guaraní captioning results across baseline machine translation and RAG-based prompting configurations. **Bold** indicates the best result, and Underline indicates the second-best result.

method (Gow-Smith and Sánchez Villegas, 2023). In Section 5, we investigate the effect that synthetic exemplars may have had on Guaraní target language performance versus the other languages for which no MultiScript30k data exists. However, as the final column in Table 1 attests, each language target outperforms the baseline method whenever the baseline is available. In particular, for the Bribri and Orizaba Nahuatl targets, respectively, we achieve 164.1% and 122.6% improvements over the baseline. As we highlight in Table 1, our improvement for Bribri is the greatest relative improvements over the baseline method for any target language.

The test results are slightly different. Bribri is replaced by Orizaba Nahuatl as most improved language. The languages achieve 155.2% and 166.9% improvement over the testing baseline, respectively (Bui et al., 2026). Guaraní performance drops by more than half in the test evaluation and most of the dev set performance gain is not reproduced in the official evaluation as a result. Wixárika performance improvement falls by a similar amount.

5 Ablations

In addition to the results we list in the previous section, we provide a number of ablations to demonstrate the superiority of the final submission over numerous alternative approaches to various facets of the captioning pipeline. In the experiments reported here, we hold the Spanish captions fixed and optimize only the Guaraní translation stage. This lets us analyze ablation effects independent of the image captioner and makes the ablations directly comparable to our dev set results. We focus on the dev set evaluation here because ablations are all performed on the dev set before the release of final shared task results.

5.1 Machine Translation Architecture

We allude in the first section to how the initial approach for the machine translation step in Guaraní utilizes an mBART based MT model (Dhawan et al., 2026). Table 2 shows Guaraní captioning performance across several MT architectures, including mBART, GPT-4o, and Gemini 2.5 Flash. Where appropriate, we adopt the retrieval-augmented approach in our final submission, and

we vary the values of r and d within architectures.

The results show that the best configuration for Guaraní translation is the Gemini 2.5 Flash LLM prompted with $r = 80$ and $d = 49$ exemplars. We traverse much of the r, d search grid with OpenAI models. The best OpenAI setting used GPT-4o-mini with 28 retrieved examples and 4 development exemplars, reaching 23.60 chrF++. This performance is reflected in Table 2. Removing or increasing development exemplars reduced performance. GPT-4.1-mini and GPT-4.1 were both competitive but did not surpass GPT-4o-mini. This pattern suggests that, for the OpenAI models we tested, the in-context regime has a narrow optimum. Too little context leaves the model underconstrained; too much context appears to add noise or dilute the style signal.

For its superior performance over GPT models and its positive receptivity to context, we adopt Gemini 2.5 Flash for Es-LoRes MT in our final submission.

5.2 Hyperparameter Search

For each language, we sweep values of r and d that we list in Section 3. Table 1 of final submission results shows performance at the submittal configuration featuring its r, d pair. To be thorough, Table 6 reports the results of a partial grid search for each target language in our appendix.

Though this ablation shows that the r and d exemplar counts from Table 1 are optimal within our search grid, there is no accounting for values outside of it. Additionally, performance changes as one scans the grid indicate differing impacts of the r and d hyperparameters for different languages and data sources. Thus, we stress in this ablation the importance of a thorough search for the optimal r, d pair for any new target language or data configuration in retrieval-augmented Es-LoRes MT.

5.3 Synthetic Exemplars

We notice in Section 4 that our pipeline performs Guaraní image captioning much more effectively in absolute terms versus any of the other target languages, even though we achieve similar percent improvement over the applicable baseline for the Bribri and Orizaba Nahuatl targets. We also note that Guaraní is the only target language for which we include synthetic exemplars from MultiScript30k. Testing for the effect of this synthetic data, we ablate the original Guaraní captioning submission by only using AmericasNLP 2023 (Ebrahimi et al., 2023) training data for retrieval exemplars (r). Table 3 compares performance with and without MultiScript30k (Driggers-Ellis et al., 2025) synthetic exemplars for three r, d pairs and provides the greatest performance overall without synthetic exemplars.

Data	r	d	Ret. Pairs	chrF++
ANLP2023 + MS30k	40	49	53,183	51.34
ANLP2023 + MS30k	40	20	53,183	48.38
ANLP2023 + MS30k	80	49	53,183	48.24
ANLP2023	80	0	26,032	22.65
ANLP2023	40	49	26,032	21.03
ANLP2023	40	20	26,032	20.50
ANLP2023	80	49	26,032	20.75

Table 3: Dev chrF++ results for Guaraní captioning with differing retrieval exemplar sets. Data includes AmericasNLP2023 (ANLP2023) training data (Ebrahimi et al., 2023) and/or MultiScript30k (MS30k) (Driggers-Ellis et al., 2025) synthetic exemplars as noted in the column *Data*.

The results are clear. Controlling for our retrieval hyperparameters by fixing r and d to three pairs, we observe that the original configuration with both genuine AmericasNLP 2023 training data and synthetic MultiScript30k exemplars outperforms the ablation with AmericasNLP 2023 alone by more than 100% relative improvement in chrF++ in each case. Additionally, the best Guaraní captioning performance without synthetic exemplars is 55.9% less than the best performance with them. These results mirror the comparison of our submission's Guaraní performance to the other target languages in the shared task and attribute much of our improvement in Guaraní captioning to synthetic exemplars in the retrieval-augmented Es-Grn translation step.

5.4 Alternative Prompting and Reranking

We also test several prompting ablations that looked reasonable but are consistently negative in their impact on performance, with the exception of a specific strategy we give additional attention to in Section 5.5. In this section, we quickly summarize other alternative prompts, and in Table 5, we produce all of the relevant prompts for completeness.

A prompt rewrite aimed at suppressing generic scene-introduction phrases reduces performance substantially, and a retrieval reranker that attempts to prefer short caption-like pairs also reduces performance, both for OpenAI and Gemini. Likewise, using 49 same-pool development exemplars with

GPT-4o-mini degraded performance instead of improving it. These results matter because they show that this task does not respond well to aggressive heuristic control. The most effective systems are built by keeping the instruction stable and varying only model choice and amount of context supplied.

For Wixárika in particular, we devise special prompts with cultural context in the form of a glossary. These glossaries contain the names of common objects from the Wixárika culture and their definitions in Spanish, with the hope that this additional context will help the LLM MT architectures translate intermediate Spanish captions into the Wixárika target language. We utilize two versions of the cultural glossary, but the results do not improve for either one. We include the additional prompts in Table 5 to document this alternative prompting.

5.5 Morphological Considerations for Bribri

We ablate our submission for the Bribri (Bzd) language by considering performance with and without morphological prompting and post-processing for the Es-Bzd MT. Bribri has complex tonal marking in orthography (circumflexes, underlines, multiple diacritics) which makes the deduplication logic harder (Coto-Solano, 2021). The tokenizer splits differently on tonal characters.

In our final submission, we utilize a complex postprocessing and additional prompting for the Gemini MT model to improve performance. Table 4 shows the evolution of our strategies to account for the morphological complexity of the Bribri language. First, careful examination of our pipeline's output reveals that Bribri captions are initially in NFC encoding, which combines base letter and diacritic encodings into one character. This does not match the dev set examples, which use NFD encoding. For instance, our model outputs *ë* as a single unit, but the dev set's NFD encoding would separate the letter *e* from the umlaut above it. Because chrF++ is a character-level lexical metric, this mismatch depresses the score. We therefore perform NFD-Normalization (NFD-Norm.) to account for the difference. Secondly, we believe the morphological considerations of Es-Bzd translation significant enough that they deserve special prompting. To this end, we formulate a Morphological Prompt (Morph.) for Bribri that includes special instructions about Subject-Object-Verb (SOV) word order, tonal diacritics, common consonant clusters, verb-final clauses, and possessive noun prefixes.

Language	Prompting	Postprocessing	chrF++
Bribri	Standard	Standard	11.50
Bribri	Standard	NFD-Norm.	17.02
Bribri	Morph.	NFD-Norm.	19.99

Table 4: Morphological ablations for our submission in the Bribri captioning task. We either utilize NFD-Normalization (NFD-Norm.) as a postprocessing step, both NFD-Norm. and Morphological Prompting (Morph.), or we apply the same prompting and postprocessing as other target languages (Standard). Rightmost column shows Dev ChrF++.

As in previous ablations, the results in Table 4 largely speak for themselves. NFD-normalization overcomes the encoding mismatch, and morphological prompting provides additional improvements.

6 Discussion

We now proceed to a discussion of our method and the results achieved. We analyze the available data and investigate the potential sources of performance improvement over the baseline method.

6.1 In-Context Retrieval

Though limited to the Guaraní target, we show in our ablation of MT architectures that long-context retrieval-based MT using state-of-the-art LLMs greatly improves image captioning performance at the MT stage versus the dedicated mBART model (Dhawan et al., 2026). While we cannot say for certain whether this relationship holds for other language targets, from comparisons to baseline performances, it appears that additional context from real and synthetic exemplars significantly boosts LLM translation in Es-LoRes tasks.

6.2 Synthetic Exemplars

For the Guaraní language, we include approximately 30k synthetic exemplars for in-context retrieval from the MultiScript30k (Driggers-Ellis et al., 2025) dataset. We observe in Section 4 that Guaraní achieves substantially higher absolute dev chrF++ than the other target languages.

Remembering that synthetic exemplars apply only to Guaraní and its significantly positive effect on Es-Grn MT in related work (Dhawan et al., 2026), we ablate for the synthetic exemplars' effect on Guaraní captioning.

For Guaraní captioning without the MultiScript30k synthetic retrieval pairs, the results we elaborate in Section 5.3 and Table 3 show a comparison similar to Guaraní captioning versus the other target languages. Results improve over 100%, approximately 28 total chrF++ or more, for Guaraní captioning with the synthetic exemplars for the same r, d pair. These results also mirror observations from (Dhawan et al., 2026) on the effect of synthetic exemplars on Es-Grn MT. We therefore conclude that synthetic exemplars drive Guaraní performance gains and speculate that synthetic retrieval pairs may further improve captioning performance for other target languages.

6.3 Morphology

Despite favorable results in the Guaraní target, we acknowledge in Section 4 that our greatest relative improvement is for the Bribri language. The ablation in Section 5.5 clarifies that much of this performance improvement stems from the morphological considerations we take for the Bribri language target via specialized prompting. We observe that explicitly prompting for Bribri's morphological features accounts for over 10% of the performance gain in our final submission over the shared task's baseline (Gow-Smith and Sánchez Villegas, 2023). We conclude morphological prompting has potential for LoRes MT in morphologically complex languages.

7 Future Work

The most impactful next step is improving the visual captioning stage. Error analysis indicates that approximately 54% of remaining Guaraní errors originate in the vision model rather than the translator. A stronger VLM will likely yield larger gains than further translation tuning. Beyond captioning, we may extend retrieval to incorporate visual similarity rather than Spanish text overlap alone, which would help when the intermediate caption is noisy or culturally ambiguous. The lowest-resource languages (Wixárika and Bribri languages) are bottlenecked by corpus domain mismatch rather than model choice. Even small caption-style parallel data corpora for these languages would likely produce larger gains than any prompting improvement.

8 Conclusion

We present the University of Florida Gators team system for the AmericasNLP 2026 shared task on cultural image captioning for Indigenous languages (Bui et al., 2026). Our two-stage pipeline, VLM captioning in Spanish followed by retrieval-augmented many-shot translation with Gemini 2.5 Flash, substantially outperforms fine-tuned baseline models across all five target languages, culminating at 48.24 chrF++ for Guaraní in the submission. We find that retrieval behavior is highly language-dependent. Large retrieval windows help Guaraní but hurt Yucatec Maya, where Gemini's pre-training knowledge is sufficient and BM25 retrieval adds noise. We also find that development exemplars are useful for matching caption register, but their use requires careful interpretation because they can inflate development-set scores when drawn from the same evaluation pool. For the lowest-resource languages, the ceiling is domain mismatch, not model capacity.

Limitations

Our system is a cascade: errors in the Spanish captions propagate into translation with no recovery mechanism. Because most ablations hold the Spanish captions fixed and vary only the translation stage, they likely underestimate the contribution of the visual captioning model to final performance.

A second limitation is the use of development examples as in-context exemplars. These examples are useful for matching the expected caption register, especially when little caption-style target-language data is available, but they can also inflate development-set scores when exemplar selection and evaluation draw from the same small pool. We therefore use dev-assisted results primarily primarily model-selection and submission-configuration. A more rigorous and comprehensive evaluation would report held-out or cross-split dev results.

Finally, evaluation relies primarily on chrF++, which is useful for low-resource and morphologically rich languages but cannot fully capture fluency, cultural appropriateness, or naturalness for native speakers. Although the shared task includes human evaluation for finalist systems, our ablations include no native-speaker evaluation.

Acknowledgments

The authors gratefully acknowledge Arnold and Lisa Goldberg, whose financial support helped to make this work possible. We also acknowledge the Gatorade, whose support provided the compute resources necessary for this research.

References

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025b. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

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

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Rolando Coto-Solano. 2021. Explicit tone transcription improves ASR performance in extremely low-resource languages: A case study in Bribri. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 173–184, Online. Association for Computational Linguistics.

Aashish Dhawan, Christopher Driggers-Ellis, Christan Grant, and Daisy Zhe Wang. 2026. Improving indigenous language machine translation with synthetic data and language-specific preprocessing. In *Proceedings for the Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2026)*, pages 119–126, Rabat, Morocco. Association for Computational Linguistics.

Christopher Driggers-Ellis, Detravious Brinkley, Ray Chen, Aashish Dhawan, Daisy Zhe Wang, and Christan Grant. 2025. Multiscript30k: Leveraging multilingual embeddings to extend cross script parallel data.

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 et al. 2024. Findings of the americasnlp 2024 shared task on machine translation into indigenous languages. In *Proceedings of the AmericasNLP Workshop*.

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 Third Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*.

Manuel Mager, Arturo Oncevay, et al. 2021. Findings of the americasnlp 2021 shared task on open machine translation for indigenous languages of the americas. In *Proceedings of the AmericasNLP Workshop*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, et al. 2024. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*. Association for Computational Linguistics.

Qwen Team. 2025. Qwen3 technical report.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Atnafu Lambebo Tonja, Hellina Hailu Nigatu, Olga Kolesnikova, Grigori Sidorov, Alexander Gelbukh, and Jugal Kalita. 2023. Enhancing translation for indigenous languages: Experiments with multilingual models. In *Proceedings of the Third Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*.

Appendix

Here, we provide additional data for validation of our method versus various ablations with particular focus on alternative prompts for various target languages and the r, d grid search. In the following sections, we give tables of alternative prompts and performance at different r, d pairs for each target language.

Alternative Prompts

Table 5 yields the prompts that we utilize in our ablations and the prompts present in our final submission for each target language as indicated in the *Standing* column. The *Language(s)* column shows which languages the prompt applies to.

r, d Hyperparameter Search

We frequently refer to a search for optimal r and d retrieval hyperparameters in the main body of this paper but reserve detailed communication of the sweep for this appendix due to the number of configurations we consider. For the Gemini 2.5 Flash MT architecture and for each target language, we partially sweep a grid consisting of r, d pairs such that $r \in \{0, 10, 20, 40, 80\}$ and $d \in \{0, 10, 20, 30, 40, 49\}$. Table 6 reports the chrF++ scores for the final captioning pipeline for each combination of r and d tested for each target language.

Short Desc.	Description	Prompt URL(s)	Standing	Language(s)
Captioning	Prompts for VLM in first stage of captioning pipeline.	v1_submission/guarani_caption_prompt.txt	**Final**	All
Many-Shot	Includes r and d in-context exemplars from previous parallel text corpora and from the dev set, respectively.	v1_submission/<target>_system_prompt.txt	**Final**	All
Morph. Bribri	Accounts for Bribri target's morphological complexity. Portion shown appended to general prompt.	v1_submission/bribri_system_prompt.txt	**Final**	Bribri
Wixárika Gloss. *v2*	Provides definitions for culturally situated nouns in the Wixárika language.	wixarika/caption_es_wixarika_v2.txt	Ablation	Wixárika
Wixárika Gloss. *v3*	Provides definitions for culturally situated nouns in the Wixárika language.	wixarika/caption_es_wixarika_v3.txt	Ablation	Wixárika

Table 5: Prompts we utilize in our ablations and final submission for each of the target languages. The *Standing* column indicates whether a prompt is an Ablation or part of our **Final** submission. The final *Language(s)* column indicates what languages the prompt applies to.

Bribri				
Language	Retrieval (r)	Dev Exemplars (d)	Notes	chrF++
Bribri	80	20	Uses Morphological Bribri prompting and NFD-normalization.	**19.99**
Bribri	80	20	Uses Regular Prompting.	11.50
Bribri	40	20	...	11.41
Bribri	10	10	...	11.41
Bribri	20	20	...	11.16
Bribri	20	10	...	10.95
Bribri	40	10	...	10.95
Bribri	10	10	...	10.63
Bribri	80	10	...	10.17
Bribri	80	0	...	6.53
Bribri	40	0	...	5.93
Bribri	20	0	...	5.30
Bribri	10	0	...	4.75
Guaraní				
Language	Retrieval (r)	Dev Exemplars (d)	Notes	chrF++
Guaraní	40	49	Includes synthetic exemplars. Not submitted because ablation was incomplete at submission deadline.	51.34
Guaraní	40	20	...	48.38
Guaraní	80	49	Includes synthetic exemplars.	**48.24**
Guaraní	80	20	...	42.61
Guaraní	0	49	...	20.80
Orizaba Nahuatl				
Language	Retrieval (r)	Dev Exemplars (d)	Notes	chrF++
Orizaba Nahuatl	40	20	–	**25.67**
Orizaba Nahuatl	40	10	–	25.59
Orizaba Nahuatl	80	20	–	25.25
Orizaba Nahuatl	80	10	–	25.16
Orizaba Nahuatl	20	20	–	24.16
Orizaba Nahuatl	20	10	–	23.91
Orizaba Nahuatl	40	0	–	16.56
Orizaba Nahuatl	80	0	–	16.37
Orizaba Nahuatl	20	0	–	15.61
Wixárika				
Language	Retrieval (r)	Dev Exemplars (d)	Notes	chrF++
Wixárika	40	20	–	**18.99**
Wixárika	40	10	–	17.74
Wixárika	20	20	–	17.56
Wixárika	80	10	–	17.48
Wixárika	80	20	–	17.13
Wixárika	20	10	–	16.81
Wixárika	40	0	–	16.59
Wixárika	80	0	–	16.25
Wixárika	20	0	–	13.80
Yucatec Maya				
Language	Retrieval (r)	Dev Exemplars (d)	Notes	chrF++
Yucatec Maya	–	49	Fixes $r = 0$ for lack of retrieval exemplars.	**26.29**
Yucatec Maya	–	20	...	26.29
Yucatec Maya	–	40	...	25.07
Yucatec Maya	–	30	...	25.05
Yucatec Maya	–	0	Fixes $r = 0$ for lack of retrieval exemplars. At $d = 0$, the model receives no signal from the target language.	20.25

Table 6: Dev chrF++ results across languages for the proposed system. (...) Indicates previous *Notes* column entry applies. (–) Indicates no *Notes*. **Bold** entries *ChrF++* column are submission scores for each target language.