Culturally-Aware Image Captioning for Guaraní with Multimodal Prompting: IUHoosiers at AmericasNLP 2026

Wenchen Shi*, **Phakphum Artkaew***, **Luke Gessler**
Indiana University Bloomington
{wencshi, partkaew, lgessler}@iu.edu
*Equal contribution

Abstract

The AmericasNLP 2026 Shared Task challenges systems to generate culturally grounded image captions in indigenous languages of the Americas, a setting that demands both cultural awareness and linguistic accuracy for severely under-resourced languages. We present IUHoosiers, Indiana University's system for the Guaraní track. Rather than fine-tuning, our approach centers on inference-time knowledge injection: for each test image, we retrieve relevant Guaraní grammatical and cultural resources using BM25 and inject them into a large vision-language model's prompt alongside the image, enabling language-specific cultural and linguistic grounding without any parameter updates. IUHoosiers placed first for Guaraní in both automatic evaluation (24.67 chrF++) and human evaluation (3.45/5), outperforming all other participating systems.

1 Introduction

The AmericasNLP 2026 Shared Task challenges systems to generate culturally grounded image captions in indigenous languages of the Americas (Bui et al., 2026). We focus on Guaraní, a Tupi-Guaraní language that is spoken widely across Paraguay, Brazil, Bolivia, and Argentina and is the co-official language along with Spanish in Paraguay (∼6.5M speakers). However, it is still considered an under-resourced language across all NLP resources (Chiruzzo et al., 2020).

Cultural captioning for Guaraní presents two intertwined challenges. First, captions must be *culturally accurate*: correctly distinguishing closely related cultural items, such as *mate* from its cold counterpart *tereré*, requires culturally grounded knowledge that generic vision-language models often lack. Second, the target language itself is challenging: Guaraní is agglutinative with active/inactive voice morphology and pervasive Spanish borrowing in everyday *jopara* speech (Estigarribia, 2020), making linguistically correct output non-trivial even for large language models.

Since the dev set contains only 50 image-caption pairs and the gold labels follow a specific format, parameter-efficient fine-tuning (PEFT) methods (Mangrulkar et al., 2022) such as low-rank adaptation (LoRA; Hu et al., 2022) may be prone to overfitting without substantial data augmentation. We therefore focus on inference-time methods instead. Recent work suggests that grammar-book and parallel-text context can improve low-resource language generation without gradient updates (Tanzer et al., 2023; Aycock et al., 2025; Zhang et al., 2025), and we build on this finding with **retrieval-augmented prompting**. For each test image, we generate a text description, query four Guaraní knowledge pools with BM25 (Robertson and Zaragoza, 2009), and inject the top-k retrieved items into the system prompt of Gemma 4 31B (Google DeepMind, 2025) to generate captions in a single forward pass. Additionally, we explore a *visual few-shot* mode that injects cultural images from Diccionario audiovisual multilingüe del Paraguay (DAMPY)[1] as multimodal context. Figure 1 gives an overview of the full pipeline.

2 Approach

Why in-context learning? The AmericasNLP 2026 Shared Task spans five indigenous languages (Guaraní, Bribri, Yucatec Maya, Wixárika, and Nahuatl), each with very distinct typological properties, morphological systems, and cultural contexts. We believe the linguistic and cultural distances among these languages are too vast for any single unified system to address all of them effectively, and we take the view that in low-resource settings, building one dedicated system per language is the more principled approach. A language-specific pipeline allows much more targeted injec-

[1] https://spl.gov.py/dampy/index.html

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 236–242

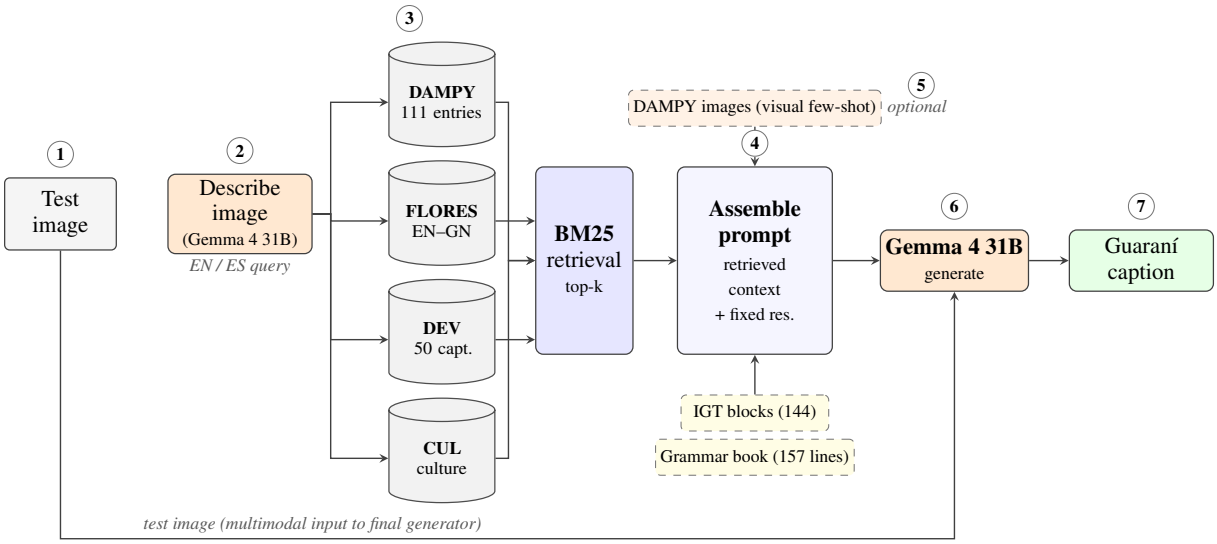

Figure 1: Overview of the IUHoosiers pipeline for Guaraní image captioning. Given a test image **(1)**, Gemma 4 31B generates an EN/ES description **(2)** that is used as a BM25 query over four Guaraní knowledge pools **(3)**: cultural entries (DAMPY), parallel sentences (FLORES), dev-set gold captions (DEV), and a curated cultural knowledge base (CUL). Step **(3)** operates in one of two modes: *RAG* dynamically selects the top-k entries per image via BM25, while *static* mode takes the first n examples from the dataset regardless of image content (used in v6 and v8; see Table 1). The retrieved context is combined with fixed resources (a 157-line grammar excerpt and 144 interlinear-glossed example blocks) into the system prompt **(4)**. In visual few-shot mode **(5)**, DAMPY images with their captions are additionally injected as multimodal shots (dashed). Gemma 4 31B **(6)** consumes the assembled prompt together with the test image and produces a Guaraní caption **(7)**. EN: English, ES: Spanish, GN: Guaraní.

tion of relevant grammatical structure and cultural knowledge than a one-size-fits-all system could provide. We therefore build a Guaraní-specific pipeline and adopt in-context learning (ICL) as our core method. Recent work on extremely low-resource languages shows that ICL, especially when paired with explicit linguistic signals, can outperform parameter-efficient fine-tuning when the target language is poorly represented in the base model (Li et al., 2025). The feasibility of injecting many resources simultaneously is further supported by recent long-context language models (Gemini Team et al., 2024), which make it practical to include large amounts of in-context material within one context window. At the same time, work on culturally aware vision-language modeling has shown that standard VLMs often miss fine-grained cultural distinctions, and that improving cultural grounding may require either specialized data construction or model adaptation (Huang et al., 2025). Our approach therefore tests whether carefully selected grammatical and cultural knowledge, injected at inference time, can close this gap without any parameter updates. [2]

Model. We use Gemma 4 31B (Google DeepMind, 2025), a 31-billion-parameter vision-language model with a 128K-token multimodal context window. The original baseline prompt (Spanish-language Guaraní captioning guidelines) is used for all nine submitted configurations.

Fixed resources. Two resources are injected into every prompt regardless of the test image: a 157-line condensed grammar-book excerpt from Estigarribia (2020) covering Guaraní morphosyntax, voice alternations, and postpositional structure, and 144 interlinear glossed example blocks (surface form / morpheme break / gloss / translation). Both resources were adapted from Aycock et al. (2025), who study grammar-book prompting across multiple low-resource languages including Guaraní. Their preprocessed materials served as our starting point.

We note that these distilled resources were generated with the assistance of Claude-Opus-4.7 in a two-step process. First, we provided Claude with the raw resource files together with the competition setting and asked it to reason about the most effective format for injecting such resources into a language model for in-context learning. Through iterative discussion, Claude then produced prompts

[2]Code is available at https://github.com/victorshi119/Cultural_Image_Captions_2026

designed to reorganize and condense the original materials into more model-friendly representations. In a second, separate session, we supplied Claude with the raw text resources together with the generated prompt in order to produce the final distilled resources.

Dynamically retrieved pools. For each test image a text description is generated and used as the BM25 query (Robertson and Zaragoza, 2009). Four resource pools are indexed and queried per image:

1. **DAMPY**: 111 culturally grounded entries from DAMPY, Paraguay's official audiovisual dictionary (42 food, 47 fauna, and 22 flora entries), each with bilingual Spanish/Guaraní labels. We note that the DAMPY dataset we scraped from the internet does not provide captions. We therefore generated bilingual captions in two steps: first, a VLM produced baseline Guaraní captions for each image; then, Claude-Opus-4.7 in Adaptive Thinking mode refined those captions for cultural accuracy and idiomatic Guaraní usage.

2. **FLORES**: English-Guaraní parallel sentence pairs from FLORES-200 (NLLB Team et al., 2022), a publicly available multilingual benchmark dataset providing in-domain Guaraní sentence structure across a broad topical distribution (1,012 sentence pairs used).

3. **DEV**: the 50 gold Guaraní dev-set captions released by the shared task organizers, retrieved by BM25 text matching per image to provide in-domain captioning style. Since the competition only provides Guaraní labels, we augmented the dataset with parallel English captions generated by Claude, which serve as the BM25 query surface for retrieval.

4. **CUL**: a curated Guaraní cultural knowledge base consisting of 22 thematic sections written in Spanish, covering food, drink, dress, architecture, household objects, flora and fauna, festivals, children's games, occupations, crafts, music, religion, mythology, and idiomatic Guaraní caption style. Each section is oriented toward visually identifiable entities and includes Guaraní vocabulary, "visual cue" sub-entries, and an explicit anti-stereotype list. The base file was compiled by the authors from Paraguayan web sources, then expanded via a two-step LLM-assisted process described in Appendix A. Since CUL is small enough to fit in the context window, it is always injected in full into the system prompt; BM25 retrieval is applied additionally to surface the most image-relevant sections.

To be specific, the retrieval works as follows: we first build a BM25 index over either grammar sections for CUL, which is organized by section, or paired entries for DAMPY, FLORES, and DEV. Then, depending on the dataset, we ask a language model (Gemma 4 31B) to first describe the image in the relevant language, using English for FLORES and Spanish for DAMPY, CUL, and DEV. We then use BM25 to retrieve the corresponding relevant Guaraní examples.

For the visual few-shot run (v7), instead of injecting text-retrieved DAMPY entries, DAMPY images and their Guaraní captions are injected as multimodal few-shot examples in the prompt. In the submitted run, these are selected statically (the first 10 images from the dataset); BM25-driven visual retrieval, where exemplars would be dynamically matched per test image, was not implemented in time (see Appendix B).

Prompt structure. Each request follows a four-step process:

1. **System prompt:** Contains the captioning guidelines, fixed grammar resources, and dynamically retrieved context blocks from CUL, DAMPY, FLORES, and the dev captions.

2. **Multimodal few-shot prompting:** Optional multimodal few shot examples are injected depending on the retrieval setting. For visual few shot runs, the retrieved DAMPY image caption pairs are inserted here as multimodal demonstrations.

3. **Test image:** The query image is provided to the model.

4. **Generation:** The model generates a caption in Guaraní.

Configurations. Table 1 summarizes the nine submitted system configurations evaluated in the shared task. Runs v0-v5 and v8 use text+RAG mode, where examples from all four pools are dynamically selected using retrieval-augmented generation (RAG) based on similarity scores. Run v6 instead uses text+static mode, where a fixed set

Ver	Mode	DMP	FLR	DEV	CUL
v0	text+RAG	5	100	5	5
v1	text+RAG	10	100	10	5
v2	text+RAG	15	100	10	5
v3	text+RAG	20	100	10	6
v4	text+RAG	15	120	10	8
v5	text+RAG	15	150	10	10
v6	text+static	3	30	0	0
v7	visual+static	10	30	0	0
v8	text+RAG	10	100	10	5

Table 1: Nine submission configurations. **DMP**: DAMPY shots. **FLR**: FLORES shots. **DEV**: dev-set shots. **CUL**: top-k CUL sections additionally surfaced by BM25 (full CUL is always injected). In **RAG** mode, shots are selected by retrieval score; in **static** mode, the first n examples from the dataset are used. In **visual** mode, DAMPY shots are injected as multimodal image–caption pairs; in **text** mode, they are injected as text only.

Ver	Mode	chrF++
v0	text+RAG	22.40
v1	text+RAG	22.02
v2	text+RAG	21.59
v3	text+RAG	21.90
v4	text+RAG	21.44
v5	text+RAG	20.43
v6	text+static	22.49
v7	visual+static	24.17
v8	text+RAG	22.02
baseline	—	20.82

Table 2: Nine submission configurations and their dev-set chrF++ scores, alongside the official baseline for reference.

of examples is selected statically from DAMPY and FLORES without retrieval. Run v7 uses visual+static mode, replacing DAMPY text examples with visual shots and using the first 10 images from the dataset as fixed demonstrations. Across all runs, the grammar book (157 lines) and interlinear glossed text (144 blocks) are always included as fixed context.

3 Results

3.1 Dev-set Results

Table 2 shows the results for each configuration on the dev set. The dev set contains only 50 image-caption pairs, which creates a tension in evaluation: using dev-set captions as few-shot examples (as in the full pipeline) would contaminate the evaluation, so DEV shots are excluded when scoring on the dev set.

Team	Ver	chrF++	CIDEr
IUHoosiers	v4	**24.67**	0.149
IUHoosiers	v5	24.42	**0.167**
IUHoosiers	v1	24.41	0.135
IUHoosiers	v8	24.41	0.135
IUHoosiers	v2	24.41	0.143
IUHoosiers	v0	24.39	0.124
IUHoosiers	v3	24.16	0.128
IUHoosiers	v6	24.04	0.112
IUHoosiers	v7	22.43	0.074
gators	v0	23.10	0.124
baseline	v0	20.14	0.005
Mila	v1	19.77	0.031
usp	v0	19.73	0.024
NAIST	v0	19.41	0.046

Table 3: Official test-set leaderboard (selected entries). Eight of nine IUHoosiers submissions outperformed all other teams. v5 achieved the highest CIDEr despite not having the best chrF++.

3.2 Test-Set Results

Table 3 shows official test-set chrF++ and CIDEr for all IUHoosiers submissions alongside competing teams. The best submission (v4) achieves 24.67 chrF++ and 0.149 CIDEr, a 4.53-point gain over the official baseline (20.14) and a 1.57-point margin over the next-best team (gators, 23.10). IUHoosiers placed first for Guaraní in human evaluation with a mean annotator rating of 3.45/5, versus 3.30 for the second-ranked team; eight of our nine submissions outscore every other team's best result.

4 Analysis

4.1 Qualitative Analysis

Submission selection. We selected these nine configurations from a larger candidate pool, with final choices informed by chrF++ scores alongside our personal judgement and our team's working knowledge in Guaraní and Jopara. While chrF++ was a useful signal for tracking progress during the pilot phase, final selection was not driven by it alone, as the metric measures character n-gram overlap against a single gold reference and is therefore blind to semantic similarity and cultural correctness. A caption that correctly identifies a cultural item but uses different surface forms than the reference can score no better than one that gets the cultural content wrong while sharing incidental n-grams. Instead, selection applied two criteria. First, *output style*: captions should match the concise, visually specific register of the dev-set gold labels and avoid formulaic boilerplate appended regardless of image content. Second, *cultural accuracy*:

captions should correctly identify culturally specific items, especially when the discriminating cue is visual and subtle. The BM25 retrieval pipeline supports both: FLORES parallel context shapes linguistic register and fluency, while DAMPY retrieval anchor culturally grounded vocabulary.

Figure 2: Test image grn_052. v4 (text+RAG) misidentifies the drink as *tereré* and offers a generic cultural gloss, while v7 (visual few-shot) correctly identifies *mate* and provides richer cultural context as this version situates it as a morning drink shared in the cool breeze and reflecting its role in Paraguayan relational life.

[wrong] v4 (text+RAG)

Tereré ekípo oĩva poyvi isañyju ári: guámpa yvyrágui, vombílla ha término. Ko tembi'u ha y'u ha'e hína ñande reko imbaretevéva ha omojoaju ñande reko Paraguáipe.

[correct] v7 (visual few-shot)

Mate ha guampakuéra oĩva peteĩ pano pytã ha hovy ári. Ko tembi'u ha imombe'upy oñembopyre hína ñane reko Paraguáipe, oje'u pyhareve yvytu ro'ysã jave.

Case study: mate vs. tereré. One notably culturally specific image among the first ten test examples is **grn_052**, depicting a *mate* setup (gourd, bombilla, *pava* kettle). Most text+RAG configurations label it *tereré*; only v2 primarily identifies it as *mate*, while v0 mentions both drinks. The text+static run (v6) and the visual few-shot run (v7) both correctly identify *mate*, but differ in descriptive accuracy. v6 lists the objects present and appends a generic cultural phrase (*"Mokõi guampa ha peteĩ pava oĩ peteĩ ao póipe. Ko'ãva ojeipuru mate pyahúpe, ñane rembi'u ha y'u rekojera. Ohechauka ñane ñemoirũ ha ñane tekopytã paraguáigua."* — "Two guampa and a pava are in a pocket; these are used in new mates, our food and drink recipes; it shows our solidarity and our Paraguayan culture"), while v7 visually grounds the scene (noting the red and blue cloth), provides the correct ideal weather setting (a morning drink taken in the cool breeze), and articulates its social significance (reflecting relationships and the surrounding environment). That is to say, v7 shows more depth in understanding than v6.

Text runs also tend to add repeated generic closing phrases (e.g., *"Ko tembi'u ha'e peteĩ rembiapokue tee Paraguáigui"*, "this dish is a true cultural product of Paraguay") regardless of image content, while gold captions use more varied, image-specific language. The visual run produces more image-specific descriptions, which explains their higher qualitative character despite lower chrF++ on the test set. Overall, the human evaluation's preference for IUHoosiers over competitors reflects the generally higher Guaraní output quality compared to translation-based baselines.

5 Conclusion

IUHoosiers achieved first place for Guaraní at AmericasNLP 2026 using Gemma 4 31B with BM25 augmented retrieval from four Guaraní knowledge sources. Beyond the result itself, the most practically useful contribution is the pipeline design: language specific knowledge injection at inference time without fine tuning rather than any particular hyperparameter choice. The broader takeaway is that carefully curated grammar and cultural resources, injected at inference time, can remain competitive without parameter updates. That is the durable result worth carrying forward. Visual few shot retrieval remains a promising but underexplored direction: dev set calibration (Appendix B) showed visual shots outperforming text injection by 1.41 chrF++ points, and BM25 driven visual retrieval, where exemplars are dynamically matched to each test image, was not implemented in time for submission and warrants future investigation.

Acknowledgments

We thank Indiana University Research Technologies for REALLMS API access and the BigRed 200 HPC cluster, and the AmericasNLP 2026 organizers for the dataset and evaluation infrastructure.

References

Seth Aycock, David Stap, Di Wu, Christof Monz, and Khalil Sima'an. 2025. Can LLMs really learn to translate a low-resource language from one grammar book? arXiv preprint arXiv:2409.19151.

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, Santiago Castro, Mariella Cardenas, Gustavo Gimenez González, Yliana Gimenez, and Dina Wonsever. 2020. Development of a Guaraní-Spanish parallel corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 893–898, Marseille, France. European Language Resources Association.

Bruno Estigarribia. 2020. *A Grammar of Paraguayan Guaraní*. UCL Press.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Google DeepMind. 2025. Gemma 4 technical report. https://blog.google/technology/google-deepmind/gemma-4/.

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. *Preprint*, arXiv:2106.09685.

Yuchen Huang, Zhiyuan Fan, Zhitao He, Sandeep Polisetty, Wenyan Li, and Yi R. Fung. 2025. CultureCLIP: Empowering CLIP with cultural awareness through synthetic images and contextualized captions. *Preprint*, arXiv:2507.06210.

Yue Li, Zhixue Zhao, and Carolina Scarton. 2025. It's all about in-context learning! teaching extremely low-resource languages to LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29544–29559, Suzhou, China. Association for Computational Linguistics.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. PEFT: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2023. A benchmark for learning to translate a new language from one grammar book. arXiv preprint arXiv:2309.16575.

Chen Zhang, Jiuheng Lin, Xiao Liu, Zekai Zhang, and Yansong Feng. 2025. Read it in two steps: Translating extremely low-resource languages with code-augmented grammar books. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3977–3997, Vienna, Austria. Association for Computational Linguistics.

A Construction of the CUL Cultural Knowledge Base

The CUL resource was built in three LLM-assisted steps using Claude Opus 4.

Step 1 — Seed file. The authors compiled a 16-section seed knowledge file in Spanish, drawing on their own working knowledge of Guaraní culture, Wikipedia, and the *Every Culture* encyclopedia entry on Guaranís. The seed covered cosmovision, food, drink, traditional medicine, flora and fauna, music, religion, and mythology, but lacked visual grounding, ethnic distinctions between mestizo Paraguayan and Indigenous Guaraní communities, and Guaraní vocabulary for common visual entities.

Step 2 — Prompt generation. We first prompted Claude Opus 4 to design a specialized research prompt for expanding the seed file, targeting gaps most relevant to image captioning: traditional dress, architecture, festivals, children's games, household utensils, and a Guaraní visual vocabulary. The generated prompt instructed the model to (1) audit the existing file for visual grounding and accuracy, (2) search authoritative Paraguayan sources to identify gaps, (3) draft additions in the same Spanish-language bullet-point style as the seed, and (4) consolidate everything into a single deliverable.

Step 3 — Expansion. In a separate Claude Opus 4 session, the generated prompt was supplied together with the seed file. The model queried sources including ABC Color, Portal Guarani, Última Hora, IWGIA, and Tierraviva, and produced the consolidated CUL file containing 22 thematic sections with explicit visual-grounding cues throughout. An anti-stereotype section warns the model against projecting folkloric cues onto ordinary modern Paraguayan scenes. The final document is included in the project code repository.

B Visual Few-Shot: Dev-Set Calibration and Future Directions

Table 4 shows chrF++ from pre-submission calibration experiments on the 50-image dev set using *static injection* (fixed FLORES sentence counts and DAMPY shot counts). These experiments explored a wider parameter range than the final v0–v8 submissions.

Mode	Config (FLR / DMP)	chrF++
visual	15 shots / FLR=50	**24.43**
visual	10 shots / FLR=30	24.17
visual	5 shots / FLR=30	23.49
text	FLR=100, DMP=5	**23.02**
text	FLR= 30, DMP=3	22.49
text	FLR= 80, DMP=3	22.21
text	FLR= 50, DMP=5	22.09
text	FLR= 20, DMP=2	21.80
text	FLR= 50, DMP=3	21.43

Table 4: Dev-set calibration results (static injection, 50 images). **FLR**: FLORES sentences injected. **DMP**: DAMPY shot count. For visual runs, DAMPY images are injected as few-shot context. The 10-shot/FLR=30 visual configuration corresponds to submitted run v7.

Visual few-shot consistently outperformed text injection across all shot counts on the dev set. We see that the best visual configuration (15 shots, FLR=50) achieved 24.43 chrF++, a 1.41-point lead over the best text configuration (FLR=100, DMP=5: 23.02). These findings motivated including visual few-shot among the nine official submissions.

Two stronger visual directions were not pursued due to time constraints. First, the 15-shot/FLR=50 configuration (24.43 on dev) was not submitted; the submitted visual run (v7, 10 shots/FLR=30) achieved 24.17 on dev and 22.43 on the test set. Second, BM25-driven visual retrieval, which dynamically selects the most image-relevant DAMPY exemplars per test image rather than using static selection, was not implemented.

The submitted v7 trailed text BM25 (v4) by 2.24 chrF++ on the test set (22.43 vs. 24.67). We hypothesize three contributing factors. First, the 50-image dev set is small, so the best static shot count may overfit to it. Second, the 111-entry DAMPY pool covers only food, flora, and fauna; when a test image falls outside this distribution, static visual selection may return misleading exemplars. Third, broad FLORES context (k=120/150) generalizes across diverse test topics in a way that static visual selection cannot when no relevant DAMPY entry exists. BM25-driven visual retrieval would address all three by dynamically matching exemplars to each image, and is a natural direction for future work.