Findings of the AmericasNLP 2026 Shared Task on Cultural Image Captioning for Indigenous Languages

Minh Duc Bui[1] **David Guzmán**[2] **Abteen Ebrahimi**[3] **Franklin Morales**[4,5] **Marvin Agüero-Torales**[6,7,8] **Raquel Insfrán**[6,9] **Cecilia González**[6,9] **Ramón Araujo**[6,9] **Luca Cernuzzi**[6,9] **Carlos Raul Noh Chi**[24] **Carlos Eduardo Tec Cahun**[24] **Sindi Estrella Poot Cohuo**[24] **Daniel Ricardo Benítez Chi**[24] **Santos Natanael Palomo Arévalo**[24] **Jessica Elizabeth Canul Canche**[24] **Deysi Aracely Poot Poot**[24] **Wendy Marleny Dzib Dzib**[24] **Eduardo José Ake Pool**[24] **Reynaldo Alexander Couoh Martin**[24] **Silvia Fernandez Sabido**[25] **Luis Samuel Santiago Melchor**[11] **Sotero Silverio**[4] **Robert Pugh**[12,13] **Raúl Vázquez**[14] **John E. Ortega**[15] **Arturo Oncevay**[16] **Rubén Manrique**[17] **Luis Chiruzzo**[18] **Rolando Coto-Solano**[19] **Elisabeth Mager**[20] **Shruti Rijhwani**[21] **David Ifeoluwa Adelani**[2,22] **Manuel Mager**[23] **Katharina von der Wense**[3,1]

[1]Johannes Gutenberg University Mainz [2]Mila-Quebec AI Institute, McGill University [3]University of Colorado Boulder [4]Independent Researcher [5]Mẽkichawak [6]Centro Tecnológico en Ingeniería (CIDIT), Paraguay [7]Universidad de Granada, Spain [8]Fujitsu, Spain [9]Universidad Católica Nuestra Señora de la Asunción, Paraguay [11]Ximomachti [12]Indiana University [13]Mozilla Data Collective [14]University of Helsinki [15]Northeastern University [16]Pontificia Universidad Católica del Perú [17]Universidad de Los Andes [18]Universidad de la República, Uruguay [19]Dartmouth College [20]Universidad Nacional Autónoma de México [21]Google DeepMind [22]Canada CIFAR AI Chair [23]Universidad Iberoamericana, Mexico [24]Universidad Intercultural Maya de Quintana Roo, Mexico [25]CentroGeo, Mexico

Abstract

Indigenous languages of the Americas face severe endangerment, and the scarcity of culturally grounded resources remains a critical barrier to revitalization efforts. We present the AmericasNLP 2026 Shared Task on Cultural Image Captioning for Indigenous Languages, the first shared task dedicated to generating captions for images depicting Indigenous cultures of the Americas, written in the Indigenous languages themselves. To support this, we introduce and publicly release a newly constructed dataset spanning five cultures and their dominant languages: Bribri, Guaraní, Yucatec Maya, Central Veracruz Nahuatl, and Wixárika. Evaluation follows a two-stage process, combining automatic evaluation using ChrF++ with human evaluation of the top-performing systems for each language. Eight teams participate, submitting 27 systems in total. Results indicate that the task remains largely unsolved: while the strongest systems produce understandable captions, they fall short on descriptive detail and, critically, cultural grounding.

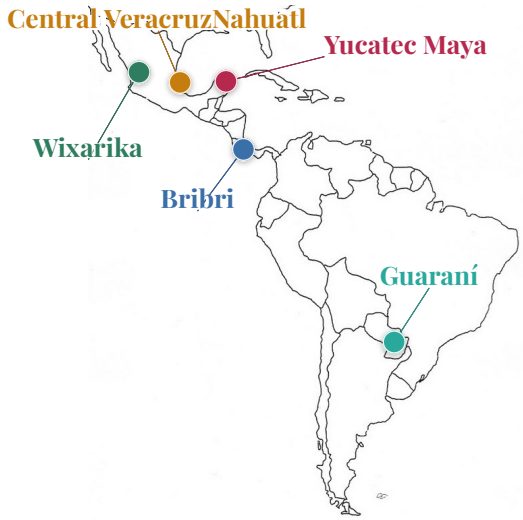

Figure 1: Approximate geographic distribution of the five Indigenous cultures covered in this work

1 Introduction

Many Indigenous languages of the Americas are endangered, spoken by small communities and at high risk of extinction. Revitalization depends on teaching materials that are costly and slow to produce (Chiruzzo et al., 2024). Image captioning is unusually well-suited to address this gap: pairing a culturally specific image with a caption in the target language teaches language and culture at once, since the learner takes in not only the words but the practices, objects, and settings they name. Yet producing accurate captions is hard on two fronts. First, even in the text-only setting, the most novel

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 279–295

Bribri
Ù tsg̱bita g̱ tg̱ idir yirîrî. E' yawèke ùkö wa, tsi̱ne̱t a̱ñì ǿmi̱k.
Palm segments are woven together at the top of the conical house. The suita palm leaves are woven really tight together.

Guaraní
Peteĩ kuimba'e oguata hína tatapỹi ári; "Tata ári jehasa" niko peteĩ ritúal [...].
A man walking barefoot on burning embers; "Tata ári jehasa" is a ritual [...].

Yucatec Maya
Junkúul puut jach ya'ab u yich beyxan yaan tu paachil ya'abach u le' ja'as.
A papaya tree with many fruits and many banana leaves behind it.

Central Veracruz Nahuatl
Se siuatsintli ika itlaken xochijyo itlanpa se aueuetl uejkapantik.
A girl in an embroidered dress under a huge ahuehuete tree.

Wixárika
Ukáa tú mati tatsiu h+xi in+net+ hipame yu iwama wa há mat+a.
The woman is harvesting green tomatillos with several colleagues.

Figure 2: Representative examples from each language. Each image is shown with the sentence in the target language (italicized) and its English translation. Note that the English translation is not released.

NLP techniques still struggle with low-resource languages (Mager et al., 2024; Weerasinghe et al., 2025; Hettiarachchi et al., 2025); adding the visual, multimodal dimension only raises the bar. Second, accurate captions demand not just linguistic competence but cultural knowledge—models frequently exhibit substantial limitations in visual cultural understanding, defaulting to Western-centric depictions and interpretations (Nayak et al., 2024; Winata et al., 2025; Liu et al., 2025; Bui et al., 2025). Both challenges are thus barriers to building systems that serve Indigenous communities.

We introduce the AmericasNLP Shared Task on Cultural Image Captioning for Indigenous Languages, the first effort to develop systems that generate accurate, culturally grounded captions for culturally relevant images, written in the Indigenous languages themselves. To make this possible, we contribute a new dataset of images and captions from five cultures and their dominant languages: Bribri, Guaraní, Yucatec Maya, Central Veracruz Nahuatl, and Wixárika; see Figure 1. Evaluation goes beyond automatic metrics: we complement them with a human evaluation to ensure that the cultural and linguistic quality of generated captions is assessed rigorously.

Eight teams participate, submitting 27 systems. Our results show that the task remains largely unsolved: while the strongest systems produce understandable captions, they fall short on detail and, critically, cultural grounding. These findings highlight the need for continued investment in culturally aware, multimodal NLP for Indigenous languages.

The remainder of the paper covers the dataset creation (Section 2), our evaluation process (Section 3.1), the submitted systems (Section 3.2), as well as results and additional insights (Section 4 and Section 5).

2 Dataset Creation

We describe the construction of our dataset, covering the cultures and languages represented, the image collection and caption annotation procedures, data collector recruitment, and dataset statistics. Representative examples are shown in Figure 2. We publicly release the development data in our Github repository[1] under the Creative Commons NonCommercial license (CC NC-BY).[2]

2.1 Cultures and Languages

The shared task features 5 cultures and their dominant languages: Bribri, Guaraní, Yucatec Maya, Central Veracruz Nahuatl, and Wixárika.

Bribri The Bribri are an Indigenous people of southern Costa Rica. They speak Bribri (BZD), a Chibchan language used by an estimated 7,000 people (INEC, 2011). It is a vulnerable language (Sánchez Avendaño, 2013), in that increasingly fewer children speak the language. The language has been documented through dictionaries (Margery, 2005; Krohn, 2020), grammars (Jara Murillo, 2018a), collections of oral literature (Jara Murillo, 2018b; García Segura, 2016; Constenla, 1996, 2006), digital corpora

[1] https://github.com/AmericasNLP/americasnlp2026

[2] https://creativecommons.org/licenses/by-nc/4.0/

Split	Metric	BZD	GRN	YUA	NLV	HCH
Dev (Train)	Instances	50	50	50	50	70
	Avg. Words	$14.9_{\pm 5.0}$	$21.6_{\pm 7.7}$	$11.3_{\pm 5.5}$	$8.3_{\pm 4.4}$	$16.0_{\pm 5.5}$
	Avg. Characters	$93.3_{\pm 30.7}$	$154.6_{\pm 51.1}$	$71.1_{\pm 40.9}$	$61.0_{\pm 31.3}$	$92.7_{\pm 28.1}$
Test	Instances	267	101	212	200	201
	Avg. Words	$13.4_{\pm 4.1}$	$21.9_{\pm 7.3}$	$15.2_{\pm 9.1}$	$5.9_{\pm 3.1}$	$12.0_{\pm 4.3}$
	Avg. Characters	$83.1_{\pm 22.3}$	$146.1_{\pm 48.9}$	$95.4_{\pm 59.4}$	$43.2_{\pm 22.5}$	$69.8_{\pm 22.7}$

Table 1: Dataset statistics for captions across our five Indigenous languages. Note that the dev split was originally intended solely for evaluation, but subsequently released for training as the task proved challenging without any in-language resources.

(Flores Solórzano, 2017), and learning materials for adults (Constenla et al., 2004; Jara Murillo and García Segura, 2013) as well as children (Sánchez Avendaño, 2020).

Guaraní The Guaraní are an Indigenous people of South America, primarily associated with Paraguay but also present in parts of Bolivia, Argentina, and Brazil. They speak Guaraní (GRN), a Tupi–Guaraní language with approximately 6.5 million speakers. Guaraní is a co-official language of Paraguay and is spoken by the vast majority of the population throughout the country, not being confined only to certain regions or social groups. This has resulted in many varieties of the language with different levels of code-switching and borrowing from Spanish, Portuguese, and other European languages.

Yucatec Maya The Yucatec Maya are an Indigenous people of the Yucatán Peninsula of Mexico, northern Belize, and parts of Guatemala. They speak Yucatec Maya (YUA), a Mayan language with approximately 800,000 speakers (INEGI, 2020). The Maya civilization of the Yucatán Peninsula represents a living ancestral legacy whose worldview deeply integrates nature, mathematical knowledge, and a cyclical conception of time. Far from being reduced to an archaeological past, contemporary Maya culture remains fully active within the daily practices and community dynamics of the region, where the language acts as the backbone of cultural identity and collective memory, as well as the social fabric. Essential architectural and environmental elements, such as the traditional Maya house, valued as a symbolic and sacred space, along with detailed linguistic expressions used to describe the surroundings and emotional states, reflect a unique way of inhabiting and understanding the world (Universidad Autónoma de Yucatán, n.d.).

Thus, far from being a static system, Yucatec Maya (*maayat'aan*) operates as a dynamic vehicle through which communities conceptualize their current environment, traditional agricultural knowledge, and everyday experiences. Nevertheless, the transition of this language toward modernity demands going beyond mere patrimonialization or discursive recognition; it requires a critical and effective exercise of refunctionalization that grants it shared practical and technological spaces, thereby guaranteeing its vitality in the face of contemporary challenges (Briceño Chel, 2021).

Central Veracruz Nahuatl Central Veracruz Nahuatl (alternatively Orizaba or Zongolica Nahuatl, *Náhuatl central de Veracruz*, ISO 639-3 NLV) is one of approximately 30 formally recognized varieties of Nahuatl, a Uto-Aztecan language (Valiñas Coalla, 2020). Many aspects of this language, spoken in several municipalities in the state of Veracruz, Mexico, in and around Orizaba, have been discussed in linguistics research (Goller et al., 1974; Tuggy, 1992, 1998). The Nahuas, the Indigenous ethnic group associated with the Nahuatl language, have diverse cultural practices that reflect their wide geographic distribution throughout Mexico. In Veracruz, many Nahuas continue to dress in traditional clothing and cultivate *milpa* in the traditional Mesoamerican manner, in both rocky/mountainous and marshy terrain.

Wixárika The Wixáritari (or Huichol) are an Indigenous people of Mexico. They live in the mountainous regions between the states of Jalisco, Nayarit, Durango, and Zacatecas (Gómez, 1999). They speak Wixárika (HCH), a polysynthetic language belonging to the Uto-Aztecan language family, spoken by approximately 50,000 people (INEGI, 2020). This ethnic group possesses a strong cultural identity, which is expressed through their clothing, speech, and ceremonies, such as the drum,

deer, and toasted corn ceremonies, among many others (Anguiano, 1978; Neurath, 2002). Their spiritual strength resides in the trinity of the peyote (*hik+ri*), deer (*kayumari*), and maize (esquite). The Wixaritari "hunt" the *hik+ri* during their pilgrimage to Wirikuta (Lumholtz, 1902), near the town of Real de Catorce in the state of San Luis Potosí (Martínez, 2006). From this plant, the *marakate* receive the spiritual strength to guide the Wixárika people and perform healings. Their primary economic activities are agriculture—involving seasonal migration to agricultural fields in Mexico's coastal regions—and the creation of handicrafts (Zingg, 1982).

2.2 Images

This section describes the image collection process, including general guidelines shared across all languages and culture-specific details for each community.

General Instruction For each language and culture pair, we recruit members of the respective Indigenous communities for data creation and annotation (see Section 2.4). We inform them about the goals, methodology, and reasons why the data collection is done and ask them to photograph everyday life in their communities, spanning a broad range of domains—food, work, ceremony, and nature. This community-driven approach ensures that the images themselves are culturally authentic, reflecting the lived experience of each community. We ask for pictures where humans are absent or not recognizable, to maintain the privacy of the community members. In the cases where people are shown, we either anonymize the images removing faces, or get permissions from the individuals to be part of the dataset.

Culture-Specific Details For Bribri, photographs are taken by cellphones or sourced from publicly available collections (see Appendix B.1). Images focus on important elements of Bribri culture (e.g., agriculture, architecture, crafts and material culture, and Talamancan landscapes) while also depicting aspects of Indigenous life elsewhere in Costa Rica and archaeological objects.

For Guaraní, images are drawn from the daily life of community members, supplemented by open-source material from the web. As the language is spoken throughout all society in Paraguay, the images present a mix of indigenous and criollo cultures, including food, activities, flora, fauna, clothing, and some archaeological items.

For Yucatec Maya, community members photograph everyday elements and activities, spanning local food, domestic and *milpa* agriculture, transportation, regional buildings, endemic plants, and campus life. Data collectors show a shared interest in documenting traditional and identity-defining aspects of their communities, such as hammock use, flowers along pathways, domestic animals, and local cuisine.

For Central Veracruz Nahuatl, a community member travels to multiple neighboring Nahuatl-speaking communities and collaborates with local contributors to photograph daily life. Images are captured using a Samsung Galaxy A53 smartphone.

For Wixárika, the collection reflects day labor, life and nature in the sierra, and elements of everyday community life.

2.3 Captions

For all images, we ask the same members of the respective Indigenous communities to provide captions in their Indigenous languages. We additionally collect Spanish translations to make the information about the images easier accessible. However, we do not pass the Spanish captions on to the shared task participants. Annotators are provided with annotation guidelines and an illustrative example (see Appendix B.2 for the full guidelines). The guidelines encourage annotators to produce *culturally enriched captions* where appropriate: rather than only describing the most salient object, captions should elaborate on the function, purpose, or significance of objects, clothing, gestures, or settings—but only where such elaborations are grounded in what is visually present in the image.

2.4 Recruitment

The dataset is constructed with active involvement of Indigenous community members, who contribute both images and captions. We briefly describe each group below.

Bribri The Bribri annotations come from an L1 speaker of the language, who has worked as a school teacher in the community.

Guaraní The annotators of the Guarani dataset are four fluent native speakers who also participate in other NLP projects. They are two women and two men, and their ages range between 24 and 39.

Yucatec Maya For the construction and evaluation of the Yucatec Maya corpus, nine native-speaking annotators are recruited from various towns across the Maya region of Quintana Roo. The team comprised both current students and alumni from the Language and Culture, as well as the Information and Communication Technologies programs at the Universidad Intercultural Maya de Quintana Roo (UIMQROO), with the assistance of relatives and neighbors.

Central Veracruz Nahuatl The Nahuatl dataset collection is organized and carried out largely by a Nahuatl teacher and translator in the area of Rafael Delgado, who has worked extensively on pedagogical material, cultural communication, and linguistic resources. He travels to multiple neighboring Nahuatl-speaking communities and works with other local collaborators.

Wixárika The dataset is created with the help of the Zoquipan community members. All pictures are either collected by the authors or by the recruited community members.

2.5 Dataset Statistics

We report the dataset statistics in Table 1. Each language has 50 instances in the development split—originally intended solely for evaluation, but subsequently released for training and inference as the task proved challenging without any in-language resources—except for HCH, which has 70 as it served as our pilot language. Test sets are considerably larger, ranging from 101 instances for GRN to 267 for BZD. Sentence length varies across languages: GRN exhibits the longest sequences on average (21.6 words in training, 21.9 in test), while NLV is the most concise (8.3 and 5.9 words, respectively).

3 Experimental Setup

3.1 Evaluation Process

We conduct a two-step evaluation process: (1) automatic evaluation and (2) human judgment of the top-5 systems per language.

Automatic Evaluation We use chrF++ (Popović, 2017) to automatically evaluate generated captions against our ground truth references. The top-5 systems for each language then proceed to the second stage, the human evaluation.

Score	Description
5	Fluent, natural, and culturally accurate. No significant errors.
4	Well-written with minor flaws in detail or cultural vocabulary.
3	Understandable, but inaccurate, incomplete, or too vague.
2	Serious grammatical errors; description mostly inaccurate.
1	Wrong language, incomprehensible, or unrelated to the image.

Table 2: Human evaluation scoring rubric (shortened, see full description in Appendix C.1). We capture two dimensions with the score: (1) *language quality* and (2) *fidelity to the image and correct use of cultural terminology*.

LANGUAGE	NUMBER OF RATINGS	IMAGES
BZD	320	267
GRN	228	101
YUA	212	212
NLV	200	200
HCH	201	201

Table 3: Number of ratings and images per language used in human evaluation.

Human Evaluation Automatic metrics such as chrF++ cannot capture the multifaceted ways in which an image can be described. Moreover, since our annotations contain only a single reference caption per image, human judgment provides a more robust and nuanced assessment of system outputs.

To keep the annotation process lightweight, we design a 1–5 rating scale intended to capture two dimensions: (1) *language quality* and (2) *fidelity to the image and correct use of cultural terminology*. Our scoring rubric is presented in Table 2. Prior to annotation, each annotator is shown a calibration example containing five captions representative of each score level. When multiple annotators rate the same example, their scores are averaged per example before computing the overall mean. All test-set samples receive at least one annotation. We report the number of ratings per language in Table 3. We further report the full guidelines with details about the annotators in C.1.

To facilitate annotation, we develop a platform which presents annotators with an image alongside five captions submitted by the top-performing teams, displayed in randomized order to avoid position bias. The reference caption is also provided to assist annotators in their judgment. Figure 7 presents a screenshot of the annotation interface

through which evaluators rate the generated descriptions.

Overall Winner Points are awarded based on each team's rank in the human evaluation for each language: 1st place receives 5 points, 2nd place 4 points, 3rd place 3 points, 4th place 2 points, 5th place 1 point, and teams not selected for human evaluation receive 0 points. A team's total score is the sum of their points earned across all five languages.

3.2 Baseline

We provide participants with a baseline that follows a two-stage *generate-then-translate* pipeline: a vision–language model (VLM) first produces a caption in Spanish, which is then translated into the target Indigenous language. We adopt Spanish as a pivot language because the machine translation resources available for Indigenous languages of the Americas are predominantly paired with Spanish; generating first in a high-resource language lets the baseline leverage existing translation systems.

Stage 1: Captioning in Spanish We use Qwen3-VL-8B-Instruct (Bai et al., 2025a) to generate a Spanish caption for each image. The model is conditioned on a culturally-informed system prompt—drawn from publicly available encyclopedic sources—that is specific to each culture. The prompt instructs the model to first describe what is visually present, add only essential and visually grounded cultural context, include target-language terms where possible, and keep the caption concise.

Stage 2: Translation into the Target Language The Spanish caption is translated into the target language using the winning system (Gow-Smith and Sánchez Villegas, 2023) from the Americas-NLP 2023 Shared Task on Machine Translation into Indigenous Languages (Ebrahimi et al., 2023). As this system does not cover Yucatec Maya, we report no baseline for that language.

3.3 Submitted Systems

We summarize each participating team's approach below. All results are shown in Tables 4, 5 and 6.

Gators (Dhawan et al., 2026) uses a two-stage retrieval-augmented translation pipeline. A VLM, either Qwen2.5-VL-72B (Bai et al., 2025b) or Qwen3-VL-8B (Bai et al., 2025a), generates a Spanish caption, which is then translated by Gemini 2.5 Flash (Comanici et al., 2025) using retrieval-augmented many-shot in-context prompting: BM25 retrieves similar Spanish–target pairs from per-language parallel banks and supplies them as in-context examples alongside development examples, with the number of retrieved and development examples tuned per language.

Mila (Lara and Raval, 2026) post-trains Aya Vision 32B (Dash et al., 2025) in multiple stages: supervised fine-tuning on Spanish–Indigenous-language machine translation, optional reinforcement learning with verifiable rewards, and a final fine-tuning stage on image captioning, such that the model generates captions directly in the target language rather than via a Spanish pivot. They additionally submit a zero-shot GPT-5.5 (OpenAI, 2026) direct-captioning system.

IUHoosiers (Shi et al., 2026) submits for Guaraní only, using inference-time knowledge injection, without any fine-tuning. For each image, Gemma 4 31B (Farabet and Lacombe, 2026) produces a description that is used as a BM25 (Robertson and Zaragoza, 2009) query over four Guaraní knowledge sources. The retrieved items, together with a fixed grammar-book excerpt and interlinear-glossed examples, are injected into the prompt to generate the caption in a single pass.

6fanle (Wang and Yang, 2026) submits for Wixárika only, using a modular Spanish-pivot pipeline: CLIP (Radford et al., 2021) retrieves visually similar images to provide grounding examples, Qwen3-VL-8B-Instruct (Bai et al., 2025a) generates Spanish caption candidates, the Sheffield 2023 MT system (Gow-Smith and Sánchez Villegas, 2023) translates these candidates into Wixárika, and a character 5-gram language model reranks the translations to select the final caption.

InclusionVLM (Bueno and Garg, 2026) compares two approaches. Their cascaded system pairs a VLM—Gemini 2.5 Flash (Comanici et al., 2025)—, using concise persona-based cultural prompting, with the Sheffield 2023 MT system (Gow-Smith and Sánchez Villegas, 2023) as well as a separate Spanish–Maya model for Yucatec Maya. Their single-stage system adapts PaliGemma 2 3B (Steiner et al., 2024) end-to-end via LoRA fine-tuning (Hu et al., 2021), continued pretraining, and multilingual joint training.

Yaduha (Cuadros et al., 2026) uses a schema-constrained, LLM-assisted rule-based approach in

Team \ Lang.	BZD	GRN	YUA	NLV	HCH
baseline	7.01	20.14	–	9.52	16.91
6fanle	–	–	–	–	19.16†
IUHoosiers	–	**24.67**†	–	–	–
InclusionVLM	7.94	16.48	16.97†	14.06†	18.37†
Mila	11.73†	19.77†	15.99†	20.66†	19.01†
NAIST	**19.37**†	19.41†	15.80†	20.93†	**19.84**†
gators	17.90†	23.10†	21.11†	**25.42**†	17.58†
usp	10.95†	19.73†	10.83	9.49	13.68
yaduha	10.03†	16.90	**23.41**†	21.00†	15.61

Table 4: Automatic evaluation results for all participating teams. We report the best ChrF++ score per team across all submitted systems. † marks the top-5 teams per language, as those are selected for human evaluation.

which the VLM never emits target-language text directly. For each language, a coding agent—Claude Opus 4.7 (Anthropic, 2026)—authors a *language package*—a Python module with a closed vocabulary, Pydantic sentence schemas, and a deterministic renderer—based on the development split and public linguistic references. At inference, a VLM—GPT-5 (Singh et al., 2026)—sees the image and schema and emits a structured representation under constrained decoding, which the renderer converts into the surface caption.

USP (Fernandes, 2026) uses a two-stage cascade pipeline in which Qwen3-VL-8B-Instruct (Bai et al., 2025a) generates a culturally-prompted Spanish caption and a fine-tuned NLLB-200-distilled-600M (NLLB et al., 2022) model, one per language, translates it into the target language, trained on AmericasNLP 2023 data augmented with public parallel corpora. The team documents a failure mode in which NLLB-200 lacks vocabulary entries for Bribri and Maya and silently produces English output.

NAIST (Vasselli et al., 2026) explores two strategies. Their primary system performs nearest-neighbor retrieval: it embeds the test image with CLIP (Radford et al., 2021), finds the most similar development image and returns its caption directly. Their second system is a generation pipeline that analyzes the scene, grounds the identified concepts in dictionary entries, and retrieves gloss templates alongside interlinear gloss representations to constrain generation in low-resource settings.

Team \ Lang.	BZD	GRN	YUA	NLV	HCH
6fanle	–	–	–	–	2.48
gators	2.758	3.390	3.175	3.375	2.90
IUHoosiers	–	**3.448**	–	–	–
InclusionVLM	–	–	1.108	1.185	2.33
Mila	1.994	1.764	**3.203**	1.560	2.21
NAIST	2.219	1.978	1.934	1.220	**3.79**
usp	1.086	2.410	–	–	–
yaduha	**2.895**	–	2.892	**3.465**	–

Table 5: Aggregated human evaluation scores. Each entry reports the mean rating on a 1–5 scale, where higher scores indicate better language quality and greater image fidelity. Bold indicates the highest score per language. Underline indicates the second highest score per language.

Team \ Lang.	BZD	GRN	YUA	NLV	HCH	Total
gators	4	4	4	4	4	**20**
NAIST	3	2	2	2	**5**	14
yaduha	**5**	0	3	**5**	0	13
Mila	2	1	**5**	3	1	12
IUHoosiers	0	**5**	0	0	0	5
InclusionVLM	0	0	1	1	2	4
usp	1	3	0	0	0	4
6fanle	0	0	0	0	3	3

Table 6: Points per language and total. Bold indicates the highest score per column; underline indicates the second highest.

4 Results

We first report the automatic evaluation, and then human evaluation.

Automatic Evaluation We report the per-language ChrF++ scores for the best systems per team and language in Table 4. Complete results can be found in Appendix D.

Overall, most participating teams surpass the baseline for at least one language. Notably, three teams—Mila, NAIST, and gators—rank in the top-5 across all five languages, reflecting strong and consistent performance. Among these, NAIST achieves the highest scores for BZD (19.37) and HCH (19.84), while Gators leads for NLV (25.42). Outside this group, Yaduha achieves the best score for YUA (23.41), and IUHoosiers comes first for GRN (24.67).

Human Evaluation Table 5 presents the human evaluation results for the top-5 systems for each of the five languages, and Table 6 the derived points used to determine the overall winner.

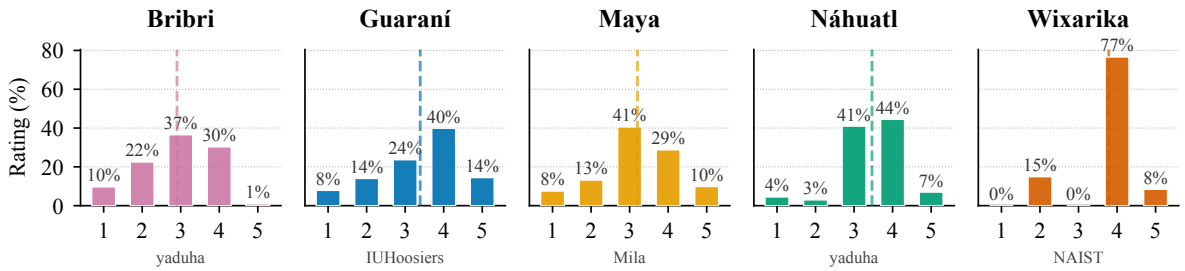

Figure 3: Rating distributions (1–5) of the best-performing system per language based on human evaluation. The dashed vertical line indicates the mean rating for each system.

The gators team demonstrates the most consistent performance, finishing second for every language, which translates into the highest total point score of 20. Per-language winners vary considerably: Yaduha achieves the highest scores for BZD (2.895) and NLV (3.465), Mila leads for YUA (3.203), IUHoosiers for GRN (3.448), and NAIST for HCH (3.79), where the margin over second place is the largest, with 0.89 points. With 14 points, NAIST finishes as runner-up overall.

Discussion Human evaluation scores suggest that the task remains largely unsolved across all five languages. Even the strongest per-language systems score between 2.895 (BZD) and 3.79 (HCH), indicating that outputs are at best understandable but lack the detail and cultural grounding that accurate captioning demands. The gators team, our overall winner, reflects this pattern well: despite finishing second across all five languages, their ratings range only from 2.76 (BZD) to 3.38 (NLV), reflecting consistent but imperfect outputs. Performance varies considerably across languages, though lower scores may reflect more challenging images or greater difficulty of generation in that language for the models, rather than being a property of the language itself: Wixárika (HCH) sees the most reliable system, with NAIST achieving a notably strong score of 3.79, while Bribri (BZD) and Yucatec Maya (YUA) prove most challenging, with lower means.

Taken together, these results indicate that no system achieves robust, culturally grounded captioning across all five languages. The gap between current outputs and fluent, culturally accurate captions underscores both the difficulty of the task and the need for continued investment in culturally aware, multimodal NLP for Indigenous languages.

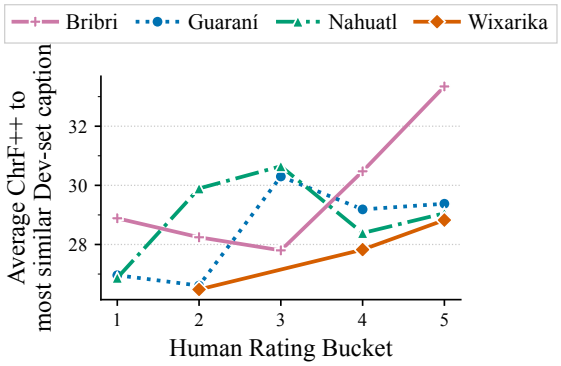

Figure 4: Average similarity between test captions and their most similar development-set captions across human rating buckets. Similarity is measured using ChrF++.

5 Analysis

5.1 Quantitative Analysis

We further provide a quantitative analysis of the outputs of the best performing system per language according to our human evaluation (Table 5).

Rating Distribution Figure 3 shows the score distributions of the best-performing system per language. Wixárika (NAIST) stands out with a strongly peaked distribution at rating 4 (77%) and near-zero low-quality outputs. Note that NAIST performed better than the second team by a substantial amount (3.79 vs. 2.90). Systems for Bribri (Yaduha) and Maya (Mila) exhibit broader distributions centered around rating 3, reflecting more variable output quality. Guaraní (IUHoosiers) and Náhuatl (Yaduha) fall in between, with moderate concentrations at ratings 3–4. Overall, the distributions suggest that caption quality is lower for Bribri and Maya while Wixárika benefits from the most reliable system.

Development–Test Caption Similarity vs. Human Ratings We examine whether human evaluation scores on the test set correlate with the similarity between test captions and captions in the development set. To measure this, we compute ChrF++ between each test caption and the most similar caption from the development set. We exclude Maya (Mila) from this analysis, since their system did not make use of the development data.

Figure 4 reports the average ChrF++ score between most similar captions within each human rating bucket. Overall, we observe a weak but consistent positive trend: captions receiving higher human ratings tend to be slightly more similar to a development-set caption. This effect is more pronounced for Wixárika and Bribri, where the increase in similarity across rating buckets is clearer, whereas other languages show a flatter, noisier pattern. Notably, for Wixárika, the top-ranked system (NAIST) directly returns development-set captions, making this relationship explicit: its high ratings are by construction tied to the development set.

Taken together, these results suggest that test-set examples that are more similar to the development split elicit higher-rated predicted captions.

5.2 Qualitative Analysis

We further present a qualitative overview based on annotator comments collected following the human evaluation.

Bribri The higher scoring descriptions for the Bribri content focus on producing oversimplified but grammatically plausible descriptions. These are not capable of capturing the cultural intricacies of the images, but they are at least able to provide a readable description of some part of the picture (e.g., a pot with a traditional stew of pork and yucca in it, held by a woman, is described as *Chkà tso' ù a. Pë' tö chkà alòk* "The food is in the house. People prepare food"). The submitted systems also produce numerous hallucinations. For example, some systems report seeing canoes on a river, where the picture merely shows a river with ripples on the water. In a picture of the important tradition of the *Ák kuk* "pulling the stone" (Brenes Mora, 2024), numerous men are seen wading in a river and carrying a large stone on a mesh of trunks similar to a palanquin. Here the same hallucination shows up again: one of the systems describes this as "*Pë' dàmi taîë kanò kï*" meaning "Lots of people are coming by canoe."

Guaraní Systems in general had a good grasp of what they needed to generate and were mostly comprehensible, but often lost important details that could be easily seen by humans.

Yucatec Maya Our annotators express surprise at how accurately some models are able to describe the images in their native language. They also note instances where certain systems produce comical descriptions of the visual content.

Central Veracruz Nahuatl Many of the captions are impressive in their specificity and naturalness. However, there are also many descriptions that veer off-topic, are editorialized, and exaggerated aspects of a seemingly idealized Nahua culture. In some cases, captions are provided in Spanish instead of Nahuatl.

Wixárika Overall, evaluation identifies a considerable number of duplicated phrases. The annotator reports that evaluation is hard as most descriptions are hard to read and confusing. Nevertheless, the systems also generate a set of good or excellent descriptions. Finally, an important number of generations contain code-switching, or consist mostly of Spanish words.

6 Conclusion

We present the AmericasNLP 2026 Shared Task on Cultural Image Captioning for Indigenous Languages, the first shared task dedicated to generating captions for images depicting Indigenous cultures of the Americas, written in the Indigenous languages themselves. To support this effort, we create a novel dataset for the task. Eight teams participate, submitting 27 systems in total. Results indicate that the task remains largely unsolved: while the strongest systems produce understandable captions, they fall short in descriptive detail and, critically, in cultural grounding. These findings highlight both the inherent difficulty of the task and the pressing need for continued investment in multimodal, culturally aware models for Indigenous languages.

Limitations

We acknowledge several limitations of our data collection and evaluation process. First, we collect only a single caption per image, which limits the reliability of the reference captions. Similarly, human evaluation annotations are sparse, with most languages covered by only one or two annotators

per item, which may introduce noise into the ratings. Second, while we provide annotation guidelines, the asynchronous nature of the collection process leads to inconsistencies across languages. This is particularly evident in the development set, which reflects the challenges of an initial annotation round. Future work could address these quality issues and refine the dataset.

Ethics Statement

This work contains pictures and descriptions taken in the context of Indigenous communities of the Americas. These groups have been historically subject to discrimination, oppression, and colonialism. This work is done with the aim of closing the technological gap between Indigenous languages and the majority language in NLP. That being said, we recognize the risks of working in this setting. Therefore, we take the following measures: all annotators and participants are informed about this work, the goals, and where to download and read the results; all annotators are community members and are recognized for their work with an hourly salary equivalent to that of a high school teacher in their respective region. We additionally avoid using pictures that contain human faces. In the case of the inclusion of human faces, we either anonymize the images or obtain explicit approval. Most of the dataset is publicly available—the exception is the held-out test set, which we retain to avoid data contamination (when scraped by LLMs without our permission). The data has been released under the CC-BY-NC license, upon agreement with the annotators. All annotators retain ownership of the dataset. All decisions while creating the data collections are taken based on the standards defined by our field (Bird, 2020; Mager et al., 2023).

Acknowledgments

We would like to thank all teams for participating in the shared task, all data contributors, and community members that participated in this effort.

References

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. nocaps: novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8947–8956.

Mariana Anguiano. 1978. *La endoculturación entre los huicholes*. México INI, 1978.

Anthropic. 2026. Introducing Claude Opus 4.7. https://www.anthropic.com/news/claude-opus-4-7. Accessed: 2026-05-29.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Samantha Brenes Mora. 2024. Pueblos bribri y cabécar celebran su cultura con la ancestral "jala de piedra".

Fidencio Briceño Chel. 2021. ¿ hacia dónde va la lengua maya de la península de yucatán? entre institucionalización y patrimonialización. *Maya America: Journal of Essays, Commentary, and Analysis*, 3(1):12.

Mirelle Bueno and Sushil Garg. 2026. Culturally grounded image captioning in indigenous languages with vision-language models: Cascaded and single-stage approaches. In *Proceedings of the 6th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2026)*, San Diego, California. Association for Computational Linguistics.

Minh Duc Bui, Katharina Von Der Wense, and Anne Lauscher. 2025. Multi^{3}Hate: Multimodal, multilingual, and multicultural hate speech detection with vision–language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9714–9731, Albuquerque, New Mexico. Association for Computational Linguistics.

Luis Chiruzzo, Pavel Denisov, Alejandro Molina-Villegas, Silvia Fernandez-Sabido, Rolando Coto-Solano, Marvin Agüero-Torales, Aldo Alvarez, Samuel Canul-Yah, Lorena Hau-Ucán, Abteen Ebrahimi, Robert Pugh, Arturo Oncevay, Shruti Rijhwani, Katharina von der Wense, and Manuel Mager. 2024. Findings of the AmericasNLP 2024 shared task on the creation of educational materials for indigenous languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous*

Languages of the Americas (AmericasNLP 2024), pages 224–235, Mexico City, Mexico. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Adolfo Constenla. 1996. *Poesía tradicional indígena costarricense*. Editorial Universidad de Costa Rica.

Adolfo Constenla. 2006. *Poesía bribri de lo cotidiano: 37 cantos de afecto, devoción, trabajo y entretenimiento*. Editorial Universidad de Costa Rica.

Adolfo Constenla, Feliciano Elizondo, and Francisco Pereira. 2004. *Curso Básico de Bribri*. Editorial de la Universidad de Costa Rica.

Diego Cuadros, Nicholas Leeds, Amanda Avalos, Azul Alipzar-Velazquez, Jared Coleman, Faezeh Dehghan Tarzjani, and Bhaskar Krishnamachari. 2026. Schema-constrained image captioning for five low-resource indigenous languages. In *Proceedings of the 6th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2026)*, San Diego, California. Association for Computational Linguistics.

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.

Aashish Dhawan, Christopher Driggers-Ellis, Dzmitry Kasinets, Christan Grant, and Daisy Wang. 2026. Retrieval-augmented long-context translation for cultural image captioning: Gators submission for AmericasNLP 2026 shared task. In *Proceedings of the 6th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2026)*, San Diego, California. Association for Computational Linguistics.

Hongyuan Dong, Jiawen Li, Bohong Wu, Jiacong Wang, Yuan Zhang, and Haoyuan Guo. 2024. Benchmarking and improving detail image caption. *Preprint*, arXiv:2405.19092.

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.

Clement Farabet and Olivier Lacombe. 2026. Gemma 4: Byte for byte, the most capable open models. `https://blog.google/innovation-and-ai/technology/developers-tools/gemma-4/`. Accessed: 2026-05-28.

Rafael M. Fernandes. 2026. USP at AmericasNLP 2026 shared task: Culturally-aware image captioning for indigenous languages via vision-language models and fine-tuned neural machine translation. In *Proceedings of the 6th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2026)*, San Diego, California. Association for Computational Linguistics.

Sofía Flores Solórzano. 2017. Corpus oral pandialectal de la lengua bribri.

Alí García Segura. 2016. *Ditsö Rukuö Identity of the Seeds: Learning from Nature*. IUCN.

Theodore R Goller, Patricia L Goller, and Viola G Waterhouse. 1974. The phonemes of Orizaba Nahuatl. *International Journal of American Linguistics*, 40(2):126–131.

Paula Gómez. 1999. El huichol de san andrés cohamiata. *Jalisco, Archivo de lenguas indígenas de México 22, México: El Colegio de México,*.

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.

Hansi Hettiarachchi, Tharindu Ranasinghe, Paul Rayson, Ruslan Mitkov, Mohamed Gaber, Damith Premasiri, Fiona Anting Tan, and Lasitha Uyangodage, editors. 2025. *Proceedings of the First Workshop on Language Models for Low-Resource Languages*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

INEC. 2011. X Censo Nacional de Población y VI de Vivienda 2011 - Territorios Indígenas - Principales Indicadores Demográficos y Socioeconómicos.

INEGI. 2020. Censo de población y vivienda 2020. Accedido: 2026-05-28.

Carla Victoria Jara Murillo. 2018a. *Gramática de la Lengua Bribri*. EDigital.

Carla Victoria Jara Murillo. 2018b. *I Ttè Historias Bribris*, second edition. Editorial de la Universidad de Costa Rica.

Carla Victoria Jara Murillo and Alí García Segura. 2013. *Se' ttö' bribri ie Hablemos en bribri*. EDigital.

Haakon S. Krohn. 2020. Diccionario digital bilingüe bribri.

Luis Lara and Param Raval. 2026. From machine translation to image captioning: Training vision-language models for indigenous languages of the americas. In *Proceedings of the 6th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2026)*, San Diego, California. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context. *Preprint*, arXiv:1405.0312.

Shudong Liu, Yiqiao Jin, Cheng Li, Derek F. Wong, Qingsong Wen, Lichao Sun, Haipeng Chen, Xing Xie, and Jindong Wang. 2025. Culturevlm: Characterizing and improving cultural understanding of vision-language models for over 100 countries. *Preprint*, arXiv:2501.01282.

Carl Lumholtz. 1902. El méxico desconocido [2 vols.]. *México, Editora Nacional*.

Manuel Mager, Abteen Ebrahimi, Shruti Rijhwani, Arturo Oncevay, Luis Chiruzzo, Robert Pugh, and Katharina von der Wense, editors. 2024. *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*. Association for Computational Linguistics, Mexico City, Mexico.

Manuel Mager, Elisabeth Mager, Katharina Kann, and Ngoc Thang Vu. 2023. Ethical considerations for machine translation of indigenous languages: Giving a voice to the speakers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4871–4897, Toronto, Canada. Association for Computational Linguistics.

Enrique Margery. 2005. *Diccionario Fraseológico Bribri-Español Español-Bribri*, second edition. Editorial de la Universidad de Costa Rica.

Isabel Martínez. 2006. Gutiérrez del angel, arturo. la peregrinación a wirikuta: El gran rito de paso de los huicholes. méxico: Etnografía de los pueblos indígenas de méxico, instituto nacional de antropología e historia, universidad de guadalajara, 2002, 310 p.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Karolina Stanczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5769–5790, Miami, Florida, USA. Association for Computational Linguistics.

Johannes Neurath. 2002. Las fiestas de la casa grande. *Universidad de Guadalajara, Instituto Nacional de Antropología e Historia, Guadalajara, Mexico*.

Team NLLB, 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.

OpenAI. 2026. Introducing GPT-5.5. https://openai.com/index/introducing-gpt-5-5/. Accessed: 2026-05-28.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 4(1-2):1–174.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, and 57 others. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pages 11479–11505. Curran Associates, Inc.

Carlos Sánchez Avendaño. 2013. Lenguas en peligro en Costa Rica: vitalidad, documentación y descripción. *Revista Káñina*, 37(1):219–250.

Wenchen Shi, Phakphum Artkaew, and Luke Gessler. 2026. Culturally-aware image captioning for Guaraní with multimodal prompting: IUHoosiers at AmericasNLP 2026. In *Proceedings of the 6th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2026)*, San Diego, California. Association for Computational Linguistics.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 467 others. 2026. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, Siyang Qin, Reeve Ingle, Emanuele Bugliarello, Sahar Kazemzadeh, Thomas Mesnard, Ibrahim Alabdulmohsin, Lucas Beyer, and Xiaohua Zhai. 2024. Paligemma 2: A family of versatile vlms for transfer. *Preprint*, arXiv:2412.03555.

Carlos Sánchez Avendaño. 2020. *Se' Dalì Diccionario y Enciclopedia de la Agricultura Tradicional Bribri*. DIPALICORI.

David Tuggy. 1992. The affix-stem distinction: A cognitive grammar analysis of data from orizaba nahuatl. *Cognitive Linguistics*, 3(3):237–300.

David Tuggy. 1998. *Giving in Nawatl*, page 35–66. John Benjamins Publishing Company.

Universidad Autónoma de Yucatán. n.d. Portal de la Cultura Maya. `https://www.mayas.uady.mx/`. Accessed: 2026-05-30.

Leopoldo Valiñas Coalla. 2020. *Lenguas originarias y pueblos indígenas de México. Familias y lenguas aisladas*. Academica Mexicana de la Lengua, México.

Justin Vasselli, Arturo Martínez Peguero, Shintaro Ozaki, Frederikus Hudi, Haruki Sakajo, and Taro Watanabe. 2026. Nearest-neighbor retrieval for indigenous image captioning. In *Proceedings of the 6th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2026)*, San Diego, California. Association for Computational Linguistics.

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kuckreja, Mykola Maslych, Wafa Al Ghallabi, Mihail Mihaylov, Chao Qin, Abdelrahman M Shaker, Mike Zhang, Mahardika Krisna Ihsani, Amiel Esplana, Monil Gokani, and 50 others. 2025. All languages matter: Evaluating lmms on culturally diverse 100 languages. *Preprint*, arXiv:2411.16508.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ji Wang and Hanqi Yang. 2026. 6fanle submission to the AmericasNLP 2026 shared task on Wixarika image captioning. In *Proceedings of the 6th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2026)*, San Diego, California. Association for Computational Linguistics.

Ruvan Weerasinghe, Isuri Anuradha, and Deshan Sumanathilaka, editors. 2025. *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*. Association for Computational Linguistics, Abu Dhabi.

Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Wang Yutong, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, Anar Rzayev, Anirban Das, Ashmari Pramodya, Aulia Adila, Bryan Wilie, Candy Olivia Mawalim, Cheng Ching Lam, Daud Abolade, Emmanuele Chersoni, and 32 others. 2025. WorldCuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3242–3264, Albuquerque, New Mexico. Association for Computational Linguistics.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Robert Mowry Zingg. 1982. *Los huicholes una tribu de artistas*. México INI, 1982.

A Related Work

Image Captioning Benchmarks. Early captioning benchmarks such as MS-COCO (Lin et al., 2015) and Flickr30k (Young et al., 2014) established standard evaluation protocols for image description, but focus primarily on everyday Western scenes. Nocaps (Agrawal et al., 2019) extended this to novel object categories drawn from Open Images. Dong et al. (2024) proposed a benchmark and evaluation metric for *detailed* image captioning, addressing the shortcomings of short-caption benchmarks.

Our work extends these efforts by encouraging culturally enriched captions.

Cultural Visual Understanding. Several benchmarks assess the cultural awareness of vision-language models (VLMs). CulturalVQA (Nayak et al., 2024) probes VLM understanding across clothing, food, drinks, rituals, and traditions from 11 countries, but is restricted to English. CVQA (Romero et al., 2024) presents a multilingual VQA setting spanning 26 languages and 28 countries with human-written questions. WorldCuisines (Winata et al., 2025) provides a large-scale food-centric VQA benchmark across 30 languages sourced from Wikipedia, and ALM-bench (Vayani et al., 2025) scales further to 100 languages, targeting low-resource languages and diverse cultural aspects.

However, none of these works addresses Indigenous communities and languages, a gap our dataset fills.

B Detailed Dataset Creation

B.1 Bribri Image Creation Sources

For Bribri, we additionally source images from three Costa Rican cultural institutions: the Sistema de Información Cultural de Costa Rica (Sicultura),[3] the Dirección de Gestión Sociocultural of the Ministerio de Cultura y Juventud,[4] and the Universidad de Costa Rica.[5].

B.2 Caption Guidelines

The full annotation guidelines are provided in Figure 5. We further provide one correct and one incorrect example caption. The correct caption reads: "A wooden house, the so-called *carretón*, built specifically to store food such as corn, also serves as living quarters for people." The incorrect caption reads: "A Wixárika woman shelling corn to make nixtamal"—chosen because the image does not clearly convey that the corn is being prepared for nixtamal.

C Detailed Experimental Setup

C.1 Human Evaluation Detail

For Yucatec Maya, the human evaluation phase involved two computer science alumni who are native Maya speakers from the Yucatán Peninsula, one from the Universidad Autónoma de Yucatán (UADY) and the other from UIMQROO. For the remaining languages, the same annotators who contributed to data collection also conducted the human evaluation.

We report the full annotation guideline in Figure 6. Note that the guideline is in Spanish originally. Furthermore, we show a screenshot of our annotation tool in Figure 7.

D All Systems Results

We report the performance of all submitted systems in Table 7. Alongside chrF++, we also report CIDEr (Vedantam et al., 2015), a metric originally proposed for image captioning and widely adopted in vision-language benchmarks. To our knowledge, CIDEr has not previously been applied to Indigenous language evaluation; we therefore adopt chrF++ as our primary metric, given its stronger track record in low-resource and morphologically rich language settings (Ebrahimi et al., 2024).

[3] https://si.cultura.cr
[4] https://www.dircultura.go.cr
[5] https://www.ucr.ac.cr

Team	**Ver.**	Bribri		Guaraní		Maya		Nahuatl		Wixárika	
		chrF	CIDEr	chrF	CIDEr	chrF	CIDEr	chrF	CIDEr	chrF	CIDEr
6fanle	v0	—	—	—	—	—	—	—	—	19.16	0.0214
IUHoosiers	v0	—	—	24.39	0.1238	—	—	—	—	—	—
IUHoosiers	v1	—	—	24.41	0.1351	—	—	—	—	—	—
IUHoosiers	v2	—	—	24.41	0.1428	—	—	—	—	—	—
IUHoosiers	v3	—	—	24.16	0.1284	—	—	—	—	—	—
IUHoosiers	v4	—	—	**24.67**	0.1489	—	—	—	—	—	—
IUHoosiers	v5	—	—	24.42	**0.1668**	—	—	—	—	—	—
IUHoosiers	v6	—	—	24.04	0.1122	—	—	—	—	—	—
IUHoosiers	v7	—	—	22.43	0.0738	—	—	—	—	—	—
IUHoosiers	v8	—	—	24.41	0.1351	—	—	—	—	—	—
InclusionVLM	v0	7.94	0.0059	16.48	0.0465	16.97	0.0100	14.06	0.0031	18.37	0.0084
InclusionVLM	v1	2.54	0.0002	7.61	0.0055	9.14	0.0059	—	—	10.52	0.0014
InclusionVLM	v2	—	—	—	—	—	—	—	—	12.34	0.0078
InclusionVLM	v3	—	—	—	—	—	—	—	—	13.93	0.0079
Mila	v0	11.73	0.0152	19.63	0.0314	—	—	19.42	0.0332	18.81	0.0246
Mila	v1	10.89	0.0086	19.77	0.0310	—	—	19.85	0.0372	18.85	0.0249
Mila	v2	11.31	0.0140	19.42	0.0520	—	—	20.66	0.0415	19.01	0.0299
Mila	v3	4.56	0.0058	12.80	0.0510	15.99	0.0950	19.72	0.0615	10.98	0.0096
NAIST	v0	**19.37**	**0.1167**	19.41	0.0458	15.80	0.0424	20.93	0.0711	**19.84**	**0.0422**
NAIST	v1	4.21	0.0007	9.71	0.0350	12.27	0.0918	10.09	0.0123	—	—
NAIST	v2	5.29	0.0018	8.33	0.0270	10.20	0.0555	15.34	0.0597	—	—
baseline	v0	7.01	0.0005	20.14	0.0050	—	—	9.52	0.0015	16.91	0.0066
gators	v0	10.64	0.0212	23.10	0.1243	21.11	**0.1765**	**25.42**	**0.1795**	17.58	0.0326
gators	v1	17.90	0.1081	21.63	0.1271	—	—	—	—	—	—
usp	v0	10.95	0.0007	19.73	0.0236	10.83	0.0100	9.49	0.0001	13.68	0.0027
yaduha	v0	10.03	0.0084	16.90	0.0379	**23.41**	0.1116	21.00	0.0284	15.61	0.0266

Table 7: Automatic evaluation results per team and submission version. Best result per language in **bold**.

Annotation Guidelines

Please follow these rules when writing image descriptions:

1. Start with only describing what is visually present.
- Focus on people, objects, animals, settings, and clearly visible actions.
- Focus on the most salient visual element(s).

2. IMPORTANT: Add cultural explanations
- Cultural enrichment: You must elaborate on/explain the function, purpose, or typical use of objects, clothing, gestures, or settings, but only when these explanations are grounded in what is visually observable.
- Visually observable: The function, purpose, or typical use must be immediately clear, without ambiguity, to any member of the community.
- Do not infer identity, background, profession, or cultural group unless it is explicitly indicated by visible, unambiguous cues. Better describe it with words, e.g. a person wearing traditional ⟨Culture X⟩ clothes vs. a ⟨Culture X⟩ person.

3. Be objective and neutral.
- Use factual, descriptive language.
- Avoid opinions or value-laden terms such as "beautiful," "cute," "poor," "aggressive", or "angry."
- Avoid emotional interpretations ("sad situation", "exciting moment", . . .) unless the emotion is visually undeniable (e.g., a person visibly crying).

4. Description Format
- Sentences should neither be too short nor too long. Try to be concise.
- Each sentence must contain a verb.
- Please pay attention to grammar and spelling.
- The description should be 1–2 sentences long.
- Additional comments are welcome.

Figure 5: Annotation guidelines provided to annotators for writing enriched image captions. Guidelines emphasize visual grounding, cultural context, objectivity, and concise formatting.

Human Evaluation Guidelines

How to Rate

For each system description, assign a single overall score from 1 to 5. Evaluate the description along two dimensions:

1. Language Quality: Is it written in the target language? Is it grammatically correct, fluent, and natural? If the description is poorly written or illegible, it cannot be considered a good description.

2. Image Fidelity and Correct Use of Cultural Terms: If the language quality is good—does the description reflect what is actually seen in the image? Does it use the correct cultural terms? Would it seem respectful and accurate to a member of the community?

Rating Scale

5 — Excellent: Fluent and natural language with an accurate, culturally grounded description of the image. Uses correct cultural and technical terms. Nothing significant to correct.

4 — Good: Clear, well-written language with a correct description, but with minor flaws, e.g., minor language errors, missing details in the description, or imprecise cultural vocabulary.

3 — Mixed: Language is mostly understandable, but the description is mistaken about the image, omits important content, or is too vague to be clearly useful.

2 — Poor: Language has serious problems (incorrect grammar, frequent errors, hard to follow) and the description is largely inaccurate. Still recognizable as an attempt at a description.

1 — Unusable: Not in the target language, not understandable as language, or bears no relation to the image.

Figure 6: Rating guidelines provided to annotators for evaluating enriched image captions (originally in Spanish).

Imagen 69

Descripción de referencia *(sólo como guía)*
Ojehecha kesu apu'a oñemboaty peteĩ mba'e ári. Ko tembi'u ha'e kesu paraguái, peteĩ kesu pyahu Paraguáipegua.

*Para la **Imagen 69** de arriba, otorgue a cada descripción **una sola puntuación general del 1 al 5**. Evalúe primero *la calidad del idioma* y, si esta es buena, *la fidelidad a la imagen y el uso correcto de términos culturales*.

	1 - Inutilizable	2 - Deficiente	3 - Mixto	4 - Bueno	5 - Excelente
Tres ñandutí (queso fresco paraguayo) oñevendéva tendápe, ojejapóva tuju renondépe, orekóva mba'erechaukaha hũ pukukue.	○	○	○	○	○
Peteĩ ajaka henyhe "chipa argolla"guí, tembi'u ojehayhuvéva ñane retã Paraguáipe, oĩnte va'erã jehomombyrýpe ha aty guasu rogayguákuéra ndive, akói hyakuãvurei hese kesu ha aramirõ.	○	○	○	○	○

Figure 7: Partial screenshot of the human evaluation annotation tool used in our study.