Nearest-Neighbor Retrieval for Indigenous Image Captioning

Justin Vasselli, Arturo Martínez Peguero, Shintaro Ozaki,
Frederikus Hudi, Haruki Sakajo, Taro Watanabe
Nara Institute of Science and Technology
vasselli.justin_ray.vk4@is.naist.jp

Abstract

This paper describes the NAIST submission to the AmericasNLP 2026 Shared Task on Indigenous Language Image Captioning. We investigate two approaches for generating captions in Bribri, Guaraní, Nahuatl, Wixárika, and Yucatec Maya. The first is a nearest-neighbor retrieval system that uses CLIP image embeddings to retrieve the most similar image from the development set and directly reuse its caption. The second is a generation pipeline that combines scene analysis, dictionary-grounded lexical planning, retrieved gloss templates, and interlinear gloss representations to constrain generation in low-resource settings.

The retrieval-based approach substantially outperformed the gloss-based pipeline under chrF++ evaluation and was competitive across all submitted systems, achieving first-place automated system rankings for Bribri and Wixárika and third place for Nahuatl. The gloss-based pipeline produced weaker automatic evaluation results and exposed problems with dictionary coverage, orthographic mismatches between resources, and unstable grammatical generation. Our results suggest that retrieval-based methods provide a strong baseline for low-resource captioning tasks when high-quality examples are available.[1]

1 Introduction

The AmericasNLP 2026 Shared Task on Indigenous Language Image Captioning focuses on generating captions for images in low-resource Indigenous languages. The task is challenging for current multimodal systems due to the limited amount of training data and the varying quality of generation available for these languages in current translation systems and large language models (Bui et al., 2026).

[1]Code available at https://github.com/JVasselli/americasnlp2026-naist

Our submission explores two approaches based on the retrieval of development set captions. The first uses nearest-neighbor retrieval. Given a query image, we retrieve the most similar image from the development set using CLIP embeddings (Radford et al., 2021) and return its caption directly. The second approach uses GPT-5.4 (Singh et al., 2026) to generate an interlinear gloss using retrieved captions and a dictionary-grounded lexical planning stage, then converts the gloss into the target language using language-specific conversion rules.

Our experiments showed that nearest-neighbor retrieval substantially outperformed the gloss generation pipeline on automatic evaluation metrics across all the languages we tested. Retrieval-based captioning produced fluent captions, while the pipeline often produced short or constrained outputs. We also observed differences across languages, suggesting that the difficulty of generating fluent captions varies depending on language support and dataset characteristics.

These results show that retrieval remains effective for low-resource captioning tasks, particularly under surface-overlap metrics such as chrF++ (Popović, 2015). However, the human evaluation results were more mixed across languages.

2 Nearest-Neighbor Retrieval

2.1 System Overview

We implement a nearest-neighbor retrieval approach that retrieves the caption associated with the most similar development image.

We evaluate four similarity strategies on the development set:

- CLIP similarity
- DINOv2 similarity
- English caption similarity
- Spanish caption similarity

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 272–278

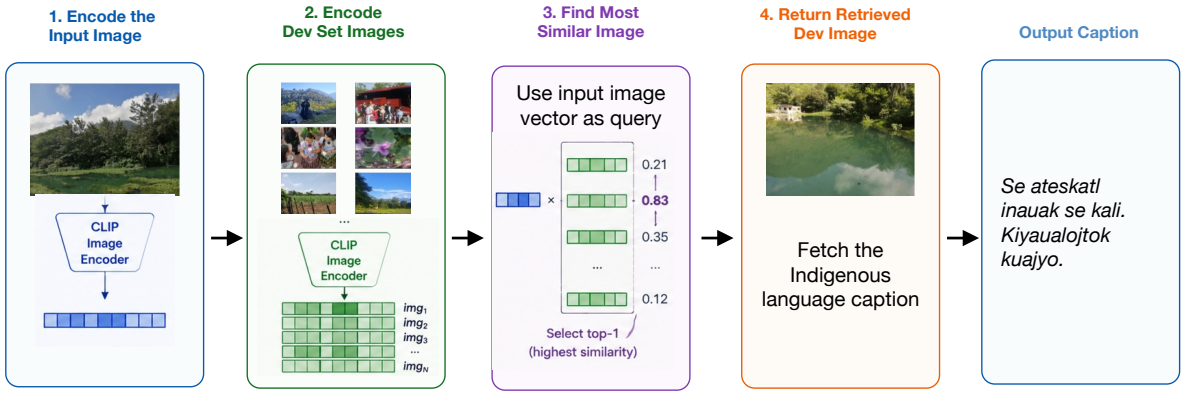

Figure 1: Overview of the nearest-neighbor retrieval system

Language	CLIP	DINOv2	Caption EN	Caption ES	Baseline
Bribri	**20.52**	19.91	18.86	18.61	7.57
Maya	**21.94**	21.65	18.68	19.66	—
Guaraní	**22.97**	22.66	22.33	21.16	20.82
Nahuatl	**25.08**	24.02	23.62	23.52	11.53
Wixárika	19.71	**19.83**	18.69	19.01	17.77

Table 1: Development set chrF++ scores for nearest-neighbor retrieval using different similarity strategies. English and Spanish denote caption similarity for each language. The official baseline did not include Maya results.

For image-based retrieval, we compute embeddings for all images and select the nearest neighbor using cosine similarity.

For CLIP-based similarity, we use OpenCLIP with a ViT-B/32 architecture pretrained on the LAION-2B dataset (Cherti et al., 2023). The resulting image embeddings have a dimension of 512, and images are processed using the default OpenCLIP preprocessing pipeline.

For DINOv2-based similarity, we use the DINOv2 large model (ViT-L/14) (Oquab et al., 2024). This model is trained with self-supervision. The resulting embeddings have a dimension of 1024, and images are processed using the default DINOv2 preprocessing pipeline (resize, center crop, and normalization).

We include caption-based retrieval to test whether the generated English and Spanish descriptions capture scene similarities missed by image embeddings. We first generate English and Spanish captions for each image in both the development and test sets using `gpt-4o-mini-2024-07-18`. These captions are then encoded using LaBSE (Feng et al., 2022), a multilingual sentence embedding model designed for cross-lingual retrieval. We compute cosine similarity between embeddings to identify the nearest neighbor.

For each query image, we select the caption associated with the most similar image in the development set. In the development set experiments, the query image is excluded from the candidate pool.

2.2 Results

As shown in Table 1, CLIP-based retrieval achieved the strongest development set performance for four of the five languages. DINOv2 retrieval remained competitive, particularly for Wixárika, where it slightly outperformed CLIP. Caption-based retrieval using generated English and Spanish descriptions consistently underperformed image-based retrieval, though the gap was smaller for Guaraní and Nahuatl. All retrieval approaches outperformed the official baseline on the languages where baseline scores were available. Based on development set performance, we selected the CLIP-based retrieval system for submission.

3 Gloss-based Pipeline Approach

3.1 System Overview

In addition to nearest-neighbor retrieval, we explored a gloss-based generation pipeline intended to constrain generation through dictionary grounding and intermediate gloss representations. Instead

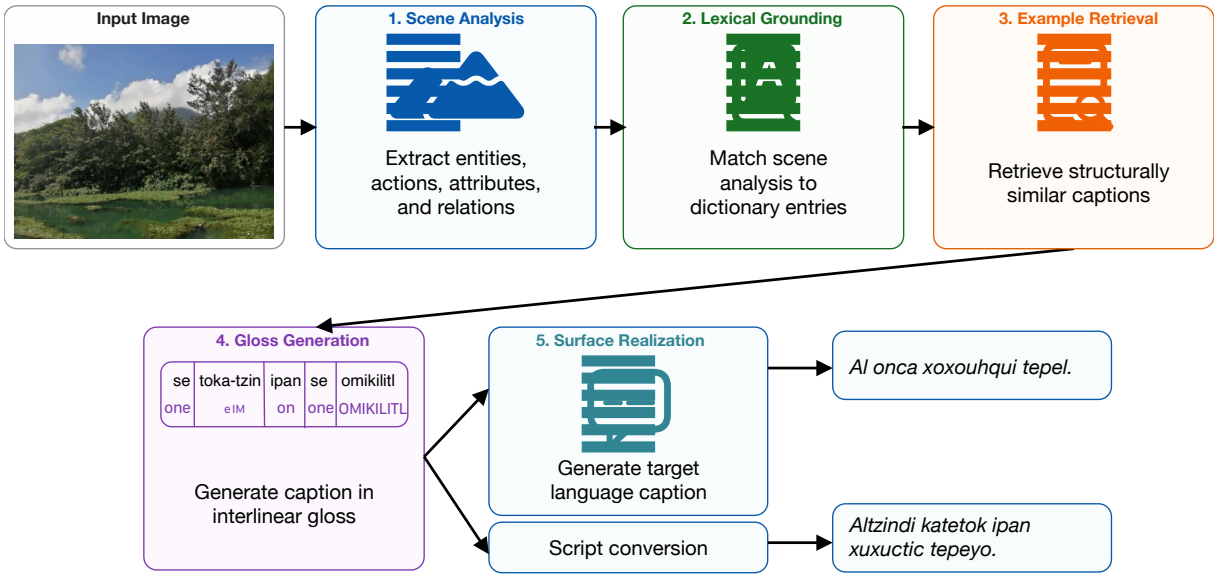

Figure 2: Overview of the gloss-based generation pipeline.

of generating captions directly from the image, the system first extracts structured scene information and maps concepts to attested lexical items before generating a caption gloss. This approach was designed to reduce hallucinations in low-resource Indigenous languages. The pipeline consists of five stages:

1. Scene analysis
2. Lexical grounding
3. Example retrieval
4. Interlinear Gloss generation
5. Surface Realization

Resources. To support the gloss-based pipeline, we constructed several synthetic linguistic resources for each language. First, we manually created a glossed caption resource from the provided target captions. Each entry contains a morphological segmentation, interlinear gloss, and approximate English translation. We created these glosses using publicly available dictionaries and grammar references, including grammar references for Guaraní (Estigarribia, 2020), Bribri (Jara, 2018), Yucatec Maya (Bolles and Bolles, 2001), and Nahuatl (Tuggy, 2004); they were intended for experimental generation support rather than authoritative linguistic annotation. Below is a sample entry:

```
"Guarani": "Chemandu'áta nderehe.",
"morphology": "che-mandu'a-ta nde=rehe",
"gloss": "1SG.INACT-remember-FUT 2SG=on",
"Spanish": "Me acordaré de ti.",
"English": "I will remember you."
```

From these glossed examples, we automatically extracted gloss-to-morpheme alignment tables that record how gloss units were morphologically realized in the examples.

```
...
3SG.front tu_táan 1
3SG.on y-óok'ol 1
ADD tak_xan 1
CL.tree kúul 2
DET le 26
...
```

For Guaraní and Nahuatl, which contain richer agglutinative morphology, we additionally created subword-level alignment resources.

```
...
3SG.OBJ ki 14
3SG.POSS i 5
3SG.POSS tlan 2
ABS tli 12
APPL li 1
APPL lia 2
CAUS lti 1
COND skia 1
...
```

We also collected bilingual dictionaries for each language and used them during lexical planning and gloss conversion, including the Bribri dictionary released by Vasselli et al. (2024), the Guaraní dictionary *Avañe'ẽ del Taragui* (Ministerio de Educación de la Provincia de Corrientes, 2022), the Yucatec Maya dictionary *Diccionario de Uso del Maya Yucateco* (Yoshida and Ucan Dzul, 2025),

Language	Systems	Nearest Neighbor	Gloss Generated	Gloss Converted	Baseline
Bribri	14	**19.37 (1)**	4.21 (13)	5.29 (11)	7.01 (10)
Maya	9	**15.80 (5)**	12.27 (6)	10.20 (8)	-
Guaraní	24	19.41 (18)	9.71 (22)	8.33 (23)	**20.14 (12)**
Nahuatl	12	**20.93 (3)**	10.09 (10)	15.34 (8)	9.52 (11)
Wixárika	14	**19.84 (1)**	-	-	16.91 (8)

Table 2: Test set results on automated metric. Number in the parentheses is the system ranking of all submitted systems.

Language	CLIP NN	Generated	Converted
Bribri	20.52	4.66	5.38
Maya	21.94	12.52	11.79
Guaraní	22.97	7.47	6.97
Nahuatl	25.08	10.13	14.23

Table 3: Development set chrF++ scores for the gloss-based pipeline variants (Generated, Converted) and CLIP-based nearest-neighbor (NN) retrieval.

the Wixárika dictionary by SIL International (SIL International, 2012), and the Nahuatl dictionary by SIL International (SIL International, 2002).

Step 1: Scene analysis. We first prompt GPT (`gpt-4o-mini-2024-07-18`) to analyze an input image and produce a structured scene representation. The representation contains entities, actions, attributes, and spatial relations, together with corresponding Spanish lexical items. An example scene representation is shown below:

```
{
  "attributes": ["green", "cloudy"],
  "main_entity": "water",
  "predicate_type": "existential",
  "secondary_entities":
    ["trees", "mountain", "sky"]
}
```

The system also produces Spanish lexical mappings, such as *green* → *verde* and *water* → *agua*.

Step 2: Lexical grounding. The extracted concepts are matched against digitized dictionaries. Concepts without dictionary matches are pruned from the scene representation. This stage produces a lexical plan containing only attested vocabulary. For example, the scene representation above may be reduced to concepts such as *green*, *water*, and *mountain* if no dictionary matches are found for the remaining items.

Step 3: Example retrieval. We retrieve candidate captions from the development set using manually assigned structure labels such as `ENTITY + STATE + LOCATION`. Retrieved examples provide interlinear glosses and morphological patterns for generation.

Step 4: Interlinear gloss generation. GPT-5.4 (`gpt-5.4-2026-03-05`) is then prompted to generate an interlinear gloss caption, together with a target-language caption. The prompt includes the image, the lexical plan, and retrieved gloss examples. The model is instructed to use only lexical items appearing in the dictionaries or retrieved examples.

For example, the system may generate the gloss:

```
water be-PROG on green mountain
```

Step 5: Surface realization. Finally, we convert the gloss into Indigenous language captions in two ways: LLM generation and rule-based script. The LLM is prompted to generate a caption immediately after the gloss and has all of the information from the captioning stage (image, scene plan, examples, dictionary entries). In the above example GPT-5.4 generated:

```
Al onca xoxouhqui tepel.
```

We additionally apply a deterministic gloss conversion to produce a second version of the caption. This stage combines dictionary matches, rule-based morphology, and morpheme alignments extracted from the synthetic gloss resources. For the previous example, the conversion stage produced:

```
Altzindi katetok ipan xuxuctic tepeyo.
```

3.2 Results

Table 3 shows the development set results for the gloss-based pipeline. Across all languages, both variants scored well below nearest-neighbor retrieval. Performance also varied across languages. The deterministic conversion stage improved results for Bribri and Nahuatl but slightly reduced performance for Maya and Guaraní. Due to incomplete language-specific resources and conversion rules, the gloss-based pipeline was not finalized for Wixárika and was therefore not included in these experiments.

Language	Reference	Pipeline
Bribri	14.92	7.40
Maya	11.28	6.02
Guaraní	21.64	6.52
Nahuatl	8.34	5.26

Table 4: Average number of tokens per caption for the reference captions and the gloss-based pipeline outputs.

Lang.	Avg rating	Sys. Ranking
Bribri	2.219	3
Maya	1.934	4
Guaraní	1.978	4
Nahuatl	1.220	4
Wixárika	3.790	1

Table 5: Test set results of nearest neighbor retrieval after human evaluation.

Lang.	Improved	Avg Δ	Min Δ	Max Δ
Bribri	32/50	0.72	-3.48	10.32
Maya	22/50	-0.85	-16.43	10.38
Guaraní	27/50	-0.24	-6.63	4.77
Nahuatl	43/50	4.88	-7.52	17.90

Table 6: Change in chrF++ score after deterministic gloss conversion relative to direct caption generation on the development set. "Improved" indicates the number of examples for which gloss conversion increased the chrF++ score.

4 Results and Discussion

4.1 Automatic Evaluation

Table 2 shows the official test set chrF++ results for our submitted systems. The nearest-neighbor retrieval system achieved the strongest performance of our submissions across all languages. The system ranked first overall for Bribri and Wixárika, third for Nahuatl, fifth for Maya, and eighteenth for Guaraní.

The gloss-based pipeline consistently produced lower chrF++ scores than nearest-neighbor retrieval. One reason is that the pipeline produced much shorter captions than the references, as shown in Table 4. The lexical planning stage aggressively pruned concepts without reliable dictionary matches, which reduced hallucinated vocabulary but also removed many scene details. As a result, generated captions often contained only a small subset of the entities and actions present in the reference captions.

We also observed several additional issues during gloss generation and lexical planning. The system relied heavily on dictionary matches and often ignored development set vocabulary. In some cases, valid dictionary matches were missed due to inflectional or orthographic variation, such as plural forms or differences in accent marking between the dictionaries and the shared task data. The gloss generation stage also occasionally produced grammatical constructions that did not appear in the retrieved examples or alignment resources.

Despite its simplicity, nearest-neighbor retrieval consistently produces fluent captions. Under surface-overlap metrics such as chrF++, this translates to stronger results than the constrained gloss generation pipeline.

4.2 Human Evaluation

Table 5 shows that the nearest-neighbor retrieval system achieved strong rankings on the automatic evaluation metric. The system ranked first for Wixárika, third for Bribri, and fourth for Maya, Guaraní, and Nahuatl. However, the corresponding human evaluation ratings were more mixed across languages.

This difference suggests that strong chrF++ performance does not always correspond to strong human evaluation performance. Retrieved captions are fluent and grammatical because they originate from attested examples in the development set, but the retrieved image may still differ from the query image in important ways. Surface-overlap metrics, such as chrF++, do not strongly penalize these semantic mismatches.

4.3 Gloss Conversion versus Direct Generation

Table 6 compares the direct caption generation against the rule-based gloss conversion script output on the development set. The effect of gloss conversion varied by language.

Gloss conversion improved the majority of examples for Bribri, Nahuatl, and Guaraní. However, only Nahuatl showed a substantial increase in the average chrF++ score after conversion. Guaraní and Maya both showed slight decreases on average despite improvements on many individual examples.

These results suggest that deterministic gloss conversion can improve lexical and morphological consistency when the generated gloss aligns well with the target language resources. However, errors introduced during lexical planning or gloss gen-

eration often propagate into the conversion stage. Since the conversion system relied heavily on dictionary lookups and synthetic alignment resources, that meant that mismatches in orthography, morphology, or gloss structure could lead to degraded outputs.

The relatively strong improvement for Nahuatl may also reflect differences in language support within current language models. Vasselli et al. (2026) found that large language models demonstrate comparatively stronger understanding of Guaraní and Nahuatl than several other Indigenous languages. However, the shared task uses Orizaba Nahuatl (`nlv`), while many multilingual resources and evaluations focus on broader Nahuatl varieties such as `nah`. These differences may have affected both gloss generation and gloss conversion quality.

5 Related Work

Early image captioning systems explored retrieval-based approaches that reused captions from visually similar images rather than generating captions directly. Farhadi et al. (2010) mapped images and captions into a shared semantic representation for caption retrieval, while Ordonez et al. (2011) demonstrated that nearest-neighbor image retrieval combined with direct caption transfer could produce competitive image descriptions using large captioned image collections.

More recent work has explored linguistic resources and explicit grammatical structures for low-resource language generation. Ginn et al. (2024) investigates gloss generation for endangered languages using large language models, demonstrating that intermediate linguistic representations can support generation for low-resource languages. Taguchi and Sproat (2026) shows that large language models can generalize grammatical patterns from linguistic descriptions and curated examples, motivating our use of retrieved gloss templates and structured prompting. Vasselli et al. (2024) applies dictionary-guided prompting and retrieved linguistic examples for educational material generation in Indigenous languages, demonstrating how lexical grounding and example retrieval can help constrain generation in low-resource settings. Our gloss-based pipeline combines lexical grounding, retrieved caption structures, and intermediate gloss representations to constrain generation.

6 Conclusion

We presented two approaches for the AmericasNLP 2026 Indigenous language image captioning shared task: a nearest-neighbor retrieval system based on CLIP image similarity and a gloss-based generation pipeline using lexical grounding and intermediate interlinear gloss representations.

Our experiments showed that nearest-neighbor retrieval substantially outperformed the gloss-based pipeline across most languages under chrF++ evaluation. Despite its simplicity, retrieval produced fluent captions drawn directly from attested examples in the development set and achieved strong rankings on the shared task leaderboard. In contrast, the gloss-based pipeline often produced short and overly constrained captions due to aggressive lexical pruning and limitations in the available linguistic resources.

The gloss-based pipeline nevertheless highlighted several challenges for low-resource Indigenous language generation, including orthographic variation across resources, sparse dictionary coverage, and difficulty in constraining grammatical generation. The mixed results from deterministic gloss conversion further suggest that improvements in lexical alignment and morphological normalization may be necessary before rule-based conversion can reliably improve generated outputs.

Overall, our results suggest that retrieval-based approaches remain competitive for low-resource captioning tasks, particularly when high-quality attested examples are available. However, the gap between automatic metrics and human evaluation shows that fluent retrieved captions are not always accurate. Future work may benefit from combining the fluency advantages of retrieval with stronger semantic grounding and more robust linguistic resource integration.

References

David Bolles and Alejandra Bolles. 2001. *A Grammar of the Yucatecan Mayan Language*. Foundation for the Advancement of Mesoamerican Studies, Crystal River, FL.

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.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829.

Bruno Estigarribia. 2020. *A Grammar of Paraguayan Guarani*. Grammars of World and Minority Languages. UCL Press.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *Computer Vision – ECCV 2010*, pages 15–29, Berlin, Heidelberg. Springer Berlin Heidelberg.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Michael Ginn, Mans Hulden, and Alexis Palmer. 2024. Can we teach language models to gloss endangered languages? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5861–5876, Miami, Florida, USA. Association for Computational Linguistics.

Carla Victoria Jara. 2018. *Gramática de la lengua bribri*. éditeur non identifié.

Ministerio de Educación de la Provincia de Corrientes. 2022. *Avañe'ẽ del Taragui: Diccionario guaraní-español, español-guaraní*. Ministerio de Educación de la Provincia de Corrientes, Corrientes, Argentina. Coordinación de Educación Intercultural Bilingüe.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, and 7 others. 2024. Dinov2: Learning robust visual features without supervision. *Preprint*, arXiv:2304.07193.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Neural Information Processing Systems (NIPS)*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. 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. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

SIL International. 2002. Diccionario náhuatl de la sierra norte de puebla.

SIL International. 2012. Diccionario huichol–español.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, and 1 others. 2026. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Chihiro Taguchi and Richard Sproat. 2026. Creating conlangs to probe the metalinguistic grammatical knowledge of llms. *Preprint*, arXiv:2510.07591.

David H. Tuggy. 2004. Náhuatl: Lecciones para principiantes. Originally published in 1991. Electronic edition copyright 2004.

Justin Vasselli, Arturo Martínez Peguero, Junehwan Sung, and Taro Watanabe. 2024. Applying linguistic expertise to LLMs for educational material development in indigenous languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 201–208, Mexico City, Mexico. Association for Computational Linguistics.

Justin Vasselli, Arturo Mp, Frederikus Hudi, Haruki Sakajo, and Taro Watanabe. 2026. Measuring linguistic competence of LLMs on indigenous languages of the Americas. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 287–296, Rabat, Morocco. Association for Computational Linguistics.

Shigeto Yoshida and Angel Abraham Ucan Dzul. 2025. *Diccionario de Uso del Maya Yucateco*, primera edición corregida edition. Publicación independiente, México. Open Educational Resource (REA).