CaMMT: Benchmarking Culturally Aware Multimodal Machine Translation

Anonymous ACL submission

Abstract

Translating cultural content poses challenges for machine translation systems due to the differences in conceptualizations between cultures, where language alone may fail to convey sufficient context to capture region-specific meanings. In this work, we investigate whether images can act as cultural context in multimodal translation. We introduce CaMMT, a human-curated benchmark of over 5,800 triples of images along with parallel captions in English and regional languages. Using this dataset, we evaluate five Vision Language Models (VLMs) in text-only and text+image settings. Through automatic and human evaluations, we find that visual context generally improves translation quality, especially in handling Culturally-Specific Items (CSIs), disambiguation, and correct gender usage. By releasing CaMMT, we aim to support broader efforts in building and evaluating multimodal translation systems that are better aligned with cultural nuance and regional variation.

1 Introduction

Translation brings cultures into contact. It usually involves deciding how much of the foreignness to keep in the resulting translation, and invariably involves blending cultures to some degree (Aixela, 1999). As pointed out by Hershcovich et al. (2022), part of the difficulty arises from the different conceptualizations that each culture holds. Translators must, therefore, choose suitable strategies for adapting vocabulary as well as deciding whether to conserve or substitute foreign elements. Conforming the source text to the target culture by substituting unknown elements with familiar ones can ease comprehension, yet it simultaneously erases traces of the original culture (Venuti, 2003). Conversely, ignoring an adequate vocabulary choice that accounts for regional variation in the target language risks misinterpretation, as lexical choice directly shapes how readers understand a text (Szymańska, 2017).

Text-only machine translation inherits this dilemma with limited contextual knowledge to ground these translation decisions. However, images can supply that missing extra-linguistic information; visual reference may act as a cultural proxy, revealing a region's set of values (Yadav et al., 2025) as well as social practices and material culture, such as clothing, architecture, and food. With photography being thought of as a form

Figure 1: Examples of CaMMT dataset

of translation from reality into images (Gagliano, 2008), we hypothesize that images can capture additional information that language alone may struggle to encode.

Multimodal Machine Translation (MMT) (Specia et al., 2016) attempts to embed this information by grounding source sentences with images. CoMMuTe (Futeral et al., 2022) provides an evaluation framework for MMT centered on lexical disambiguation, but does not address broader cultural nuances, leaving questions about how visuals influence translation in culturally grounded settings largely unanswered.

In this work, we present CAMMT (**C**ulturally-**A**ware **M**ultimodal **M**achine **T**ranslation Benchmark), the first human-curated MMT corpus with triples across 19 languages of culture-related captions spanning 23 regions worldwide. Additionally, we study the impact of visual grounding for culture-aware multimodal machine translation in Vision–Language Models (VLMs).

To frame our study, we pose the following **research questions**:

- **RQ1** : How does visual grounding impact translation quality and native speakers' preferences across different languages in culturally-relevant settings?

- **RQ2** : What reasons drive preferences between text-only and multimodal translations?

- **RQ3** : How do VLMs perform in MT compared to each other and to state-of-the-art machine translation models?

- **RQ4** : Which translation strategies do native speakers prefer in the case of Culturally-Specific Items (CSIs)?

Our contributions are as follows:

- **Culturally-Specific MMT Dataset**: We present CAMMT, a human-curated corpus of 5,817 image-captions triples, where the captions are collected for both English and regional languages. The dataset spans 19 languages and 23 regions. Additionally, a dedicated dataset of over 1,550 image-captions tuples is collected, covering different translation strategies for CSIs[1].

- **Insights into visual grounding for culture-aware translation**: We evaluate five VLMs on CAMMT to assess the impact of visual grounding on human preferences and performance in automatic metrics. Through these experiments, we find that visual context improves translation outputs. Native speakers tend to prefer multimodal translations because they better preserve CSIs, resolve lexical ambiguities, and reflect correct gender usage, highlighting aspects of quality that standard metrics often fail to fully reflect.

2 Related Work

In translation studies, CSIs (Aixela, 1999) refer to words or concepts that lack direct equivalents or carry different connotations in the target culture. These often arise when cultural references embedded in the source language do not directly exist or are understood differently in the target language. When translating CSIs, translators typically adopt one of two strategies: *substitution*, which adapts the foreign element into a culturally familiar counterpart to reduce its strangeness; or *conservation*, which preserves the original cultural reference, maintaining the source text's foreignness and exposing readers to its original context (Aixela, 1999; Venuti, 2003).

Efforts to incorporate cultural awareness into machine translation have been addressed in specific domains such as cultural adaptation in recipe translation (Cao et al., 2024; Zhang et al., 2024). Yao et al. (2023) generalized beyond this scope by constructing an evaluation dataset by automatically extracting CSIs from Wikipedia to study how LLMs and MT systems handle cultural references. However, the dataset is restricted to a smaller number of languages, automatically generated without input from regional speakers, and does not consider the effect of visual context on translation decisions.

Recent benchmarks such as CVQA (Romero et al., 2024), CulturalVQA (Nayak et al., 2024), ALM-bench (Vayani et al., 2025), and FoodieQA (Li et al., 2024) demonstrate growing progress in regional image understanding within VLMs. However, none of these works study how imagery can affect translation across cultures.

Together, these studies motivate our evaluation on the multimodal translation ability of VLMs, where we introduce the first human-curated benchmark for culturally grounded MMT.

[1]We will make the corpus publicly available.

Figure 2: Examples where the text+image translation was marked as preferred over the text-only setting. Image (a) is generated by Gemma3 27B, while (b) and (c) are from Qwen2.5-VL 32B. Examples (a) and (c) illustrate translations preferred because of CSI-preservation, while (c) was preferred as the correct gender of "athlete" was used when translating from English to Arabic (a gender-marking language).

3 CaMMT Dataset

CVQA (Romero et al., 2024) is a visual question answering dataset comprising more than 10,000 questions across 39 country-language pairs. The questions within CVQA are formulated in both regional languages and English, classified into 10 distinct categories. To develop CaMMT, we utilized CVQA's question-answer pairs and transformed them into declarative statements using Gemini 2.0 Flash (Team et al., 2024) to generate parallel caption pairs in English and regional languages.

Human Annotations To ensure the correctness of the generated caption pairs, we involved native speakers (annotators) for each of the languages that participated in the original data curation and are co-authors of this paper. The annotators were asked to complete three tasks: **(1)** evaluate and ensure the grammatical correctness and parallelism of the generated pairs in English and regional language by correcting captions when needed, **(2)** ensure Culturally-Specific Items in regional language captions are preserved and **(3)** categorize each of the pairs into three categories: **(a)** Not culturally relevant sentences, **(b)** do not contain any Culturally-Specific Items (Non-CSI) or **(c)** contain Culturally-Specific Items (CSI).

We followed the work of Aixela (1999) to provide the annotators with a definition of CSIs. To achieve a better coverage of translation strategies for CSIs (as previously discussed in Section 2), we asked them to further categorize sentence pairs with CSI into **(i)** CSI with possible translation - captions containing CSIs that have culturally equivalent terms that can convey an equivalent meaning when translated into English and **(ii)** CSI forced translation - captions containing CSIs that do not have any equivalent translation in English. For each sentence containing CSI items, we asked the annotators to provide both *conserved* (retaining CSIs) and *substituted* (using familiar equivalents) English translations, then select their preferred version as native speakers.

For example, in the possible translation category, the Mexican term *tianguis* can be translated as flea market, as in: "The name for this type of Mexican informal market is *tianguis*" (*conserved*) or "The name for this type of Mexican informal market is flea market" (*substituted*). In contrast, a forced translation case is: "The name of the Egyptian food in the glass plate in the picture is *Hawawshi*" (*conserved*) and "The name of the Egyptian food in the glass plate in the picture is *minced meat sandwich*" (*substituted*), where the original term lacks an exact English equivalent. For forced translations, they provide the closest possible English approximation. We provide the annotation guidelines in Appendix G.

Dataset Statistics In total, CaMMT comprises 23 regions with 19 different languages, with a total of 5,817 triples with additional 1,550 with *conserved* and *substituted* CSIs for targeted analysis. We present representative samples in Figure 1, and report the number of triples per language included in the corpus in Appendix A.

4 Methodology

This section outlines the experimental methodology, including model selection and evaluation setup of both human and automatic assessments used to measure translation quality across text-only and multimodal conditions.

4.1 VLMs for Multimodal Machine Translation

As discussed in Section 2, task-specific MMT models are limited by their training data, often lacking coverage for many languages. On the other hand, LLMs have demonstrated strong performance in machine translation across multiple language pairs (Hendy et al., 2023; Zhu et al., 2024). As the paradigm shifts from text-only to multimodal LLMs which can process both text and images (VLMs), we explore their potential for multimodal translation, particularly in culturally grounded scenarios.

Model	Setting	De	Fr	Ru
mBART+MT	T	25.9	38.2	
VGAMT	T+I	29.3 (+3.4)	32.2 (-2.3)	
NLLB-600M	T	36.2	39	19.4
NLLB-3.3B	T	40.8	41.4	23.1
Gemma3 27B	T	39.1	41.7	23.2
Gemma3 27B	T+I	44.9 (**+5.8**)	49.6 (**+7.9**)	31.7 (**+8.4**)
Qwen2.5 VL 32B	T	32.8	33.1	21.4
Qwen2.5 VL 32B	T+I	37.0 (**+4.2**)	41.7 (**+8.6**)	24.1 (**+2.7**)
Gemini 2.0 Flash	T	42.6	43.1	26.8
Gemini 2.0 Flash	T+I	49.9 (**+7.3**)	55.2 (**+12.1**)	32.3 (**+5.5**)

Table 1: BLEU scores reported on CoMMuTe for text-only (T) and text+image (T+I) settings. The scores from mBART+MT and VGAMT (an MMT system based on BART) are as reported by Futeral et al. (2022), who does not evaluate Russian.

To initially assess the ability of VLMs in grounding translations using images, we conduct a control experiment on the CoMMuTe dataset, comparing VLMs against strong task-specific MT and MMT baselines. In the text-only setting, models are prompted to translate from English to the target language. In the text+image setting, they are additionally provided with an image and prompted to use it as context for the translation (see Appendix D for prompt details). Importantly, no further instructions are given regarding the nature of the disambiguation task. We evaluate five VLMs: Gemma 3 27B and 12B (Team et al., 2025), Qwen 2.5-VL 32B and 8B (Bai et al., 2025), and Gemini 2.0 Flash (Team et al., 2024).

Results presented in Table 1 demonstrate consistent and significant improvements in the performance of VLMs in the text+image over text-only setting. Moreover, the BLEU scores achieved by VLMs match or surpass those of NLLB-600M and NLLB-3.3B, strong baselines, as well as dedicated MMT systems. These findings confirm that VLMs can indeed leverage visual context to guide translation decisions. Based on this validation, we continue with VLMs as our testbeds to probe how visual grounding influences translation choices in our culturally relevant dataset, particularly in the handling of CSIs and region-sensitive translation.

4.2 Experimental Setup

We evaluate the five VLMs discussed in the previous section, which cover both closed- and open-weights models at different parameter scales, in both text-only and text+image setups. In the text+image setting, we do not instruct models to use images as a cultural reference, only as additional context, allowing us to observe their default effect in translation. To evaluate the impact of visual input on translation quality, we conduct both *human preference evaluation* and *automatic evaluation* using standard machine translation metrics.

Human Preference Evaluation For 21 of the CAMMT regions, native speakers are presented with anonymized translations from three models—Qwen2.5-VL 32B, Gemma 3 27B, and Gemini 2.0 Flash—generated under both text-only and text+image settings. For each instance, they select the preferred translation and specify the reason for their preference from a predefined set: "*CSI is preserved*," "*Correct gender*," "*Disambiguates word*," or "*Regionally appropriate phrasing*". We identified this set of reasons based on an analysis carried out in preliminary experiments on a subset of languages. Annotators are also allowed to specify *other* reasons if none of the previous reasons explain the preference. In Appendix H, we present the instructions provided for this evaluation.

Automatic Evaluation We automatically evaluate translation quality using BLEU (Papineni et al., 2002), chrF++ (Popović, 2017), and BERTScore (F1) (Zhang et al., 2019). BLEU and chrF++ are calculated using SacreBLEU (Post, 2018a).

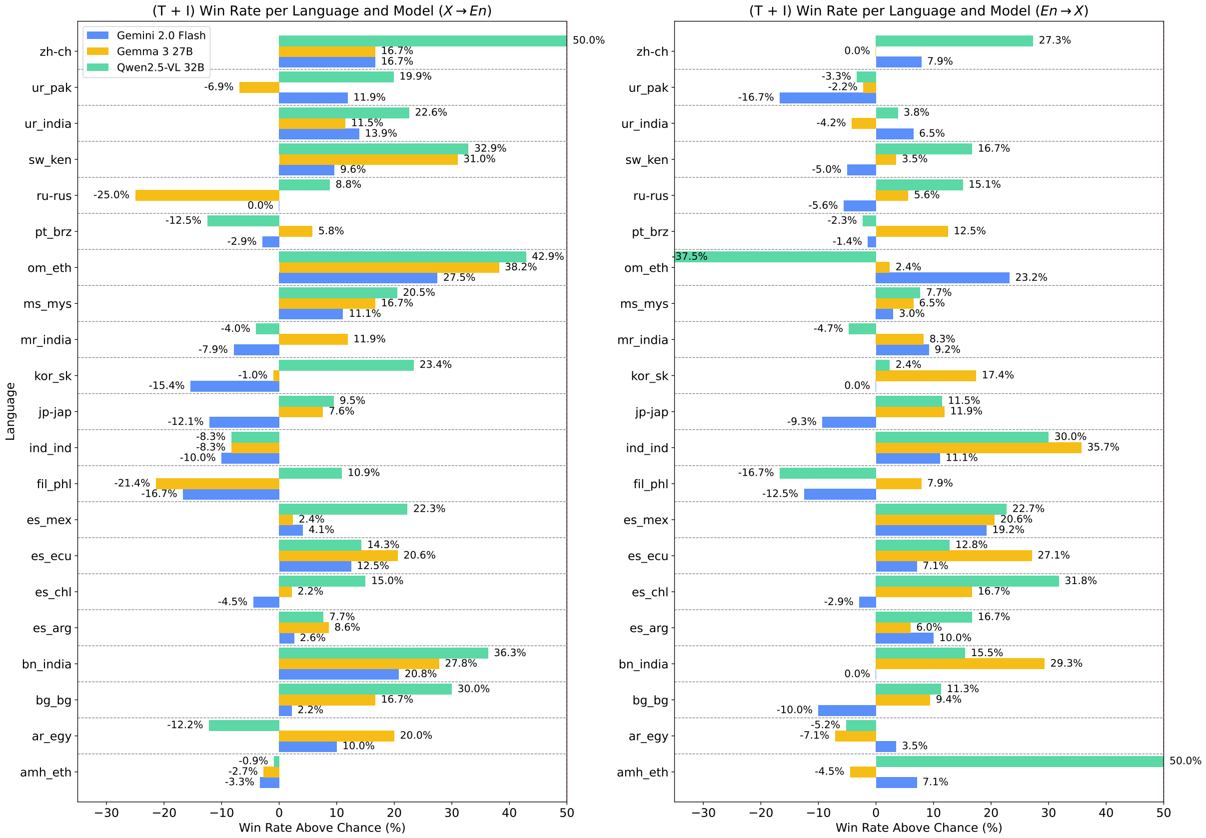


Figure 3: Win rates in human preference evaluation of text+image (T+I) translations over text-only (T) across languages and models. Each bar represents the win rate *above chance* (i.e., over 50%) for cases where native speakers expressed a preference between the two translation conditions. The left plot corresponds to the $X \rightarrow En$ direction, and the right to $En \rightarrow X$.

5 Evaluation

Building on our experimental setup, this section presents the results of our multimodal translation evaluations.

5.1 Effect of Visual Grounding

We begin by assessing translation quality and the effect of visual grounding using both human preference and automatic evaluations.

Human Preferences Evaluation Figure 3 shows native speaker preferences across 21 languages, comparing translations from text-only and text+image settings. We report win rates in instances where a preference was expressed between the two. Overall, translations with visual context are preferred above chance (50%) in the majority of language–model combinations. Specifically, in the $X \rightarrow En$ direction, multimodal outputs are favored in 43 out of 63 experiments. A similar trend holds in the $En \rightarrow X$ direction, where text+image translations are preferred in 42 out of 63 cases. We observe that the text-only output was preferred in 37 out of the 126 total comparisons (29.4%), while 4 out of 126 show a tie in preferences between modalities. These results suggest that visual grounding generally leads to translations that are more aligned with native speaker preferences, *regardless of translation direction*.

Automatic Evaluation We base our main analyses on chrF++ as it has shown higher correlation with human judgments over BLEU (Popović, 2017; Kocmi et al., 2021). Figure 4 reports chrF++ for 23 regions across 19 language pairs. In the $X \rightarrow En$ direction, most regions show improvements with image-grounded translations, with a few exceptions (e.g., Japan, Indonesia, and China). In the $En \rightarrow X$ direction, the benefit of multimodality is less consistent: while Gemini demonstrates clear gains, other models show mixed trends, with no systematic advantage or degradation from adding images. We present the results on BLEU and BertScore in Appendix C, which reflect a similar pattern. Additionally, in Appendix E we report

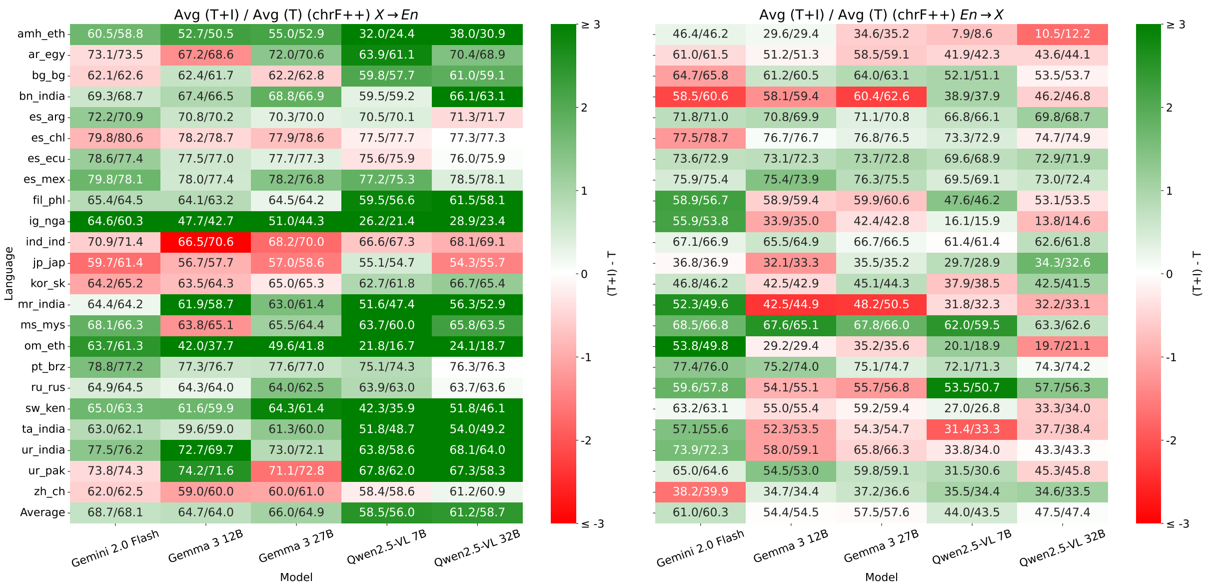


Figure 4: Heatmaps showing average chrF++ scores for text+image (T+I) and text-only (T) settings. Left: Regional-to-English translation. Right: English-to-regional. Each cell shows (T+I) / (T) scores, with color indicating the difference, green shades represent improvements from image input.

		$X \rightarrow En$	$En \rightarrow X$
Model	Setting	chrF++	chrF++
NLLB-600M	T	56.9	50.3
NLLB-3.3B	T	58.9	54.9
Gemini 2.0	T	68.1	60.3
Gemini 2.0	T+I	68.7 (**+0.7**)	61.0 (**+0.7**)
Gemma3 12B	T	64.0	54.5
Gemma3 12B	T+I	64.7 (**+0.7**)	54.4 (-0.1)
Gemma3 27B	T	64.9	57.6
Gemma3 27B	T+I	66.0 (**+1.1**)	57.5 (-0.1)
Qwen2.5 VL 7B	T	56.0	43.5
Qwen2.5 VL 7B	T+I	58.50 (**+2.5**)	44.0 (**+0.5**)
Qwen2.5 VL 32B	T	58.7	47.4
Qwen2.5 VL 32B	T+I	61.2 (**+2.5**)	47.5 (**+0.1**)

Table 2: chrF++ scores averaged across languages for text-only (T) vs multimodal (T+I) settings in both directions ($X \rightarrow En$ and $En \rightarrow X$). The difference (T+I - T) is shown in parentheses.

average chrF++ scores per CVQA-category.

In Table 2, we report the average chrF++ scores across languages (for BLEU and BERT scores, refer to Appendix C). Notably, the addition of image context consistently improves performance across most VLMs, with gains most pronounced in the $X \rightarrow En$ direction. Both evaluations support the conclusion that visual grounding improves translation quality for most languages, particularly when translating from regional languages to English. For the reverse direction, benefits are model-specific: native speakers still tend to prefer image-grounded translations from open-weight models, but this is not always reflected in automatic metrics.

5.2 Reasons Behind Preferences

To better understand how visual input influences translation decisions, we analyze the reasons provided by annotators during the human preference evaluation. Table 3 reports the number of preferences for text-only (T) versus text+image (T+I) translations, broken down by reason.

Across both directions, the primary factors driving preferences toward multimodal translation include *CSI preservation*, *correct gender*, and *lexical disambiguation*. These effects are more pronounced in the $X \rightarrow En$ direction, where VLMs appear better at resolving gender and lexical ambiguities when images are available. The most common reason for preference is more regionally appropriate phrasing. While the T+I setting is still generally favored here, the margin over text-only is smaller, suggesting that visual input has a more modest impact on phrasing compared to other factors.

In preferences explained by annotators with *other* reasons, which include reasons such as grammatical correctness, plural forms, and capitalization, the difference between T and T+I is also minimal, suggesting that images have a greater impact in resolving cultural or semantic ambiguity than in improving general linguistic quality.

We also examine native speakers' preferences in

	$X \rightarrow En$		$En \rightarrow X$		$En_{Sub} \rightarrow X$		$En_{Cons} \rightarrow X$	
	# (T+I / T)	%(T+I)	# (T+I / T)	%(T+I)	# (T+I / T)	%(T+I)	# (T+I / T)	%(T+I)
CSI-preserved	**380** / 277	57.8	**304** / 203	60.0	**223** / 147	60.3	**139** / 79	63.8
Gender	33 / 2	94.3	45 / 36	55.6	10 / 12	45.5	10 / 6	62.5
Disambiguation	**432** / 174	71.3	**239** / 170	58.4	**92** / 78	54.1	**67** / 58	53.6
Phrasing	**1329** / 1046	56.0	**1238** / 1152	51.8	**402** / 394	50.5	**370** / 343	51.9
Others	**368** / 320	53.5	**301** / 289	51.0	56 / **74**	43.1	86 / **98**	46.7

Table 3: Breakdown of human preference reasons across translation directions. For each category, we report the number of times across all languages where a translation with image (T+I) or without image (T) was preferred, as well as the percentages for preferred (T+I). Numbers in bold indicate the modality with the highest preference. Results are shown for both directions and aggregated across languages and models.

the *conserved* and *substituted* splits of CAMMT ($En_{Cons} \rightarrow X$ and $En_{Sub} \rightarrow X$), where the CSI in the source sentence has either been preserved or substituted. In these preferences (labeled with the CSI-preserved reason), speakers more frequently prefer translations from the T+I setting, implying that visual input helps models recover or preserve relevant cultural content.

Models' Behavior on CSIs Beyond human preferences, we further analyze VLMs' ability to handle CSIs in translation. Specifically, we compute the average proportion of translations in which a CSI is preserved when the source sentence contains a substituted ($En_{Sub} \rightarrow X$) or conserved ($En_{Cons} \rightarrow X$) CSI. To do this, we use GPT-4o in a two-step process: (1) extract the CSI from the conserved version of each sample, and (2) check whether it appears in the model-generated translation. Details of this procedure are provided in Appendix I. We then compute the percentage of CSI preservation by dividing the number of retained instances by the total number of samples.

Results presented in Table 4 show that, in the *substituted* setting, the inclusion of images leads to a higher rate of CSI preservation, indicating the model's ability to retrieve appropriate region-specific concepts with visual grounding. In contrast, for the *conserved* setting, the image's influence is less consistent, though it still appears to affect retention patterns to some extent.

5.3 Comparisons of VLMs' Performance

We assess the overall MT performance of VLMs. Firstly, as shown in Figure 4 and Table 2, the best performance is achieved by Gemini, followed by Gemma and Qwen models. Secondly, we compare their performance against a strong text-based MT baseline. As shown in Table 2, compared to NLLB-3.3B, the best-performing VLMs (Gemini 2.0 and Gemma3-27B) achieve comparable or superior translation performance in most metrics, particularly in the $X \rightarrow En$ direction, where they show considerable advantages.

	$En_{Sub} \rightarrow X$		$En_{Cons} \rightarrow X$	
Model	T	T+I	T	T+I
Qwen2.5 VL 32B	20.27	23.05	80.70	77.83
Gemma3 27B	32.49	36.03	90.32	89.33
Gemini 2.0 Flash	41.72	44.05	91.24	90.91

Table 4: Average percentage of preserved CSIs across languages. A value of 100 indicates that all CSIs are retained in the translation; 0 indicates none are preserved. Appendix I reports per-language differences and the average impact of images.

5.4 Human Translation Preferences for CSIs

This section examines native speakers' preferred translation strategies when handling CSIs at the moment of curating CAMMT, where we examine their patterns across languages with different script types. Table 5 presents the percentage distribution of human preferences for *conserved* versus *substituted* translations for Latin and non-Latin scripts under two distinct conditions: when the CSI has a similar equivalent in English (*conserved*), against the case in which there is no equivalent (*forced)*.

For **forced translations**, annotators with Latin script languages strongly favored conservation. Annotators with non-Latin scripts also leaned towards conservation, but were more open to substitution. When possible translations existed, both script types demonstrated a more balanced choice between the two strategies.

6 Discussion

In this section, we revisit our research questions in light of the experimental findings.

RQ1 & RQ2: What is the impact of visual grounding on translation quality, and what factors explain this effect? Visual grounding gener-

	Forced-C	Forced-S	Possible-C	Possible-S
Latin	94±6.7	6±6.7	63±29.4	37±29.4
Non-Latin	75±36.2	25±36.2	53±20.2	47±20.2

Table 5: Translation preferences when curating CaMMT. Annotators classified each CSI as either having a 'Forced' translation or having a 'Possible' translation. 'C' and 'S' represent *conserved* and *substituted* translations, respectively.

ally improves translation quality, particularly in ways that are meaningful to human evaluators. While gains in automatic metrics such as BLEU and chrF++ may appear modest, human preference evaluations tell a richer story: In 85 out of 126 model–language–direction comparisons (67.5%), native speakers preferred multimodal translations, underscoring the value of images for improving cultural and semantic alignment of translations.

Reasons for preference, shown in Table 3, reveal that images are particularly helpful in preserving CSIs, correcting gender, and improving disambiguation. These improvements often involve small textual changes that can significantly impact perceived quality, but may not strongly affect automatic metrics. We conclude that, **visual grounding seems to strengthen translation quality primarily by supporting semantic precision and cultural retention**, benefits that are better captured by human judgments than by traditional MT metrics.

That said, in 37 out of 126 comparisons (29.4%), text-only translations were preferred, indicating that visual input can occasionally degrade translation quality. Understanding why this occurs remains an open question and is an important direction for future work. Moreover, the relatively small gains in automatic metrics are consistent with patterns observed in earlier multimodal MT studies (Futeral et al., 2022), underscoring the need for improved evaluation methods that more accurately reflect the contribution of visual context, particularly in multicultural scenarios.

RQ3 : How do VLMs perform in MT compared to each other and to specialized systems? In terms of Machine Translation performance, all evaluated VLMs matched or exceeded the performance of strong baselines like NLLB-600M and 3.3B, where the closed-source model (Gemini 2.0 Flash) outperformed open-weight models (Qwen2.5 and Gemma3 families). Notably, **we do not observe an evident tradeoff when using VLMs for translation**: they offer competitive performance in standard metrics while simultaneously providing the ability to leverage visual context. This highlights their potential as general-purpose translation systems capable of steering translations using multimodal inputs without sacrificing textual quality.

RQ4 : Which translation strategies do native speakers prefer in the case of CSIs? Contrary to the predominant research direction in NLP on substitution strategies for unfamiliar CSIs, our findings suggest that native speakers often prefer conservation, especially when no culturally equivalent term exists in English. This trend holds across both Latin and non-Latin scripts, although the latter group shows greater variability, possibly due to transliteration. When equivalents are available, preferences are more balanced, but still do not lean completely toward substitution. These results point to the importance of incorporating script-aware translation strategies regarding CSIs in future research, highlighting the need for MT systems to better align with native speaker preferences by adapting conservation and substitution choices to regional and linguistic contexts.

7 Conclusions

We present CaMMT, a human-curated dataset for Multimodal Machine Translation that encompasses 19 languages across 23 regions. We evaluated five VLMs at different scales on CaMMT and observed that providing images as auxiliary context generally improves translation quality in ways that native speakers find meaningful. When translations incorporate visual context, they tend to better preserve cultural elements, use correct gender forms, and resolve ambiguities—improvements that automatic MT metrics often miss. However, we also observe a non-trivial number of cases where visual input negatively affects translations. Understanding when and why this occurs remains an important direction for future research.

Our findings also show that annotators tend to favor conserving CSIs, particularly when no clear equivalent exists in English, underscoring the importance of culturally sensitive translation strategies. Future work should incorporate such speaker-aligned choices when designing models and datasets for grounded, culturally aware translation.

Limitations

While CAMMT provides broad language and regional coverage, the number of samples is constrained by the original CVQA dataset. Due to design choices inherited from CVQA, some samples are marked as non-culturally relevant; however, we retain them as they remain useful for evaluating general multimodal machine translation. When curating CAMMT, we relied on a single annotator per region for human annotations, which may introduce subjectivity in CSI assessments and translation preferences. Expanding annotator diversity would likely improve the reliability and objectivity of these judgments. On the evaluation side, we do not evaluate specialized MMT systems, as most lack training data for the 19 languages included. To keep human evaluation feasible across three models, we restrict evaluation to pairwise preferences between text-only and text+image outputs. We do not include Likert-scale judgments of translation quality, relying primarily on automatic metrics for this purpose. Future work should explore how visual grounding affects human perception of translation quality, as well as expand the dataset with more samples per region and involve multiple annotators to improve coverage and objectiveness of cultural relevance and CSI judgments.

References

Javier Franco Aixela. 1999. *4 Culture-specific Items in Translation*, page 52–78. Multilingual Matters.

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. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Yong Cao, Yova Kementchedjhieva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. 2024. Cultural adaptation of recipes. *Transactions of the Association for Computational Linguistics*, 12:80–99.

Paulo Cavalin, Pedro H Domingues, and Claudio Pinhanez. 2025. Sentence-level aggregation of lexical metrics correlates stronger with human judgements than corpus-level aggregation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23532–23540.

Matthieu Futeral, Cordelia Schmid, Ivan Laptev, Benoît Sagot, and Rachel Bawden. 2022. Tackling ambiguity with images: Improved multimodal machine translation and contrastive evaluation. *arXiv preprint arXiv:2212.10140*.

Maurizio Gagliano. 2008. Photography as translation. visual meaning, digital imaging, trans-mediality. *Recherches sémiotiques*, 28(1):29–42.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *Preprint*, arXiv:2302.09210.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, and 1 others. 2022. Challenges and strategies in cross-cultural nlp. *arXiv preprint arXiv:2203.10020*.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Conference on Machine Translation*, pages 478–494.

Wenyan Li, Xinyu Zhang, Jiaang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, and 1 others. 2024. Foodieqa: A multimodal dataset for fine-grained understanding of chinese food culture. *arXiv preprint arXiv:2406.11030*.

David Orlando Romero Mogrovejo, Chenyang Lyu, Haryo Akbarianto Wibowo, Santiago Góngora, Aishik Mandal, Sukannya Purkayastha, Jesus-German Ortiz-Barajas, Emilio Villa Cueva, Jinheon Baek, Soyeong Jeong, and 1 others. Cvqa: Culturally-diverse multilingual visual question answering benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd van Steenkiste, Lisa Anne Hendricks, Karolina Stańczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. *Preprint*, arXiv:2407.10920.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Matt Post. 2018a. A call for clarity in reporting BLEU scores. In *Proceedings of the Conference on Machine Translation: Research Papers*, pages 186–191.

Matt Post. 2018b. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, and 1 others. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *arXiv preprint arXiv:2406.05967*.

Lucia Specia, Stella Frank, Khalil Sima'An, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *First Conference on Machine Translation*, pages 543–553. Association for Computational Linguistics (ACL).

Izabela Szymańska. 2017. The treatment of geographical dialect in literary translation from the perspective of relevance theory. *Research in Language*, 15(1):61–77.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1331 others. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

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.

Lawrence Venuti. 2003. *The Translator's Invisibility*. Routledge.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

Srishti Yadav, Zhi Zhang, Daniel Hershcovich, and Ekaterina Shutova. 2025. Beyond words: Exploring cultural value sensitivity in multimodal models. *Preprint*, arXiv:2502.14906.

Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. 2023. Benchmarking machine translation with cultural awareness. *arXiv preprint arXiv:2305.14328*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with bert. In *Proceedings of ICLR*.

Zhonghe Zhang, Xiaoyu He, Vivek Iyer, and Alexandra Birch. 2024. Cultural adaptation of menus: A fine-grained approach. *arXiv preprint arXiv:2408.13534*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. *Preprint*, arXiv:2304.04675.

A CaMMT Statistics

In Table 6, we report the number of samples per region in CaMMT, their language and writing script. In addition, we include number of samples that are: CSIs (Forced translation or Has possible translations), Culturally Relevant (non-CSI), or Not culturally relevant.

We use statistics of this dataset (specifically, scripts of each language), to understand translations choices of annotators when it comes to conserving or substituting CSIs.

B Experimental Setting

We employ the *transformers* library (Wolf et al., 2020) for all the experiments conducted on open-weight models. The specific identifiers for each model are shown in Table 7. All experiments are run on single NVIDIA A100 80G card. We set temperature to 0.0 for generating the translations.

Following Cavalin et al. (2025), we evaluate chrF++ and BLEU scores at sentence-level using SacreBLEU (Post, 2018b). BERTScore is calculated using *bert-base-multilingual-cased* model for all languages[1] at corpus-level.

[1]https://github.com/Tiiiger/bert_score

Model	Hugging Face Identifier
Gemma 3 27B[2]	`google/gemma-3-27b-it`
Gemma 3 12B[3]	`google/gemma-3-12b-it`
Qwen2.5-VL 32B[4]	`Qwen/Qwen2.5-VL-32B-Instruct`
Qwen2.5-VL 7B[5]	`Qwen/Qwen2.5-VL-7B-Instruct`
AyaVision 32B[6]	`CohereForAI/aya-vision-32b`
AyaVision 8B[7]	`CohereForAI/aya-vision-8b`

Table 7: HuggingFace identifiers for models used in our experiments.

C BLEU and BertScore metrics across models

In Table 8, we calculate BLEU and BERTScore metrics for both MMT and text-based translations averaged across languages for all models. We also present heatmaps in Figure 8 showing the results for each language, providing a comparison between the performance of MMT and text-based settings.

D Translation Prompts

In our experiments, we use two types of prompts for translation tasks: text-only translation (MT) and multimodal translation (MMT). The prompts are defined as follows:

```
PROMPT_MT = '''Translate the following
    sentence from {source} to {target}.
    Provide ONLY the translated text,
    with no additional information,
    explanation, or context.
"{sentence}"
'''
```

```
PROMPT_MMT = '''Translate the following
    sentence from {source} to {target}
    using the provided image as
    additional context. Provide ONLY the
     translated text, with no additional
     information, explanation, or
    context.
"{sentence}"
'''
```

Where PROMPT_MT was used for text-only translation (T) and PROMPT_MMT was used for multimodal translation with text and image (T+I).

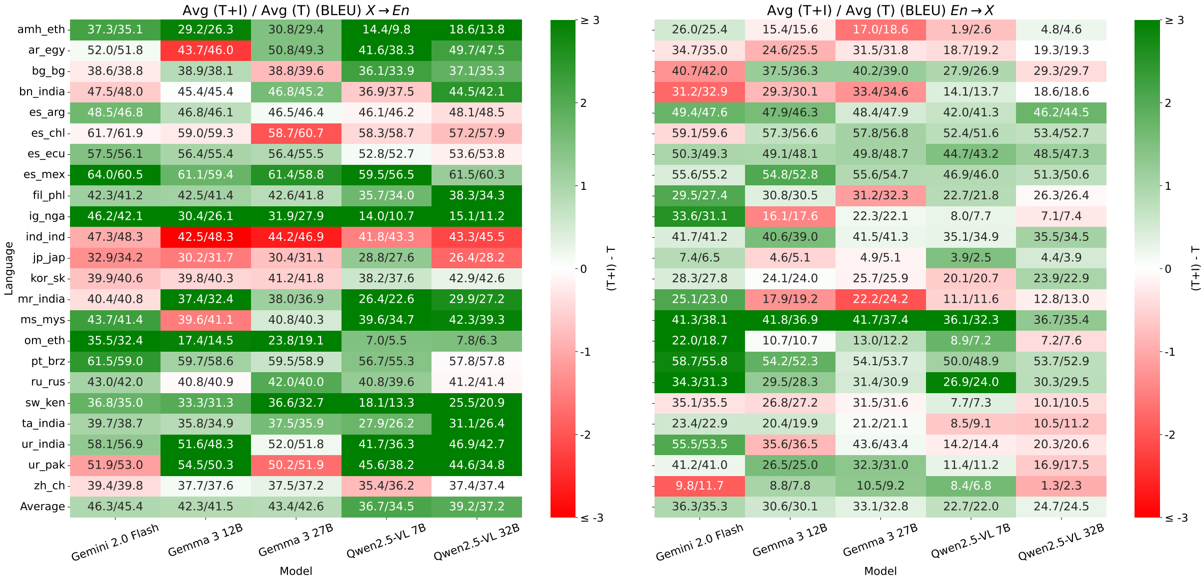
[2]https://huggingface.co/google/gemma-3-27b-it
[3]https://huggingface.co/google/gemma-3-12b-it
[4]https://huggingface.co/Qwen/Qwen2.5-VL-32B-Instruct
[5]https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct
[6]https://huggingface.co/CohereForAI/aya-vision-32b
[7]https://huggingface.co/CohereForAI/aya-vision-8b

E Comparison between categories measured by chrF scores

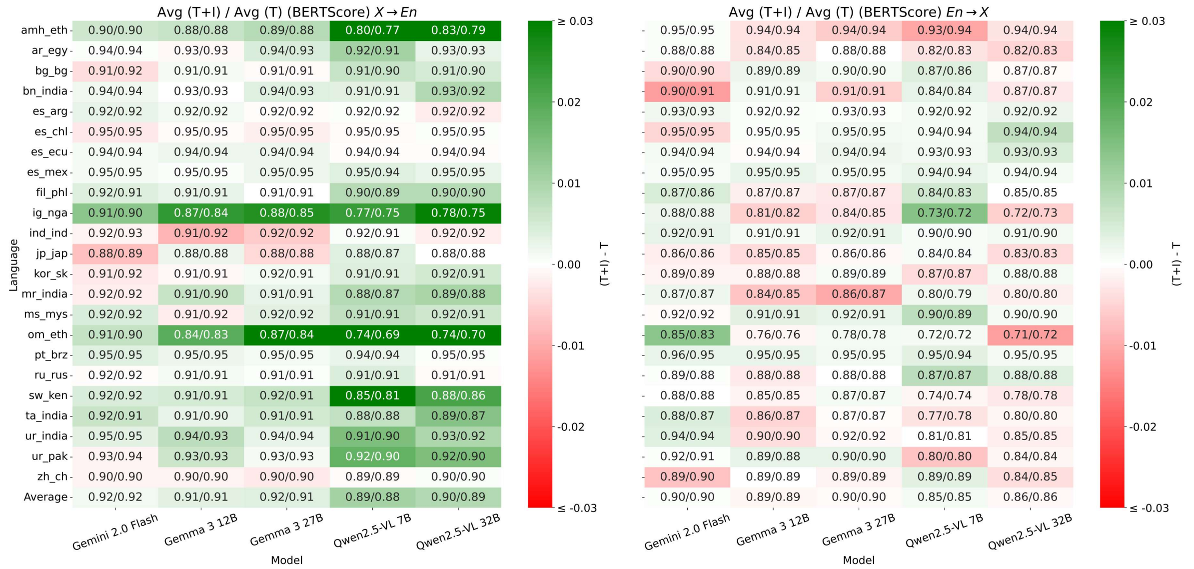
The original CVQA dataset encompasses questions across 10 diverse categories: vehicles, food, people, sports, plants & animals, objects, brands, geography, tradition, and pop culture. Figure 6 shows automatic evaluation using CHRF++ scores across models and CVQA categories.

Language-Region	Script(s)	Size	CSI		Culturally Relevant (non-CSI)	Not culturally relevant
			Forced	Possible		
Amharic-Ethiopia	Ge'ez	234	31	49	97	57
Arabic-Egypt	Arabic	203	16	8	95	84
Bengali-India	Bengali	286	54	31	61	140
Bulgarian-Bulgaria	Cyrillic	369	8	19	90	252
Chinese-China	Hanzi	308	26	18	152	112
Filipino-Philippines	Latin (Rumi)	203	26	29	20	128
Igbo-Nigeria	Latin	200	22	41	62	75
Indonesian-Indonesia	Latin (Rumi)	202	29	7	81	85
Japanese-Japan	Kanji	203	46	26	51	80
Korean-South Korea	Hangul	290	51	11	103	125
Malay-Malaysia	Latin (Rumi)	315	48	40	196	31
Marathi-India	Devanagari	202	27	25	99	51
Oromo-Ethiopia	Latin	214	51	70	93	0
Portuguese-Brazil	Latin	284	46	31	203	4
Russian-Russia	Cyrillic	200	31	26	31	112
Spanish-Argentina	Latin	265	32	50	55	128
Spanish-Chile	Latin	234	34	49	73	78
Spanish-Ecuador	Latin	362	12	60	70	220
Spanish-Mexico	Latin	323	12	67	94	150
Swahili-Kenya	Latin	271	43	99	124	5
Tamil-India	Tamil	213	32	16	44	121
Urdu-India	Perso-Arabic	220	27	22	97	74
Urdu-Pakistan	Perso-Arabic	216	24	28	120	44

Table 6: **Languages covered in CAMMT and Dataset statistics**: including writing script, region, number of samples, and CI counts. Each region was annotated by native speaker.



(a) BLEU scores comparison



(b) BERT scores comparison

Figure 5: Heatmaps showing the difference in average BLEU and BERT scores for text+image (T+I) and text-only (T) settings. Left: Regional-to-English translation. Right: English-to-regional. Each cell shows (T+I) / (T) scores, with color indicating the difference, green shades represent improvements from image input.

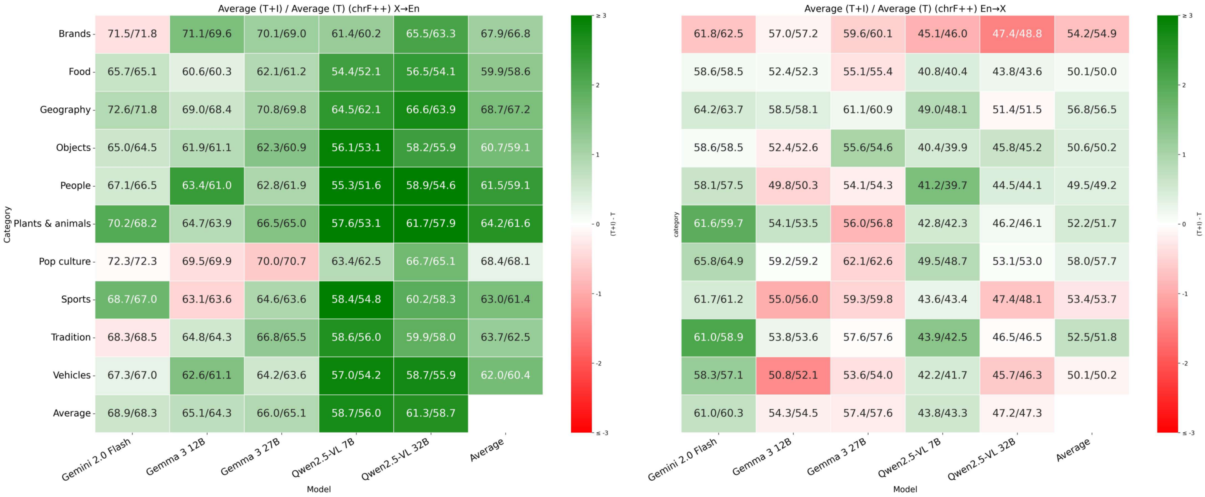


Figure 6: Heatmaps showing the difference in average chrF++ scores for text+image (T+I) and text-only (T) across categories and models. Left: Regional-to-English translation. Right: English-to-regional. Each cell shows (T+I) / (T) scores, with color indicating the difference, green shades represent improvements from image input.

		$En \rightarrow X$		$X \rightarrow En$	
Model	Setting	BLEU	BERT	BLEU	BERT
NLLB-3.3B	T	28.98		36.01	
Gemini 2.0	T	35.70	0.9	45.60	0.92
Gemini 2.0	T+I	36.56 (**+0.87**)	0.9	46.51 (**+0.91**)	0.92
Gemma3 12B	T	30.03	0.89	41.46	0.91
Gemma3 12B	T+I	30.62 (**+0.59**)	0.89	42.33 (**+0.87**)	0.91
Gemma3 27B	T	32.78	0.9	42.60	0.91
Gemma3 27B	T+I	33.17 (**+0.38**)	0.9	43.45 (**+0.85**)	0.92 (**+0.01**)
Qwen VL 7B	T	21.96	0.85	34.57	0.88
Qwen VL 7B	T+I	22.68 (**+0.72**)	0.85	36.71 (**+2.14**)	0.89 (**+0.01**)
Qwen VL 32B	T	24.39	0.86	37.32	0.89
Qwen VL 32B	T+I	24.65 (**+0.26**)	0.86	39.21 (**+1.89**)	0.90 (**+0.01**)

Table 8: BLEU and BERT scores averaged across languages for text-only (T) vs multimodal (T+I) settings in both directions ($En \rightarrow X$ and $X \rightarrow En$). The difference (T+I - T) is shown in parentheses.

In the $En \rightarrow X$ direction, the impact of visual input is notably selective. Only the *geography* and *traditions* categories consistently benefit from multimodal input across all models. The $X \rightarrow En$ direction presents a different pattern, where visual context provides substantial benefits across most categories. Interestingly, two categories consistently show minimal benefits from visual input in $X \rightarrow En$ direction: *brands* and *pop culture*.

F License

CVQA (Mogrovejo et al.) allows using their QA data for research purposes, which is the aim of this work. We do not include the images in our release, and instead include their ID in CVQA. Refer to Mogrovejo et al. for the licenses of the images, as each has a specific license.

The CaMMT corpus is exclusively for academic research, under the Creative Commons Attribution-NonCommercialShareAlike 4.0 International (CC BY-NC-SA 4.0) license.

G CaMMT Data Curation Guideline

Guidelines for cleaning captions

Thank you for participating in this project!

You will receive items from the CVQA dataset specific to your region. Each item includes two *automatically generated* captions:

- One caption in English
- One caption in your regional language

Each caption describes an image depicting culturally-specific content. Your task is to review and correct these captions as needed. You have one week to complete this task.

Task Guidelines:

For each item, fix `regional_corrected` and `English_corrected`, ensuring the following:

1. **Grammatical Correctness and Parallelism:**
   - Ensure **both captions** (English and regional language) are grammatically correct.
   - Ensure **both captions** are as parallel as possible.
   - Correct grammatical errors, awkward phrasing, and unclear meanings.
2. **Regional Language Caption (** `regional_corrected` field**):**
   - Retain the **cultural specificity** of the original QA pair accurately.
   - Preserve culturally-specific items (CSIs) clearly.
   - Avoid unnecessary naturalization or cultural substitution.

After fixing `regional_corrected` and `English_corrected`**,** you need to do the following.

3. **English Caption Categories:**

   Categorize each English caption into one of the highlighted categories by selecting it in the *Category* column and take action accordingly:

   - **Not culturally-relevant sentence**
     - Example: *"This bank was founded in 1898."*
     - Only ensure grammatical correctness and parallelism. *(Leave* `Conserved_translation` *and* `Substituted_translation` *fields blank.)*
   - **Non-CSI** (Does not contain a Culturally-Specific Item)**;**
     - Includes widely borrowed words (e.g., *"falafel"*), named entities (e.g., *"El Santo"*), or well-known equivalents (e.g., *"Great Pyramids"*).
     - Ensure grammatical correctness and parallelism only. *(Leave* `Conserved_translation` *and* `Substituted_translation` *fields blank.)*
   - **CSI (Culturally-Specific Item)**
     These are culturally-specific terms with no direct equivalent or carrying different connotations in English (See Appendix). Categorize them further as:
     1. **CSI with possible translation:** Has a culturally-equivalent that can convey an *equivalent* meaning.
     2. **CSI forced translation:** Does not have any equivalent in English, to translate it we would need to use another concept which may have an impact on the meaning

Guidelines for cleaning captions 1

Figure 7: Annotation guideline

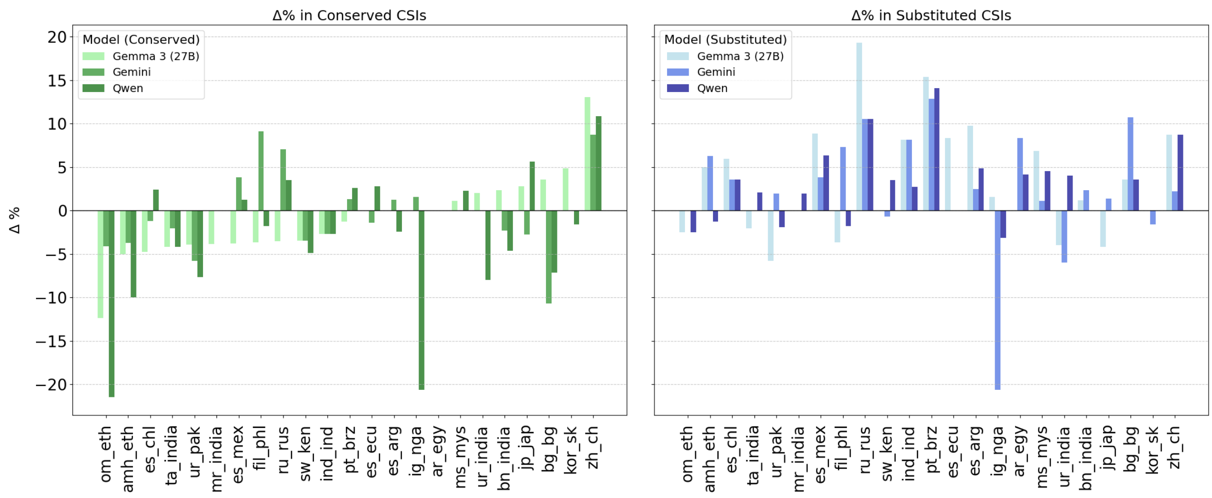


Figure 8: Differences in CSI retention percentages between text-only and text+image settings for Gemma 3 (27B), Gemini 2.0 Flash, and Qwen2.5-VL 32B across languages. Left: conserved CSIs; right: substituted CSIs.

H Human Preference Evaluation Instructions

Instructions for Translation Evaluation Task

You are tasked with selecting your preferences on the provided evaluation sheet. Each item includes:

- A **source sentence**
- Two **model translations** (Model A and Model B)
- The **target translation** you previously created
- A **reference image** to help you disambiguate or contextualize cultural elements

Please fill the following columns:

1. **Translation Quality**:
   - Indicate whether one translation is better, both are good, or both are bad/unintelligible.
2. **Translation Preference**:
   - Choose **A** or **B** based on which translation you prefer.
   - Try to select one even if both are equally good or bad.
3. **Reason for Preference**:
   - If you selected one translation as better, choose a reason from the predefined list.
   - If no reason applies, explain briefly in the "Other Reasons" column (a few words are enough).
4. In the case of 'both are good':
   - **If both translations are essentially identical and equally good (e.g., differing only in word order), you may leave the preference entry blank.**

I CSI Retention Evaluation

In this section, we report the per-language analysis of the impact of visual input on the retention of CSIs across languages (comparing text-only and text+image settings) and describe the algorithm for CSI identification in translations.

For each language and model, we compute the difference in CSI preservation rates using translations from the conserved and substituted splits. As shown in Table 4, and further illustrated in Figure 8, visual input tends to help models recover CSIs in the substituted setting—where the original term is not present in the source sentence, by providing complementary visual cues. In contrast, when translating from the conserved split, where the CSI is explicitly present in the source, we observe no consistent effect from the image across models or languages.

CSI extraction and identification We developed a two-stage approach to evaluate how well machine translation systems preserve CSIs. This methodology leverages large language models to first identify CSIs and then evaluate their preservation in different translation outputs.

Our methodology consists of two key stages:

1. **CSI Extraction**: Automatically identifying culturally specific items using the prompt shown in Box I, which compares conserved translations (containing the CSI) against substituted translations (where the CSI is replaced with a more general term).

2. **CSI Preservation Evaluation**: Determining which of two competing translation systems better preserves the identified CSI when compared to a gold reference, following the evaluation setup in Box I.

For both CSI extraction and evaluation, we utilized GPT-4o with $temperature = 0.0$ to ensure deterministic outputs. The CSI extraction was limited to $max\_tokens = 50$, while we used default token limits for the evaluation task. All processing was performed through the OpenAI API, maintaining consistent parameters across all language pairs and translation systems.

CSI Extraction Prompt

Given two versions of a sentence:

1. A sentence with a culturally specific item (conserved_translation)
2. A sentence where that item has been replaced with a more general term (substituted_translation)

Your task is to identify the culturally specific item (CSI) that appears only in the conserved translation.
Compare the two sentences and extract only the specific culturally-significant word or phrase that was replaced in the substituted version.
Return ONLY the culturally specific item as a single word or phrase, without any explanations, quotation marks, or additional text.
Example:
Conserved: "The person in the picture is a famous **charro** from the state of Jalisco."
Substituted: "The person in the picture is a famous **cowboy** from the state of Jalisco."
Output: **charro**

...

CSI Evaluation Prompt

Given two translations (0 and 1), a gold reference sentence (y), and a culturally specific item (CSI), your task is to:
Evaluate which translation better preserves the CSI from the reference.
Output the results strictly as a JSON list of dictionaries with the following exact structure:

```
[
  {
    "word": [word_in_0, word_in_1, word_in_y],
    "type": "CSI",
    "aligned_translation": "0" | "1" | "None" | "both"
  }
]
```

Where "aligned_translation" values mean:

- **"0"**: Translation 0 better preserves the CSI
- **"1"**: Translation 1 better preserves the CSI
- **"both"**: Both translations include the provided CSI
- **"None"**: None of the translations includes the original CSI (it is replaced by another term)

Example 1:
Input:
y: Este personaje es un **charro** famoso
0: Este personaje es un **vaquero** famoso
1: Este personaje es un **charro** famoso
csi: **charro**
Output:

```
[{"word": ["vaquero", "charro", "charro"], "type": "CSI", "aligned_translation": "1"}]
```

...