Is a Picture Worth a Thousand Words? Exploration and Implementation Considerations for Visual Context in Translation Workflows

Vera Senderowicz Guerra
Welocalize
vera.senderowicz@welocalize.com

Olesia Khrapunova
Welocalize
olesia.khrapunova@welocalize.com

Abstract

Vision-language models (VLMs) have the potential to enhance machine translation (MT) by leveraging visual context alongside text, yet their real utility for production workflows remains unclear. We conduct a unified, multi-condition evaluation of six leading VLMs–both open and proprietary–on two benchmarks (CoMMuTE and CaMMT), targeting lexical and cultural disambiguation respectively. We complement this with a domain-style case study simulating technical documentation localization. Results show that model performance varies widely, and the benefit of relevant images does not transfer uniformly across use cases. Proprietary models are notably sensitive to irrelevant images while open-source models are generally more stable. Contradicting visuals, by contrast, degrade translation across all models. Taken together, our findings show that rigorous evaluation is a necessary precondition for production deployment: metric gains can mask real accuracy losses, model sensitivity to irrelevant images should inform model selection, and avoiding contradicting visuals is a hard requirement for any pipeline.

1 Introduction

The extent to which translation choices depend on the various conceptions of *context* has been widely studied (Dimitriu, 2015; House, 2006; Damaskinidis, 2016). In Machine Translation (MT), context retrieval is especially challenging, as it is inherently constrained by the model's input, with the available context limited to the information the model is exposed to. This has motivated a growing body of work on context-aware MT (Voita et al., 2019; Fernandes et al., 2021; Castilho and Knowles, 2025).

Organizations deploying MT in production workflows increasingly work with content that includes visual context: technical documentation, UI strings, marketing assets. In theory, visual context can inform translation decisions –from guiding lexical choice to signaling register or style– but a natural question arises: should available images be fed to the model? And if so, what happens when the image attached to a segment is not the most relevant to translate its content? Before visual context can be reliably deployed in production MT, we need to understand precisely when and why images help or hurt.

Our results show that relevant images consistently improve lexical disambiguation, but this benefit does not necessarily transfer to other use cases, which is precisely why rigorous evaluation across conditions and scenarios is necessary prior to production deployment. We focus mainly on disambiguation, a task that provides a straightforward-to-measure signal, and evaluate six VLMs from both open and proprietary providers. We make the following contributions:

- A **multi-condition experimental design** (no image, correct image, random image, and contradicting image when available) that disentangles the impact of *image content* versus *image presence*;
- A **unified evaluation protocol** across both *lexical* and *cultural disambiguation*, enabling direct comparison under consistent conditions

and overlapping languages to assess robustness transference across tasks;

- An illustrative **practitioner-oriented case study** testing whether benchmark findings hold in a realistic localization pipeline, illustrating how to measure visual context impact in production-like conditions.

2 Related Work

Recent work has benchmarked visual context for machine translation, including lexical and cultural disambiguation tasks. The CoMMuTE benchmark (Futeral et al., 2023; Futeral et al., 2025) evaluates multimodal translation on sentences with ambiguous words, but compares only text-only and text+image conditions, without establishing whether models exploit image *content* or merely react to image *presence*.

The CaMMT benchmark (Villa-Cueva et al., 2025) performs evaluation on cultural disambiguation and includes controls with unrelated images, finding that they can degrade quality, supporting that gains under the correct image reflect content rather than mere visual input. However, no previous study has jointly assessed both benchmarks, so transfer of visual grounding across these task types remains an open question.

Concurrently, Liu et al. (2025) propose a reasoning-based framework that detects ambiguity before invoking visual context, but do not evaluate model behavior under irrelevant or contradicting images. Multimodal MT has a longer lineage in the WMT shared tasks (Specia et al., 2016; Elliott et al., 2017; Barrault et al., 2018) and the Multi30K dataset (Elliott et al., 2016). Shen et al. (2024) survey the resulting body of work across architectures, training strategies, and evaluation paradigms. We depart from this line by evaluating modern instruction-tuned VLMs with native multimodal interfaces, and by systematically controlling for image content versus mere image presence, a confound noted in prior work (Elliott, 2018) but not part of the shared-task evaluation protocol itself. We additionally bridge benchmark evaluation to production deployment through a domain-style case study; a step that, to our knowledge, remains unaddressed in prior multimodal MT work.

	CoMMuTE	**CaMMT**
Disambiguation type	Lexical	Cultural
Translation direction	EN→X	EN→X
Languages used	FR, DE, AR, ZH, RU	RU, ZH
Items per language	150–162 pairs	44 (ZH); 57 (RU)
Images per item	2 (disambiguating)	1
Reference translations	Correct + incorrect	Regional only

Table 1: Overview of the two evaluation benchmarks.

3 Experimental Design

3.1 Datasets

We perform our evaluations on the two benchmarks that target complementary forms of visual disambiguation in translation, summarized in Table 1.

Each item in **CoMMuTE** consists of an ambiguous English sentence paired with two images that trigger different (unambiguous) target-language translations. For each image, a correct reference translation is provided; each image's correct reference translation constitutes an incorrect translation for the other image in that specific item. See Appendix A for a sample item.

In **CaMMT**, each item pairs a source sentence in the regional language with English reference translations and an image depicting a culturally-specific context. For items containing Culturally-Specific Items (CSIs), the dataset provides two English translation variants: a *conserved* version that preserves the original cultural term, and a *substituted* version that replaces it with a target-language generic equivalent. Since our goal is to perform a direct comparison with CoMMuTE and be able to assess the transfer of capabilities across tasks, we adapt CaMMT by reversing the translation direction and using only the substituted source condition with the generic term, analogous to the CoMMuTE's lexical ambiguity case (see Appendix B for a sample item).

3.2 Conditions and Research Questions

Each sentence is translated under multiple visual conditions, framed around specific research questions:

RQ1 (CoMMuTE + CaMMT): Does a relevant image improve translation? We compare a *correct image* (matching the intended meaning)

Model	Provider	Access
Qwen3-VL-8B	Alibaba	Open
Aya Vision 8B	Cohere	Open
InternVL3-8B	OpenGVLab	Open
GPT-4o	OpenAI	API
Claude Sonnet 4.6	Anthropic	API
Gemini 2.5 Flash	Google	API

Table 2: Models evaluated. Open-source models are all 8B parameters; proprietary model sizes are undisclosed.

against a text-only *no image* baseline.

RQ2 (CoMMuTE + CaMMT): Does any image affect translation? We compare a *random image* (an unrelated image from the same dataset) against the *no image* baseline.

RQ3 (CoMMuTE): Does a misleading image actively degrade translation? We compare a *contradicting image* (encoding the opposite meaning) against the *no image* baseline.

3.3 Models

We evaluate six VLMs (Table 2), spanning multiple providers and model families, both open-source and proprietary. Open-source models were fixed at 8B parameters due to local compute constraints and use greedy decoding (`do_sample=False`), while proprietary models are queried with each provider's default sampling configuration. We deliberately retained these defaults to reflect out-of-the-box practitioner behavior. We acknowledge that this introduces a determinism asymmetry between the two groups, and that proprietary results are subject to additional run-to-run variance not captured in our single-run point estimates. We therefore frame open-vs-proprietary differences as descriptive rather than causal.

All models receive the same prompt, and the image (when applicable) is provided as visual input alongside the text in a single request, using each model's native multimodal interface. The full prompt template is given in Appendix C. Beyond the structured output instruction, we deliberately avoid further prompt engineering, as our goal is to isolate the effect of the image as much as possible.

3.4 Evaluation Metrics

Reference-based translation quality. We compute chrF (Popović, 2015) against both the correct-meaning and incorrect-meaning references (for CoMMuTE) and against the regional reference (for CaMMT), across all conditions. We use it as our primary metric because character n-gram overlap is robust to segmentation and morphological variation across scripts. We additionally report BLEU (Papineni et al., 2002) in the Appendix for comparability with prior MT literature and industrial practice, where it remains a familiar benchmark metric. Given per-language sample sizes (150–162 for CoMMuTE, 44–57 for CaMMT), small absolute differences (within ± 2 chrF) should be interpreted with caution.

Disambiguation gap (CoMMuTE only). Futeral et al. (2023) measure disambiguation via *pair accuracy*. We instead report the *disambiguation gap*, defined as chrF(correct reference) $-$ chrF(incorrect reference), which preserves the magnitude and direction of the effect, enabling finer-grained comparison across models, languages, and conditions, while also disentangling sense selection from overall translation quality, a distinction that raw chrF against a single reference cannot make.

4 Results

We organise results by benchmark: CoMMuTE (lexical; RQ1-RQ3) in Section 4.1, then CaMMT (cultural; RQ1, RQ2) alongside a cross-task comparison in Section 4.2. Runtime and API cost relative to chrF performance are reported in Appendix D, providing guidance for model selection based on both translation quality and time/cost efficiency.

4.1 Lexical Disambiguation (CoMMuTE)

Figure 1 reports chrF deltas averaged across all five languages. Relevant images do improve translation (RQ1), consistent with Futeral et al. (2023): all six models gain between $+2.1$ and $+5.8$ chrF, with GPT-4o and Claude 4.6 benefiting the most. However, not just any image helps (RQ2). Random images leave most models within ± 1 chrF of the baseline, confirming that the RQ1 gains stem from image *content*, not mere image *presence*. Only two models are negatively affected by irrelevant visual input: Gemini 2.5 Flash loses -9.2 chrF on average, and Claude 4.6 drops moderately (-2.4). Misleading images do degrade translation (RQ3): all models lose chrF, with losses generally proportional to their correct-image gains, though not perfectly symmetric. Notably, Gemini 2.5 loses more from contradicting images (-6.4) than it gains from correct ones ($+2.1$), consistent with chrF cap-

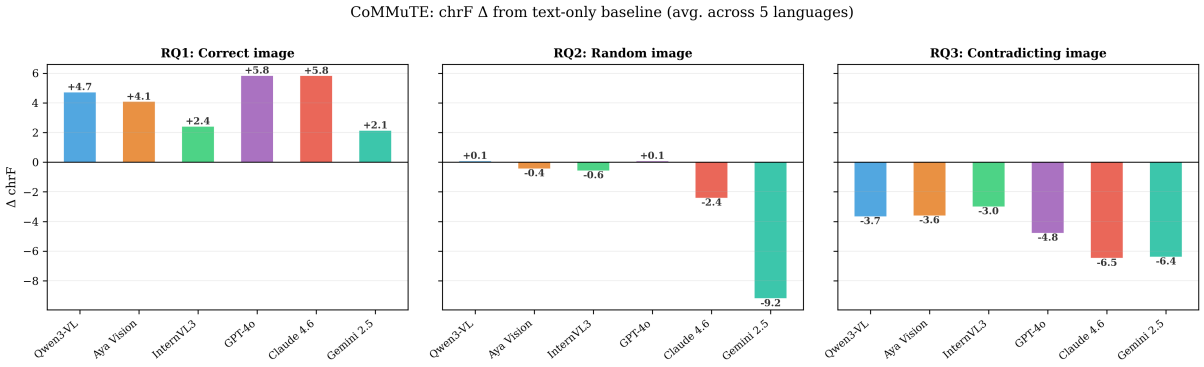

Figure 1: CoMMuTE: mean chrF change from the text-only baseline, averaged across five languages (FR, DE, AR, RU, ZH). Each panel corresponds to one visual condition (RQ1–RQ3). BLEU deltas follow the same pattern (Appendix E); per-language breakdowns and absolute values can be found in Appendix F.

turing translation-wide effects beyond the disambiguated term alone.

Per-language breakdowns (Appendix F) reveal that effect sizes are language- and model-dependent. Correct-image gains are largest on French for all three proprietary models, but open-source models show a different pattern: Qwen3-VL and InternVL3 gain most on Chinese, and Aya Vision peaks on French and German equally. In relative terms, gains tend to be proportionally larger for lower-baseline languages, though the pattern is not uniform across models; the clear exception is Gemini 2.5 Flash, which gains almost exclusively on French. Open-source models are largely robust to irrelevant images, whereas proprietary models diverge: GPT stays mostly robust across languages, while Claude shows random-image penalty and Gemini degrades sharply across all languages. The disambiguation gap heatmap in Appendix G corroborates these. Notably, Gemini's near-zero random-image gaps suggest that its sharp raw chrF drops are not accompanied by a systematic shift in sense selection.

4.2 Cultural Disambiguation (CaMMT) and Cross-Task Comparison

Figure 2 places CoMMuTE and CaMMT chrF deltas side by side on Russian and Chinese, the two languages shared between benchmarks. The correct-image benefit does *not* generalise from lexical to cultural disambiguation (RQ1): CoMMuTE gains are positive on both languages, whereas CaMMT deltas are slightly negative on Russian for all models except GPT-4o, and mixed on Chinese: marginally positive for most (Qwen $+2.5$ being the largest) but negative for Gemini (-2.7). Random-image sensitivity amplifies on CaMMT (RQ2): all three proprietary models degrade on both languages –Claude 4.6 up to -17.1 chrF (Russian) and Gemini 2.5 up to -27.1 (Chinese)– while open-source models remain largely stable on both tasks.

5 Domain-Style Case Study

Benchmark segments are cleaner and shorter than production strings (Vijayan et al., 2024), so understanding where benchmark behavior transfers and where it differs is precisely what makes systematic domain-grounded evaluation a prerequisite for production deployment. To probe feasibility in more realistic settings without requiring proprietary data, we add a small illustrative case study built around a fictitious company, *Nordic-Trans Motors* (heavy-vehicle service documentation). We use the synthetic figure in Appendix I and attach it to two English fragments of fairly comparable length: one **matched** text, whose translation can benefit from the figure and table, and one **mispaired** text that belongs to the same domain but is not related to the image. This setup mirrors a likely production trade-off: without a robust method for automatically verifying image–string relevance, available visuals would need to be linked to multiple segments, resulting in potentially non-helpful –or even harmful– pairings.

Each text is translated as a single unit, without an image and with the synthetic image, under the same prompt as in Section 3. We limit this case study to two languages and a small number of samples, both for practical constraints and to keep the qualitative analysis transparent and easy to follow. Source texts, hypotheses, and human-reviewed reference translations appear in Appendix J. We contrast Qwen and GPT only, as they rank among the

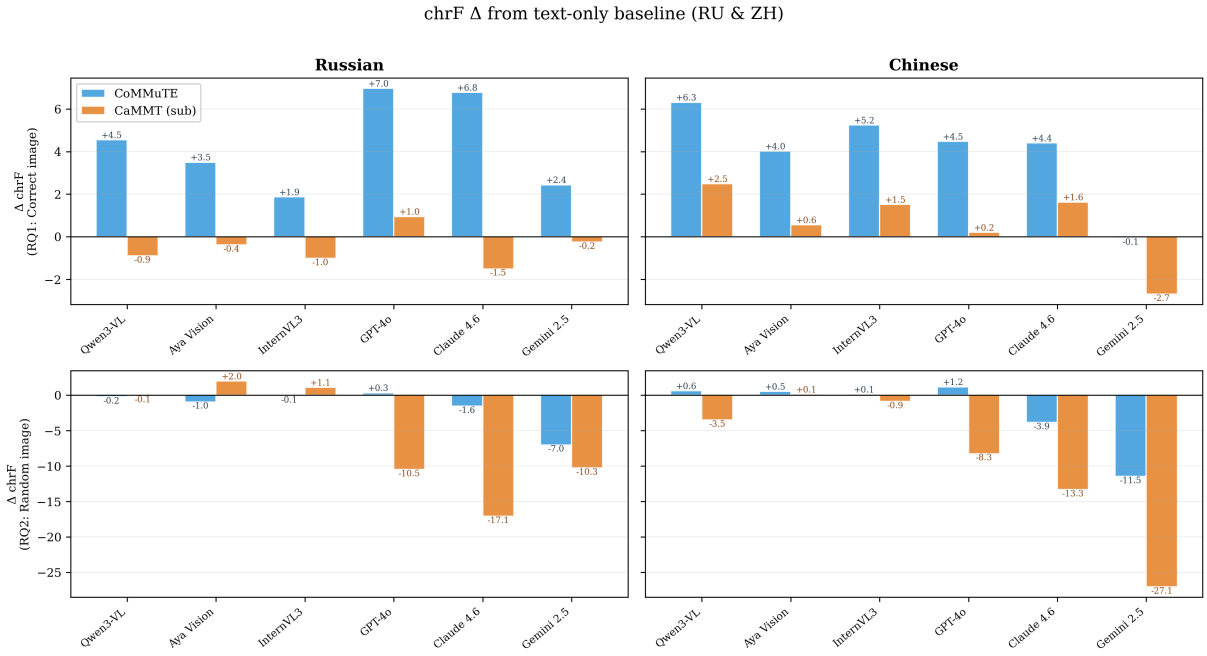

Figure 2: CaMMT and cross-task comparison: mean chrF change from the text-only baseline on Russian and Chinese for correct and random images (RQ1–RQ2). CoMMuTE (blue) vs. CaMMT substituted-source (orange). BLEU deltas (in Appendix H) follow a similar, although not identical, pattern.

strongest correct-image lift for open vs. proprietary families on CoMMuTE (Section 4.1) while remaining comparatively stable under misleading vision inputs. Tables 3 and 4 summarise chrF and BLEU computed over each text as a whole against our synthetic references, for the matched and mispaired conditions, respectively. Each table also reports reference-free COMET-QE (Rei et al., 2020) as a sanity check on fluency and adequacy relative to the English source.

COMET-QE mostly agrees with chrF/BLEU across matched conditions. However, across mispaired conditions, COMET-QE frequently diverges from chrF and BLEU. A human assessment provides the following contextualizing insights:

French. The matched image boosts translation adequacy in aspects beyond disambiguation: GPT moves from an incorrect/non-standard translation for *threadlocker* and an awkward *callout* calque to workshop-credible wording (*frein-filet*, *repère*, *routage*); Qwen similarly restructures phrasing and avoids the *callout* calque. Under the mispaired condition, outputs stay on topic –no accuracy issues stem from the unrelated drawing–, and relevantly, the image's presence still nudges surface wording for French: Qwen corrects a garbled term (*télématicie* → *télémétrie*) and aligns the English *forced*, from *obligatoire* to *forcé*, while GPT shows minor lexical swaps in French and significant metrics' boosts. These shifts suggest that even domain-grounding from an irrelevant image can improve output quality, a hypothesis that only surfaced when moving beyond benchmark conditions. However, COMET-QE edges downward despite chrF/BLEU gains, suggesting these surface improvements do not register as adequacy gains from a source-fidelity perspective.

Russian. Both conditions serve as a cautionary case: with the matched image, GPT shows significant drops in MT metrics, and while Qwen's metrics rise with the figure, the model shifts from one incorrect term to another (*cuff* → *cylinder head* instead of *manifold* in the output). With the mispaired image, neither model's translation meaningfully improves or degrades on human inspection, yet metrics diverge sharply: chrF and BLEU drop for Qwen but rise substantially for GPT, while COMET-QE moves in the opposite direction for both. These disagreements illustrate how metric gains under visual context do not reliably indicate actual translation improvement.

6 Conclusion

Our study finds potential for VLMs in MT workflows, not only for disambiguation but also for broader quality improvements in production-like contexts. Our benchmark evaluation was designed not as an end in itself, but to establish conditions

Lang	Model	chrF		BLEU		COMET–QE	
		text	+fig.	text	+fig.	text	+fig.
FR	Qwen3-VL-8B	58.66	63.37	32.83	36.00	−0.403	−0.266
	GPT-4o	55.96	64.66	14.91	34.66	−0.618	−0.424
RU	Qwen3-VL-8B	38.70	41.84	7.56	11.19	−0.243	−0.083
	GPT-4o	57.66	52.68	23.11	13.72	−0.114	−0.153

Table 3: Matched (on-topic) practical segment. **text** = translation without the image; **+fig.** = translation with that figure. COMET-QE (`Unbabel/wmt20-comet-qe-da`) is a reference-free quality estimator with unbounded scores; only relative comparisons within the same source text are meaningful. Higher values indicate higher estimated quality.

Lang	Model	chrF		BLEU		COMET–QE	
		text	+fig.	text	+fig.	text	+fig.
FR	Qwen3-VL-8B	76.26	77.87	62.07	62.49	−0.227	−0.265
	GPT-4o	83.03	88.73	66.87	76.97	−0.238	−0.258
RU	Qwen3-VL-8B	68.24	63.14	31.82	25.50	−0.147	−0.139
	GPT-4o	62.60	69.55	17.70	41.17	−0.236	−0.137

Table 4: Mispaired (off-topic) practical segment. **text** = translation without the image; **+fig.** = translation with unrelated figure. COMET-QE (`Unbabel/wmt20-comet-qe-da`) is a reference-free quality estimator with unbounded scores; only relative comparisons within the same source text are meaningful. Higher values indicate higher estimated quality.

under which visual context can be trusted: a necessary step before any deployment decision.

Our cross-task comparison shows that a model's ability on one task does not predict its performance on another: models that benefit from images in lexical disambiguation may see limited gains, or even losses, on cultural or domain-specific content. Model robustness also varies: open-source models and GPT remain largely stable when images are irrelevant, whereas Claude and Gemini show significant sensitivity to loosely matched inputs. On the benchmarks, this sensitivity causes quality drops, yet in the French case study, even a mispaired image improved surface quality for both models tested, suggesting that domain-grounding effects can emerge even without direct image–text relevance. Contradicting images were not tested in the case study, but on CoMMuTE they consistently harm outputs across all models. This suggests that reliable image–text matching is a strict requirement for any production pipeline.

Future directions include dedicated, production-scale benchmarks that span domain complexity, content types, and segment lengths across multiple conditions, as well as for advanced, scalable methods to reliably match images with the right source texts (e.g., via multimodal embeddings or hybrid retrieval). Until such tooling matures, deploying visual-context MT at scale will require care, especially when irrelevant or misleading images could negatively impact translation quality.

Overall, our results suggest that the benefits of vision-augmented MT are real but contingent. For practical adoption, organizations should not treat visual context as a universal improvement: they should carefully analyze the content types in their workflows, understand the specific challenges they expect visually-encoded information to address, verify image–text relevance, select models that provide robustness to irrelevant inputs, and design pipelines and evaluation methods targeting quality degradation signals that automatic metrics alone may miss.

7 Limitations

All results come from a single inference run; reported deltas are point estimates without confidence intervals. The domain-style case study uses a single synthetic image, two source texts, two languages, and two models; its findings are intended as illustrative hypotheses rather than generalisable conclusions.

8 Sustainability Statement

All local inference ran on a single Apple M4 Max (14-core, 36 GB unified memory) over approximately 20 h. Proprietary models were queried via API for 9.6 h. Using the Green Algorithms calculator (Lannelongue et al., 2021), we estimate a carbon footprint of 387.93 gCO_2e (2.27 kWh) for the local runs, equivalent to 0.42 tree-months, assuming European energy mix.

References

Barrault, Loïc, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, Belgium, Brussels, October. Association for Computational Linguistics.

Castilho, Sheila and Rebecca Knowles. 2025. A survey of context in neural machine translation and its evaluation. *Natural Language Processing*, 31(4):986–1016.

Damaskinidis, George. 2016. The visual aspect of translation training in multimodal texts. *Meta*, 61(2):299–319.

Dimitriu, Rodica. 2015. The many contexts of translation (studies). *Linguaculture*, 6(1):5–23, Jun.

Elliott, Desmond, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74. Association for Computational Linguistics.

Elliott, Desmond, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In Bojar, Ondřej, Christian Buck, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, and Julia Kreutzer, editors, *Proceedings of the Second Conference on Machine Translation*, pages 215–233, Copenhagen, Denmark, September. Association for Computational Linguistics.

Elliott, Desmond. 2018. Adversarial evaluation of multimodal machine translation. In Riloff, Ellen, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2978, Brussels, Belgium, October-November. Association for Computational Linguistics.

Fernandes, Patrick, Kayo Yin, Graham Neubig, and André F. T. Martins. 2021. Measuring and increasing context usage in context-aware machine translation. In Zong, Chengqing, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6467–6478, Online, August. Association for Computational Linguistics.

Futeral, Matthieu, Cordelia Schmid, Ivan Laptev, Benoît Sagot, and Rachel Bawden. 2023. Tackling ambiguity with images: Improved multimodal machine translation and contrastive evaluation. In Rogers, Anna, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5394–5413, Toronto, Canada, July. Association for Computational Linguistics.

Futeral, Matthieu, Cordelia Schmid, Benoît Sagot, and Rachel Bawden. 2025. Towards zero-shot multimodal machine translation. In Chiruzzo, Luis, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 761–778, Albuquerque, New Mexico, April. Association for Computational Linguistics.

House, Juliane. 2006. Text and context in translation. *Journal of Pragmatics*, 38(3):338–358. Special Issue: Translation and Context.

Lannelongue, Loïc, Jason Grealey, and Michael Inouye. 2021. Green algorithms: Quantifying the carbon footprint of computation. *Advanced Science*, 8(12):2100707.

Liu, Danyang, Fanjie Kong, Xiaohang Sun, Dhruva Patil, Avijit Vajpayee, Zhu Liu, Vimal Bhat, and Najmeh Sadoughi. 2025. Detect, disambiguate, and translate: On-demand visual reasoning for multimodal machine translation with large vision-language models. In Chiruzzo, Luis, Alan Ritter, and Lu Wang, editors, *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 1559–1570, Albuquerque, New Mexico, April. Association for Computational Linguistics.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Isabelle, Pierre, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Popović, Maja. 2015. chrF: character n-gram F-score for automatic MT evaluation. In Bojar, Ondřej, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In Webber, Bonnie, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020*

Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2685–2702, Online, November. Association for Computational Linguistics.

Shen, Huangjun, Liangying Shao, Wenbo Li, Zhibin Lan, Zhanyu Liu, and Jinsong Su. 2024. A survey on multi-modal machine translation: Tasks, methods and challenges.

Specia, Lucia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In Bojar, Ondřej, Christian Buck, Rajen Chatterjee, Christian Federmann, Liane Guillou, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Aurélie Névéol, Mariana Neves, Pavel Pecina, Martin Popel, Philipp Koehn, Christof Monz, Matteo Negri, Matt Post, Lucia Specia, Karin Verspoor, Jörg Tiedemann, and Marco Turchi, editors, *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553, Berlin, Germany, August. Association for Computational Linguistics.

Vijayan, Vipin, Braeden Bowen, Scott Grigsby, Timothy Anderson, and Jeremy Gwinnup. 2024. The case for evaluating multimodal translation models on text datasets.

Villa-Cueva, Emilio, Sholpan Bolatzhanova, Diana Turmakhan, Kareem Elzeky, Henok Biadglign Ademtew, Alham Fikri Aji, Israel Abebe Azime, Jinheon Baek, Frederico Belcavello, Fermin Cristobal, Jan Christian Blaise Cruz, Mary Dabre, Raj Dabre, Toqeer Ehsan, Naome A Etori, Fauzan Farooqui, Jiahui Geng, Guido Ivetta, Thanmay Jayakumar, Soyeong Jeong, Zheng Wei Lim, Aishik Mandal, Sofia Martinelli, Mihail Minkov Mihaylov, Daniil Orel, Aniket Pramanick, Sukannya Purkayastha, Israfel Salazar, Haiyue Song, Tiago Timponi Torrent, Debela Desalegn Yadeta, Injy Hamed, Atnafu Lambebo Tonja, and Thamar Solorio. 2025. Cammt: Benchmarking culturally aware multimodal machine translation.

Voita, Elena, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In Korhonen, Anna, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy, July. Association for Computational Linguistics.

A CoMMuTE Sample Item

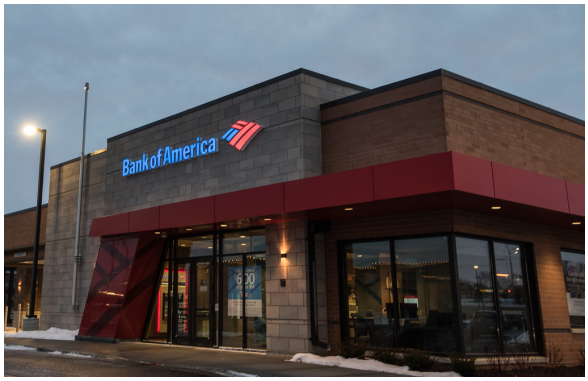

Figure 3: CoMMuTE sample item (Futeral et al., 2023). Source: *"He finally made it to the bank."* Image A (left): *rive* 'river bank'; Image B (right): *banque* 'financial bank'.

B CaMMT Sample Item

Figure 4: CaMMT sample item (Villa-Cueva et al., 2025). Substituted source: *"This dish is called a beet soup."* Conserved source: *"This dish is called borsch."* Russian reference: "Это блюдо называется борщ." The image provides the cultural context needed to recover the specific term борщ from the substituted source. Under the conserved source, the term is already present in the input; the image may help the model preserve it rather than paraphrase.

C Prompt Template

All models receive the following prompt. When an image condition applies, the image is provided as visual input alongside the text in a single request, using each model's native multimodal interface (chat-template processing for local models; base64 encoding for API models).

```
Translate the following English sentence into [LANGUAGE]: `[SENTENCE]'. Return your
translation as a JSON string as follows: {"translation": "<your translation>"}.
```

Output is parsed by extracting the `translation` field from the returned JSON. When a model does not produce valid JSON, the raw output is used as a fallback.

D Efficiency

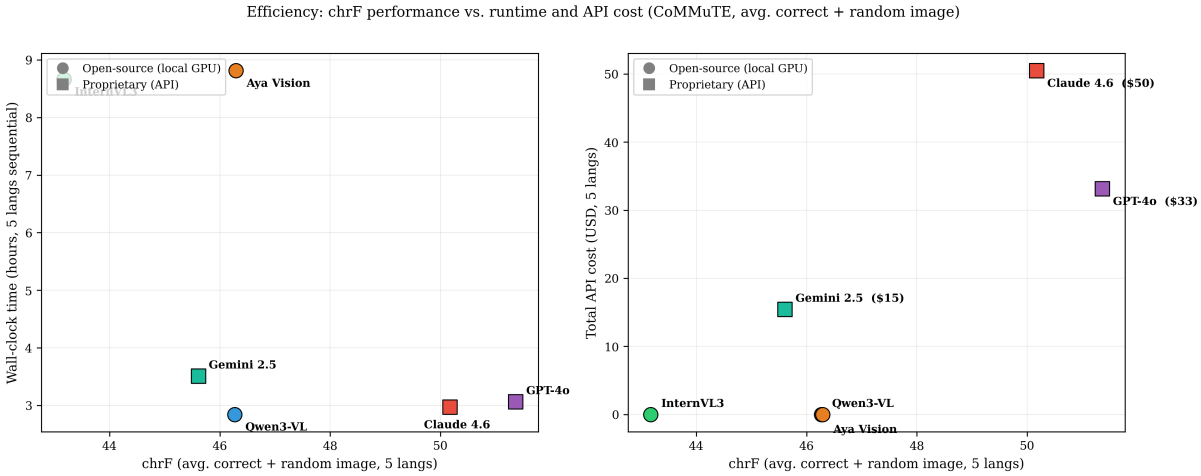

Figure 5: Efficiency trade-offs on CoMMuTE. The x-axis reports chrF averaged over the *correct-image* and *random-image* conditions (5 languages), reflecting realistic mixed-relevance usage. Left: wall-clock runtime. Right: total API cost (open-source models sit at $0). Squares denote proprietary (API) models; circles denote open-source (local GPU) models.

E CoMMuTE: BLEU Deltas

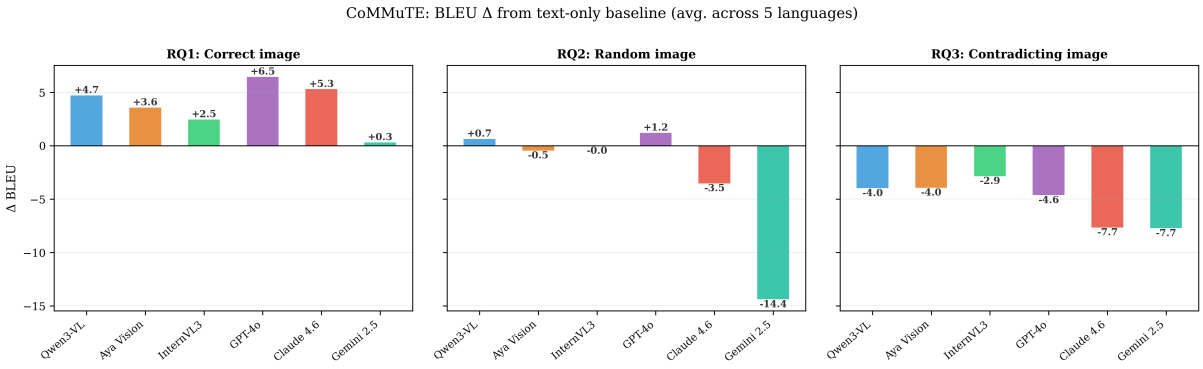

Figure 6: CoMMuTE: mean BLEU change from the text-only baseline, averaged across five languages (FR, DE, AR, RU, ZH). Layout mirrors Figure 1.

F CoMMuTE: Per-Language Breakdowns

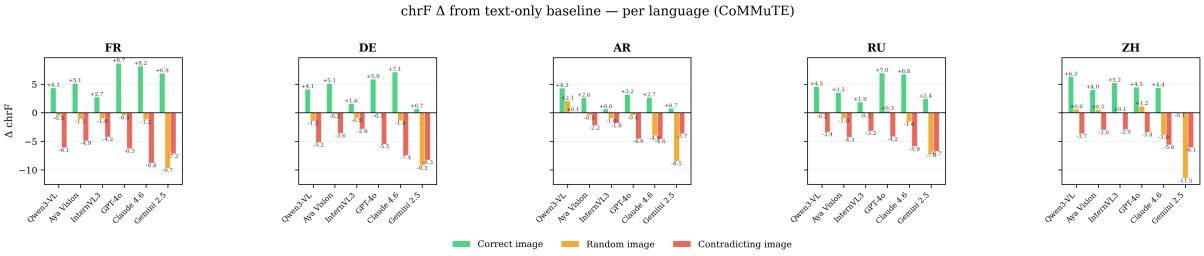

Figure 7: CoMMuTE: chrF deltas from text-only baseline, per language.

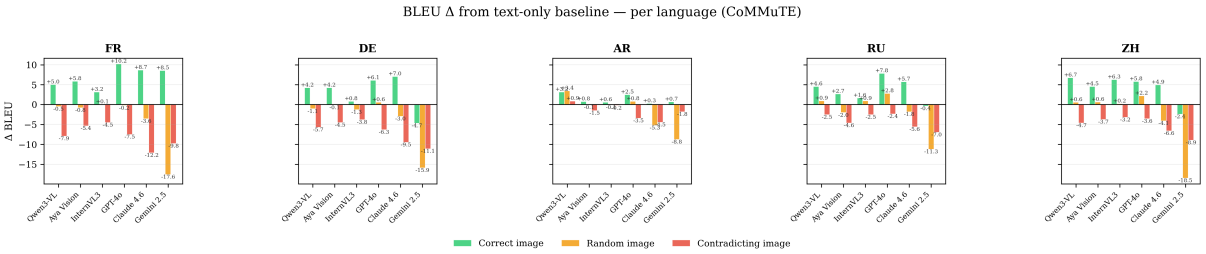

Figure 8: CoMMuTE: BLEU deltas from text-only baseline, per language.

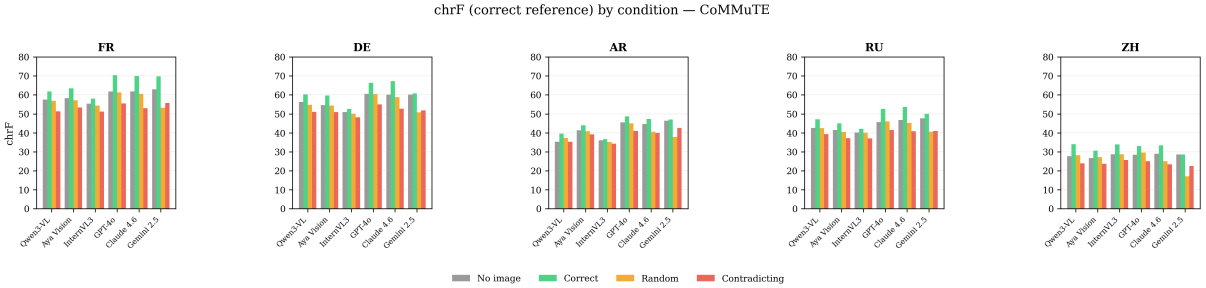

Figure 9: CoMMuTE: absolute chrF scores (correct reference) by condition and language.

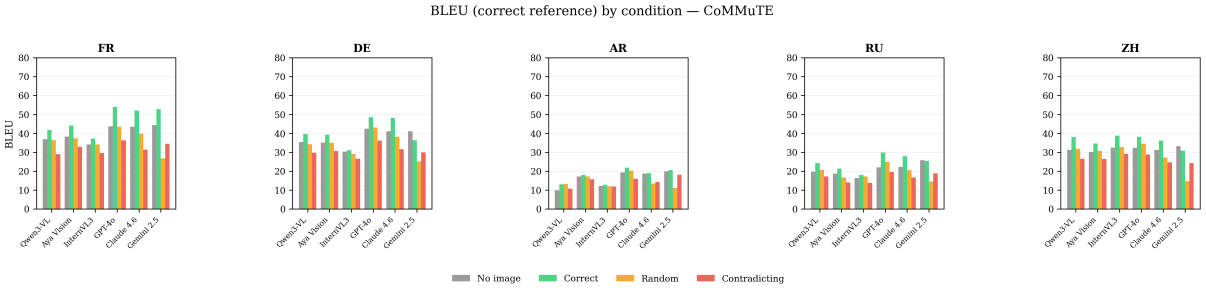

Figure 10: CoMMuTE: absolute BLEU scores (correct reference) by condition and language.

G CoMMuTE: Disambiguation Maps

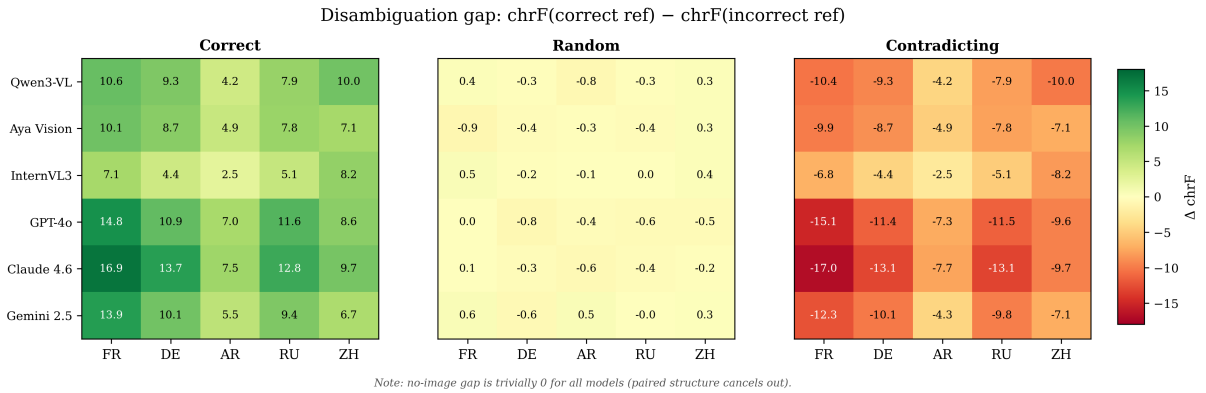

Figure 11: CoMMuTE disambiguation gap by visual condition and language. Positive values indicate the translation is closer to the intended meaning; negative values indicate it is closer to the wrong one.

H CaMMT and cross-task comparison: BLEU Deltas

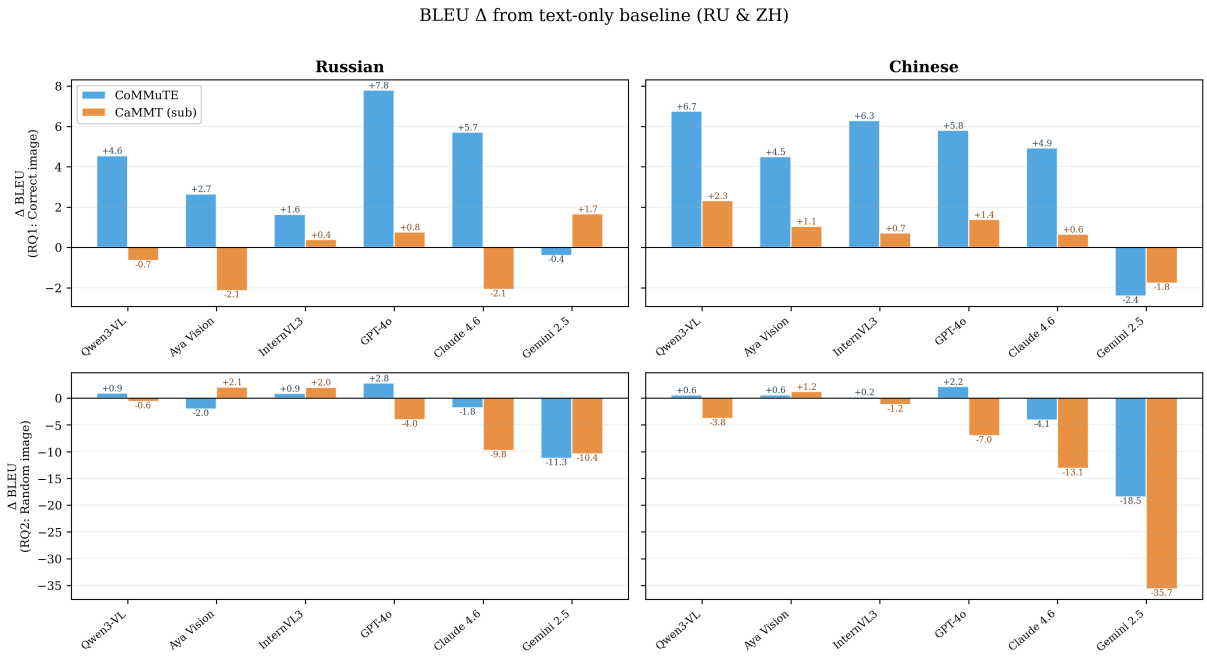

Figure 12: CaMMT and cross-task comparison: BLEU delta from the text-only baseline on Russian and Chinese.

I Synthetic *NordicTrans Motors* service illustration

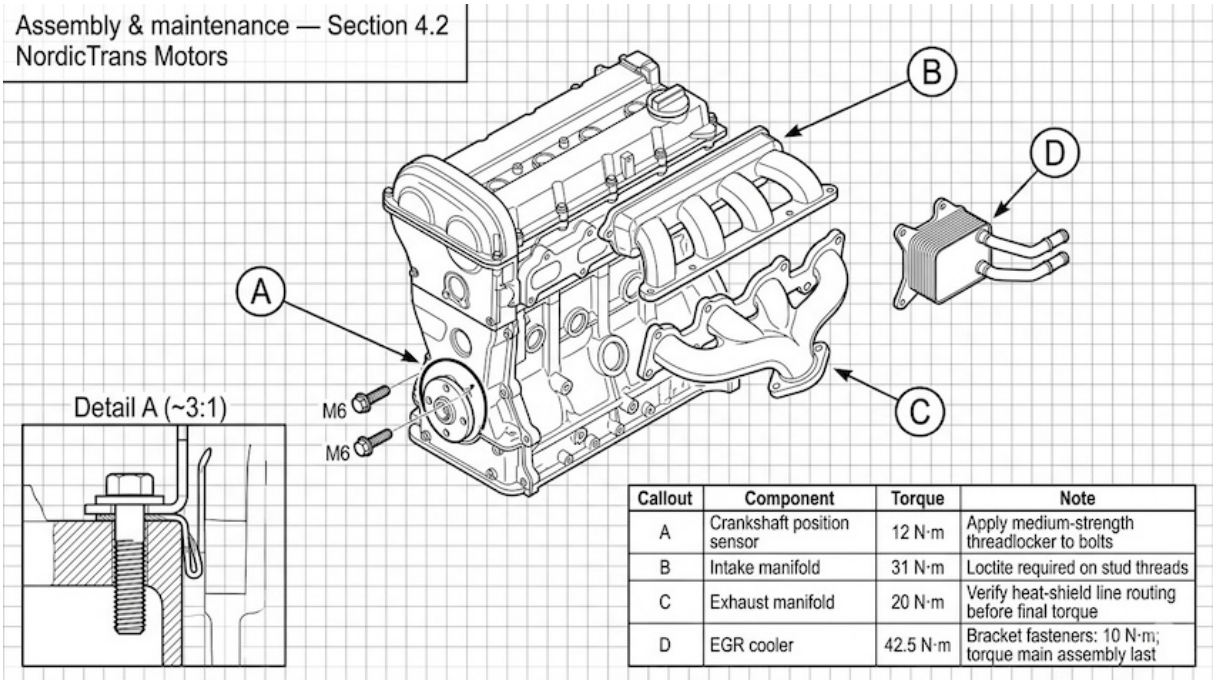

Figure 13: Synthetic image used as the shared visual for both the on-figure and off-figure English procedures in the practical assessment.

J Practical Assessment: Source, Hypotheses and Reference Translations

Condition	Source (EN)	Reference	Qwen3-VL-8B, no img	Qwen3-VL-8B, w/ img	GPT-4o, no img	GPT-4o, w/ img
Matched (FR)	Refer to Figure 4.2 and the Section 4.2 assembly table. Apply threadlocking compound to manifold studs only for the manifold callout with the higher torque value, as specified in the note for that row, and do not apply final tightening to the other manifold callout until the routing-related note for that row has been met.	Se reporter à la figure 4.2 et au tableau d'assemblage de la section 4.2. Appliquer un composé frein-filet sur les goujons du collecteur uniquement pour le repère de collecteur présentant la valeur de couple la plus élevée, conformément à la note figurant sur la ligne correspondante, et ne pas effectuer le serrage final sur l'autre repère de collecteur tant que la note relative au routage figurant sur la ligne de ce repère n'a pas été respectée.	Veuillez vous référer à la figure 4.2 et au tableau d'assemblage de la section 4.2. Appliquez uniquement un composé anti-décollement des filets aux boulons du collecteur correspondant à l'appel de collecteur avec la valeur de couple la plus élevée, tel que spécifié dans la note de cette ligne, et n'appliquez pas le serrage final aux autres appels de collecteur jusqu'à ce que la note liée au routage de cette ligne soit respectée.	Veuillez consulter la figure 4.2 et le tableau d'assemblage de la section 4.2. Appliquez uniquement un composé anti-décollement des filets aux boulons du collecteur pour le collecteur dont la valeur de couple est la plus élevée, conformément à la note spécifiée dans la ligne correspondante, et n'appliquez pas le serrage final au collecteur correspondant jusqu'à ce que la note relative au routage pour cette ligne soit respectée.	Consultez la Figure 4.2 et la table d'assemblage de la Section 4.2. Appliquez un composé de freinage de filetage uniquement aux goujons du collecteur pour l'appel de collecteur avec la valeur de couple plus élevée, comme spécifié dans la note pour cette rangée, et n'effectuez pas le serrage final de l'autre appel de collecteur jusqu'à ce que la note relative à l'acheminement pour cette rangée ait été respectée.	Consultez la figure 4.2 et le tableau d'assemblage de la section 4.2. Appliquez un **frein-filet** uniquement aux goujons du collecteur pour le **repère** du collecteur avec la valeur de couple la plus élevée, comme spécifié dans la note pour cette ligne, et n'effectuez pas le serrage final pour l'autre **repère** du collecteur tant que la note relative au **routage** pour cette ligne n'est pas respectée.
Matched (RU)	Refer to Figure 4.2 and the Section 4.2 assembly table. Apply threadlocking compound to manifold studs only for the manifold callout with the higher torque value, as specified in the note for that row, and do not apply final tightening to the other manifold callout until the routing-related note for that row has been met.	Обратитесь к рис. 4.2 и таблице сборки раздела 4.2. Нанесите фиксатор резьбы на шпильки крепления коллектора только для той позиции коллектора, которой в таблице соответствует большее значение момента затяжки, в соответствии с примечанием к строке этой позиции; не выполняйте окончательную затяжку для другой позиции коллектора, пока не будут выполнены требования примечания о прокладке линий, указанного для строки этой позиции.	См. Рисунок 4.2 и таблицу сборки раздела 4.2. Применяйте клей для **нитей** только к болтам **манжеты** с более высоким значением крутящего момента, как указано в примечании к этой строке, и не применяйте окончательную затяжку к другому **обозначению манжеты** до тех пор, пока не будет выполнено примечание, связанное с маршрутизацией, для этой строки.	См. рис. 4.2 и таблицу сборки раздела 4.2. Применяйте клей для **резьбы** только к болтам **головки блока цилиндров**, указанным в строке с наибольшим значением крутящего момента, как указано в примечании к этой строке, и не применяйте окончательную затяжку к другому **болту головки блока цилиндров**, пока не будет выполнено примечание, связанное с маршрутизацией, для этой строки.	Обратитесь к Рисунку 4.2 и таблице сборки Раздела 4.2. Нанесите фиксатор резьбы только на **шпильки** коллектора для той позиции коллектора, где указано более высокое значение крутящего момента, как указано в примечании для этой строки, и не выполняйте окончательное затягивание другой позиции коллектора, пока примечание относительно маршрута для этой строки не будет выполнено.	Обратитесь к Рисунку 4.2 и таблице сборки в Разделе 4.2. Наносите фиксатор резьбы на **штифты** коллектора только для обозначения коллектора с более высоким крутящим моментом, как указано в примечании для этой строки, и не производите окончательную затяжку для другого обозначения коллектора, пока не будет выполнено примечание, связанное с маршрутизацией, для этой строки.
Mispaired (FR)	If the telematics gateway logs J1939 DM1 frames indicating SCR system fault SPN 3364 while DEF tank level reads above 10%, perform a forced Dosing Air Pressure Test via the diagnostics menu before replacing the DEF pump assembly.	Si la passerelle télémétrique enregistre des trames DM1 J1939 signalant un défaut du système SCR SPN 3364 alors que le niveau du réservoir DEF reste supérieur à 10%, effectuez un test forcé de pression d'air de dosage via le menu de diagnostic avant de remplacer l'ensemble de la pompe DEF.	Si la passerelle de télématicie enregistre des cadres J1939 DM1 indiquant une défaillance du système SCR SPN 3364 pendant que le niveau du réservoir de DEF est supérieur à 10 %, effectuez un test obligatoire de pression d'air de dosage via le menu de diagnostic avant de remplacer l'ensemble de la pompe DEF.	Si la passerelle de **télémétrie** enregistre des cadres J1939 DM1 indiquant une défaillance du système SCR SPN 3364 pendant que le niveau du réservoir DEF est supérieur à 10 %, effectuez un test forcé de pression d'air de dosage via le menu de diagnostic avant de remplacer l'assemblage de pompe DEF.	Si la passerelle télématique enregistre des trames J1939 DM1 indiquant une défaillance du système SCR SPN 3364 alors que le niveau du réservoir de DEF est supérieur à 10 %, effectuez un test de pression d'air de dosage forcé via le menu de diagnostic avant de remplacer l'ensemble de la pompe DEF.	Si la passerelle télématique enregistre des trames DM1 J1939 indiquant un défaut du système SCR SPN 3364 alors que le niveau du réservoir DEF est supérieur à 10 %, effectuez un test de pression d'air de dosage forcé via le menu de diagnostic avant de remplacer l'ensemble de la pompe DEF.
Mispaired (RU)	If the telematics gateway logs J1939 DM1 frames indicating SCR system fault SPN 3364 while DEF tank level reads above 10%, perform a forced Dosing Air Pressure Test via the diagnostics menu before replacing the DEF pump assembly.	Если телематический шлюз фиксирует кадры DM1 J1939 с индикацией неисправности системы SCR SPN 3364 при уровне мочевины в баке выше 10%, выполните принудительную проверку давления воздуха дозирования через меню диагностики перед заменой узла насоса подачи мочевины.	Если телематический шлюз фиксирует кадры J1939 DM1, указывающие на неисправность системы SCR SPN 3364, при условии, что уровень топлива в баке DEF превышает 10%, выполните принудительный тест давления воздуха для дозирования через меню диагностики перед заменой агрегата насоса DEF.	Если шлюз телематики фиксирует кадры J1939 DM1, указывающие на неисправность системы SCR SPN 3364, при условии, что уровень топлива в баке DEF превышает 10%, выполните принудительный тест давления воздуха для дозирования через меню диагностики перед заменой агрегата насоса DEF.	Если телематический шлюз регистрирует кадры J1939 DM1, указывающие на неисправность системы SCR с SPN 3364, в то время как уровень DEF в баке превышает 10%, выполните принудительный тест давления воздуха для дозирования через диагностическое меню перед заменой узла насоса DEF.	Если телематический шлюз регистрирует кадры J1939 DM1, указывающие на неисправность системы SCR SPN 3364, при этом уровень в баке DEF показывает выше 10%, выполните принудительный тест давления воздуха дозирования через меню диагностики перед заменой узла насоса DEF.

Table 5: *NordicTrans Motors* practical stimuli: English source, synthetic human reference, and model outputs (Qwen3-VL-8B / GPT-4o) under **no-image** vs. **with-image** conditions. One row per scored text unit (matched vs. mispaired × FR / RU).