

Multi-Candidate Synthesis in Multimodal Machine Translation

Szymon Niewiadomski

*Faculty of Mathematics and Computer Science
Adam Mickiewicz University
Poznań, Poland
ORCID: 0009-0007-6587-9633*

Jan Kolwicz

*Faculty of Mathematics and Computer Science
Adam Mickiewicz University
Poznań, Poland
ORCID: 0009-0001-5523-8927*

Kamil Guttman

*Faculty of Mathematics and Computer Science
Adam Mickiewicz University; Lanigo
Poznań, Poland
ORCID: 0009-0007-4050-4966*

Michał Ciesiołka

*Faculty of Mathematics and Computer Science
Adam Mickiewicz University; Lanigo
Poznań, Poland
ORCID: 0009-0009-8283-1118*

Krzysztof Jassem

*Faculty of Mathematics and Computer Science
Adam Mickiewicz University; Lanigo
Poznań, Poland
ORCID: 0000-0001-7122-9206*

Abstract—Multimodal Machine Translation (MMT) typically relies on passing text and corresponding images to a model to resolve ambiguities. However, the direct presence of an image can occasionally mislead the model, causing regressions compared to text-only translation. To address this, we propose a 5-step pipeline for evaluating and refining MMT. Instead of single-pass translations, our approach generates four diverse candidate translations leveraging different subsets of multimodal context (text-only, text+image, text+caption, text+image+caption). A final synthesis step acts as a judge and editor to produce the optimal translation. Evaluated on the ConECT and Multi30k datasets using large language models (Qwen3.6-35B and Gemma 4 26B), our synthesis pipeline consistently outperforms text-only and standard multimodal approaches. We observe up to a +1.34 improvement in COMET scores.

Index Terms—Multimodal Machine Translation, Large Language Models, Machine Translation, Visual Language Models

I. INTRODUCTION

MULTIMODAL Machine Translation (MMT) uses images to help models translate ambiguous words. However, adding visual information does not always improve results. Research by Caglayan et al. [1] shows that MMT models often ignore images and rely entirely on text. Even worse, directly adding an image can sometimes lower the translation quality. This happens when images contain irrelevant backgrounds, text on packaging, or brand logos that confuse the model.

This problem is highly visible in e-commerce translation tasks, as demonstrated by the ConECT dataset [2]. Product names in catalogs are often short and lack grammatical structure. While they need visual context for disambiguation, the

noise in product photos makes naive MMT approaches prone to errors.

To address this issue, we propose a multi-candidate inference pipeline. Instead of relying on a single translation attempt, our system generates four different translations for each sample. These candidates use different combinations of inputs: text-only, text with the image, text with a generated image description (caption), and text with both the image and caption. Finally, the model acts as a judge to synthesize the best final translation from these four options.

We tested this approach using two modern Mixture-of-Experts (MoE) models across four dataset-language pairs, generating a total of 132,950 translations. Our findings show that this synthesis step consistently improves translation quality, increasing COMET scores by up to +1.34 compared to the text-only baseline.

II. RELATED WORK

MMT is an ongoing research topic in the machine translation community. One of the use-cases for it is to leverage both textual and visual inputs for translating image descriptions [3]. However, the actual contribution of the visual modality remains a subject of research. Previous studies demonstrated that MMT models often ignore visual features when the source text is adequate, relying on images primarily when the textual context is degraded or masked [1]. Furthermore, recent surveys highlight that while semantically aligned images can resolve textual ambiguities, the direct integration of noisy or irrelevant visual information frequently disrupts the translation process and causes semantic shifts [4].

This vulnerability to visual noise is especially pronounced in domain-specific applications, such as e-commerce translation [2]. In these settings, short and unstructured product names heavily depend on visual context for disambiguation, yet they are frequently paired with cluttered or misleading product photos. Existing baselines in this domain reveal the limitations of naive multimodal integration, showing that unconditionally appending image features can lead to quality regressions compared to text-only translation.

To address these limitations, we introduce a multi-candidate synthesis approach. It builds upon recent test-time synthesis methods, such as the Hunyuan-MT-Chimera approach, which generates multiple textual translation candidates and fuses them into a single, high-quality output using a specialized evaluator model [5]. While the original Chimera model operates strictly on text by varying generation parameters, we extend this best-of-n synthesis paradigm to the visual modality, enabling the model to filter noisy candidates.

III. METHODOLOGY

In this section, we describe the datasets, models, and evaluation metrics employed to comprehensively assess our multimodal translation pipeline.

A. Data

Our experiments were conducted on two established datasets covering different language pairs and translation scenarios.

1) *ConECT*: The ConECT dataset [2] contains 10,295 samples of Czech-to-Polish (cs→pl) translations. It is highly domain-specific, featuring product names, e-commerce terminology, and product descriptions. Such data is notoriously difficult to translate without context, as product names and concise catalog texts lack the syntactical structure of standard sentences, making visual context highly relevant for disambiguation.

2) *Multi30k*: The Multi30k dataset [6] is a standard benchmark for Multimodal Machine Translation. We utilize the `test_2016_flickr` split, encompassing 1,000 samples for three translation directions: English-to-German (en→de), English-to-French (en→fr), and English-to-Czech (en→cs). This dataset primarily contains descriptive sentences of everyday scenes, offering a contrast to the highly specific e-commerce domain of ConECT.

B. Models

We evaluated two vision-language models:

- **Qwen3.6-35B-A3B-AWQ-4bit**: A Mixture of Experts (MoE) model [7] with 35 billion total parameters and 3 billion active parameters per token. For this task, we disabled the model's default reasoning mode to reduce inference cost and avoid exhausting token limits on reasoning traces.
- **Gemma 4 26B-A4B-AWQ-4bit**: A MoE model [8] comprising 25.2 billion total and 3.8 billion active parameters.

Both models were quantized to INT4 using AWQ to reduce memory requirements and accelerate inference. They were

served using vLLM with greedy decoding (temperature = 0.0) to guarantee reproducibility.

C. Evaluation

We evaluated the generated translations using three primary metrics to capture different aspects of translation quality. To ensure standard and reproducible evaluations, we provide exact signatures using SacreBLEU [9].

- **BLEU**: The standard n-gram precision metric for machine translation, computed using SacreBLEU (with `effective_order=True`). Its word-level granularity makes it particularly sensitive to single-word differences, which is relevant in our setting where candidate translations may differ only in one visually-disambiguated term.
- **chrF**: A character n-gram F-score metric implemented via SacreBLEU. It is particularly effective for morphologically rich languages like Polish and Czech, where word-level metrics penalize valid inflected forms.
- **COMET**: We utilized the `wmt22-comet-da` model [10], which provides a neural evaluation of translation quality based on a pre-trained cross-lingual encoder. COMET captures semantic similarities and nuances that traditional lexical metrics often miss.

D. Preliminary Observations

Before designing the full pipeline, we ran baseline experiments comparing text-only and direct multimodal (text+image) translation. These confirmed that providing the image does not consistently help: in several configurations, image-only mode degraded COMET relative to text-only (e.g., Gemma on ConECT dropped from 91.44 to 91.00; see Table I). This is consistent with prior findings that MMT models often fail to leverage visual context effectively [1]. At the same time, some configurations did benefit from the image, indicating that the visual signal carries useful information that is simply not being reliably exploited. These observations motivated a synthesis-based approach: instead of committing to a single input modality, we generate candidates under multiple context configurations and delegate the final selection to a judge model.

IV. APPROACH

Our core contribution is the 5-step MMT pipeline. Instead of a standard end-to-end translation, the system generates four distinct candidate translations and synthesizes them into a final output. This process comprises a single caption generation call, four parallel translation generation calls, and one final translation synthesis call.

A. Mode 1: Text-Only

The baseline control group. The model is prompted to translate the source text strictly based on its textual content, with no visual information provided.

```
You are reviewing candidate translations
to {lang_name}. Produce the best final
translation. You may pick one of the
candidates verbatim, or improve any of
them. Output only the final translation,
nothing else.
```

```
Source text: {source_text}
```

```
Candidate 1 (text only): {t_text_only}
Candidate 2 (with image): {t_multimodal}
Candidate 3 (with image description):
  {t_caption_only}
Candidate 4 (with image + description):
  {t_caption_multimodal}
```

Listing 1. Synthesis prompt template. {lang_name} is the expanded target language name; {source_text} is the input; candidates are listed in mode order.

B. Mode 2: Text + Image

The model receives the source text and the image directly, passed as-is without any resizing or preprocessing (internal resizing is handled by the model’s vision encoder). This approach assesses the raw value of visual context. While it frequently helps resolve ambiguities, it can also introduce noise or prompt the model to hallucinate details present in the image but absent in the source text.

C. Mode 3: Text + Caption

In an initial preparatory step, the model is issued a separate inference call in which it is instructed to produce a 1–2 sentence caption for a given image. The translation is then performed in an independent call that receives only the source text and the generated caption text — the image is not included. Because the two calls share no context, the translation model is strictly isolated from the visual input. This tests whether a textual proxy of the image is sufficient for disambiguation.

D. Mode 4: Text + Caption + Image

The model receives the source text, the image, and the caption generated in Mode 3. This mode checks whether combining both the raw image and its textual description improves over using either one alone.

E. Synthesis

The final step acts as a judge and editor. The model is provided with the original source text and all four previously generated translation candidates. The model is instructed to produce the single best translation by either selecting one candidate verbatim or composing a refined translation based on the strengths of the provided options. This agentic layer aims to filter out noise and hallucinations sometimes introduced by direct visual context while retaining disambiguation benefits. The full prompt used for synthesis is shown in Listing 1.

V. RESULTS

Our evaluation encompasses a total of 132,950 translations across all dataset pairs, models, and modes.

A. Translation Quality

Table I shows absolute COMET scores for all four candidate modes across both models and datasets. Table II reports the corresponding absolute BLEU and chrF scores for the text-only baseline and synthesis. Table III summarises the improvement of synthesis over the text-only baseline.

Notably, the image-only mode occasionally caused regressions compared to text-only (e.g., Gemma on ConECT dropped from 91.44 to 91.00), confirming the motivation for the synthesis step. The synthesis mode consistently outperformed the text-only baseline across all 8 configurations (4 datasets $\times$ 2 models), with an average COMET improvement of +0.65. The largest gain was observed for Gemma on Multi30k (en $\rightarrow$ cs) (+1.34 COMET), indicating that the pipeline yields the highest impact on weaker language pairs. Gemma consistently achieves higher absolute scores than Qwen on ConECT (cs $\rightarrow$ pl), while the two models are more competitive on Multi30k.

B. Synthesis Behavior: Verbatim Selection vs. Novel Generation

To understand the internal behavior of the synthesis step, we analyzed how often it produces a novel translation versus selecting one translation candidate verbatim. Table IV shows the results across all datasets and models.

The synthesis step is predominantly a selector, not a generator: across all configurations, it picks an existing candidate verbatim in 84.9–96.0% of cases. When selecting verbatim, the *text-only* candidate is favored most frequently (48–66% of all samples), followed by the *text+image* candidate (16–23%). This is consistent with the overall ranking of standalone modes: synthesis mirrors the quality signal from the underlying candidates.

Critically, despite the high rate of verbatim selection, synthesis still improves upon *text-only* by up to +1.34 COMET (see Table III). This apparent paradox is resolved by the nature of the improvement: synthesis does not need to rewrite the majority of samples. It only needs to correctly identify and reject the minority of cases where text-only fails — specifically, the cases where visual context from Modes 2–4 provides the correct translation. Even a 15–35% rate of non-text-only selections, concentrated in the hardest samples (brand name ambiguities, language confusion, repetition loops), is sufficient to produce a consistent metric improvement across all 8 dataset–model pairs.

The highest novel generation rate — 15.1% for Gemma on Multi30k (en $\rightarrow$ cs) — coincides with the largest COMET gain (+1.34), suggesting that on pairs that are more linguistically complex, the synthesis model more actively fuses and repairs candidates rather than relying on any single one.

VI. CONCLUSION

We proposed a five-step agentic pipeline for multimodal machine translation that generates four candidate translations under different combinations of visual context — text-only, text with image, text with a generated caption, and text with

TABLE I
ABSOLUTE COMET SCORES (0–100 SCALE) FOR THE FOUR CANDIDATE MODES, EXCLUDING THE SYNTHESIS STEP.

Model	Dataset	Text-Only	Text+Image	Text+Caption	Text+Caption+Image
Qwen3.6-35B	ConECT (cs→pl)	89.67	89.69	89.92	89.70
Qwen3.6-35B	Multi30k (en→de)	87.12	87.33	86.77	87.12
Qwen3.6-35B	Multi30k (en→fr)	88.24	88.29	87.74	88.09
Qwen3.6-35B	Multi30k (en→cs)	89.85	90.47	89.80	90.03
Gemma 4 26B	ConECT (cs→pl)	91.44	91.00	91.54*	91.18
Gemma 4 26B	Multi30k (en→de)	87.18	87.74	87.23	87.79
Gemma 4 26B	Multi30k (en→fr)	88.78	88.75	88.55	89.14
Gemma 4 26B	Multi30k (en→cs)	89.07	89.69	89.08	89.74

*Text+Caption exceeds Text-Only here, but remains below Synthesis (92.09).

TABLE II
ABSOLUTE BLEU AND chrF SCORES FOR THE TEXT-ONLY BASELINE AND SYNTHESIS MODE.

Model	Dataset	BLEU (Base)	BLEU (Synth)	chrF (Base)	chrF (Synth)
Qwen3.6-35B	ConECT (cs→pl)	46.02	47.22	67.25	68.26
Qwen3.6-35B	Multi30k (en→de)	41.91	43.94	69.20	70.13
Qwen3.6-35B	Multi30k (en→fr)	48.84	50.03	71.76	72.28
Qwen3.6-35B	Multi30k (en→cs)	33.59	35.49	59.90	61.15
Gemma 4 26B	ConECT (cs→pl)	54.35	54.63	72.46	73.85
Gemma 4 26B	Multi30k (en→de)	42.12	43.33	70.04	70.72
Gemma 4 26B	Multi30k (en→fr)	57.69	57.22	76.19	76.14
Gemma 4 26B	Multi30k (en→cs)	35.16	35.80	60.35	61.15

TABLE III
TEXT-ONLY BASELINE COMET AND IMPROVEMENT (Δ) OF SYNTHESIS MODE. ALL COMET VALUES ON 0–100 SCALE.

Model	Dataset	Δ BLEU	Δ chrF	Base COMET	Synth COMET	Δ COMET
Qwen3.6-35B	ConECT (cs→pl)	+1.20	+1.01	89.67	90.30	+0.63
Qwen3.6-35B	Multi30k (en→de)	+2.03	+0.93	87.12	87.62	+0.50
Qwen3.6-35B	Multi30k (en→fr)	+1.19	+0.51	88.24	88.51	+0.27
Qwen3.6-35B	Multi30k (en→cs)	+1.89	+1.25	89.85	90.73	+0.88
Gemma 4 26B	ConECT (cs→pl)	+0.27	+1.39	91.44	92.09	+0.65
Gemma 4 26B	Multi30k (en→de)	+1.21	+0.68	87.18	87.80	+0.62
Gemma 4 26B	Multi30k (en→fr)	−0.47	−0.04	88.78	89.09	+0.31
Gemma 4 26B	Multi30k (en→cs)	+0.64	+0.80	89.07	90.41	+1.34

TABLE IV
SYNTHESIS OUTPUT COMPOSITION: PERCENTAGE OF CASES WHERE THE SYNTHESIS STEP SELECTS A CANDIDATE VERBATIM (BROKEN DOWN BY SOURCE MODE) VERSUS GENERATES A NOVEL TRANSLATION. ROWS SUM TO 100%.

Model	Dataset	Text-Only	Text+Image	Text+Caption	Text+Caption+Image	Verbatim total	Novel
Qwen3.6-35B	ConECT (cs→pl)	65.5%	16.0%	8.7%	4.4%	94.6%	5.4%
Qwen3.6-35B	Multi30k (en→de)	64.9%	17.1%	8.8%	4.0%	94.8%	5.2%
Qwen3.6-35B	Multi30k (en→fr)	65.6%	18.0%	7.8%	4.6%	96.0%	4.0%
Qwen3.6-35B	Multi30k (en→cs)	63.2%	19.0%	9.2%	3.2%	94.6%	5.4%
Gemma 4 26B	ConECT (cs→pl)	57.8%	18.5%	10.1%	4.3%	90.7%	9.3%
Gemma 4 26B	Multi30k (en→de)	57.7%	18.4%	10.4%	7.0%	93.5%	6.5%
Gemma 4 26B	Multi30k (en→fr)	55.9%	22.5%	9.2%	6.0%	93.6%	6.4%
Gemma 4 26B	Multi30k (en→cs)	48.0%	18.9%	11.3%	6.7%	84.9%	15.1%

both — and passes them to a final synthesis step that selects or composes the best output. The key insight is that no single input modality is universally optimal: direct image input sometimes degrades translation quality relative to the text-only baseline, yet the same visual signal can be beneficial in other cases. The synthesis step resolves this tension by delegating

the final decision to a judge model that has access to all candidate outputs at once.

Evaluated on the ConECT and Multi30k datasets with two vision-language models (Qwen3.6-35B and Gemma 4 26B), the synthesis mode consistently outperforms the text-only baseline across all 8 dataset–model configurations, improving

COMET by an average of +0.65 and up to +1.34 on the hardest language pair. Importantly, the synthesis step is predominantly a selector rather than a generator — in 85–96% of cases it picks one of the four candidates verbatim — yet this selective behavior is sufficient to produce consistent metric gains. The gains are concentrated in the minority of samples where the text-only baseline fails and visual context provides a better translation.

These results suggest that agentic multi-candidate approaches are a practical and effective way to exploit visual context in machine translation without committing to a single modality. The approach is model-agnostic and orthogonal to improvements in the underlying vision-language models, making it a promising direction for robust multimodal translation systems.

VII. LIMITATIONS

The primary limitation of the proposed pipeline is its computational overhead. Generating four candidates and a final synthesis requires six separate model calls per sample, significantly increasing inference cost and time compared to a single-pass translation.

A further limitation concerns the evaluation metrics. BLEU, chrF, and COMET are all text-to-text metrics: they compare the generated translation against a textual reference and do not account for the visual modality. As a result, they cannot directly measure whether the pipeline correctly leverages image content — for example, they cannot reward a translation that accurately reflects the visual scene when the reference translation does not. A vision-aware evaluation metric would more precisely capture the benefit of multimodal context.

REFERENCES

- [1] O. Caglayan, P. Madhyastha, L. Specia, and L. Barrault, "Probing the need for visual context in multimodal machine translation," in *Proceedings of the 2019 Conference of the North American*

- [1] *Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4159–4170. [Online]. Available: <https://aclanthology.org/N19-1422/>
- [2] M. Pokrywka, W. Kusa, M. Rutkowski, and M. Koszowski, “Conect dataset: Overcoming data scarcity in context-aware e-commerce mt,” 2025. [Online]. Available: <https://arxiv.org/abs/2506.04929>
- [3] L. Specia, S. Frank, K. Sima’an, and D. Elliott, “A shared task on multimodal machine translation and crosslingual image description,” in *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, O. Bojar, C. Buck, R. Chatterjee, C. Federmann, L. Guillou, B. Haddow, M. Huck, A. J. Yepes, A. N  v  l, M. Neves, P. Pecina, M. Popel, P. Koehn, C. Monz, M. Negri, M. Post, L. Specia, K. Verspoor, J. Tiedemann, and M. Turchi, Eds. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 543–553. [Online]. Available: <https://aclanthology.org/W16-2346/>
- [4] Y. Feng, C. Li, J. He, Z. Hou, and V. Ng, “Multimodal neural machine translation: A survey of the state of the art,” in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, Eds. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 22 130–22 147. [Online]. Available: <https://aclanthology.org/2025.emnlp-main.1125/>
- [5] M. Zheng, Z. Li, B. Qu, M. Song, Y. Du, M. Sun, and D. Wang, “Hunyuan-mt technical report,” 2025. [Online]. Available: <https://arxiv.org/abs/2509.05209>
- [6] D. Elliott, S. Frank, K. Sima’an, and L. Specia, “Multi30k: Multilingual english-german image descriptions,” *arXiv preprint arXiv:1605.00459*, 2016.
- [7] Q. Team, “Qwen3 technical report,” 2025. [Online]. Available: <https://arxiv.org/abs/2505.09388>
- [8] G. Team, “Gemma 3 technical report,” 2025. [Online]. Available: <https://arxiv.org/abs/2503.19786>
- [9] M. Post, “A call for clarity in reporting BLEU scores,” in *Proceedings of the Third Conference on Machine Translation: Research Papers*. Belgium, Brussels: Association for Computational Linguistics, Oct. 2018, pp. 186–191. [Online]. Available: <https://www.aclweb.org/anthology/W18-6319>
- [10] R. Rei, J. G. C. de Souza, D. Alves, C. Zerva, A. C. Farinhas, T. Glushkova, A. Lavie, L. Coheur, and A. F. T. Martins, “COMET-22: Unbabel-IST 2022 submission for the metrics shared task,” in *Proceedings of the Seventh Conference on Machine Translation (WMT)*. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 578–585. [Online]. Available: <https://aclanthology.org/2022.wmt-1.52>