Escaping the Probability Trap: Mitigating Semantic Drift in Cantonese-Mandarin Translation

Yuzhi Liang[*] **and Fangqi Chen**
School of Information Science and Technology
Guangdong University of Foreign Studies, China
yzliang@gdufs.edu.cn, 20221003110@mail.gdufs.edu.cn

Abstract

Fine-tuning multilingual models for low-resource dialect translation frequently encounters a "plausibility over faithfulness" dilemma, resulting in severe semantic drift on dialect-specific tokens. We term this phenomenon the "Probability Trap," where models prioritize statistical fluency over semantic fidelity. To address this, we propose MVS-Rank (Multi-View Scoring Reranking), a generate-then-rerank framework that decouples evaluation from generation. Our method assesses translation candidates through three complementary perspectives: (1) Source-Side Faithfulness via a Reverse Translation Model to anchor semantic fidelity; (2) Local Fluency using Masked Language Models to ensure syntactic precision; and (3) Global Fluency leveraging Large Language Models to capture discourse coherence. Extensive experiments on Cantonese-Mandarin benchmarks demonstrate that MVS-Rank achieves state-of-the-art performance, significantly outperforming strong fine-tuning baselines by effectively rectifying hallucinations while maintaining high fluency.

1 Introduction

Cantonese is one of the most widely spoken varieties of Chinese, serving as a vital linguistic link for over 86.6 million native speakers globally (Suen et al., 2024). Despite sharing a writing system with Standard Mandarin, Cantonese exhibits significant diglossia, characterized by distinct divergence in lexicon, syntax, and phonology. For instance, Cantonese retains unique items such as "佢" (he/she) and specific post-verbal adverbs that are absent in Mandarin. Consequently, Mandarin-centric NLP models often fail to generalize to Cantonese. Given the scarcity of high-quality parallel corpora, the dominant paradigm involves fine-tuning multilingual Pre-trained Language Models like mBART (Liu et al., 2020).

However, we observe a critical limitation in this paradigm: a trade-off between fluency and faithfulness, often referred to as the "plausibility over faithfulness" trap. When encountering dialect-specific tokens absent from Mandarin-dominant pre-training data, models frequently suffer from semantic drift. Driven by the objective of maximizing next-token probability, models tend to generate words that are statistically plausible in the target context but contradict the source meaning—a phenomenon we term the ”Probability Trap.”

Consider the Cantonese sentence: "Ni go man tai tai lai zung jau dak aau" (呢個問題睇嚟仲有得拗), which translates to "This issue seems debatable." The character "拗" (argue/debate) is rare in Standard Mandarin. Our analysis reveals that mBART-based models typically mistranslate this as "This issue seems hopeless" (這個問題看來還是沒有希望了). This error stems from the model's reliance on the language prior: in the general training corpus, the completion "hopeless" follows "This issue seems..." with significantly higher probability than the faithful translation. The model thus defaults to the most "fluent" hallucination, sacrificing semantic fidelity.

To address this hallucination problem caused by the "Probability Trap," we propose Multi-View Scoring Reranking (MVS-Rank[1]) , a framework that shifts from single-view generation to a comprehensive, multi-dimensional assessment. Our method evaluates candidate translations through three distinct yet complementary perspectives to ensure holistic quality. First, we establish a *Source-Side Faithfulness View* using a Reverse Model to calculate the conditional likelihood $P(x|y)$. Rather than generating back-translations explicitly,

[*] Corresponding author.

[1] Our code is available at `https://github.com/fangtuyu/yue-cmn-rerank`.

this metric measures how effectively the candidate reconstructs the source information, serving as a rigid anchor against semantic drift. Second, we incorporate a *Local Fluency View* via Masked Language Models to correct fine-grained syntactic errors and phrase collocations. Finally, we integrate a *Global Fluency View* using Large Language Models (LLMs) to capture long-range dependencies and discourse-level coherence.

By synthesizing these views, MVS-Rank effectively decouples generation from evaluation. Experiments on Cantonese-Mandarin benchmarks demonstrate that our approach achieves state-of-the-art performance, significantly outperforming strong fine-tuning baselines by mitigating semantic drift while maintaining high fluency.

Our main contributions are summarized as follows:

- We identify the “Probability Trap” in low-resource dialect translation, providing concrete evidence of how Mandarin-dominant language priors override semantic content on rare dialectal tokens.

- We propose MVS-Rank, a generate-then-rerank framework that enforces semantic consistency. Specifically, we utilize the reconstruction likelihood from a Reverse Model as a rigid penalty for fluent but unfaithful hallucinations.

- We introduce a novel hybrid fluency scoring mechanism that combines the local precision of BERT with the global coherence of LLMs. This multi-granularity approach ensures translations are both grammatically sound and contextually natural.

2 Related Work

Cantonese machine translation presents unique challenges due to the substantial lexical, syntactic, and stylistic divergence between Cantonese and Standard Mandarin. Furthermore, the scarcity of high-quality parallel corpora characterizes it as a quintessential low-resource and dialectal MT problem.

Early approaches primarily relied on rule-based or example-based methods, necessitating extensive handcrafted linguistic resources and expert knowledge (e.g., Zhang, 1998; Wu et al., 2006; Mak and Lee, 2021; Wang et al., 2007). With the advent of statistical machine translation, research shifted toward leveraging limited parallel data. For instance, Hong et al. (2024) explored data augmentation via back-translation to construct synthetic parallel corpora. Similarly, Dare et al. (2023) investigated unsupervised machine translation by exploiting cross-lingual word embeddings to mitigate the dependency on large-scale parallel data.

In the era of pre-trained models, the dominant paradigm involves fine-tuning multilingual models like mBART (Liu et al., 2020). Kozhirbayev (2024) demonstrated that such models could transfer knowledge to unseen dialects through fine-tuning on related language pairs. Salazar et al. (2020) demonstrated leveraging full bidirectional context to assess sentence fluency and grammaticality more holistically. More recently, efforts have focused on mining potential parallel sentences from Wikipedia to scale up training resources (Liu, 2022) . Despite these advancements, standard fine-tuning often struggles with dialect-specific idioms, leading to the ”Probability Trap” where models prioritize fluency over adequacy.

Most recently, Large Language Models have opened new avenues for this task. Dai et al. (2025) enhanced Cantonese comprehension in LLMs by fine-tuning on syntactic and POS data, followed by post-processing refinement. Jiang et al. (2025) systematically evaluates the performance of large models of the mainstream in Cantonese understanding and reasoning. Chang et al. (2025) proposed integrating Retrieval-Augmented Generation with LLMs to incorporate external knowledge. However, while LLMs exhibit strong generalization, they still suffer from hallucinations when handling rare dialectal tokens without explicit constraints, highlighting the need for robust reranking mechanisms.

3 Methodology

To address the scarcity of Cantonese data and the semantic drift caused by language priors, we propose **MVS-Rank** (Multi-View Scoring Reranking). As illustrated in Figure 1, our framework operates in a coarse-to-fine retrieve-then-rerank paradigm. It comprises two core stages: (1) Candidate Generation, where a fine-tuned mBART model serves as the forward backbone to generate an N-best hypothesis set; and (2) Multi-View Reranking, where we evaluate candidates through three distinct perspectives to ensure holistic quality.

3.1 Candidate Generation

We adopt the multilingual pre-trained model mBART as our backbone forward model. Given a source Cantonese sentence x, the model seeks the target Mandarin sentence y that maximizes the conditional probability $P(y|x)$. In the inference phase, we employ Beam Search to generate a set of N best candidate sequences $\mathcal{Y} = \{y_1, y_2, ..., y_N\}$. For each candidate $y \in \mathcal{Y}$, its base forward score S_{fwd} is defined as the length-normalized log-probability calculated by mBART:

$$S_{\text{fwd}}(y|x) = \frac{1}{|y|} \sum_{t=1}^{|y|} \log P_{\text{mBART}}(y_t|y_{<t}, x) \quad (1)$$

where $|y|$ denotes the length of the target sequence.

3.2 Source-Side Faithfulness View

To explicitly model the semantic consistency (adequacy) between the source text and the translation, and to counter the ”Probability Trap” where fluent hallucinations are preferred, we establish the Source-Side Faithfulness View.

We train a reverse (Target-to-Source) translation model on the same parallel corpus. For each candidate translation y, we calculate the probability of reconstructing the source text x. This reconstruction likelihood serves as the reverse score S_{rev}:

$$S_{\text{rev}}(x|y) = \frac{1}{|x|} \sum_{j=1}^{|x|} \log P_{\text{rev}}(x_j|x_{<j}, y) \quad (2)$$

Functionally, this score is the negative cross-entropy loss of the reverse model given the pair (y, x). A higher S_{rev} implies that the candidate y retains sufficient information to restore the source x, thereby penalizing candidates with semantic omissions or deviations.

3.3 Local Fluency View

Traditional language models often struggle to balance local precision with global coherence. To ensure local syntactic correctness and idiomatic phrasing, we employ the Local Fluency View using Masked Language Models (such as BERT).

We utilize BERT to calculate Pseudo-Log-Likelihoods, which capture bidirectional context dependencies. By masking each token in the candidate sentence sequentially and predicting its original probability:

$$S_{\text{mlm}}(y) = \frac{1}{|y|} \sum_{i=1}^{|y|} \log P_{\text{BERT}}(y_i|y_{\setminus i}) \quad (3)$$

where $y_{\setminus i}$ denotes the sequence with the i-th token masked. This view effectively penalizes fine-grained syntactic errors and unnatural collocations (e.g., grammatical particles or word order issues) that may be produced by the forward model.

3.4 Global Fluency View

To compensate for the limited context window of BERT and to enhance sentence-level coherence, we introduce the Global Fluency View. We utilize a LLM to estimate the generation probability of the entire candidate sentence.

To align with the log-probability scale of the other modules (where higher scores indicate better quality), we define the LLM score S_{llm} as the negation of the model's cross-entropy loss:

$$S_{\text{llm}}(y) = \frac{1}{|y|} \sum_{t=1}^{|y|} \log P_{\text{llm}}(y_t \mid y_{<t}) \quad (4)$$

Leveraging its extensive pre-training knowledge, the LLM effectively identifies long-range dependencies and discourse-level patterns. This view ensures that the translation conforms to natural Mandarin linguistic patterns at the whole-sentence level, which is particularly crucial for longer, complex sentences.

3.5 Scoring and Optimization

We compute an adjustment score to refine the N-best candidates via a linear combination of the scores from the three views:

$$\begin{aligned} S_{\text{adjust}}(y, x) = \lambda_r \cdot S_{\text{rev}}(x|y) + \lambda_m \cdot S_{\text{mlm}}(y) \\ + \lambda_l \cdot S_{\text{llm}}(y) \end{aligned} \quad (5)$$

where λ_r, λ_m, and λ_l are hyperparameters controlling the weights of the faithfulness, local fluency, and global fluency features, respectively.

The final score for each candidate is defined as the combination of the original forward score and the adjustment score:

$$S_{\text{final}}(y, x) = S_{\text{fwd}}(y|x) + S_{\text{adjust}}(y, x) \quad (6)$$

Finally, we select the candidate with the highest aggregated score as the final translation output $\hat{y}$:

$$\hat{y} = \arg \max_{y \in \{y_1, ..., y_N\}} S_{\text{final}}(y, x) \quad (7)$$

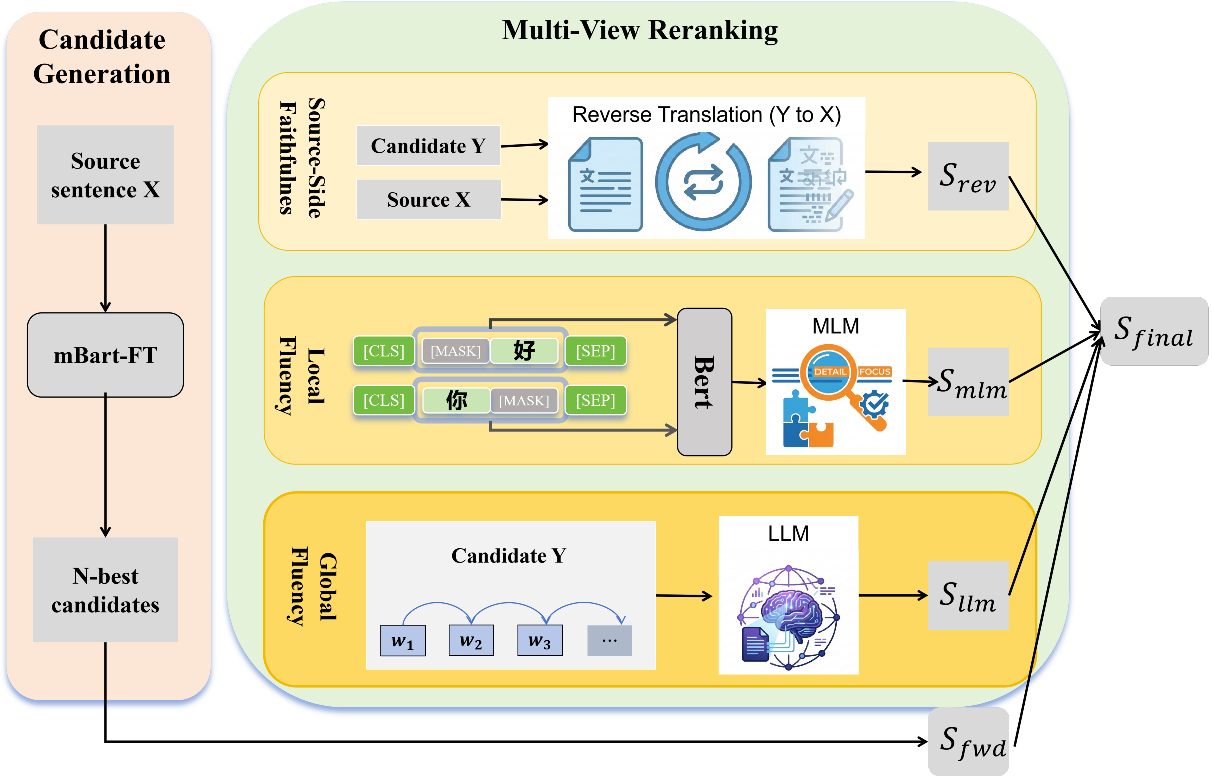

Figure 1: Overview of the MVS-Rank framework. We generate N-best candidates using mBART and rerank them by integrating scores from three views: Source-Side Faithfulness (Reverse Model), Local Fluency (BERT), and Global Fluency (LLM).

4 Experiment

4.1 Dataset

In our experiments, we use two datasets, Dare-Cantonese (Dare et al., 2023) and Cantonese-Chinese-Parallel. The dataset descriptions are provided below, and their statistics are reported in Table 1.

- **Dare-Cantonese** (Dare et al., 2023) : including: (1) 1,004 manually transcribed and aligned sentence pairs from the Cantonese-HK/Chinese-HK Universal Dependencies Treebank; and (2) 13,004 bilingual sentences from the Kaifangcidian[2] online dictionary.

- **Cantonese-Chinese-Parallel**: A parallel dataset of Cantonese and Traditional Chinese, extracted from a publicly available corpus.[3]

Split	Dare-Cantonese	Cantonese-Chinese Parallel
Train	10,085	10,800
Validation	1,121	1,200
Test	2,802	3,000

Table 1: Dataset Statistics

4.2 Baseline

We compare our approach with five baseline methods, spanning traditional models to state-of-the-art LLMs. Detailed descriptions are provided in Appendix A.

4.3 Experimental Setup

We employ the fine-tuned `mBART-large-50`[4] as the generator to produce the top-$N = 10$ candidate translations. For the re-ranking components, we utilize `bert-base-chinese`[5] to calculate local fluency (Eq. 3) and `Qwen2.5-7B`[6] to measure global fluency (Eq. 4). All experiments are conducted on NVIDIA RTX 3090 (24GB) GPUs; fine-tuning the mBART model requires approximately 1 hour, while the re-ranking process takes about 50

[2] https://kaifangcidian.com/han/yue/

[3] https://huggingface.co/datasets/raptorkwok/cantonese-traditional-chinese-parallel-corpus

[4] https://huggingface.co/facebook/mbart-large-50

[5] https://huggingface.co/google-bert/bert-base-chinese

[6] https://huggingface.co/Qwen/Qwen2.5-7B

Dataset	Method	SacreBLEU	BERTscore	ChrF	ChrF++	COMET
Dare-Cantonese	NLLB	11.29	0.7096	11.29	9.73	0.6736
	Dic-based	21.44	0.7841	19.14	16.53	0.7825
	Qwen2.5-ZS	23.41	0.7977	21.48	18.67	0.7960
	Llama3.1-ZS	14.61	0.7824	14.82	12.87	0.7932
	mBART-FT	40.11	0.8504	34.23	30.50	0.8469
	MVS-Rank	**42.52**	**0.8584**	**36.26**	**32.40**	**0.8597**
Cantonese-Chinese-Parallel	NLLB	13.75	0.6950	13.18	9.92	0.6675
	Dic-based	29.57	0.7890	27.03	23.93	0.7959
	Qwen2.5-ZS	25.68	0.7849	22.49	17.26	0.7907
	Llama3.1-ZS	19.32	0.7709	17.79	13.88	0.7846
	mBART-FT	39.07	0.8013	35.74	29.95	0.8000
	MVS-Rank	**40.62**	**0.8052**	**37.00**	**30.92**	**0.8096**

Table 2: Main results on Dare-Cantonese and Cantonese-Chinese-Parallel datasets. The best performance in each column is highlighted in bold. 'ZS' denotes Zero-Shot setting.

	SacreBLEU	**BERTscore**	**ChrF**	**ChrF++**	**COMET**
MVS-Rank	**42.52**	**0.8584**	**36.26**	**32.40**	**0.8597**
w/o Reverse Translation	40.93	0.8496	34.98	31.16	0.8511
w/o MLM	42.10	0.8570	35.90	32.02	0.8581
w/o LLM	42.24	0.8572	36.00	32.17	0.8575
S_{fwd} only (mBART)	40.11	0.8504	34.23	30.50	0.8469

Table 3: Ablation study of the MVS-Rank framework on the Dare-Cantonese dataset, systematically removing the Source-Side Faithfulness (Reverse), Local Fluency (MLM), and Global Fluency (LLM) components.

minutes. The hyperparameters λ_m, λ_r, and λ_l are determined via a grid search on the validation set, selected to maximize the SacreBLEU score.

4.4 Evaluation Metrics

To comprehensively assess translation quality, we employ five metrics: SacreBLEU, ChrF, ChrF++, BERTScore, and COMET. Detailed descriptions of these evaluation metrics are provided in Appendix B.

4.5 Main result

Table 2 presents the evaluation results on the *Dare-Cantonese* and *Parallel* datasets. Our proposed MVS-Rank consistently achieves the best performance across all metrics.

Baseline Analysis. The performance of baselines highlights their respective limitations: NLLB performs poorly, yielding the lowest scores across all metrics, as its capacity is diluted across many languages and lacks the dialect-specific depth required for Cantonese. The dictionary-based model is heavily constrained by lexicon coverage, failing to translate out-of-vocabulary terms. Notably, Zero-shot LLMs exhibit a divergence between respectable COMET and low SacreBLEU scores. This confirms they generate fluent Mandarin paraphrases, preserving semantics but lacking the specific Cantonese lexical accuracy. mBART-FT serves as the strongest baseline due to domain adaptation, significantly outperforming other methods.

Our Method. MVS-Rank further improves upon the strong mBART baseline by +2.41 and +1.55 SacreBLEU on the two datasets, respectively. This demonstrates that our reranking framework —which enforces faithfulness through the reverse model and ensures multi-granularity fluency via hybrid language models—effectively mitigates semantic drift and the "Mandarinization" issues observed in standard fine-tuning.

4.6 Ablation Study

To assess the individual contributions of each component in the proposed framework, we perform an ablation study on the Dare-Cantonese dataset by systematically removing the Reverse Translation model (Source-Side Faithfulness View), the MLM scorer (Local Fluency View), and the LLM scorer (Global Fluency View). The results are summarized in Table 3.

ID	Model	Sentence	Remark / Error Analysis
1	*Source*	食得啦，大家**埋位**吖	*Gloss: Ready to eat, everyone take your seats.*
	Ref	可以吃了，大家入席吧	
	NLLB	請大家坐下來,	**Omission**: Misses the ”eating” context.
	Dic-based	食得啦，大家埋位啊	**Copying**: Output is identical to Cantonese source.
	Qwen2.5-ZS	吃得了，大家讓個位置吧。	**Mistranslation**: Interprets “埋位” (take seat) as “give up seat”.
	LLama3.1-ZS	吃饱了，大家都坐好吧	**Mistranslation**: Translates “Ready to eat” as “Full”.
	mBART-FT	吃吧，大家同席吧	**Grammar**: “ ” is not a commonly used expression in Mandarin..
	MVS-Rank	可以吃啦，大家**入席**吧	**Correct**.
2	*Source*	攞嚿石**責住**佢	*Gloss: Take a stone and press it down.*
	Ref	拿塊石頭壓著它	
	NLLB	拿著石頭責他。	**Literal**: Copies “責” without translating meaning.
	Dic-based	拿塊石責住他	**Mixed**: Partially Cantonese (“責住”).
	Qwen2.5-ZS	拿些石頭砸住他	**Nuance**: “砸” (smash) implies violence.
	LLama3.1-ZS	責怪他一通	**Mistranslation**: Mistakes “責” for “blame/scold”.
	mBART-FT	拿塊石頭罩著它	**Mistranslation**: “罩” means cover, not press.
	MVS-Rank	拿塊石頭**壓著**它	**Correct**.
3	*Source*	**頭先個人**真係好心	*Gloss: That person just now was really kind.*
	Ref	剛才那個人真是好心	
	NLLB	我以前是個很有禮貌的人。	**Hallucination**: Completely unrelated to source.
	Dic-based	剛剛個人真的好心	**Grammar**: “個人” does not follow standard classifier usage in Mandarin.
	Qwen2.5-ZS	先前本人真是很用心。	**Entity Error**: Changes “that person” to “myself”.
	LLama3.1-ZS	頭先的人民是很好的心意	**Mistranslation**: Interprets “個人” as “people”.
	mBART-FT	剛才個人真好心	**Grammar**: “個人” does not follow standard classifier usage in Mandarin.
	MVS-Rank	剛才**那個人**真好心	**Correct**.
4	*Source*	**成間**學校**得一個**球場。	*Gloss: The whole school only has one court.*
	Ref	整個學校只有一個足球場。	
	NLLB	學校有個球場。	**Omission**: Misses “whole” and “only”.
	Dic-based	成間學校得一個球場。	**Copying**: Output is identical to Cantonese source.
	Qwen2.5-ZS	所有學校都有一個球場。	**Mistranslation**: “All schools” contradicts “one school”.
	LLama3.1-ZS	該所學校有個運動場。	**Omission**: Misses the constraint “only”.
	mBART-FT	整個學校都有一個體育場。	**Mistranslation**: “都有” (all have) vs “只有” (only have).
	MVS-Rank	整個學校**只有**一個球場。	**Correct**.

Table 4: Comprehensive qualitative analysis of translation outputs. We compare our proposed method against multiple baselines including NLLB, Dictionary-based Translation, Qwen, LLama, and fine-tuned mBART.

As shown in Table 3, the full MVS-Rank method outperforms all ablated variants, validating the effectiveness of our multi-view scoring mechanism. Removing the Reverse Translation module results in the most significant performance degradation across all metrics (e.g., SacreBLEU drops by 1.59). This underscores the critical role of source-side faithfulness in ensuring semantic adequacy and filtering out hallucinations. The removal of the MLM scorer leads to a clear drop in ChrF/ChrF++, which relies on character n-grams. This aligns with the role of MLM in ensuring Local Fluency, as it helps refine morphology and local word patterns. The LLM scorer (Global Fluency) provides further marginal gains, ensuring the final output is not only grammatically correct but also coherent at the sentence level. All ablated variants still outperform the S_{fwd} only baseline (mBART), demonstrating that our re-ranking strategy—even with partial components—consistently selects better translations than the greedy decoding of the base model.

4.7 Case Study

4.7.1 Comparison with Baselines

To qualitatively compare the translation outputs of various baseline methods with our proposed method MVS-Rank, we select several representative examples from the dataset. The reference sentences are the annotated translations provided in the dataset. Detailed examples are presented in Table 4.

	Sentence (Chinese)	Remark / Gloss	mBART	MVS-Rank	w/o	w/o	w/o
					MLM	LLM	Rev.
Source	呢個問題睇嚟仲有得詉	*Gloss: This issue seems still open for debate.*	-	-	-	-	-
Ref	這個問題看來還有一番爭論		-	-	-	-	-
10-best Cands.	這個問題看來還是沒有希望了	**Mistranslation**: Interprets “can argue” as “hopeless”.	**1**	5	7	6	**1**
	這個問題看來還是沒有希望	**Mistranslation**: Interprets “can argue” as “hopeless”.	2	8	10	8	2
	這個問題看來還有得爭論	**Correct**.	3	**1**	**1**	**1**	5
	這個問題看來還能有打開策	**Hallucination**: Invented term “opening strategy”.	4	10	9	10	9
	這個問題看來還有得及辦	**Mistranslation**: “Manageable” instead of “debatable”.	5	3	3	2	6
	這個問題看來還能有打開的方案	**Mistranslation**: “Open solution” instead of “debate”.	6	7	8	7	4
	這個問題看來還能有打開的了	**Hallucination**: Ungrammatical meaning.	7	9	5	9	7
	這個問題看來還有些沒有完	**Mistranslation**: “Unfinished” vs “Debatable”.	8	4	4	4	3
	這個問題看來還有得及而爭	**Hallucination**: Ungrammatical phrasing.	9	6	6	5	10
	這個問題看來還有得挽回	**Mistranslation**: “Salvageable” vs “Debatable”.	10	2	2	3	8

Table 5: Ablation case study demonstrating the impact of the Reverse Translation model. Numbers indicate the rank of the corresponding sentences within the candidate translations.

	Sentence (Chinese)	Remark / Gloss	mBART	MVS-Rank	w/o	w/o	w/o
					MLM	LLM	Rev.
Source	碎料嚟啫	*Gloss: It’s just a trivial matter / a piece of cake.*	-	-	-	-	-
Ref	小意思而已		-	-	-	-	-
10-best Cands.	都是些小碎料	**Literal**: Retains Cantonese term “碎料”, unnatural in Mandarin.	**1**	2	**1**	2	3
	都是些小碎兒	**Grammar**: Unnatural phrasing (“small bits”).	2	5	7	6	7
	小兒科而已	**Acceptable**: Idiomatic meaning (“child’s play”).	3	6	8	4	5
	這只是小兒科而已	**Acceptable**: Natural phrasing.	4	3	3	5	2
	只是小兒科而已	**Correct**.	5	**1**	2	**1**	**1**
	這只是小提琴	**Mistranslation**: Translates as “violin” (小提琴).	6	7	4	8	6
	這只是小提檔	**Hallucination**: Nonsense output.	7	9	9	9	8
	都是些碎料	**Literal**: Retains Cantonese term “碎料”	8	8	5	7	9
	都是些小碎的東西	**Literal**: Misses the metaphorical meaning.	9	4	6	3	4
	算些零碎	**Literal**: Misses the metaphorical meaning.	10	10	10	10	10

Table 6: Ablation case study demonstrating the impact of the MLM scorer. Numbers indicate the rank of the corresponding sentences within the candidate translations.

First, we observe that NLLB suffers from significant *information omission*. For instance, in Case 1, the core action “eating” is missing, and in Case 4, the critical quantifier “only one” (得一個) is dropped, leading to an incomplete translation.

Second, the Dictionary-based method is heavily constrained by lexicon coverage. When encountering Cantonese-specific terms not present in the dictionary, it resorts to copying the source text directly. A clear example is found in Case 2, where the verb “責住” (press down) is left untranslated, resulting in a code-mixed output.

Third, LLMs (Qwen and LLaMA) demonstrate a tendency to generate highly fluent but *semantically unfaithful* sentences (hallucinations). They often fail to grasp specific Cantonese semantics, leading to divergence from the original meaning. Taking Case 2 as an example, the Cantonese character “責” implies “to press.” However, LLaMA misinterprets it as the Mandarin homograph for “blame” (責怪), while Qwen interprets it as “smash” (砸), both deviating from the source intent.

Finally, while the mBART-FT baseline achieves the highest semantic similarity to the reference among all baselines, it still struggles with precise nuances and occasional unnatural phrasing (e.g., mistranslating “only have” as “all have” in Case 4).

In contrast, MVS-Rank effectively addresses these issues by incorporating the Multi-View Scoring Reranking framework. By leveraging the reverse model to ensure semantic faithfulness and the hybrid language models to guarantee multi-granularity fluency, our method successfully corrects the lexical ambiguities and omissions found in baselines, yielding translations that are both accurate and natural.

	Sentence (Chinese)	**Remark / Gloss**	**mBART**	**MVS-Rank**	**w/o MLM**	**w/o LLM**	**w/o Rev.**
Source	佢去咗睇醫生	*Gloss: He went to see a doctor.*	-	-	-	-	-
Ref	他看病去了		-	-	-	-	-
10-best Cands.	他去了看病	**Slightly Unnatural**: Word order "去了..." is less idiomatic.	**1**	2	3	**1**	3
	他去看病去了	**Correct**.	2	**1**	**1**	4	**1**
	他去了看醫生	**Acceptable**: "Went to see the doctor".	3	5	7	3	4
	他去看了看醫生	**Mistranslation**: "Took a look at the doctor".	4	3	4	2	2
	他去看了看病	**Unnatural**: "Took a look at the illness".	5	4	2	5	10
	他去醫院看病	**Added Info**: Explicitly mentions "hospital".	6	8	8	9	6
	他去看病去了	**Acceptable**: Slightly unnatural.	7	6	6	6	7
	他到醫院看病	**Added Info**: "Arrived at hospital".	8	10	9	10	8
	他去看了醫生	**Acceptable**: "Went and saw the doctor".	9	9	10	8	5
	他去看病了	**Acceptable**: "Has gone to see a doctor".	10	7	5	7	9

Table 7: Ablation case study demonstrating the impact of the LLM scorer. Numbers indicate the rank of the corresponding sentences within the candidate translations.

4.7.2 Ablation Case Study

To qualitatively assess the contributions of the Reverse Translation model, the MLM scorer, and the LLM scorer, we select several representative cases from the Dare-Cantonese dataset to highlight their respective effects.

Ablation case study highlighting the role of the Reverse Translation model: As shown in Table 5, the character "諚" is a rare character in Mandarin. The source phrase "有得諚" means "open for debate." The mBART baseline incorrectly predicts 沒有希望" ("hopeless"), likely due to a semantic hallucination. By contrast, MVS-Rank, leveraging the faithfulness constraint from the reverse model, correctly produces the semantically accurate translation "還有得爭論" ("debatable").

Ablation case study highlighting the role of the MLM scorer: Table 6 shows an ablation case study on translating Cantonese slang. The source "碎料" (fragments) metaphorically means "trivial matter". The mBART baseline (Rank 1) produces a literal translation "小碎料", retaining the dialect term which is incorrect in Standard Chinese. The variant w/o MLM fails to correct this lexical issue (Rank 1 remains unchanged). However, MVS-Rank successfully identifies the idiomatic Mandarin equivalent "小兒科" (child's play). This highlights the role of the MLM scorer in filtering out non-standard lexical usages.

Ablation case study highlighting the role of the LLM scorer: As shown in Table 7, the mBART baseline produces "他去了看病", which suffers from slightly awkward word order. The variant w/o LLM fails to distinguish this nuance and assigns it Rank 1. In contrast, MVS-Rank, leveraging the LLM's capability in modeling global fluency and long-range dependencies, correctly reranks the more natural phrasing "他去看病去了" to the top position.

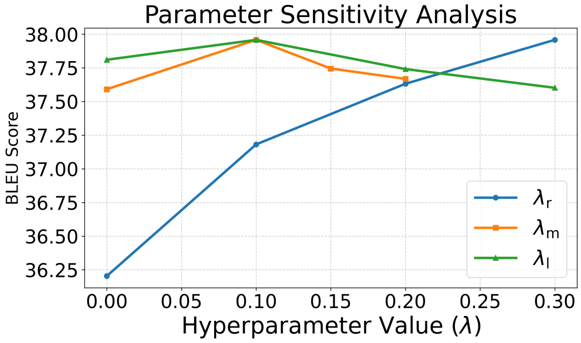

Figure 2: Parameter sensitivity analysis of the weighting coefficients λ_r, λ_m, and λ_l on SacreBLEU score.

Additional ablation case studies are provided in Appendix D.

4.8 Parameter Sensitivity

As shown in Figure 1, λ_r demonstrates a monotonic improvement in SacreBLEU score as it increases from 0.0 to 0.3, highlighting its strong contribution to model performance. On the other hand, both λ_m and λ_l exhibit an inverted U-shaped behavior, peaking at $\lambda = 0.1$. Beyond this threshold, further increases may degrade performance, likely due to excessive interference with the primary training objective. These results suggest that while λ_r can be safely increased within the tested range, careful tuning is required for λ_m and λ_l to avoid negative impacts.

Limitations

A primary limitation of our method lies in its dependence on the search space defined by the forward model (mBART). As a reranking approach, our method selects the optimal hypothesis from a generated N-best list but lacks the capability to synthesize new translations from scratch. Therefore, the improvement ceiling is determined by the recall of the N-best list; if the mBART model fails to generate a valid translation within the top-N candidates, our reranker is unable to produce the correct output.

Acknowledgments

This research is supported by the Joint Project of Philosophy and Social Sciences Planning Discipline in Guangdong Province (Grant No.: GD23XZY07), and the Ordinary University Characteristic Innovation Project of Guangdong Province (Grant No.: 2023KTSCX031).

References

Chen-Chi Chang, Chong-Fu Li, Chu-Hsuan Lee, and Hung-Shin Lee. 2025. Enhancing low-resource minority language translation with llms and retrieval-augmented generation for cultural nuances. *CoRR*, abs/2505.10829.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, and 19 others. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Yuqian Dai, Chun Fai Chan, Ying Ki Wong, and Tsz Ho Pun. 2025. Next-level cantonese-to-mandarin translation: Fine-tuning and post-processing with llms. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025 - Workshops, Abu Dhabi, UAE, January 19-24, 2025*, pages 427–436. Association for Computational Linguistics.

Megan Dare, Valentina Fajardo Diaz, Averie Ho Zoen So, Yifan Wang, and Shibingfeng Zhang. 2023. Unsupervised mandarin-cantonese machine translation. *CoRR*, abs/2301.03971.

Kung Hong, Lifeng Han, Riza Batista-Navarro, and Goran Nenadic. 2024. CantonMT: Cantonese to English NMT platform with fine-tuned models using real and synthetic back-translation data. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 590–599, Sheffield, UK. European Association for Machine Translation (EAMT).

Jiyue Jiang, Pengan Chen, Liheng Chen, Sheng Wang, Qinghang Bao, Lingpeng Kong, Yu Li, and Chuan Wu. 2025. How well do llms handle cantonese? benchmarking cantonese capabilities of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 4464–4505. Association for Computational Linguistics.

Zhanibek Kozhirbayev. 2024. Enhancing neural machine translation with fine-tuned mbart50 pre-trained model: An examination with low-resource translation pairs. *Ingerie des systs d information*, 29(3):831–838.

Evelyn Kai-Yan Liu. 2022. Low-resource neural machine translation: A case study of Cantonese. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 28–40, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pretraining for neural machine translation. *Trans. Assoc. Comput. Linguistics*, 8:726–742.

Hei Yi Mak and Tan Lee. 2021. Low-resource NMT: A case study on the written and spoken languages in hong kong. In *NLPIR 2021: 5th International Conference on Natural Language Processing and Information Retrieval, Sanya, China, December 17 - 20, 2021*, pages 81–87. ACM.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

King Yiu Suen, Rudolf Chow, and Albert Y.S. Lam. 2024. Leveraging Mandarin as a pivot language for low-resource machine translation between Cantonese and English. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 74–84, Bangkok, Thailand. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *CoRR*, abs/2008.00401.

Llama Team. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Wen Wang, Andreas Stolcke, and Jing Zheng. 2007. Reranking machine translation hypotheses with structured and web-based language models. In *IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU 2007, Kyoto, Japan, December 9-13, 2007*, pages 159–164. IEEE.

Yan Wu, Xiukun Li, and Caesar Lun. 2006. A structural-based approach to cantonese-english machine translation. *Int. J. Comput. Linguistics Chin. Lang. Process.*, 11(2).

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *CoRR*, abs/2407.10671.

Xiaoheng Zhang. 1998. Dialect MT: A case study between cantonese and mandarin. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, COLING-ACL '98, August 10-14, 1998, Université de Montréal, Montréal, Quebec, Canada. Proceedings of the Conference*, pages 1460–1464. Morgan Kaufmann Publishers / ACL.

A Appendix A: Baseline Details

The baselines are described as follows:

- **NLLB** (Costa-jussà et al., 2022): A Transformer-based multilingual translation model optimized for low-resource languages, which controls translation direction via target language identifiers. We use the `nllb-200-distilled-600M` variant supporting both Cantonese and Mandarin.

- **Dictionary-based Translation (Dic-based)**: A traditional hybrid system that combines dictionary lookup for lexical mapping with language model-based rules for reordering Cantonese syntax to Mandarin.

- **mBART Fine-tuned (mBART-FT)** (Tang et al., 2020): A sequence-to-sequence denoising autoencoder pre-trained on large-scale multilingual corpora is used as the backbone model. We perform full-parameter fine-tuning.

- **Qwen2.5 Zero-shot (Qwen2.5-ZS)** (Yang et al., 2024): We evaluate the `Qwen2.5-7B-Instruct` variant. The model is instructed to translate Cantonese input into Mandarin using the specific prompt illustrated in Figure 3.

- **Llama3.1 Zero-shot (Llama3.1-ZS)** (Team, 2024): We evaluate the `Llama3.1-8B-Instruct` variant by applying the same zero-shot prompting strategy illustrated in Figure 3 to generate translations

B Appendix B: Evaluation Metrics Details

We further elaborate on the evaluation metrics used in our experiments. SacreBLEU measures lexical overlap by computing the geometric mean of n-gram precisions, scaled by a brevity penalty. ChrF and ChrF++ evaluate translation quality using character n-gram F-scores, which are particularly robust for morphologically rich languages. While ChrF relies solely on character n-grams, ChrF++ additionally incorporates word n-grams. BERTScore captures semantic similarity using contextualized embeddings; specifically, we employ the `bert-base-chinese` model to calculate BERTScore. Finally, COMET utilizes cross-lingual pre-trained language models to encode semantic information, measuring the semantic consistency between the generated translation and the reference text. We employ the `wmt22-comet-da` variant[7] for our evaluation.

C Appendix C: Prompt

The prompts used for LLM-based translations are presented in Figure 3.

D Appendix D: More Ablation Case Study

Additional ablation case studies are presented in Tables 8, 9, and 10.

[7] `https://huggingface.co/Unbabel/wmt22-comet-da`

Qwen prompt

You are a professional Cantonese translation expert. Please accurately translate the following Cantonese sentences into Traditional Written Mandarin, using formal and natural Chinese expressions without adding any explanations, titles, or extra content. Only output the translated Traditional Mandarin sentences.
Cantonese:
Traditional Mandarin:

(a) Qwen prompt

Llama prompt

You are a professional Cantonese translation expert. Please accurately translate the following Cantonese sentences into Traditional Written Mandarin, using formal and natural Chinese expressions without adding any explanations, titles, or extra content. Only output the translated Traditional Mandarin sentences. Do not include prefixes such as ’Traditional Mandarin:’ or similar phrases.
Cantonese:
Traditional Mandarin:

(b) Llama prompt

Figure 3: Prompts used for LLM-based translation.

	Sentence (Chinese)	Remark / Gloss	mBART (Rank)	MVS-Rank Rank	w/o MLM	w/o LLM	w/o Rev
Source	我尋晚成晚冇瞓	*Gloss: I didn’t sleep all night last night.*	-	-	-	-	-
Ref	我昨晚一夜沒睡		-	-	-	-	-
10-best Cands.	我昨晚整夜沒合適	**Hallucination**: “Didn’t suit/fit” makes no sense here.	**1**	2	**1**	2	3
	我昨晚整夜沒合蹤	**Hallucination**: Nonsense.	2	5	4	5	6
	我昨晚整夜沒合宿	**Hallucination**: Nonsense.	3	4	5	3	5
	我昨晚整夜沒睡覺	**Acceptable**: “Didn’t sleep”.	4	3	3	4	**1**
	我昨晚整夜沒睡	**Correct**.	5	**1**	2	**1**	2
	我昨天晚上沒合過眼	**Idiomatic**: “Didn’t close my eyes”.	6	8	8	9	4
	我昨晚整夜沒合乎	**Hallucination**: Nonsense.	7	7	7	7	9
	我昨晚整夜沒有合適	**Collocation Error**: Same as Rank 1.	8	6	6	6	7
	我昨晚睡大覺	**Mistanslation**: Opposite meaning (“Slept soundly”).	9	10	10	8	8
	我昨晚沒睡覺	**Omission**: Misses “all night” (整夜).	10	9	9	10	10

Table 8: Ablation case study highlighting the role of the MLM scorer. The numbers in the table indicate the ranking of the corresponding sentences in the candidate translations. The mBART baseline (Rank 1) generates “沒合適” (didn’t suit), which is a clear collocation error likely intended to be “沒合眼” (didn’t close eyes). **w/o MLM** fails to correct this error (assigning it Rank 1). However, MVS-Rank, leveraging BERT’s bidirectional context modeling, identifies this awkward phrasing and correctly reranks the fluent translation “整夜沒睡” to the top position.

	Sentence (Chinese)	Remark / Gloss	mBART	MVS-Rank	w/o	w/o	w/o
			(Rank)	Rank	MLM	LLM	Rev.
Source	佢會好快好翻啩	*Gloss: He will get better soon, I suppose.*	-	-	-	-	-
Ref	他會很快好起來吧		-	-	-	-	-
10-best Cands.	他會很快的回來	**Mistranslation**: “Come back” (回來) vs “Recover” (好翻).	**1**	4	4	4	**1**
	他會很快好回來的	**Mistranslation**: Blends “recover” and “come back”.	2	3	3	3	6
	他會很快很快的回來	**Mistranslation**: Blends “recover” and “come back”.	3	7	9	7	2
	他會很快好起來	**Acceptable**: “He will get better soon”.	4	2	2	2	5
	他會很快好起來的	**Correct**.	5	**1**	**1**	**1**	3
	他會很快的回來的	**Mistranslation**: Blends “recover” and “come back”.	6	5	5	6	4
	他會很快好 회복	**Hallucination**: Contains Korean text.	7	9	8	9	10
	他會很快的復歸	**Unnatural**: “Revert/Return” is archaic/ technical.	8	8	7	8	8
	他會很快回來的	**Mistranslation**: Blends “recover” and “come back”.	9	10	10	10	7
	它會很快好起來	**Wrong Pronoun**: “It” (它) vs “He” (他).	10	6	6	5	9

Table 9: Ablation case study demonstrating the impact of the Reverse Translation model. Numbers indicate the rank of the corresponding sentences within the candidate translations. The source term 好翻” means “recover.” The mBART baseline mistranslates it as 回來” (“come back”), likely due to a hallucination. The **w/o Rev.** variant fails to detect this semantic error, leaving it at Rank 1. In contrast, MVS-Rank, leveraging the faithfulness constraint from the Reverse Model, correctly identifies the semantic mismatch and reranks the accurate translation “好起來” (“recover”) to the top.

	Sentence (Chinese)	Remark / Gloss	mBART	MVS-Rank	w/o	w/o	w/o
			(Rank)	Rank	MLM	LLM	Rev.
Source	佢冇做到功課	*Gloss: He didn’t do his homework.*	-	-	-	-	-
Ref	他沒有做功課		-	-	-	-	-
10-best Cands.	他做不到功課	**Mistranslation**: “Inability” (做不到) vs. “Did not” (沒有).	**1**	2	2	**1**	**1**
	他沒功課	**Mistranslation**: He has no homework.	2	7	3	8	8
	他沒做功課	**Acceptable**: (Correct meaning).	3	**1**	**1**	2	3
	他沒有功課	**Mistranslation**: He has no homework.	4	4	4	5	5
	他不功課	**Hallucination**: Nonsense.	5	10	9	10	10
	他未能功課	**Hallucination**: Ungrammatical.	6	6	8	4	7
	他沒有做功課	**Correct**.	7	3	5	3	2
	他還沒做功課	**Added Info**: Added “yet”.	8	5	6	7	4
	他不做功課	**Mistranslation**: “did not” was translated as “refuse” Refusal.	9	9	10	6	6
	他沒功課做	**Mistranslation**: “did not” was translated as “has not”.	10	8	7	9	9

Table 10: Ablation case study demonstrating the impact of the LLM scorer and the Reverse Translation model. Numbers indicate the rank of the corresponding sentences within the candidate translations. The Cantonese source “冇做到” simply means “did not do”. However, the mBART baseline (Rank 1) hallucinates capability, translating it as “做不到” (unable to do). Both **w/o LLM** and **w/o Rev** fail to correct this semantic error, keeping the incorrect candidate at Rank 1. MVS-Rank effectively identifies the nuance and ranks the acceptable translation “沒做功課” as the top candidate.