

Hy-MT2-1.8B for WMT26: Staged Multilingual Adaptation for Chinese–Southeast Asian Translation

Anonymous Submission

Abstract

We describe a training and evaluation pipeline for Hy-MT2-1.8B in the WMT26 Chinese–Southeast Asian multilingual machine translation task. The system uses one instruction-following model for fourteen directions between Chinese and Indonesian, Thai, Khmer, Malay, Vietnamese, Burmese, and Lao. Our pipeline separates three responsibilities: monolingual continued pretraining (CPT), bilingual supervised fine-tuning (SFT), and evaluation-time serving. We construct leakage-controlled data with a complete chat-template token-length check, train a shared CPT checkpoint, and compare parameter-efficient LoRA SFT against full-parameter SFT under matched data, seed, and evaluation conditions. On the completed 26,474-example validation set, full-parameter SFT achieves higher macro-average sacreBLEU, COMET, and Translation Quality Score (the official WMT26 metric) than LoRA SFT, although the magnitude of improvement varies across directions.

1 Introduction

The WMT26 Chinese–Southeast Asian multilingual translation task (WMT26 Chinese–Southeast Asian Task Organizers, 2026) requires a single system to translate in both directions between Chinese and seven Southeast Asian languages. This setting combines two practical constraints. First, the language directions are heterogeneous: the model must handle different scripts, tokenization behavior, and resource levels without routing to a separate model for each pair. Second, the final system must be evaluated not only for translation quality but also for serving efficiency.

We study this setting with Hy-MT2-1.8B (Zheng et al., 2026), a compact instruction-following multilingual model in the broader text-to-text and multilingual pretraining lineage (Raffel et al., 2020; Xue et al., 2021; Liu et al., 2020; NLLB Team, 2022). Rather than building a separate model for each

direction, we express translation as a controlled chat interaction. Each example contains an explicit source language, target language, and instruction to output only the translation. The same model and prompt family are used for all fourteen directions.

The main engineering question in this work is how much adaptation is required after monolingual continued pretraining (Gururangan et al., 2020). We therefore construct two supervised adaptation branches from the same CPT checkpoint. The first uses LoRA (Hu et al., 2022) on the attention projections, while the second updates all model parameters. This comparison isolates the trade-off between parameter-efficient adaptation and full-model specialization while keeping the data construction, context length, seed, and evaluation protocol fixed.

Our contributions are:

- a reproducible, leakage-controlled data pipeline for fourteen Chinese–Southeast Asian translation directions;
- a staged adaptation design that uses monolingual CPT before bilingual SFT and keeps the two SFT branches comparable;
- an implementation-level comparison of LoRA and full-parameter SFT for Hy-MT2-1.8B; and
- a direction-level evaluation protocol that reports sacreBLEU, COMET, and the official WMT26 Translation Quality Score without mixing validation data with the organizer-held blind test set.

The quantitative comparison is completed on the shared validation set. Full-parameter SFT improves the macro-average BLEU, COMET, and Translation Quality Score over LoRA SFT, while the direction-level analysis shows that the gains are not uniform across languages and directions.

Direction family	Directions
Southeast Asian → Chinese	ID-ZH, TH-ZH, KM-ZH, MS-ZH, VI-ZH, MY-ZH, LO-ZH
	ZH-ID, ZH-TH, ZH-KM, ZH-MS, ZH-VI, ZH-MY, ZH-LO

Table 1: The fourteen translation directions covered by the system.

2 Task and System Scope

The task covers Chinese and seven Southeast Asian languages: Indonesian (ID), Thai (TH), Khmer (KM), Malay (MS), Vietnamese (VI), Burmese (MY), and Lao (LO). We evaluate the fourteen directions shown in Table 1. The system uses one shared Transformer-style model (Vaswani et al., 2017), one tokenizer, and one instruction format for all directions.

The development validation split and the organizer-held blind-test split are kept separate. The blind-test split is not used for training, check-point selection, prompt design, or decoding-parameter tuning.

3 Data Construction

3.1 Source resources

The data builder reads the WMT26 bilingual CSV files for each Southeast Asian language and the corresponding monolingual training files. Bilingual pairs are expanded into two instruction examples: one from the Southeast Asian language to Chinese and one from Chinese to the Southeast Asian language. The two directions are generated from the same original split, which prevents placing opposite directions of one parallel pair into different splits. The tokenizer and subword behavior are treated as part of the translation design because subword units are important for rare words and mixed-resource directions (Sennrich et al., 2016).

The resulting data resources contain bilingual supervised examples, monolingual continuation-pretraining text, and audit summaries used internally to verify counts, rejection reasons, and split separation. These internal construction records support the data analysis but are not part of the manuscript’s scientific claims.

After expanding both directions and applying length filters, the processed validation set contains 26,474 records across all fourteen directions. Six directions have 2,000 examples each: ID-ZH, TH-ZH, MS-ZH, ZH-ID, ZH-TH, and ZH-MS. KM-ZH and ZH-KM have 1,999 examples; VI-ZH and

Language	Records	Avg. chars/line	Tokens/char	Avg. tokens/line
Indonesian (ID)	98,001	931.5	0.31	286.82
Khmer (KM)	97,998	1,020.3	1.56	1,589.30
Lao (LO)	97,991	1,642.5	2.05	3,362.66
Malay (MS)	96,179	1,095.8	0.30	333.76
Burmese (MY)	97,252	1,662.0	1.21	2,014.41
Thai (TH)	98,001	869.6	0.54	468.88
Vietnamese (VI)	98,001	1,130.2	0.36	411.29

Table 2: Monolingual CPT corpus size and tokenizer efficiency per Southeast Asian language. *Records* is the number of segments in the official monolingual training file after length filtering. *Avg. chars/line* is the mean number of characters per segment. *Tokens/char* is measured by re-tokenizing 2,000-record random samples from each file with the shared Hy-MT2-1.8B tokenizer, and *Avg. tokens/line* is the per-segment token count derived from the same sample. The seven languages differ in tokenizer efficiency by up to a factor of seven, which is the empirical input to the script-driven analysis in Section 6.

ZH-VI have 1,996; MY-ZH and ZH-MY have 1,989; and LO-ZH and ZH-LO have 1,253. Both LoRA SFT and full-parameter SFT are evaluated on this same reconstructed validation set.

3.2 Normalization and leakage control

Before tokenization, each text is normalized to Unicode NFC, carriage returns and tabs are canonicalized, control characters are removed, repeated whitespace is compressed, and leading or trailing whitespace is stripped. Empty fields and exact duplicates are rejected. Validation examples are also checked against the training split at both the pair and segment level.

We count the complete chat-formatted example rather than counting only the source sentence. The prompt, source, target, and model-specific chat markers must fit within a fixed 4,096-token training context. Examples that exceed this limit are rejected during data construction instead of being silently truncated. This is important for translation because truncation near the end of a target can create incomplete supervision.

The construction pipeline applies explicit split-separation checks and rejects any example that would mix development data with the organizer-held blind-test split.

3.3 Instruction format

For a source language s , target language t , and source text x , the user message is instantiated as:

Translate the following s into t . Output only the translation and do not add explanations.

x

The actual implementation uses the Chinese-language equivalent of this instruction and the model’s native chat template. The assistant message in training contains only the reference translation. At inference time, the assistant response is the model prediction. This contract discourages explanations, language identification, and chat-style preambles that are harmful in a machine translation evaluation.

4 Training Method

4.1 Monolingual continued pretraining

We first adapt the base Hy-MT2-1.8B model with monolingual text from the seven Southeast Asian languages. The monolingual streams are tokenized without special chat formatting and packed into fixed-length blocks of 4,096 tokens. The training configuration uses 1,500 optimization steps, a learning rate of 2×10^{-5} , cosine decay, a 1% warmup ratio, weight decay of 0.01, bf16 computation, gradient accumulation of four, and a global seed of 2026. The model is trained with DeepSpeed ZeRO-2 offload and gradient checkpointing (Rajbhandari et al., 2020).

The resulting CPT 1-epoch checkpoint is the common starting point for both supervised branches. Sharing this checkpoint means that the LoRA and full-SFT comparison does not conflate adaptation method with different monolingual initialization.

4.2 Bilingual supervised fine-tuning

Both supervised branches use the same bilingual train and validation files, the same Hy-MT2 chat template, and a 4,096-token cutoff. Evaluation loss is measured periodically during training, and the checkpoint with the lowest validation loss is selected rather than assuming that the final checkpoint is optimal. The complete branch configurations are given in Tables 3 and 4.

LoRA SFT. The parameter-efficient branch applies LoRA to the query, key, value, and output projections of the attention layers. The LoRA rank is 64, the scaling factor is 128, and dropout is 0.05. Training uses a learning rate of 2×10^{-4} , three epochs, per-device batch size 8, gradient accumulation of two, cosine scheduling, 1% warmup, and bf16 computation.

Full-parameter SFT. The comparison branch updates all model parameters. It uses the same

Configuration	LoRA SFT
Initialization	CPT 1 epoch
Trainable parameters	Attention LoRA adapters
LoRA rank / alpha / dropout	64 / 128 / 0.05
LoRA targets	q, k, v, o projections
Learning rate	2×10^{-4}
Epochs	3
Per-device batch size	8
Gradient accumulation	2
Context cutoff	4,096
Precision	bf16
Seed	2026

Table 3: LoRA supervised fine-tuning configuration.

Configuration	Full-parameter SFT
Initialization	CPT 1 epoch
Trainable parameters	All model parameters
Learning rate	1×10^{-5}
Epochs	3
Per-device batch size	4
Gradient accumulation	4
Context cutoff	4,096
Precision	bf16
Seed	2026

Table 4: Full-parameter supervised fine-tuning configuration.

three-epoch schedule and seed, but a lower learning rate of 1×10^{-5} , per-device batch size 4, and gradient accumulation of four. The lower learning rate is intended to reduce destructive updates when the CPT checkpoint is further specialized on bilingual supervision.

4.3 Model preparation

A LoRA checkpoint is not used as a standalone generation model. We load the selected adapter on the CPT 1-epoch base model and merge the adapter weights into a complete model directory containing the configuration, tokenizer, custom model code, and model weights. The full-SFT branch is used directly when its selected checkpoint contains complete model weights. This distinction prevents confusing training-resume artifacts, such as optimizer or ZeRO states, with deployable model files.

5 Inference and Evaluation Protocol

5.1 Generation

The two final models are evaluated on exactly the same validation records and with the same instruction contract. The current evaluation implementation uses the Transformers ecosystem (Wolf et al.,

2020) for native inference and also supports an OpenAI-compatible vLLM path. The native Transformers run is divided into independent shards across available GPUs, while preserving shard order for later reconstruction. The vLLM path is retained for throughput experiments and is not required for the quality comparison.

The final validation generation uses a maximum of 2,048 newly generated tokens. This parameter is an upper bound rather than a target length: generation stops earlier when the model emits an end-of-sequence marker. The reconstructed prediction files contain 26,474 records for each branch, and the same records are used for both metric computations.

5.2 Automatic metrics

We report sacreBLEU (Papineni et al., 2002; Post, 2018) and COMET (Rei et al., 2020) for every direction. Tokenization is selected by target language according to the evaluation script: Chinese uses the Chinese tokenizer, Thai, Khmer, Lao, and Burmese use no tokenization, and the remaining directions use the 13a tokenizer. COMET is computed in batches in a separate evaluation environment to avoid mixing training and scoring dependencies.

For a direction d , let B_d be the sacreBLEU score and C_d be COMET multiplied by 100. We adopt the official WMT26 Translation Quality Score (WMT26 Chinese–Southeast Asian Task Organizers, 2026),

$$Q_d = \frac{B_d + C_d}{2} = 0.5 \times \text{sacreBLEU} + 0.5 \times \text{COMET}, \quad (1)$$

where both component metrics are reported on a common $[0, 100]$ scale (COMET is multiplied by 100), so their unweighted arithmetic mean is directly comparable and requires no further rescaling. To keep the comparison transparent and the metric on which any claim depends visible, every direction in Table 5 reports sacreBLEU, COMET, and Q as separate columns rather than collapsing the analysis into the composite score. BLEU and COMET are therefore independently inspectable for any direction, and Q is reported in parallel rather than as the basis of any directional ranking. The primary report includes all three quantities and a macro average across the fourteen directions. We do not use a pooled score that allows a high-resource direction to hide a severe regression in a lower-resource direction.

Direction	CPT	LoRA 26	LoRA 27	LoRA $\mu \pm \sigma$	Full 26	Full 27	Full $\mu \pm \sigma$
ID-ZH	18.30	62.28	62.31	62.30 $\pm$ 0.02	65.99	65.29	65.64 $\pm$ 0.49
KM-ZH	2.88	42.92	42.91	42.92 $\pm$ 0.01	44.51	44.17	44.34 $\pm$ 0.24
LO-ZH	0.73	27.67	27.67	27.67 $\pm$ 0.00	31.02	31.33	31.17 $\pm$ 0.22
MS-ZH	13.76	61.11	60.88	61.00 $\pm$ 0.16	63.88	64.08	63.98 $\pm$ 0.14
MY-ZH	3.47	41.88	41.88	41.88 $\pm$ 0.00	42.61	42.78	42.70 $\pm$ 0.12
TH-ZH	9.48	44.84	45.31	45.08 $\pm$ 0.33	47.74	47.62	47.68 $\pm$ 0.08
VI-ZH	21.02	59.07	58.60	58.84 $\pm$ 0.33	61.12	61.35	61.24 $\pm$ 0.16
ZH-ID	10.09	49.55	49.76	49.66 $\pm$ 0.15	52.94	52.71	52.83 $\pm$ 0.16
ZH-KM	0.29	16.29	15.82	16.06 $\pm$ 0.33	18.25	17.84	18.05 $\pm$ 0.29
ZH-LO	1.00	7.27	7.06	7.16 $\pm$ 0.15	10.09	9.83	9.96 $\pm$ 0.18
ZH-MS	6.55	44.63	44.42	44.53 $\pm$ 0.15	47.98	47.42	47.70 $\pm$ 0.40
ZH-MY	0.31	5.12	5.38	5.25 $\pm$ 0.18	6.12	6.02	6.07 $\pm$ 0.07
ZH-TH	1.81	10.03	9.53	9.78 $\pm$ 0.35	13.37	13.66	13.52 $\pm$ 0.21
ZH-VI	17.62	53.84	54.20	54.02 $\pm$ 0.25	57.37	58.43	57.90 $\pm$ 0.75
Macro avg.	7.67	37.61	37.55	37.58 $\pm$ 0.04	40.21	40.18	40.20 $\pm$ 0.02

Table 5: sacreBLEU on the 26,474-example validation set. CPT is the pre-adaptation 1-epoch checkpoint (single fixed baseline, no training seed). LoRA 26/27 and Full 26/27 are two independent training seeds; $\mu \pm \sigma$ is the per-direction mean and sample standard deviation across the two seeds. Full SFT improves over LoRA on all fourteen directions; both branches improve sharply over the CPT baseline.

5.3 Validation results

The completed validation results are reported across two independent training seeds (2026 and 2027) under identical data, prompt, and decoding conditions. We first establish the CPT 1-epoch checkpoint as the pre-adaptation baseline, then compare LoRA SFT and full-parameter SFT against it. Table 5 reports sacreBLEU, Table 6 reports COMET $\times 100$, and both list the CPT baseline, the two individual seeds, and the per-seed mean $\mu \pm \sigma$ (sample standard deviation, $n=2$, denominator $n-1$) for each branch. Table 7 reports the paired Full-minus-LoRA difference per seed, and Table 8 reports the gain of each branch over the CPT baseline. Values are rounded to two decimal places. The validation set is used for checkpoint selection (lowest validation loss, Section 4); the reported scores are therefore not an independent blind test and must not be conflated with the organizer-held test set.

6 Results and Analysis

The completed validation results support a consistent advantage for full-parameter SFT under the matched evaluation setup. Across two independent training seeds, full SFT improves over LoRA SFT on all fourteen directions in BLEU and on thirteen of fourteen in COMET $\times 100$ (Table 6); the ZH–VI exception reverses across seeds and is not claimed as a stable advantage. The macro-average BLEU mean is 37.58 ± 0.04 for LoRA and 40.20 ± 0.02 for full SFT ($\Delta+2.62$), and macro-average COMET $\times 100$ is 85.27 ± 0.03 versus 86.00 ± 0.04 ($\Delta+0.74$). Both branches improve

Direction	CPT	LoRA 26	LoRA 27	LoRA $\mu \pm \sigma$	Full 26	Full 27	Full $\mu \pm \sigma$
ID-ZH	78.14	85.52	85.43	85.48 $\pm$ 0.06	85.98	85.94	85.96 $\pm$ 0.03
KM-ZH	59.09	84.35	84.26	84.31 $\pm$ 0.06	84.66	84.70	84.68 $\pm$ 0.03
LO-ZH	46.24	81.20	80.90	81.05 $\pm$ 0.21	81.91	81.87	81.89 $\pm$ 0.03
MS-ZH	75.46	85.34	85.37	85.36 $\pm$ 0.02	85.67	85.51	85.59 $\pm$ 0.11
MY-ZH	64.96	83.42	85.43	85.43 $\pm$ 0.01	85.72	86.09	85.91 $\pm$ 0.26
TH-ZH	69.80	83.61	83.71	83.66 $\pm$ 0.07	83.87	83.89	83.88 $\pm$ 0.01
VI-ZH	83.52	87.98	88.46	88.22 $\pm$ 0.34	88.16	88.40	88.28 $\pm$ 0.17
ZH-ID	62.64	88.11	88.47	88.29 $\pm$ 0.25	88.70	88.59	88.65 $\pm$ 0.08
ZH-KM	58.11	86.03	86.16	86.10 $\pm$ 0.09	86.96	86.95	86.96 $\pm$ 0.01
ZH-LO	57.12	76.37	76.12	76.25 $\pm$ 0.18	78.61	78.67	78.64 $\pm$ 0.04
ZH-MS	61.07	87.61	87.68	87.65 $\pm$ 0.05	88.21	88.48	88.35 $\pm$ 0.19
ZH-MY	65.67	87.23	87.38	87.31 $\pm$ 0.11	88.67	88.62	88.65 $\pm$ 0.04
ZH-TH	55.71	83.77	83.66	83.72 $\pm$ 0.08	85.23	85.23	85.23 $\pm$ 0.00
ZH-VI	59.98	90.90	90.96	90.93 $\pm$ 0.04	91.22	91.52	91.37 $\pm$ 0.21
Macro avg.	64.11	85.25	85.29	85.27 $\pm$ 0.03	85.97	86.03	86.00 $\pm$ 0.04

Table 6: COMET $\times$ 100 on the same validation set, same layout as Table 5. Full SFT leads LoRA on all fourteen directions by macro-average +0.74 COMET $\times$ 100, but the per-direction lead is not uniform: ZH-VI (VI $\rightarrow$ ZH) reverses sign across seeds (Full 88.16 vs LoRA 88.46 at seed 2027; +0.06 $\pm$ 0.17 mean) and is not treated as a stable advantage.

Direction	ΔB 26	ΔB 27	$\Delta B \mu \pm \sigma$	ΔC 26	ΔC 27	$\Delta C \mu \pm \sigma$
ID-ZH	+3.71	+2.98	3.34 $\pm$ 0.52	+0.46	+0.51	0.48 $\pm$ 0.04
KM-ZH	+1.59	+1.26	1.43 $\pm$ 0.23	+0.31	+0.44	0.38 $\pm$ 0.09
LO-ZH	+3.35	+3.66	3.50 $\pm$ 0.22	+0.71	+0.97	0.84 $\pm$ 0.18
MS-ZH	+2.77	+3.20	2.98 $\pm$ 0.30	+0.33	+0.14	0.23 $\pm$ 0.13
MY-ZH	+0.73	+0.90	0.81 $\pm$ 0.12	+0.30	+0.66	0.48 $\pm$ 0.25
TH-ZH	+2.90	+2.31	2.60 $\pm$ 0.42	+0.26	+0.18	0.22 $\pm$ 0.06
VI-ZH	+2.05	+2.75	2.40 $\pm$ 0.49	+0.18	-0.06	0.06 $\pm$ 0.17
ZH-ID	+3.39	+2.95	3.17 $\pm$ 0.31	+0.59	+0.12	0.36 $\pm$ 0.33
ZH-KM	+1.96	+2.02	1.99 $\pm$ 0.04	+0.93	+0.79	0.86 $\pm$ 0.10
ZH-LO	+2.82	+2.77	2.80 $\pm$ 0.04	+2.24	+2.55	2.39 $\pm$ 0.22
ZH-MS	+3.35	+3.00	3.17 $\pm$ 0.25	+0.60	+0.80	0.70 $\pm$ 0.14
ZH-MY	+1.00	+0.64	0.82 $\pm$ 0.25	+1.44	+1.24	1.34 $\pm$ 0.14
ZH-TH	+3.34	+4.13	3.74 $\pm$ 0.56	+1.46	+1.57	1.52 $\pm$ 0.08
ZH-VI	+3.53	+4.23	3.88 $\pm$ 0.49	+0.32	+0.56	0.44 $\pm$ 0.17
Macro avg.	+2.61	+2.63	2.62 $\pm$ 0.02	+0.72	+0.75	0.74 $\pm$ 0.02

Table 7: Paired Full-SFT minus LoRA-SFT difference per seed, where ΔB is sacreBLEU and ΔC is COMET $\times$ 100. Differences are computed within the same seed before averaging. Full SFT leads on all fourteen directions in BLEU and on thirteen of fourteen in COMET; the ZH-VI exception reverses across seeds and is not claimed as a stable advantage. With only two seeds the standard deviation is an unstable estimate; mean $\pm \sigma$ is not a confidence interval and no significance test is performed.

sharply over the CPT 1-epoch baseline (macro BLEU 7.67, COMET $\times$ 100 64.11), with full SFT gaining +32.53 BLEU and +21.89 COMET $\times$ 100 over CPT at the macro level (Table 8).

The gains are present in both translation directions. For Southeast Asian languages translated into Chinese, the largest BLEU improvement of full over LoRA is observed for ID-ZH (+3.34 $\pm$ 0.52), while LO-ZH also improves substantially (+3.50 $\pm$ 0.22). For Chinese-to-Southeast-Asian translation, the largest BLEU improvement is ZH-VI (+3.88 $\pm$ 0.49). COMET gains are largest for ZH-LO (+2.39 $\pm$ 0.22), followed by ZH-TH (+1.52 $\pm$ 0.08) and ZH-MY (+1.34 $\pm$ 0.14). The MY-ZH versus ID-ZH contrast is informative: both target Chinese with identical BLEU

Direction	LoRA-CPT BLEU	Full-CPT BLEU	LoRA-CPT COMET	Full-CPT COMET
ID-ZH	+44.00	+47.34	+7.34	+7.82
KM-ZH	+40.03	+41.46	+1.46	+25.59
LO-ZH	+26.94	+30.44	+34.81	+35.65
MS-ZH	+47.24	+50.22	+9.89	+10.13
MY-ZH	+38.41	+39.22	+20.46	+20.94
TH-ZH	+35.59	+38.20	+13.86	+14.08
VI-ZH	+37.81	+40.21	+4.70	+4.76
ZH-ID	+39.57	+42.74	+25.65	+26.01
ZH-KM	+15.76	+17.75	+27.98	+28.84
ZH-LO	+6.17	+8.96	+19.12	+21.52
ZH-MS	+37.98	+41.15	+26.58	+27.28
ZH-MY	+4.94	+5.76	+21.63	+22.97
ZH-TH	+7.97	+11.70	+28.01	+29.52
ZH-VI	+36.40	+40.28	+30.95	+31.39
Macro avg.	+29.91	+32.53	+21.16	+21.89

Table 8: Gain of each SFT branch over the CPT 1-epoch baseline (mean of the two seeds). Both branches improve sharply over the pre-adaptation checkpoint, confirming that the bilingual SFT stage is the dominant source of translation quality; the remaining LoRA-versus-Full gap is reported in Table 7.

tokenization, yet ID-ZH gains more (+3.34 vs +0.81), so the difference cannot be attributed to tokenization alone. These results indicate that full-parameter adaptation improves both lexical overlap and learned semantic quality in this validation setting, although the magnitude of improvement is direction-dependent and the two-seed estimate remains noisy.

The absolute scores also reveal an asymmetry between translation directions. Chinese-to-Southeast-Asian directions generally have lower BLEU than their reverse directions, especially ZH-LO and ZH-MY. This pattern is visible for both methods and should not be interpreted as a failure of one SFT branch alone. It is consistent with the greater difficulty of producing several Southeast Asian target languages and with target-side tokenization differences. COMET remains relatively high for some of these directions, so BLEU and COMET should be considered jointly rather than reduced to a single ranking.

We attribute the direction-level variation in Table 5 primarily to script-driven tokenization efficiency rather than to CPT data volume. Each monolingual CPT file contributes between 96,179 and 98,001 training records, so corpus size is essentially balanced across the seven languages and cannot by itself explain the gains; quantitative details are reported in Table 2. To make the comparison reproducible, we re-tokenized 2,000-record random samples from each official monolingual training file with the deployment tokenizer and measured a single efficiency quantity. The tokens-per-character ratio spans a wide range: Latin-script Indonesian and Malay around 0.30, Vietnamese 0.36, Thai 0.54, Burmese 1.21, Khmer 1.56, and Lao 2.05. Latin-script directions therefore compress roughly

three characters into one token and already sit close to their semantic ceiling under LoRA, which is consistent with the small and unstable COMET movement we observe for VI-ZH ($+0.06 \pm 0.17$, reversing across seeds). Lao has the highest tokens-per-character and shows the largest COMET gain (ZH-LO, $+2.39 \pm 0.22$): under a fixed 4,096-token context, LoRA’s limited adapter capacity is largely spent representing the script, and full-parameter SFT recovers the headroom for translation quality. Burmese is an instructive exception. Although its tokens-per-character is moderate (1.21), MY-ZH shows the smallest BLEU gain ($+0.81 \pm 0.12$); because Burmese is scored with character-level BLEU (no tokenizer, per our metric setup), both adaptation branches struggle similarly with word-boundary morphology, so the lexical gap stays narrow even as COMET improves. These patterns indicate that the LoRA-versus-full-SFT gap widens exactly where subword representation is expensive and narrows where the script is already compact.

The results favor full SFT for this quality-oriented comparison, but they do not establish that full SFT is universally preferable. LoRA remains attractive when trainable-parameter count, storage, or deployment update cost is the primary constraint. The present comparison is also limited to the shared validation set; the organizer-held blind test remains separate and must not be used to retroactively alter this conclusion.

7 Efficiency and Reproducibility

Training, scoring, and serving are isolated into separate environments because DeepSpeed, Transformers, COMET, and vLLM impose different dependency constraints. The project records environment manifests for the training and evaluation environments. The evaluation scripts also record the model path, decoding limit, concurrency, COMET model, and score signatures.

For deployment-oriented measurements, the project includes a four-GPU vLLM service configuration and a concurrent benchmark. To quantify the LoRA-versus-full-SFT trade-off beyond quality, Table 9 reports a fixed-sample latency and peak GPU memory measurement on one NVIDIA H100 80GB (BF16, batch size 1, greedy decoding, up to 128 generated tokens, one fixed sample per direction plus one warm-up). Latency is the generate time after the encoded tensor reaches the GPU, excluding model load, tokenization, and network;

Model	Latency (s/sample)	Avg output tok	Peak alloc (GiB)	Peak reserved (GiB)	Adapter size
CPT (1-epoch)	4.333	109.86	3.690	3.848	—
LoRA (merged)	2.898	127.93	3.690	3.850	52.03 MiB
Full-SFT	3.346	127.86	3.690	3.848	—

Table 9: Fixed-sample efficiency on one NVIDIA H100 80GB (BF16, batch 1, greedy, ≤ 128 tokens). Peak memory is the PyTorch allocated/reserved peak, not total device or training memory. The LoRA adapter size is the unmerged weight (52.03 MiB); deployment still loads the base model. Peak allocated memory is identical across the three models, so this measurement does *not* show that LoRA reduces inference memory. LoRA’s measured average latency is lower, but this is a single small-sample run without repeated measurement, and full SFT is higher on some directions, so no stable speedup ratio or inherent latency advantage is claimed. LoRA’s efficiency advantage therefore lies in trainable-parameter count, adapter storage, and training cost, not in inference latency or memory.

peak memory is the PyTorch allocated/reserved peak, not total device memory or training peak.

8 Conclusion

We presented a reproducible Hy-MT2-1.8B adaptation pipeline for fourteen Chinese–Southeast Asian translation directions. The pipeline separates monolingual CPT, bilingual SFT, model preparation, and evaluation. Its central comparison is between LoRA SFT and full-parameter SFT from the same CPT checkpoint, under matched data and evaluation conditions.

On the shared 26,474-example validation set, full-parameter SFT outperforms LoRA SFT in every reported direction for BLEU and in thirteen of fourteen directions for COMET $\times 100$ (ZH–VI reverses across seeds). Across two training seeds, its macro-average scores are 40.20 ± 0.02 BLEU and 86.00 ± 0.04 COMET $\times 100$, compared with 37.58 ± 0.04 and 85.27 ± 0.03 for LoRA SFT; the gains are $+2.62$ BLEU and $+0.74$ COMET $\times 100$. Both branches improve sharply over the CPT 1-epoch baseline (macro BLEU 7.67, COMET $\times 100$ 64.11). These results support full-parameter SFT as the stronger choice when translation quality is the primary objective, while LoRA remains a practical alternative when parameter efficiency and update cost are more important.

This conclusion is restricted to the reconstructed validation set and the stated decoding and scoring protocol. The organizer-held blind-test data remain separate and are not used to revise the validation-based comparison.

Limitations

The evaluation compares two adaptation branches on a shared validation set and relies on automatic metrics. These metrics cannot fully capture adequacy and fluency for all seven Southeast Asian languages, especially for low-resource scripts and domain-specific terminology. A model can obtain a reasonable similarity score while omitting a clause, changing a named entity, or producing the wrong script. The direction-level results therefore should be read as evidence for the observed validation behavior, not as a complete human-quality assessment.

The prediction files and metric outputs are retained for reproducibility, but the current manuscript does not report a separate human evaluation or a systematic qualitative annotation study. The organizer-held final test data are kept outside model selection and hyperparameter tuning. This separation is necessary because using the final test set to choose a decoding limit, prompt, or checkpoint would invalidate the reported comparison.

Ethical Considerations

The training data contain public-facing texts that may include personal names, political content, health information, financial terminology, and other sensitive material. Data cleaning reduces formatting noise and split leakage but does not guarantee that every harmful or incorrect statement is removed. The system should not be used without qualified human review for medical, legal, financial, or other safety-critical translation.

Acknowledgments

Acknowledgments omitted for double-blind review.

References

- Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. [Don't stop pretraining: Adapt language models to domains and tasks](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360. Association for Computational Linguistics.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Sheel Wang, Lu Wang, and Weizhu Chen. 2022. [Lora: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. [Multilingual denoising pre-training for neural machine translation](#). *Transactions of the Association for Computational Linguistics*, 8:726–742.
- NLLB Team. 2022. [No language left behind: Scaling human-centered machine translation](#). *Transactions on Machine Learning Research*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [BLEU: A method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. [Zero: Memory optimizations toward training trillion parameter models](#). In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*.
- Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. [COMET: A neural framework for MT evaluation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2685–2702, Online. Association for Computational Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Neural machine translation of rare words with subword units](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30.
- WMT26 Chinese–Southeast Asian Task Organizers. 2026. [Wmt26 chinese–southeast asian multilingual machine translation task](#).
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara

- Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. [mt5: A massively multilingual pre-trained text-to-text transformer](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498. Association for Computational Linguistics.
- Mao Zheng, Zheng Li, Tao Chen, Bo Lv, Mingrui Sun, Mingyang Song, Jinlong Song, Hong Huang, Decheng Wu, Hai Wang, Yifan Song, Yanfeng Chen, and Guanwei Zhang. 2026. [Hy-mt2: A family of fast, efficient and powerful multilingual translation models in the wild](#). *arXiv preprint arXiv:2605.22064*.