

paramanandaAI@WMT26 Video Subtitle Translation: QLoRA Fine-Tuning with Test-Time Metadata Prompting

Shiv Ram Saud, Sundeep Dawadi, Sunil Regmi

Department of Artificial Intelligence, Kathmandu University

{saudshivram373, sundeepdwd}@gmail.com, sunilregmi233@gmail.com

Abstract

Subtitle translation poses challenges distinct from general-purpose machine translation model trained on web corpus. The outputs must fit strict display constraints as the text is predominantly spoken dialogue, local slangs, proper names, cultural nuances. In the "WMT26 Chinese $\rightarrow$ English task" all test videos belong to the historical costume drama genre, where period vocabulary cannot be resolved from a single line in isolation. We address these challenges with a two-part system. We fine-tune Hy-MT2-1.8B (Zheng et al., 2026) on 500,000 Chinese-English subtitle pairs from the TVsub corpus (Wang et al., 2018) using 4-bit QLoRA (Detrmers et al., 2023), and at test time we inject show-level metadata (title, episode name, summary) into a structured system prompt to ground character names and period terms. On an in-domain 200-pair development split, QLoRA fine-tuning raises BLEU from 6.93 to 31.29 (+24.4 points) which demonstrates the value of domain adaptation over the base multilingual model. Across four prompt configurations evaluated via LLM-as-judge on 20 sampled test translations the metadata-augmented, previous context-free variant (`structured_noctx`) wins 14 of 20 judged dialogues and is sent for submission. The fine-tuned model weights are released under Apache 2.0 at ShivRamSaud/hy-mt2-1.8b-wmt26.

1 Introduction

Subtitle translation occupies a specific need based niche within machine translation(MT)

where output must be concise enough to fit on screen, the source is predominantly spoken dialogue rather than written prose, and the meaning of individual lines is often inseparable from the show's setting. Consider the Chinese term 仵作: without domain context, a general-purpose MT system might render it as "mortician" or leave it untranslated. With knowledge that the show features a female forensic coroner in an imperial court, the same system can produce the contextually correct "coroner" (Matusov et al., 2019). The WMT26 Video Subtitle Translation task (<https://www2.statmt.org/wmt26/video-subtitle-translation.html>) makes this domain signal explicit and each test video is accompanied by a metadata record containing the video title, album title, and album description.

Our system, **paramanandaAI**, is built around two complementary components. First, we adapt the multilingual model Hy-MT2-1.8B to the subtitle register through parameter-efficient fine-tuning on a large subtitle corpus, so that the model's generation style matches the short, colloquial dialogue of the test set. Second, at inference time we condition translation on the video-level metadata via a structured system prompt for the drama-specific context needed.

Our primary contributions are as follows:

- We demonstrate that QLoRA fine-tuning on 500K in-domain subtitle pairs yields a +24.4 BLEU improvement over the base

multilingual model on a held-out development split, showing that domain adaptation yields higher gains than general purpose bigger model in constrained environment.

- We design a test-time metadata prompting scheme and found the metadata-augmented, context-free configuration as the best-performing submission strategy.
- We release the fine-tuned model weights under Apache 2.0, providing a reproducible, domain-adapted baseline for subtitle translation research.

Section 2 reviews related work. Section 3 describes our training data, fine-tuning procedure, and inference pipeline. Section 4 presents experimental results. Section 5 analyzes system behavior through qualitative examples and error patterns. Section 6 concludes, and Section 6 lists limitations.

2 Related Work

Our work sits between three research threads: subtitle machine translation, parameter-efficient fine-tuning, and prompt-based adaptation of large language models.

Subtitle machine translation. Subtitle translation imposes constraints like length, timing, and a colloquial register, making it different from document or sentence-level MT (Matusov et al., 2019). NMT models specifically customized for subtitling outperform general-purpose systems by better accounting for display length and reading speed. The TVsub corpus (Wang et al., 2018), which we use for fine-tuning, provides large-scale Chinese-English subtitle pairs captures the colloquial style and turn-taking structure typical of the WMT26 test domain.

Document-level context in MT. A natural extension of subtitle MT is to leverage cross-sentence context for consistency. Maruf et al. (2021) survey the field and find that local context can improve coherence, but also introduce noise when the surrounding lines are themselves erroneous or topically unrelated to the current line. This tension is directly relevant to our context-window experiments in Section 4.4.

Parameter-efficient fine-tuning.

LoRA (Hu et al., 2022) fine-tunes large models by injecting trainable low-rank matrices into frozen weight projections, dramatically reducing GPU memory requirements. QLoRA (Dettmers et al., 2023) extends this by additionally quantizing the frozen base model to 4-bit NF4 with double quantization, enabling fine-tuning of billion-parameter models on consumer-grade hardware. For our setup of two 12 GB GPUs using QLoRA is the only viable fine-tuning strategy; full-parameter training of a 1.8B model would exceed available memory.

LLM prompting and instruction tuning.

Instruction tuning (Ouyang et al., 2022) has become a standard approach for aligning language models with task-specific behavior, and our chat-format fine-tuning template follows this paradigm. Recent work has shown that large language models can translate effectively with few-shot in-context examples (Agrawal et al., 2023) and that adaptive prompting with domain-specific terminology improves translation accuracy in specialized settings (Moslem et al., 2023). We build on these findings by injecting show-level metadata as a structured system prompt at test time.

3 Method

3.1 Training Data

We train on the TVsub corpus (Wang et al., 2018), which contains 2.15 M Chinese-English subtitle pairs. From this corpus we randomly sample 500,000 pairs (fixed seed 42) and format each as a single chat record:

```
system: Translate the following Chinese
subtitle to English.
user: {Chinese subtitle}
assistant: {English translation}
```

Training on subtitle dialogue directly exposes the model to the colloquial style that dominates the WMT26 test set. Metadata is **not** used during training and was introduced only at test-time inference (Section 3.3).

3.2 QLoRA Fine-Tuning

We use the base Hy-MT2-1.8B (Zheng et al., 2026) model then attach LoRA (Hu et al., 2022) adapters of rank $r=16$ and scaling factor $\alpha=32$ with dropout 0.05 to all attention and

MLP projection matrices (q, k, v, o, gate, up, down projections). Only the adapter weights are updated during training; the quantized base weights remain frozen.

After training, the adapters are merged into the base weights, producing a single merged bfloat16 checkpoint (~3.58 GB) that requires no PEFT runtime or adapter loading at inference. Table 1 summarizes the training configuration.

Hyperparameter	Value
Base model	tencent/Hy-MT2-1.8B
Quantization	NF4 + double quant (4-bit)
LoRA rank r / α	16 / 32
LoRA dropout	0.05
Target modules	q, k, v, o, gate, up, down proj
Trainable params	6.5M (0.36% of 1.8B)
Training pairs	500,000
Epochs	1
Max sequence length	512 tokens
Per-device batch	8
Gradient accum.	4 (effective batch: 64)
Learning rate	2×10^{-4}
Compute precision	bf16
Hardware	2× NVIDIA RTX 3080 Ti (12 GB)
Wall-clock time	~5 hours

Table 1: Training hyperparameters of the submitted model.

3.3 Test-Time Metadata Prompting

At test-time inference, each subtitle entry is translated as a single chat turn. The system message is populated with the video’s meta-data record:

You are translating Chinese subtitles for a historical costume drama.
Show: {album_title}. Episode: {video_title}.
Synopsis: {album_desc}.
Translate naturally, preserving character names and period register.

Table 2 shows a concrete instance. The description establishes the setting like a female forensic coroner, a sickly prince, an imperial murder investigation etc which disambiguates period terms such as 仵作 (coroner) and 王爷 (prince) that are underspecified in any individual subtitle line.

Field	Content
video_title	病娇王爷仵作妻 _11
album_title	病娇王爷仵作妻
album_desc	女法医孔柒柒在血月之夜意外穿越到男友写的小说之中，成为京城相府养女，穿越后的孔柒柒丢失了穿越前的部分记忆，并卷入了一起溺水谋杀案，在案件中孔柒柒结识了病弱失势的肺癆王爷沐之曦。
subtitle	说好一起来查沈万两的案子

Table 2: Metadata example from test video a0046j15d6d. The synopsis provides domain context necessary to translate period-specific terms accurately.

Figure 1 illustrates the full inference pipeline. The metadata and Chinese subtitle are combined into a single prompt that is passed to the fine-tuned model. The output is written back to the SRT file with the original entry index and timestamp preserved.

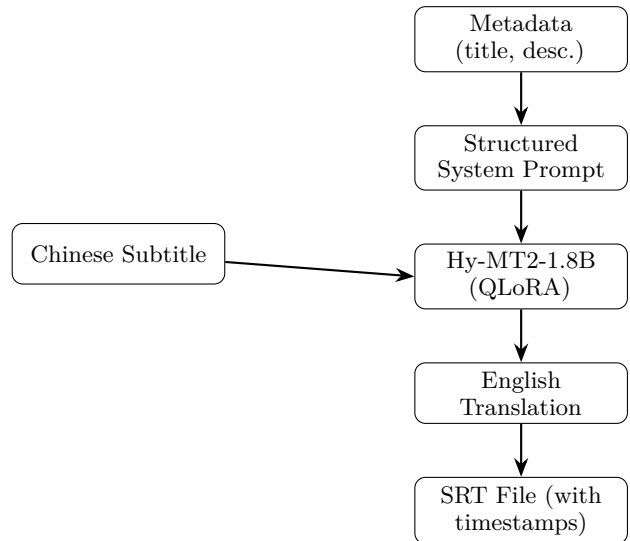


Figure 1: Test-time inference pipeline. Metadata is injected into a structured system prompt and, together with the Chinese subtitle, passed to the fine-tuned model. Each subtitle entry is translated independently.

Generation uses temperature 0.3, top- p 0.9, and a repetition penalty of 1.05, capped at 128 new tokens. We use nucleus sampling rather than beam search considering subtitle dialogue benefits from natural, varied phrases (Agrawal et al., 2023). The 128-token cap is deliberately generous relative to typical subtitle length (10-20 words) and extension later is applied uniformly across all four prompt configurations.

4 Experiments

We conduct two complementary evaluations. First, we measure the effect of QLoRA fine-tuning using automatic metrics on an in-domain development set (Section 4.1). Second, we use an LLM-as-judge protocol on test videos to select among four prompt configurations before final submission (Section 4.3).

4.1 Dev-Set Evaluation

Setup. We used held out 200 subtitle pairs from the devset in TVsub dataset. Considering this data from the TVsub dataset, the reported scores are an in-domain diagnostic and should be interpreted as an upper bound on expected test-set performance rather than wmt26 blind test estimate. We evaluate corpus-level BLEU (Papineni et al., 2002) via SacreBLEU (Post, 2018), chrF (Popović, 2015), and TER (higher BLEU and chrF indicate better quality; lower TER is better). The baseline is the unmodified Hy-MT2-1.8B model evaluated under identical conditions with same plain system prompt, same generation parameters, same dev set and without QLoRA fine-tuning.

Results. Table 3 presents the main comparison. The base model scores BLEU 6.93, reflecting a large domain gap between the model’s pretraining distribution and the subtitle register. QLoRA fine-tuning raises BLEU to 31.29, a +24.4 point improvement, with consistent gains across all three metrics. The sentence-level distribution in Table 4 further illustrates the improvement where the base model produces 22 out of 200 sentences with BLEU 0 (catastrophically wrong translations) when compared to only 3 for the fine-tuned model, indicating that fine-tuning eliminates most complete failures. There are some examples with BLEU scores of 100 which were primarily in very short translations.

Model	BLEU↑	chrF↑	TER↓
Base Hy-MT2-1.8B	6.93	29.76	87.17
Fine-tuned 1.8B(our)	31.29	42.94	76.51
Δ	+24.4	+13.2	−10.7

Table 3: Dev-set automatic metrics. QLoRA fine-tuning on TVsub yields substantial gains across all three metrics.

Statistic	Base	Fine-tuned
Median sentence BLEU	7.31	23.64
Mean sentence BLEU	10.63	34.92
Max sentence BLEU	66.87	100.00
BLEU = 0 count	22 / 200	3 / 200

Table 4: Sentence-level BLEU distribution on the 200-pair dev set.

In the WMT26 Video Subtitle Translation shared task, system outputs were evaluated using an LLM-as-a-Judge protocol with a 100-point deduction-based rubric across four dimensions (Accuracy and fidelity: 40 points maximum, Fluency and readability: 35 points maximum, Cultural and stylistic adaptation: 20 points maximum, Subtitle text conventions: 5 points maximum), which was subsequently verified and corrected by human reviewers. Participating in only the Chinese-to-English (en) direction, Ours(paramanandaAI) achieved a mean score of 55.645.

4.2 Comparison with Larger Baselines

To contextualize our fine-tuned 1.8B submission against a larger model, we evaluate Hy-MT2-7B (Zheng et al., 2026) on the same 200-pair dev split under the same plain-prompt conditions. We test both the full bfloat16 model and a 4-bit quantized variant to assess quantization degradation at the 7B scale.

Model	BLEU↑	chrF↑	TER↓
Hy-MT2-1.8B (base)	6.93	29.76	87.17
Hy-MT2-7B (bf16)	7.94	32.47	86.32
Hy-MT2-7B (4-bit)	7.57	31.83	86.14
Fine-tuned 1.8B(our)	31.29	42.94	76.51

Table 5: Comparison of base models at different scales with our fine-tuned 1.8B submission on the 200-pair dev set.

The 7B base model improves marginally over the 1.8B base (BLEU 7.94 vs. 6.93), while our fine-tuned 1.8B surpasses both by a wide margin. This shows that domain-specific fine-tuning provides suitable alternative in subtitle translation than using larger pretrained model on zero-shot configuration, and that 4-bit quantization of the 7B model causes negligible degradation relative to full precision (Dettmers et al., 2023). Yet, Applying QLoRA to the 7B model to properly isolate

the effect of model scale exceeded our hardware constraints ($2 \times 12\text{GB}$ GPUs).

4.3 Prompt Configuration Selection

For test submission we evaluate four prompt configurations along two axes: *system prompt style* (plain vs. structured with metadata) and *context window* (no context vs. three previous translations prepended).

Config	Description
simple_noctx	Plain system prompt, no context
simple_ctx3	Plain system prompt + 3 prev. translations
structured_noctx	Metadata prompt, no context
structured_ctx3	Metadata prompt + 3 prev. translations

Table 6: The four prompt configurations evaluated for test submission.

We sample 20 videos from the test set and generate translations under all four configurations. For each video, we sample 10 subtitle entries and present them as a blind 4-way comparison (labeled A/B/C/D and suffled to prevent positional bias) to an external LLM judge (llama-3.3-70b-versatile via Groq). The judge ranks the four outputs on three criteria: *fluency* (grammaticality and naturalness), *faithfulness* (semantic accuracy relative to the Chinese source), and *subtitle display suitability* (conciseness and on-screen readability). The system identities are not revealed to the judge.

structured_noctx wins 14 of 20 judged dialogues and is selected for submission. Its advantage is attributable to the metadata grounding (Section 4.4 discusses why context windows hurt).

4.4 Context Window Analysis

The **ctx3** variants prepend the three most recently translated entries to the system prompt as reference context, with the hypothesis that cross-sentence consistency would improve. The judge evaluation shows the opposite: both **ctx3** variants underperform their **noctx** counterparts.

In subtitle dialogue where consecutive lines frequently belong to different speakers. When a preceding entry is translated incorrectly, that error is fed as reference to subsequent entries, amplifying the mistake.

These findings are consistent with Maruf et al. (2021), who show that local context can degrade MT quality when it is noisy or topically discontinuous.

5 Analysis

5.1 Qualitative Examples

Table 7 shows three subtitle entries from the same test video, each illustrating how the synopsis grounds a term that would otherwise be ambiguous. All three examples are drawn from video a0046j15d6d (drama 病娇王爷件作妻), whose metadata is shown in Table 2.

Example 1	
Source	说好一起来查沈万两的案子
Synopsis	Female forensic coroner investigates imperial murder case
Output	We agreed to investigate Shen Wanliang’s case together.
Example 2	
Source	件作验过了
Synopsis	Same: forensic coroner in imperial court setting
Output	The coroner has completed the examination.
Example 3	
Source	王爷的病如何了
Synopsis	Sickly prince Mu Zhixi, imperial court drama
Output	How is the Prince’s illness?

Table 7: Three subtitle translations from the same test video, illustrating how the synopsis grounds domain-specific terms.

In Example 2, 件作 is a historical term for a court-appointed forensic examiner. Without the descriptive context, a general-purpose system would likely produce “mortician” or leave the term untranslated; with the synopsis identifying the protagonist as a female forensic coroner, the model produces the contextually accurate “coroner.” Similarly, 王爷 in Example 3 could be rendered as “lord,” “royal uncle,” or omitted, but the synopsis identifies a specific prince character, leading to the unambiguous “Prince.”

5.2 Error Analysis

We identify three recurring error patterns in the WMT26 blind submission:

1. **Post-processing needs** The model had failures that stem primarily from pipeline deficiencies. Many errors were dominated by untranslated source script retention

(43% of videos), and surface formatting leaks such as unescaped HTML entities (`apos;`), subword markers (`@-@`), and uncapitalized sentence starts. Additionally, in our dev set model produced repeated words like `ha ha`. The untranslated Chinese text were mostly in parenthesis, song lyrics with musical notes (`♪ ... ♪`), and short proper names, martial art terms, episode header and formal greeting. By category speech with spoken idioms and slang were mostly affected. During submission these errors weren't handled explicitly showing exactly how model performed on those setting.

2. **Dialogue particle loss.** Chinese dialogue particles (吧, 呢, 啊) encode conversational attitude like tentativeness, persuasion, surprise and that is difficult to preserve in English. For example, a source line ending in 吧 (a softening particle used in suggestions) may be translated as a flat declarative rather than a suggestion, losing the interpersonal nuance. This is a known challenge in Chinese-to-English MT (Matusov et al., 2019) and likely requires explicit pragmatic-awareness training to address systematically.
3. **Synopsis-absent degradation.** For test videos with a very short or absent `album_desc` field, the metadata prompt reduces to little more than the show title, providing insufficient domain signal. In those cases, the model's output qualitatively resembles that of the plain-prompt baseline, suggesting that the metadata content is the primary source of domain disambiguation for period-specific vocabulary.

6 Conclusion

We presented **paramanandaAI**, a system for Chinese-to-English subtitle translation combining QLoRA fine-tuning with test-time metadata prompting. Fine-tuning Hy-MT2-1.8B on 500K TVsub subtitle pairs yields a +24.4 BLEU improvement over the unmodified base model on an in-domain development set, and larger 7B baselines without

fine-tuning remain far behind, indicating that domain adaptation fine-tuning in resource constraint is a more effective lever than larger general purpose translation models scale for this task.

A structured metadata prompt grounds period-specific vocabulary and character names that would otherwise be ambiguous from a single subtitle line. LLM-as-judge evaluation over 20 test videos suggests that this configuration (`structured_noctx`) outperforms plain prompting and context-window variants alike. Future work should explore length-aware decoding techniques to eliminate over-long subtitle outputs, and evaluate metadata prompting against a held-out test set with automatic metrics to precisely quantify its contribution independently of fine-tuning. Additionally, post processing and analysis of punctuation errors, missing translation and repetition penalty should be explored.

Limitations

- **In-domain dev set.** The 200-pair development split is drawn from the same TVsub distribution as the training data. The reported BLEU, chrF, and TER scores therefore represent an in-domain upper bound and may overestimate performance on the WMT26 held-out test set, which consists of historical costume dramas rather than the mixed drama genres in TVsub. Additionally, lack of pre-processing of training set and post-processing of model outputs should be considered.
- **No ablation of metadata vs. fine-tuning.** The base model was evaluated under the plain prompt only. The relative contribution of the metadata prompt versus QLoRA fine-tuning on the test set cannot be disentangled without evaluating the base model under the metadata prompt, which we did not do.
- **LLM-as-judge validity.** The judge protocol was not validated against human annotator agreement, and the evaluation covers only 20 videos with 10 sampled entries each. The selection of `structured_noctx` should be interpreted

as an informed preference rather than a statistically significant finding.

- **No length or timing constraints.** The system does not enforce subtitle display-length limits or timing synchronization, both of which are important for production use. We use generation constraint of max Tokens to assume the desired length of generation to be small.
- **Metadata hallucination risk.** The metadata prompt introduces a risk that the model incorporates synopsis details not present in the source line. This failure mode was not systematically evaluated.

Open Weights

The merged bfloat16 checkpoint (~3.58 GB), tokenizer, and model card are publicly available under the Apache 2.0 license at:

<https://huggingface.co/ShivRamSaud/hy-mt2-1.8b-wmt26>

No PEFT runtime or adapter loading is required for inference.

References

- Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. [In-context examples selection for machine translation](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [QLoRA: Efficient finetuning of quantized LLMs](#). In *Advances in Neural Information Processing Systems (NeurIPS 2023)*.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *Proceedings of the International Conference on Learning Representations (ICLR 2022)*.
- Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2021. [A survey on document-level neural machine translation: Methods and evaluation](#). *ACM Computing Surveys*, 54(2):45:1–45:36.
- Evgeny Matusov, Patrick Wilken, and Yota Georgakopoulou. 2019. [Customizing neural machine translation for subtitling](#). In *Proceedings of the Fourth Conference on Machine Translation (WMT 2019)*, pages 82–93, Florence, Italy. Association for Computational Linguistics.
- Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. [Adaptive machine translation with large language models](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation (EAMT 2023)*, pages 227–237, Tampere, Finland. European Association for Machine Translation.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems (NeurIPS 2022)*.
- 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Maja Popović. 2015. [chrF: Character \$n\$ -gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation (WMT 2018)*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- Longyue Wang, Zhaopeng Tu, Shuming Shi, Tong Zhang, Yvette Graham, and Qun Liu. 2018. [Translating pro-drop languages with reconstruction models](#). In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI 2018)*, pages 4937–4945, New Orleans, Louisiana, USA. AAAI Press.
- 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](#). <https://huggingface.co/tencent/Hy-MT2-1.8B>. ArXiv:2605.22064.