

rupiah at WMT26: Training-Free Context-Aware Subtitle Translation with QE Selection

Pieter Christy Yan Yudhistira / Sayyidah Fatimah Azzahra / Dzaki Rafif Malik /
Tirana Noor Fatyanosa

Intelligent System Laboratory

Faculty of Computer Science

Brawijaya University

pieterchristyan7@gmail.com, sfatimapple@gmail.com

malikdzaki16@gmail.com, fatyanosa@ub.ac.id

Abstract

We present the training-free subtitle translation system developed by Team rupiah for the WMT26 Video Subtitle Translation shared task. The system translates Chinese subtitles into English and Indonesian using Hy-MT2-7B while preserving the source cue indices and timestamps. To support discourse and terminology consistency, each cue is conditioned on episode metadata, preceding source cues, translation history, and an automatically constructed episode glossary. At inference time, the system generates candidates at six temperatures, removes exact duplicates, filters candidates containing Han characters, and selects a hypothesis using reference-free CometKiwi quality estimation. A deterministic lexical-recovery step then replaces glossary terms that remain untranslated in the selected hypothesis. In the preliminary official evaluation over 100 videos per direction, rupiah ranks fifth of eight systems for Chinese-to-English and third of four for Chinese-to-Indonesian. Rubric-level results indicate that the largest performance gap relative to the top-ranked system is in accuracy and fidelity, whereas the gaps in fluency and subtitle conventions are smaller. Reference-based scores over 89 videos per direction provide additional diagnostics of the complete pipeline. These aggregate evaluations do not isolate the contribution of individual components.¹

1 Introduction

The WMT26 Video Subtitle Translation shared task requires systems to translate Chinese subtitles into English and Indonesian while allowing the use of associated textual metadata (WMT). Although individual subtitle cues are usually short, their interpretation often depends on information beyond the current cue. Pronouns may refer to entities introduced

in preceding dialogue, names and titles may recur throughout an episode, and the appropriate register may depend on relationships between speakers or the broader narrative setting. Independent cue-level translation can therefore produce locally plausible outputs that become inconsistent when read as a sequence. In addition to these linguistic requirements, the translated output must remain a valid SubRip (.srt) file aligned with the structure and timing of the source subtitles.

These requirements motivate a system that separates linguistic processing from subtitle-format handling. The primary problem addressed by rupiah is the lack of episode-level information available to an independent cue-level translator. The system consequently treats each video as a sequential translation episode. Source cue indices and timestamps are parsed and retained separately, while translation is performed using the current cue together with episode metadata, preceding source cues, translation history, and an episode-specific terminology mapping. The system uses cue-level translation without task-specific fine-tuning; adaptation is performed entirely at inference time through prompt construction, contextual conditioning, terminology guidance, and candidate selection.

A second consideration concerns decoding. The hypothesis assigned the highest probability by a translation model is not necessarily the hypothesis preferred under an external translation-quality criterion (Eikema and Aziz, 2020). Instead of relying on a single sampled translation, the system constructs a small candidate set using the same translation model at multiple temperatures. The resulting hypotheses are deduplicated and subjected to a source-script check before being scored by a reference-free quality-estimation model. This procedure is applicable at test time because it does not require access to a reference translation. The context mechanism provides information relevant to discourse-level decisions, while the candidate-

¹Source code available and can be accessed here: <https://github.com/Pieter414/-rupiah-training-free-video-translation-subtitle>

<https://github.com/Pieter414/-rupiah-training-free-video-translation-subtitle>

selection procedure permits comparison among several possible translations. The present evaluation measures these mechanisms as parts of the complete pipeline and does not establish their individual effects.

Terminology consistency is handled at two points in the pipeline. Before cue translation, an episode glossary is constructed and relevant entries are included in the translation prompt. After quality-estimation selection, a narrow lexical-recovery procedure checks whether a glossary source term remains verbatim in the selected non-Chinese hypothesis and, when applicable, replaces it with its mapped target rendering. These two uses of the glossary address recurring expressions without modifying the model parameters or the underlying decoding algorithm.

This paper reports the submitted systems for Chinese-to-English (zh→en) and Chinese-to-Indonesian (zh→id). First, we describe a training-free subtitle translation pipeline that conditions cue-level generation on episode metadata, preceding source cues, translation history, and episode terminology while preserving the source subtitle structure. Second, we present an inference-time candidate-selection procedure based on temperature sampling, exact deduplication, Han-character filtering, and reference-free CometKiw scoring. Third, we report the system’s preliminary official ranking and rubric-level results over all 100 videos per direction, together with reference-based diagnostics over the 89-video subset for which references were available.

2 Related Work

2.1 Context in Subtitle and Document Translation

Subtitle translation combines sentence-level translation with discourse context and presentation constraints. Earlier work on neural machine translation for subtitles has investigated contextual modeling and segmentation as part of adapting translation systems to subtitle-specific requirements (Matusov et al., 2019). Studies of professional subtitling have further examined how contextual and extra-textual information, including film metadata, contributes to translation decisions (Vincent et al., 2024). Such information is particularly relevant for short dialogue cues, for which the local text may not identify the intended referent, the relationship between speakers, the appropriate interpersonal register, or

the established rendering of a recurring entity.

The system described in this paper retains cue-level generation but augments each cue with information from the surrounding episode. Its prompt may include task-provided metadata, preceding source cues, and translation history. This design differs from strictly independent sentence translation because earlier dialogue decisions can inform subsequent cues. However, the rolling translation history does not include future cues. Episode-level information may nevertheless enter through metadata and glossary construction, which are prepared separately from the sequential cue history.

2.2 Terminology Consistency

Terminology consistency is important when names, organizations, titles, places, and other expressions recur across an episode. Lexically constrained decoding can enforce required target strings directly within the decoding process (Hokamp and Liu, 2017). Our approach does not alter the decoding search algorithm. Instead, it uses an episode-specific source-to-target terminology mapping during prompt construction and after candidate selection.

Relevant glossary entries are supplied to the translation model as target-rendering guidance. After quality-estimation selection, the system applies a narrow deterministic recovery procedure when a mapped source expression remains verbatim in a non-Chinese hypothesis. The recovery procedure does not perform general terminology correction: it may miss terminology errors expressed through paraphrases, inflectional variants, or alternative incorrect translations. The glossary therefore acts as both prompting context and a limited post-selection safeguard rather than as a hard decoding constraint.

2.3 Candidate Generation and Quality-Estimation Selection

Maximum-probability decoding does not necessarily select the translation preferred by an external quality criterion (Eikema and Aziz, 2020). Minimum Bayes risk decoding addresses this discrepancy by selecting a hypothesis according to its expected utility over a candidate distribution. Neural utility functions can improve the relationship between the decoding objective and external translation-quality measures (Freitag et al., 2022). Quality-aware decoding provides a related approach in which multiple hypotheses are generated, evaluated by an external quality model, and

reranked according to the resulting scores (Fernandes et al., 2022).

Our system follows the latter approach. It generates multiple hypotheses using one translator at different sampling temperatures and selects among the surviving candidates using CometKiwi, a reference-free quality-estimation model that requires a source sentence and translation hypothesis but no reference translation (Rei et al., 2022). Unlike minimum Bayes risk decoding, the procedure does not estimate expected utility over the candidate distribution. It directly selects the candidate with the highest CometKiwi score after deduplication and filtering.

Optimizing candidate selection against a learned metric can expose weaknesses in that metric (Amrhein and Sennrich, 2022). Moreover, quality-estimation scoring alone does not enforce a target-script constraint. Before scoring, the system therefore rejects candidates containing Han characters when producing English or Indonesian output. This filter has a deliberately narrow scope: it detects the presence of Han characters rather than the language of the complete hypothesis. Consequently, it cannot identify every off-target output and may reject a legitimate retained Chinese expression. If every candidate is removed, the system applies its documented fallback so that each source cue retains an output candidate.

2.4 Terminology consistency

Lexically constrained decoding can enforce required target strings during generation (Hokamp and Liu, 2017). Our mechanism operates through prompt instructions and post-selection replacement, without modifying the decoding search algorithm. A per-episode mapping gives one target rendering for recurring source terms such as character names, organizations, titles, places, and other repeated expressions. Relevant mappings are included in the cue prompt, and a narrow deterministic recovery is applied after candidate selection when a mapped source string remains verbatim in a non-Chinese output.

2.5 Candidate generation and QE selection

Decoding by maximum model probability can prefer hypotheses that differ from those favored by translation-quality criteria (Eikema and Aziz, 2020). Minimum Bayes risk (MBR) decoding addresses this problem by choosing a hypothesis according to expected utility over a candidate distribution,

Item	Documented configuration
Directions	zh→en; zh→id
Translator	Hy-MT2-7B; no task-specific fine-tuning
Serving Prompt context	vLLM OpenAI-compatible server Metadata, prior source cues, selected cue pairs, relevant terms
Candidates	$\tau \in \{0.6, 0.7, 0.8, 0.9, 1.0, 1.1\}$, deduplicated.
Sampling	top_p=0.6, top_k=20, repetition penalty 1.05
QE selector	Unbabel/wmt22-cometkiwi-da
Output	Source indices/timestamps preserved 1:1

Table 1: Inference configuration for the two evaluated directions.

and neural utility functions can improve the relationship between decoding and external quality measures (Freitag et al., 2022). Quality-aware decoding provides another route: generate alternatives, score them using an external quality model, and select the best-scoring hypothesis (Fernandes et al., 2022). Recent work has also studied combining rather than simply ranking QE-scored hypotheses (Vernikos and Popescu-Belis, 2024).

We use CometKiwi, a reference-free QE model requiring a source and hypothesis but no reference (Rei et al., 2022). Unlike MBR, this procedure does not estimate utility over a candidate distribution: it directly maximizes a reference-free QE score. Metric-guided decoding can expose weaknesses in the optimized metric (Amrhein and Sennrich, 2022), motivating a separate Han-character check before scoring.

3 System Components and Methodology

3.1 End-to-end architecture

The system converts Chinese subtitle cues into English or Indonesian cues. Each cue contains an index, start and end timestamps, and dialogue text. The indices and timestamps are retained separately, this approach is to make sure only dialogue and textual context reach the translator. Figure 1 shows the subtitle parsing, metadata extraction, and the cue-level translation loop. The episode glossary is used during prompting and again after QE selection. Solid arrows show cue processing and iteration process, whereas dashed arrow represented the glossary use. Earlier cue translations are fed into later prompts through a rolling history. The inference settings are listed in Table 1. The remaining reproducibility details are given in Section 6.

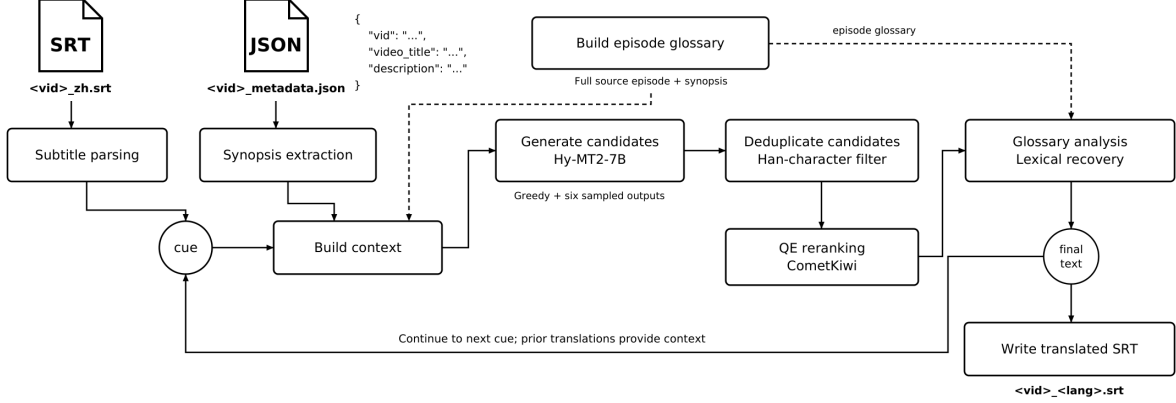


Figure 1: Subtitle translation pipeline. The context construction are provides by the Source SRT cues and JSON synopsis metadata. The episode glossary add the terminology for prompting and post-selection lexical recovery. Hy-MT2-7B generated-candidates are deduplicated and filtered before CometKiwi reranking. The feedback path shows that we use iterates over ques and use prior translation, but it does not show when we update the history store. We keep the original indices and timestamps when rebuilding the SRT.

3.2 Base translation and serving

Hy-MT2-7B is the sole translation generator and the model is served through vLLM (Zheng et al., 2026; Kwon et al., 2023). We are not performs any task-specific fine tuning, so there will be no training corpus, optimizer, or adaptation schedule for the submitted system. The user message contains the full translation instruction, and target languages are written out in full.

Generation settings. For each cue, the model generates one sampled candidate at each of the six temperatures listed in Table 1. All the candidates use $\text{top_p}=0.6$, $\text{top_k}=20$, $\text{repetition_penalty}=1.05$, and $\text{max_tokens}=512$, the difference are only in the temperature. So candidate diversity comes from a single translation model and a single cue prompt, not from an ensemble of generators.

3.3 Context-aware prompting

Let x_i be source cue i and t the target language. The cue prompt is represented as

$$p_i = [I_t; M; G_i; x_{i-h:i-1}; (x, y)_{i-r:i-1}; x_i], \quad (1)$$

where the brackets denote conceptual prompt components, I_t is the target-language instruction, M is task-provided title/synopsis metadata, G_i is the set of relevant terminology entries, $x_{i-h:i-1}$ denotes a bounded window of preceding source cues, and $(x, y)_{i-r:i-1}$ denotes preceding source cues paired with their selected translations. Unselected sampled candidates do not enter history. The exact

history update point is among the configuration details listed in Section 6.

Roles of the context sources. Each context source does a different job. Metadata gives a stable episode-level description that can disambiguate setting or character relationships. The preceding source cues includes nearby context that may resolve short or elliptical dialogue. Previously selected source-target are able to reveal target-side lexical and stylistic choices already in place, so later decisions can take them into consideration. Terminology entries give explicit preferred renderings for recurring terms. This make the two histories are capped to the prompt budget and contain earlier cue. However, the restriction are only applies into rolling history, the glossary construction are not affected by the restriction. This approach are to make sure the glossary construction could reads from the whole episode.

Target-specific instruction. For the Indonesian instruction, we additionally add a colloquial but clean register and provide guidance on interpersonal address. For example, to distinguish peer-oriented *aku/kamu* from more formal forms such as *Anda* or *Bapak/Ibu* when the relationship calls for them. The instruction exposes these choices to the models, with the the surrounding dialogue gives it evidence for choosing among them.

3.4 Episode terminology and lexical recovery

Before cue-level translation, the system performs a single LLM pass over the entire source episode.

The synopsis metadata acts as a seed, helping the model build a mapping from source terms to their preferred target renderings. This mapping is meant to cover recurring proper names, organizations, titles, places, and other repeated expressions in the episode. or a cue containing a mapped term, the relevant entry is included in G_i in Equation 1. This approach help the translator already has the preferred rendering before it starts generating.

Post-selection lexical recovery. The *Glossary Analysis* stage in Figure 1 act as a second, narrower mechanism. After QE selection, it checks whether a mapped source string presents verbatim in a non-Chinese target. When it finds an exact match, it simply replaces that string with the mapped rendering. The recovery step is intentionally limited, where it does not perform any morphological inflection, semantic entity linking, or general rewriting. As the results, it can miss terminology errors expressed through paraphrases and produce awkward morphology after the replacement. Let R_G denote this recovery operation. The emitted text is $\hat{y}_i = R_G(y_i^*)$, which may differ from the QE-selected candidate y_i^* . The selection score applies to y_i^* , so any QE score attributed to the emitted output must be computed on $\hat{y}_i$.

3.5 Candidate generation, filtering, and QE

Candidate search and deduplication. For prompt p_i , the raw candidate set contains one sampled hypothesis y_i^τ for each temperature in $T = \{0.6, 0.7, 0.8, 0.9, 1.0, 1.1\}$:

$$C_i = \{y_i^\tau \mid \tau \in T\}, \quad \tilde{C}_i = \text{dedup}(C_i). \quad (2)$$

The raw pool therefore contains at most six hypotheses, and order-preserving deduplication can reduce it further when different temperatures yield identical text.

Han-character filtering. Before QE scoring, a target-specific filter F_t removes out candidates that containing Han characters for English and Indonesian. The goal is not to judge whether a translation is semantically adequate, but to catch one narrow and easily observable failure, where source-script leakage into a non-Chinese output. Since the filter relies on script rather than acting as a general language detector, it may keep Latin-script text that is in the wrong language, and it may also reject legitimate cases where a Chinese name or quotation is intentionally retained. If applying the filter would

remove every candidate, the system falls back to the unfiltered deduplicated pool $\tilde{C}_i$, ensuring that every source cue still has at least one output candidate.

QE selection and model reuse. Each of the remaining candidate then scored with Unbabel/wmt22-cometkiwi-da using the current source cue and candidate translation (Rei et al., 2022). The checkpoint is loaded just once and then kept in memory for the entire run, so scoring candidates does not require reloading the model for each cue. With $Q(x_i, y)$ denoting the QE score, selection is

$$y_i^* = \arg \max_{y \in F_t(\tilde{C}_i)} Q(x_i, y), \quad (3)$$

Under the empty-filter fallback of $F_t(\tilde{C}_i) = \tilde{C}_i$, all candidates are temperature-sampled. There is no deterministic anchor and samples selected on QE score alone. With that, A fluency-biased QE model may thus favor fluency over adequacy. Equation 3 is a selection rule, not a measured gain, where assessing it requires a controlled comparison against a direct-translation baseline.

3.6 Subtitle reconstruction

The parser keeps every source cue index and timestamp. After translation and lexical recovery, only the text field is replaced. The cues are written back in their original order and named using the required `vid_langshort.srt` convention. This reconstruction preserves the source one-to-one in cue count, indexing, and timing. The pipeline can split a cue that exceeds length limits, where on the submitted data this did not occur, so output and source cue counts coincide. Thus, preserving this structure does not establish subtitle usability, where we still found translated cues can still exceed suitable line lengths or reading speeds.

4 Inference-Time Optimization

The inference pipeline combines the candidate sampling, translation history, terminology handling, and compute reuse. We describe how they work in Section 3 and summarize the computational consequences here.

Candidate budget and deduplication. We sample six decodes, giving at most six candidates per cue. The exact deduplication are able to reduces the number of QE evaluations, but does not recover

generation work that has already performed. After filtering, each cue requires between one and six candidates scores, the all-filtered fallback able to restore at least one candidate if every generation are filtered out.

Resident scoring and episode reuse. We keep CometKiwi loaded across cues, so we avoid reinitializing the checkpoint. The episode glossary is only constructed once and reuse for the whole episode, while translation history is bounded rather than growing with the entire output episode. Our choices are attempt to limit any redundant work, but resident QE also takes up memory and extra context, making the prompt processing more expensive.

Cost interpretation. The training-free approach we describes are not involving any task-specific parameter updates. The proposed method currently does not establish any low inference cost. A direct run needs one generation per cue, and the full pipeline adds sampled generation, QE scoring, and episode glossary construction. The optimizations goals are to reduce specific redundant work, even with the benefits are depends on the generation length, candidate duplication, batching, and the serving hardware. No measured runtime or memory advantage is analyze in the paper.

5 Evaluation and Results

5.1 Evaluation scope and metrics

We evaluate the complete `~rupiah` outputs for `zh→en` and `zh→id`. The primary evaluation uses the shared task’s preliminary official assessment, based on the subtitle quality rubric. The rubric components are Accuracy/Fidelity (40), Fluency/Readability (35), Cultural/Stylistic Adaptation (20), and Subtitle Text Conventions (5). Each video are scored and aggregated to 0–100.² These scores are produced by an automatic judge with human verification, and they apply to the actual submitted subtitle files. The secondary evaluation is a reference-based diagnostic (BLEU, chrF, chrF++) computed against ground-truth subtitles, and a reference-free CometKiwi diagnostic (because CometKiwi is also the selector components in our pipeline, we treat it as diagnostic only.

²Preliminary results and per-video judgments released by the organizers: https://github.com/jasonNLP/wmt26_video_subtitle_translation.

English		Indonesian	
Rank	System (score)	Rank	System (score)
1	ReopenAI (90.19)	1	ReopenAI (87.84)
2	LensLinguals (75.25)	2	laniqo (74.27)
3	AuditTranslate (74.70)	3	~rupiah (67.60)
4	laniqo (74.16)	4	BVSLP (62.42)
5	~rupiah (70.89)		
6	BVSLP (60.48)		
7	paramanandaAI (55.65)		
8	SubVision (49.77)		

Table 2: Preliminary official mean quality score (0–100), all systems, $n = 100$ videos per direction. Scores from the released organizer results.

The separate reference-score file covers 89 videos (34,250 cues) per direction, so BLEU/chrF/chrF++ describe 89% of the submitted videos. It also flags 1,622 English cues (4.74%) and 1,594 Indonesian cues (4.65%) with `empty_ref`, but, these flags are concern references, not as empty system outputs. BLEU measures n-gram precision with a brevity penalty (Papineni et al., 2002); chrF compares character n-grams and can be less brittle to surface variation (Popović, 2015). We report corpus-level `__ALL__` aggregates rather than averaging per-video scores (Post, 2018).

5.2 Official preliminary results

Table 2 ranks `~rupiah` on the preliminary official metric: fifth out of eight in English and third out of three in Indonesian, though the Indonesian result reflects a smaller comparison set. In both directions, the top system (ReopenAI) is well separated from the field, while `~rupiah` falls within a compact band below it.

5.3 Per-dimension analysis

Table 3 decomposes the official score by rubric dimension. `~rupiah` and the top system are close on Fluency/Readability, tied on Subtitle Text Conventions, and moderately apart on Cultural/Stylistic Adaptation. Almost all the separation is on Accuracy/Fidelity, where `~rupiah` scores about three-fifths of the top system in both directions. Consistent with this, critical mistranslation dominates `~rupiah`’s deductions, with a per-video penalty an order of magnitude larger than the top system’s. The submitted configuration therefore produces fluent, well-formed subtitles whose main limitation is adequacy, as examples of correct, names, and referents that not surface form or timing.

Dir	System	Acc/40	Flu/35	Cul/20	Conv/5
en	ReopenAI	36.25	31.29	18.35	4.36
	~rupiah	22.07	29.08	15.49	4.34
id	ReopenAI	34.83	32.80	17.76	4.65
	~rupiah	22.55	29.82	14.64	4.74

Table 3: Mean per-video rubric-dimension scores, ~rupiah vs. the top-ranked system, from the preliminary official judgments. Maxima in headers.

Target	BLEU $\uparrow$	chrF $\uparrow$	chrF++ $\uparrow$	CometKiwi $\uparrow$
English	16.49	39.11	37.10	0.7515
Indonesian	11.15	38.68	35.50	0.6882

Table 4: Reference-based diagnostics over 89 videos (34,250 cues) per direction, and reference-free CometKiwi. CometKiwi is also the selector and is not independent evidence; see Section 5.1.

5.4 Reference-based diagnostics

Table 4 reports the reference-based diagnostics. Because each direction uses its own references, these scores are not directly comparable across languages. They do support the official metric’s ordering, where English above Indonesian, however, they do not isolate any individual component.

5.5 Evaluation boundaries

The evaluations above characterize the full pipeline and, via Table 2, its standing among submitted systems. But they do not isolate the contributions of context, terminology, or QE selection. So, the no direct-translation baseline or component ablation is included, and CometKiwi serves as both selector and diagnostic. A matched greedy-baseline comparison and a small blinded human assessment, especially of Indonesian register where it remain the necessary next steps, as outlined in the released analysis scaffold.

6 Reproducibility and Limitations

6.1 Configuration and parameter accounting

We release the code behind our submission. It locks in the pipeline structure, prompt templates, glossary construction, candidate temperatures, filtering, and QE selection. For exact reproduction, read the run-specific settings from the config, where the source- and translation-history window sizes h and r , prompt-truncation rules, random seed policy, checkpoint revisions, and vLLM version. The config also states whether the rolling history keeps the

QE-selected candidate y_i^* or the recovered output $\hat{y}_i$, so you don’t have to infer it from the outputs.

We also use two checkpoints, Hy-MT2-7B for translation and glossary construction and Unbabel/wmt22-cometkiwi-da for QE selection. Glossary construction reuses the translator, so there’s no third model. That gives roughlt 7B + 0.57B, where around 7.6B parameters, within the WMT26 constrained-track budget of 20B (WMT). Both are counted because both run at inference. Revisions and licenses are in the released code.

6.2 Limitations and responsible use

We want to highlight two points, error propagation and local correction. We found a mistaken name or register in the rolling history can bleed into later cues. The Han-character check tests script, not language. Exact-match recovery misses paraphrased terminology and can break morphology. These issues help explain the adequacy gap in Section 5.3: without a deterministic anchor in the candidate pool, QE can favor fluent hypotheses over faithful ones, and the official judgments mainly penalize critical mistranslation, not fluency or formatting. Wrong titles or pronouns can also distort speaker relationships, and preserved timestamps don’t guarantee readable speed or line length. Since these failures are about adequacy and register, aggregate metrics should be paired with contextual human review (especially of Indonesian address forms) and and subtitle-usability checks.

7 Conclusion

We introduced a training-free subtitle translation system based on Hy-MT2-7B. Episode metadata, dialogue history, and terminology guide cue-level generation. Temperature sampling supplies alternatives for Han-character filtering and CometKiwi selection, while lexical recovery and SRT reconstruction yield the final subtitles without changing cue indices or timestamps. In the preliminary official evaluation, the system ranks fifth of eight in English and third of four in Indonesian. Fluency and subtitle conventions are competitive, but accuracy clearly lags Per-dimension analysis traces almost the entire gap from the top system to adequacy, not surface form. Future works could explore whether QE-guided selection without an adequacy anchor helps or hurts the faithfulness, a matched greedy baseline and a small blinded human assessment would settle it.

References

- Chantal Amrhein and Rico Sennrich. 2022. [Identifying weaknesses in machine translation metrics through minimum Bayes risk decoding: A case study for COMET](#). In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1125–1141, Online only. Association for Computational Linguistics.
- Bryan Eikema and Wilker Aziz. 2020. [Is MAP decoding all you need? the inadequacy of the mode in neural machine translation](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and André F. T. Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
- Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022. [High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics](#). *Transactions of the Association for Computational Linguistics*, 10:811–825.
- Chris Hokamp and Qun Liu. 2017. [Lexically constrained decoding for sequence generation using grid beam search](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada. Association for Computational Linguistics.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Evgeny Matusov, Patrick Wilken, and Yota Georgakopoulou. 2019. [Customizing neural machine translation for subtitling](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 82–93, Florence, Italy. Association for Computational Linguistics.
- 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*.
- Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*.
- Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*.
- Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-Unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Giorgos Vernikos and Andrei Popescu-Belis. 2024. [Don’t rank, combine! combining machine translation hypotheses using quality estimation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12087–12105, Bangkok, Thailand. Association for Computational Linguistics.
- Sebastian Vincent, Charlotte Prescott, Chris Bayliss, Chris Oakley, and Carolina Scarton. 2024. [A case study on contextual machine translation in a professional scenario of subtitling](#). In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 561–572, Sheffield, UK. European Association for Machine Translation (EAMT).
- WMT. 2026. [Shared task: Video subtitle translation](#).
- 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](#). *Preprint*, arXiv:2605.22064.