

Full-Video Context and Anonymous Candidate Selection for WMT26 Video Subtitle Translation

jianfeng Wu

Independent Researcher

Building 10, Block B, Shenzhen Bay Science and Technology Ecological Park
Yuehai Subdistrict, Nanshan District, Shenzhen, Guangdong 518057, China
1550266278@qq.com

Abstract

We describe ReopenAI’s constrained-track submission to the WMT26 Video Subtitle Translation Shared Task. The system translates Simplified Chinese subtitles into English, Thai, Indonesian, Malay, and Traditional Chinese (Taiwan). It combines Hy-MT2-7B and Gemma-4-12B-it without fine-tuning; their unique weights total approximately 19B parameters. Both models receive the complete source subtitle sequence as context and sample two cue-level candidates. Gemma selects one of the four anonymous candidates using a stable random order and its reverse. A selection is accepted only when both judgments identify the same candidate; otherwise a fixed Gemma candidate is used. Official preliminary evaluation ranks ReopenAI first overall and in every target direction, with a macro-average score of 89.357 over 100 videos. On an internal three-video diagnostic totaling 705 cues per target, fusion obtains a 93.8% pooled macro-average acceptable-translation rate, compared with 93.0% for Gemma, 90.0% for Hy-MT2, and 77.0% for Qwen. On a separate 122-cue ablation, full-video context improves Hy-MT2 in all five targets, with gains from 4.1 points in English and Traditional Chinese to 18.0 points in Malay. The system uses no task-specific parallel data, audio, or video frames.

1 Introduction

Subtitle translation requires more than independent sentence translation. Full-video textual context can help resolve the referents of pronouns and demonstratives, recover omitted arguments, and maintain discourse and terminology consistency across adjacent cues. However, it cannot fully track speakers when disambiguation requires audio cues, such as voice identity or prosody, or visual evidence. Output must also preserve cue boundaries and remain concise. The WMT26 task evaluates five directions from Simplified Chinese:

English, Thai, Indonesian, Malay, and Traditional Chinese as used in Taiwan.

Our system combines the translation-specialized Hy-MT2-7B (Zheng et al., 2026) with Gemma-4-12B-it (Gemma Team, 2026). It supplies full-video Chinese context to every cue-level request, generates four candidates using two models and two seeds, and reuses Gemma for conservative candidate selection. Reusing a checkpoint does not add unique parameters, so the system remains below the constrained track’s 20B limit. Our contributions are: (1) a context format that preserves SRT alignment; (2) heterogeneous candidate generation without training; (3) anonymous, reversed-order selection with deterministic fallback; and (4) a single-model comparison and context ablation that document both benefits and failure modes.

2 Related Work

Early context-aware NMT established that information beyond the current sentence can improve document translation. Miculicich et al. (2018) model document context with hierarchical attention, while Voita et al. (2018) show on subtitle data that contextual models learn anaphora resolution and improve pronoun translation. For production-oriented subtitling, Matusov et al. (2019) combine domain and style adaptation, inter-sentence context, and learned subtitle segmentation with explicit length and duration constraints. These studies motivate our use of cross-cue context while also highlighting the need to preserve subtitle structure.

Recent LLM-based document translation combines broad source context with segment-level decoding. Hu et al. (2025) prime an LLM with the complete source document and then translate its segments in a multi-turn conversation, retaining prior context while reusing the key-value cache.

Because useful long-range dependencies can be sparse, Peng et al. (2025) instead retrieve informative local and distant source and target context for each segment. Our method uses the simpler training-free alternative of supplying the complete source subtitle sequence to every cue-level request. This avoids a separate retrieval component, but irrelevant distant cues may still dilute the useful context.

Recent subtitle-specific studies also emphasize metadata, character style, and structure. Pramodya et al. (2025) show that movie metadata such as title, summary, and genre can improve LLM subtitle translation. Ito et al. (2026) derive character profiles and scene descriptions from context to create style-preserving augmentation for movie-specific translation. In an end-to-end speech setting, Cettolo et al. (2026) use a longer-context refinement pass and realign its output to the original subtitle timing. We likewise combine video-level text with cue-level output, but use only the supplied titles and Chinese subtitles: we perform no training, audiovisual processing, or resegmentation, and preserve the given SRT indices and timestamps exactly.

Candidate diversity and quality-aware selection provide the other basis of our system. Shyhunyuan-MT distills diverse translations and applies multi-reward reinforcement learning (Zheng et al., 2025), whereas IRB-MT uses an unchanged Gemma checkpoint for free-form self-refinement (Grubisic and Korencic, 2025). Kaze-MT ranks a large candidate set with multiple reference-free estimators (Tan, 2025), and KIKIS combines translation and reranking subsystems with long-context decisions (Iwakawa et al., 2025). We instead select verbatim from four candidates with no separately trained ranker, trading possible oracle quality for parameter-budget compliance and auditability.

3 Task and Data

The official test package contains 100 Chinese SRT files with 41,837 cues and title metadata for each video. Five target languages produce 209,185 requests per decoding run. We use only these source subtitles and metadata: no reference translations, task-specific training data, external glossary, video frames, or audio. Every output retains the source cue indices and timestamps and follows the required `vid_langshort.srt` naming convention.

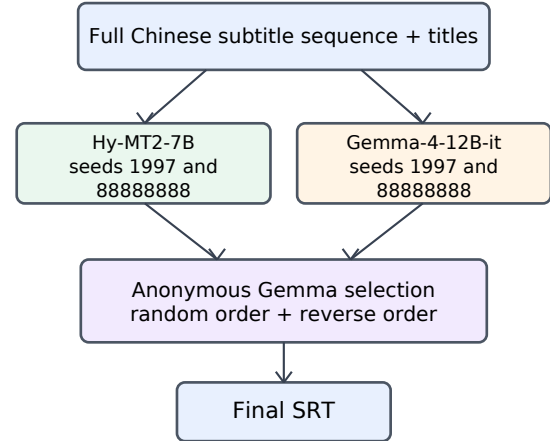


Figure 1: Submitted inference pipeline. Both generators share full-video source context; Gemma selects one anonymous candidate verbatim.

4 Method

For source cue x_i in video v , we build context c_v from title metadata and all Chinese cues in temporal order. The two submitted models $m \in \{\text{Hy}, \text{Gemma}\}$ decode with seeds $s \in \{1997, 88888888\}$:

$$y_i^{m,s} \sim p_m(y \mid x_i, c_v; s). \quad (1)$$

The four candidates are anonymized and one is copied verbatim to the final SRT. Figure 1 summarizes the pipeline.

4.1 Full-Video Context

Each SRT is parsed into cue index, timestamp, and text. Album and video titles are prepended to the complete Chinese cue sequence. Prompts describe this sequence as background for discourse, references, sentence continuation, terminology, tone, and proper names, but require translation of the current cue only. The long prefix is identical across requests for the same video and language, allowing vLLM prefix caching (Kwon et al., 2023).

4.2 Candidate Generation

Hy-MT2 receives the context and current cue in one user message. It uses BF16, temperature 0.7, top- p 0.6, top- k 20, repetition penalty 1.05, and a 32,768-token context limit. Gemma receives the context as a system message and the cue as a user message. Thinking is disabled; it uses BF16, temperature 0.7, top- p 0.95, top- k 64, and the same

context limit. Both produce at most 512 new tokens. Diversity therefore comes from model heterogeneity and two independent seeds rather than a large decoding sweep.

Qwen3.5-9B (Qwen Team, 2026) was evaluated as a single-model baseline with full-video context, thinking disabled, temperature 0.7, top- p 0.8, top- k 20, presence penalty 1.5, and at most 128 new tokens. It was not included in the final ensemble: using all three checkpoints would total 28B parameters, and Gemma had the strongest five-language macro average.

4.3 Agreement-Gated Selection

Gemma is reused as the selector for two practical reasons. In pilot tests, Hy-MT2 did not reliably complete the constrained task of returning only a candidate label, and adding a separate ranking model would risk exceeding the track’s 20B parameter limit. Reusing Gemma keeps the submitted system at 19B unique parameters.

For batches of at most eight cues, the selector sees the Chinese source and four candidates labeled A–D, without model names or seed values. Its prompt prioritizes faithfulness, negation and logical relations, cross-cue coherence, natural subtitle style, and terminology. Rewriting and explanations are forbidden; the response is strict JSON containing cue identifiers and labels.

Let π_i be a stable pseudorandom candidate order. Gemma judges π_i and its reverse. After mapping labels back to identities, the decision is

$$\hat{k}_i = \begin{cases} k, & g(\pi_i) = g(\text{rev}(\pi_i)) = k, \\ k_{\text{fb}}, & \text{otherwise,} \end{cases} \quad (2)$$

where k_{fb} is the Gemma candidate generated with seed 1997. Invalid JSON or labels also trigger fallback. Identical candidates bypass the selector. Selection uses temperature 0, and every mapping and decision is recorded in a JSONL audit file excluded from the translation submission.

5 Implementation

The Python implementation uses Transformers and vLLM. Generation is batched by video and target language; Hy-MT2 and Qwen use one-way tensor parallelism, and Gemma uses two-way parallelism. Before fusion, every SRT is checked for equal cue count, identical indices and timestamps, and non-empty text. No weights

are trained or modified. The release package includes the three baseline scripts and the anonymous fusion program; public model identifiers are tencent/Hy-MT2-7B, Qwen/Qwen3.5-9B, and google/gemma-4-12B-it.

Each generation run is separated by model and seed, so candidate SRTs remain independently inspectable. The fusion stage writes both the final SRT and a JSONL record containing the hidden candidate mapping, the forward and reverse judgments, agreement status, fallback reason, and selected identity. These records are excluded from the translation submission but permit every final cue to be traced to an unmodified model output. Failed validation stops fusion rather than silently repairing indices, timestamps, or missing cues.

6 Evaluation

6.1 Official Shared-Task Results

The official preliminary findings evaluate submitted systems with three language-aware LLM-based LQA assessments per video, followed by human verification and correction (WMT26 Video Subtitle Translation Shared Task, 2026). ReopenAI was evaluated on all 100 test videos in each target direction and ranked first overall and first in every language. Table 1 reports the top two multilingual systems. ReopenAI’s 89.357 macro average exceeds Ianiqo’s 78.265 by 11.092 points. Human-verified assessments were available for 93.36% of all shared-task evaluation units. These 100-point, video-level LQA scores are not directly comparable to the cue-level A+B rates reported below.

Team	en	th	id	ms	zh-TW	Avg.
ReopenAI	90.195	89.015	87.845	82.160	97.571	89.357
Ianiqo	74.155	79.895	74.270	69.120	93.885	78.265

Table 1: Official preliminary WMT26 shared-task scores. Each language score is the mean over 100 videos; Avg. is the language macro average. ReopenAI ranked first overall and in every direction.

6.2 Internal Evaluation Protocol

To complement the official evaluation with cue-level analysis, we report an internal diagnostic. Three test videos with 122, 199, and 384 cues (705 total) were examined in all five target languages. Cues were graded A (accurate and natural), B (correct with a minor issue), C (noticeable but recoverable error), or D (severe meaning, relation,

Model	Params	en	th	id	ms	zh-TW	Avg.
Hy-MT2	7B	81.7	86.7	91.3	91.9	98.3	90.0
Qwen3.5	9B	82.0	71.9	72.6	66.0	92.8	77.0
Gemma 4	12B	87.8	90.5	94.3	94.9	97.7	93.0
Fusion	19B	88.1	91.6	95.5	95.7	98.2	93.8

Table 2: Quality of the single-model baselines and submitted fusion. Values are pooled A+B percentages over three videos (705 cues per language); Avg. is the language macro average.

logic, or language error). A+B is the acceptable-translation rate. Manual grading was conducted by five annotators, one for each target language. All were native Chinese speakers and second-year undergraduates majoring in their assigned target language. Each annotator scored all samples for that language across the three videos while viewing the complete Chinese subtitle sequence and the translations under evaluation. Translations from the different models and methods were anonymized and presented in shuffled order, so annotators were blind to system identity. Several worked examples served as annotation guidance, but no separate calibration round was conducted. Because each language had only one annotator and no items were double-annotated, inter-annotator agreement could not be computed. All three baselines were graded cue by cue on the same sources and outputs. An existing grade was reused when an output string exactly matched a previously graded candidate; changed outputs were judged with the same rubric, and high-risk cross-cue, name, title, and translation-direction errors were manually re-adjudicated. Indonesian and Malay grading also penalizes lexical mixing; Traditional Chinese grading includes Taiwan-oriented character conversion. These scores are diagnostic rather than official task metrics.

6.3 Translation Quality

Table 2 compares the three baselines and fusion. Fusion improves over Gemma by 0.3 points in English, 1.1 in Thai, 1.2 in Indonesian, 0.8 in Malay, and 0.5 in Traditional Chinese. Its pooled macro average is 93.8%, 0.8 points higher than Gemma’s 93.0%. The direction is consistent across videos: Fusion versus Gemma macro averages are 91.3 versus 90.0 on a0046j15d6d, 95.5 versus 95.2 on j0048kesypr, and 93.8 versus 92.9 on z41019npv19. Hy-MT2 still leads pooled Traditional Chinese by 0.1 points and remains useful because it occasionally avoids Gemma errors

in names, omitted subjects, cue segmentation, and source-preserving conversion. Qwen only edges Hy-MT2 in English; its other results are weaker.

Selector behavior. Across all 209,185 cues in the official submission, forward and reverse judgments agreed for 54,030 cues (25.83%); 115,364 (55.15%) triggered the fixed Gemma fallback, and all four candidate strings were identical for 39,791 (19.02%). Among agreement cases, the selector chose a Gemma-origin candidate 45,684 times (84.55%). Excluding 1,787 agreement cases in which the selected string was produced by both model families leaves 44,401 Gemma selections among 52,243 attributable cases (84.99%), with language-specific shares from 77.73% for Traditional Chinese to 89.55% for Thai. This skew does not establish self-preference: model identities were hidden. However, model-specific lexical or stylistic cues may reveal candidate provenance, so the selector-bias concern remains.

6.4 Context Ablation

Context	Lang.	A+B	D	$\Delta A+B$
None	en	71.3	17	–
Full video	en	75.4	12	+4.1
None	th	68.0	15	–
Full video	th	77.0	11	+9.0
None	id	73.0	14	–
Full video	id	86.9	6	+13.9
None	ms	68.9	16	–
Full video	ms	86.9	6	+18.0
None	zh-TW	93.4	3	–
Full video	zh-TW	97.5	0	+4.1

Table 3: Five-language Hy-MT2 context ablation on the same 122 cues.

Table 3 shows that full-video context improves Hy-MT2 in all five languages and reduces severe errors in every target. The largest gains are 18.0 points for Malay and 13.9 for Indonesian, followed by 9.0 for Thai; English and Traditional Chinese each improve by 4.1 points. Across the five targets, D-grade errors fall from 65 to 35. The largest gains involve cross-cue syntax and omitted subjects. Context also introduces risks: a model can leak text from the next cue, drift on names or historical titles, or create overlong subtitles. A rhetorical-question error shared by all candidates also showed the limitation of selection without rewriting.

7 Conclusion

We submitted a constrained, inference-only system combining Hy-MT2-7B and Gemma-4-12B-it under the 20B limit. Full-video context improves isolated cue translation, heterogeneous generation exposes complementary outputs, and agreement-gated selection preserves auditability because it cannot invent a new translation. Official preliminary evaluation ranks ReopenAI first overall and in every target direction, with an 89.357 macro average over 100 videos. The internal baseline comparison supports Gemma as the final selector and fallback, while the three-video fusion diagnostic shows a 0.8-point pooled macro improvement (93.8% versus 93.0%) over Gemma alone and a higher macro average on every evaluated video. The 122-cue, five-language context ablation shows gains in every target while identifying remaining subtitle-specific risks.

8 Limitations

The system uses textual context only and cannot resolve ambiguities requiring visual evidence, voice identity, prosody, or sound effects. Complete context can leak neighboring content, while cue-level selection can mix terminology across a video. The Gemma selector may favor Gemma-like candidates; anonymous labels and reverse ordering reduce identity and position cues but do not guarantee unbiased estimation. The official preliminary scores use human-verified LLM-based LQA rather than exhaustive independent human ratings. Our semantic manual analysis covers all five targets on three videos totaling 705 cues, still a small subset of the 100-video test set. Each target language was evaluated by one annotator without overlapping annotations, so the scores may reflect individual judgment and no inter-annotator reliability statistic is available. Reusing grades for exact string matches also means that not every system output received an independent rating. The manual grades should therefore not be read as official corpus-level scores.

References

Mauro Cettolo, Roldano Cattoni, Matteo Negri, and Luisa Bentivogli. 2026. [The FBK sentence-aware subtitling system at the IWSLT 2026 subtitling track](#). In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*,

pages 68–77, San Diego, USA (in-person and online). Association for Computational Linguistics.

Gemma Team. 2026. [Gemma 4 technical report](#). Preprint, arXiv:2607.02770.

Ivan Grubisic and Damir Korencic. 2025. [IRB-MT at WMT25 translation task: A simple agentic system using an off-the-shelf LLM](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 753–764, Suzhou, China. Association for Computational Linguistics.

Hanxu Hu, Jannis Vamvas, and Rico Sennrich. 2025. [Source-primed multi-turn conversation helps large language models translate documents](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23702–23712, Suzhou, China. Association for Computational Linguistics.

Hitoshi Ito, Naoto Shirai, Kazutaka Kinugawa, Hideya Mino, Rei Endo, and Yoshihiko Kawai. 2026. [Context-driven and reference-guided data augmentation for subtitle translation](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 41381–41394, San Diego, California, United States. Association for Computational Linguistics.

Koichi Iwakawa, Keito Kudo, Subaru Kimura, Takumi Ito, and Jun Suzuki. 2025. [KIKIS at WMT 2025 general translation task](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 705–722, Suzhou, China. 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 29th Symposium on Operating Systems Principles*, pages 611–626.

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.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. [Document-level neural machine translation with hierarchical attention networks](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Ziqian Peng, Rachel Bawden, and François Yvon. 2025. [Self-retrieval from distant contexts for document-level machine translation](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 220–240, Suzhou, China. Association for Computational Linguistics.

- Ashmari Pramodya, Yusuke Sakai, Justin Vasselli, Hidetaka Kamigaito, and Taro Watanabe. 2025. [Translating movie subtitles by large language models using movie-meta information](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 315–330, Vienna, Austria. Association for Computational Linguistics.
- Qwen Team. 2026. [Qwen3.5-9B model card](#). Hugging Face model repository. Accessed 2026-08-06.
- Shaomu Tan. 2025. [Simple test time scaling for machine translation: Kaze-MT at the WMT25 general translation task](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 651–656, Suzhou, China. Association for Computational Linguistics.
- Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. [Context-aware neural machine translation learns anaphora resolution](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.
- WMT26 Video Subtitle Translation Shared Task. 2026. [Preliminary findings of the WMT26 video subtitle translation shared task: Context and candidate selection matter](#). Shared-task overview and findings paper.
- 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.
- Mao Zheng, Zheng Li, Yang Du, Bingxin Qu, and Mingyang Song. 2025. [Shy-hunyuan-MT at WMT25 general machine translation shared task](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 607–613, Suzhou, China. Association for Computational Linguistics.