

Findings of the WMT26 Video Subtitle Translation Shared Task: Context and Candidate Selection Matter

Zheng Li*, Mao Zheng*

Foundation Model Department, Tencent
{jasonzli, moonzheng}@tencent.com

Abstract

We present the setup, systems, evaluation, and findings of the WMT26 Video Subtitle Translation shared task. The task covers translation from Simplified Chinese into English, Thai, Indonesian, Malay, and Traditional Chinese for Taiwan, China. Eight teams submitted 2,200 SRT files across 22 team-language combinations. We evaluated 2,198 unique team-language-video units with repeated LLM-based language quality assessment followed by human checking and correction. Each video was assessed three times, with valid positive scores aggregated by the median. ReopenAI led all five directions and achieved a macro-average of 89.357; lanigo ranked second with 78.265. Corrected scores were extractable from 93.36% of the evaluation units, and 15.30% of those scores differed from the automatic scores. Stronger results were associated with contextual prompting, diverse candidate generation, constrained selection, translator-critic specialization, terminology control, and deterministic subtitle validation. These cross-system associations are not controlled ablation evidence. Persistent challenges include cross-cue reference resolution, cultural and register adaptation, multilingual judge calibration, and balancing brevity with semantic completeness.

1 Introduction

The Conference on Machine Translation has long served as a common evaluation ground for advances in machine translation. Following the shared-task reporting tradition exemplified by the WMT25 General Machine Translation findings paper (Kocmi et al., 2025), this paper documents the WMT26 Video Subtitle Translation shared task, summarizes the participating systems, reports the evaluation results, and uses the submitted systems as evidence for broader conclusions about the current state of subtitle translation.

Video subtitle translation differs from general sentence-level translation in several important ways, including length, display-duration, and segmentation constraints (Matusov et al., 2019). Subtitle cues are short, time-aligned units that do not necessarily correspond to complete sentences. A single utterance may span several cues. Chinese zero pronouns can require discourse context for recovery (Wang et al., 2019), while context-sensitive deixis, ellipsis, and lexical cohesion have also been studied in subtitle translation (Voita et al., 2019). The output must also remain concise enough for on-screen presentation, retain the source cue structure, and use consistent names, titles, and terminology throughout a video. Colloquial expressions, historical titles, humor, slang, and culturally specific language further increase the difficulty of the task.

The WMT26 shared task focuses on five translation directions from Simplified Chinese: English, Thai, Indonesian, Malay, and Traditional Chinese for Taiwan, China. Participants were provided with source SRT files and video-level metadata, including a synopsis or description. The task used a constrained setting in which final systems were required to remain below 20 billion parameters and to use releasable model weights. This setting encouraged reproducible research with compact open-weight translation and instruction models.

This paper makes four contributions. First, it gives a unified account of the eight participating systems and their language coverage. Second, it describes a scalable evaluation procedure that combines repeated LLM-based LQA assessment with human checking and correction. Third, it reports overall and language-specific rankings. Finally, it analyzes which modeling and inference strategies were associated with stronger performance and identifies remaining challenges for subtitle translation and evaluation.

The rankings reported here are derived from the submitted result files and the AI-plus-human verifi-

*Equal contribution.

cation procedure described in Section 4. They are intended as an analysis of the available submissions and evaluation artifacts.

2 Task and Data

This section defines the task, constrained setting, and submission coverage.

2.1 Task definition

The shared task required participants to translate video subtitles from Simplified Chinese into one or more of the following targets:

- English (en);
- Thai (th);
- Indonesian (id);
- Malay (ms); and
- Traditional Chinese for Taiwan, China (zh-TW).

Each test item contained a source subtitle file in SRT format and associated metadata. The metadata included a video description or synopsis and could be used to disambiguate character relations, domain terminology, and the overall narrative setting. Participants were expected to submit one translated SRT file per video and language using the naming pattern `vid_langshort.srt`.

The official task description emphasizes three properties that distinguish subtitle translation from general-purpose MT: length and timing constraints, the use of metadata, and the need to preserve style and register. Systems could process cues independently, use local context windows, or translate with broader video-level context, provided that the resulting files remained structurally valid.

2.2 Constrained setting

The task adopted a constrained setting. Systems could use different training data, models, and frameworks, but the final model or ensemble had to contain fewer than 20 billion parameters. For ensembles, the parameter counts of all members were combined, and mixture-of-experts systems were counted by their full parameter count. Final model weights were expected to be released under a license that permits unrestricted non-commercial use.¹

¹Official task description: <https://www2.statmt.org/wmt26/video-subtitle-translation.html>.

This constraint shaped system design in two ways. First, several teams relied on compact translation-specialized models from the Hy-MT2 family (Zheng et al., 2026). Second, teams combining a translator with a general instruction model had to allocate the parameter budget carefully, often using a 7B translation model with a 12B critic or selector.

2.3 Submission coverage

Eight teams participated in the evaluated submissions. ReopenAI, lanigo, and BVSLP covered all five language directions. LensLinguals submitted English and Thai; rupiah submitted English and Indonesian; SubVision, AuditTranslate, and paramanandaAI submitted English only.

The collected results contain 2,200 SRT files, corresponding to 22 team–language combinations with 100 files per submitted direction. The evaluation workbook contains 2,198 unique team–language–video units. Two submitted videos had no unique scored evaluation record. One additional duplicated evaluation entry was removed using the tuple (team, language, video ID). These missing score records do not indicate missing SRT submissions.

3 Participating Systems

This section summarizes the eight systems and the main architectural trends.

3.1 System overview

Table 2 summarizes the main design choices. Six of the eight teams used a Hy-MT model in their submitted system. The two exceptions were lanigo, which selected Gemma-4-12B-it (Gemma Team, 2026) after development-set screening, and BVSLP, which used a Qwen3.5-based system (Qwen Team, 2026). The table follows the compact system-comparison style used in WMT findings papers: it records the base model, adaptation strategy, inference-time selection, context use, and post-processing. System-specific configurations below are taken from the submitted descriptions or notebook; citations to base models and general methods do not independently verify those configurations.

3.2 Common architectural trends

The submissions reveal three broad trends. First, large instruction-tuned language models and translation-specialized decoder-only models dominated the task. Second, competitive systems in-

Team	en	id	ms	th	zh-TW	Languages
ReopenAI	100	100	100	100	100	5
laniqo	100	100	100	100	100	5
BVSLP	100	100	100	100	100	5
LensLinguals	100	—	—	100	—	2
rupiah	100	100	—	—	—	2
AuditTranslate	100	—	—	—	—	1
paramanandaAI	100	—	—	—	—	1
SubVision	100	—	—	—	—	1

Table 1: Submission coverage. Each value is the number of collected SRT files; a dash denotes an unsubmitted direction. All 22 submitted team–language combinations contain 100 files. Two files lack unique scored evaluation records.

creasingly treated translation as an inference-time search problem rather than a single generation call. ReopenAI, laniqo, AuditTranslate, and rupiah all generated multiple candidates and applied a separate selection or reranking mechanism. Third, most systems explicitly represented context, terminology, or previous translations rather than translating each cue in isolation.

Training strategies varied substantially. BVSLP used multistage parameter-efficient adaptation, while SubVision and paramanandaAI fine-tuned the smaller Hy-MT2-1.8B model. By contrast, ReopenAI, laniqo, AuditTranslate, LensLinguals, and rupiah obtained their gains primarily through prompting, candidate search, selection, and deterministic post-processing. The final rankings therefore provide an opportunity to compare weight adaptation with inference-time orchestration under the same parameter constraint.

3.3 ReopenAI

ReopenAI submitted all five target languages. Its system combined Hy-MT2-7B (Zheng et al., 2026) and Gemma-4-12B-it (Gemma Team, 2026), for a nominal total of approximately 19 billion parameters. For each cue, Hy-MT2-7B generated two candidates with different random seeds, and Gemma-4-12B-it generated two more. Gemma then acted as an anonymous four-way selector.

The selector saw candidates under labels rather than model identities. Two judgments were made: one with a stable pseudorandom candidate order and another with the reversed order. A candidate was accepted only when both decisions mapped to the same original output; otherwise the pipeline used a fixed fallback. The selector could not rewrite or merge candidates. Reversing candidate order

addresses a known evaluation concern: positional bias has been observed in LLM judges (Zheng et al., 2023). This related finding does not establish that the present selector is bias-free. Both generation and selection received the complete Chinese subtitle sequence for the video together with title metadata. The system used no task-specific fine-tuning, external glossary, or manual post-editing.

3.4 laniqo

laniqo submitted all five language directions using Gemma-4-12B-it without fine-tuning or adapters. The team screened Qwen3.5-9B, Hy-MT2-7B, and Sailor2-20B (Dou et al., 2025) on development data before selecting Gemma-4; Hy-MT was therefore evaluated during system development but was not used in the submitted pipeline.

The pipeline generated 25 translations per cue through hedged sampling and four prompt variants. After removing empty, malformed, and wrong-language outputs, it scored candidates with reference-free xCOMET (Guerreiro et al., 2024) and MetricX-24 (Juraska et al., 2024), together with subtitle-compliance measures based on characters per second and characters per line. Non-dominated candidates formed a Pareto set. A second Gemma pass, called Chimera/FusioN, then selected, rewrote, or merged candidates using multiple cues at once. A final validation pass guaranteed non-empty output in the target language and compressed or retransmitted cues that failed hard constraints.

3.5 AuditTranslate

AuditTranslate submitted Chinese-to-English. It used Hy-MT2-7B as the only translation author and an INT4 Gemma-4-12B model as an English-

Team and base model(s)	Training and candidate selection	Context and post-processing
ReopenAI Hy-MT2-7B; Gemma-4-12B-it	<i>Adaptation:</i> No task-specific training <i>Selection:</i> Four candidates; anonymous two-order agreement selection	<i>Context:</i> Full-video Chinese subtitles and title metadata <i>Post-processing:</i> Exact cue and timestamp preservation
Ianigo Gemma-4-12B-it	<i>Adaptation:</i> No fine-tuning <i>Selection:</i> 25 candidates; Pareto reranking; Chimera/FusioN	<i>Context:</i> Multi-cue context in the fusion stage <i>Post-processing:</i> Language, completeness, CPS, and CPL checks
AuditTranslate Hy-MT2-7B; Gemma-4-12B INT4	<i>Adaptation:</i> No fine-tuning <i>Selection:</i> Best-of-3; decomposed critic; retranslation callback	<i>Context:</i> Neighboring cues and persistent name memory <i>Post-processing:</i> Name normalization, title repair, 42-character wrapping
LensLinguals Hy-MT2-7B	<i>Adaptation:</i> No fine-tuning <i>Selection:</i> Single contextual generation	<i>Context:</i> Synopsis, nearby cues, and generated name glossary <i>Post-processing:</i> Leakage, length, and SRT structure checks
rupiah Hy-MT2-7B; CometKiwi	<i>Adaptation:</i> Zero-shot translation <i>Selection:</i> Multi-temperature candidates with QE reranking	<i>Context:</i> Synopsis, preceding cues, previous translations, episode glossary <i>Post-processing:</i> Source-language leakage filtering and glossary enforcement
BVSLP Qwen3.5-9B-Instruct	<i>Adaptation:</i> Multistage 4-bit QLoRA <i>Selection:</i> Revision stage	<i>Context:</i> Subtitle blocks, synopsis, previous translations, terminology memory, selected keyframes <i>Post-processing:</i> Duration-aware length control and SRT validation
paramanandaAI Hy-MT2-1.8B	<i>Adaptation:</i> QLoRA on TVSub <i>Selection:</i> Not specified	<i>Context:</i> Not specified <i>Post-processing:</i> Not specified
SubVision Hy-MT2-1.8B	<i>Adaptation:</i> 4-bit QLoRA on approximately 60k TVSub pairs <i>Selection:</i> Beam search	<i>Context:</i> No video-level context documented <i>Post-processing:</i> Standard generation and metric-based validation

Table 2: System design summary based on submitted descriptions; SubVision is summarized from its training notebook. Development-only Hy-MT use by Ianigo is not counted as use in the submitted system. CPS: characters per second; CPL: characters per line; QE: quality estimation.

side critic and selector. Hy-MT generated three diverse candidates per cue. Gemma selected among them and evaluated the selected output using decomposed yes/no checks rather than a single monolithic rubric.

The checks targeted omissions, hallucinations, negation errors, pronoun gender, context leakage, register, sentence mood, agent direction, speaker attribution, and invented names. Candidate error flags were subjected to a skeptical re-verification step. When all candidates shared a confirmed defect, Gemma sent a targeted hint back to Hy-MT for retranslation instead of translating from Chinese itself. Deterministic passes normalized names and ordinal titles and wrapped output to at most two lines of 42 characters.

3.6 LensLinguals

LensLinguals submitted English and Thai using the released Hy-MT2-7B checkpoint without fine-tuning. Its pipeline first extracted character names from metadata and translated them into a per-video glossary. Each cue was then translated with the video title, the two preceding cues, the two following cues, and the glossary as read-only context.

The implementation used vLLM (Kwon et al., 2023) batching and prefix caching to improve throughput. Post-processing cleared outputs with substantial Chinese leakage, enforced a maximum line length, and verified that cue counts, sequence numbers, and timestamps matched the source. This system illustrates a comparatively simple approach built around a strong translation model, focused

context, and deterministic structural checks.

3.7 rupiah

rupiah submitted English and Indonesian. Hy-MT2-7B was used as a zero-shot translator, while the rest of the pipeline was model-agnostic. Each prompt incorporated the video synopsis, a window of preceding source cues, and already translated source–target pairs from the current episode. The system also built a per-episode glossary to stabilize recurring proper names.

Several temperature-diverse candidates were generated for each cue. Candidates containing source-language leakage were removed, and the remaining candidates were ranked with CometKiwi reference-free quality estimation (Rei et al., 2022). The final SRT preserved the source timing and cue segmentation. The system represents a compact combination of contextual prompting, terminology control, diverse decoding, and QE reranking.

3.8 BVSLP

BVSLP submitted all five target languages with QwenSub-MT, a Qwen3.5-9B-Instruct system² adapted through 4-bit QLoRA (Dettmers et al., 2023). Training proceeded through multilingual translation adaptation, subtitle-domain adaptation, and metadata-enriched adaptation. At inference time, adjacent cues were grouped into short blocks rather than processed independently.

The system used a video synopsis, previous translations, a dynamic terminology memory, and selected visual keyframes. A revision stage checked semantic, terminology, and formatting errors. Duration-aware length control and line-break optimization were applied before reconstruction into SRT files. BVSLP was the only submission that explicitly documented selective visual grounding.

3.9 paramanandaAI

paramanandaAI submitted English using Hy-MT2-1.8B with QLoRA (Dettmers et al., 2023) fine-tuning on a corpus identified as TVSub in the submitted description. The submitted description provided the base model, adaptation method, dataset, and a model-weight link, but did not describe additional context construction, candidate selection, or

subtitle-specific post-processing. The available evidence therefore indicates that the principal contribution was parameter-efficient domain adaptation of a compact translation model.

3.10 SubVision

SubVision submitted English. No standalone system description paper was found in the submission directory, so this summary is based on the supplied training notebook. The system used Hy-MT2-1.8B with NF4 4-bit quantization and QLoRA (Dettmers et al., 2023). Approximately 60,000 Chinese–English training pairs were sampled from TVSub (Wang et al., 2018);³ the LoRA (Hu et al., 2022) rank was 16 and training ran for three epochs.

The notebook evaluated the adapted model with BLEU (Papineni et al., 2002), sacreBLEU (Post, 2018), chrF (Popović, 2015), TER (Snover et al., 2006), and COMET (Rei et al., 2020), and used beam search for generation. It did not document video-level context, multi-candidate reranking, or dedicated subtitle-structure quality control.

4 Evaluation

This section describes automatic scoring, the LQA rubric, human verification, and aggregation.

4.1 Automatic evaluation

Because the submission set contained 2,200 translated videos, exhaustive independent human scoring of every output was impractical. We therefore used an LLM-as-a-Judge procedure as the first evaluation stage and followed it with human verification. Related approaches include GEMBA for translation quality scoring (Kocmi and Federmann, 2023b) and GEMBA-MQM for error-span detection (Kocmi and Federmann, 2023a); our judge, prompting, and aggregation settings differ from these methods. The unit of evaluation was a (team, language, video) triple.

For each unit, the source and target SRT files were parsed into cue-indexed text. The evaluator aligned source and target cues by subtitle index and retained only non-empty aligned pairs. Multi-line cue text was joined with spaces. The resulting input was a source–translation table covering the complete video.

²The official post-trained checkpoint is named Qwen/Qwen3.5-9B (Qwen Team, 2026). We retain the -Instruct label used in the submission; the exact checkpoint revision should be confirmed with the team.

³The notebook points to <https://github.com/longyuewangdcu/tvsub>. Its data attribution names <http://assrt.net> and <http://www.zimuzu.tv> as original subtitle sources.

The judge was gemini-3.1-pro-preview (Google, 2026), invoked with a language-specific LQA prompt. The prompt and aligned subtitles were concatenated into a single user message because the serving marker did not reliably follow a separate system role. Generation used a maximum of 8,192 output tokens, low reasoning effort, and temperature 0.7. Requests were retried with exponential backoff, except for non-retryable content-safety failures.

Each video was independently evaluated three times. Valid positive scores were aggregated by the median to reduce sampling variance. If all three attempts failed, the video received an automatic score of zero.

4.2 LQA rubric

The automatic evaluation used a 100-point deduction-based rubric with four dimensions. Error-based evaluation has a broader precedent in MQM (Lommel et al., 2014); however, the dimensions and the 40/35/5/20 weights here are task-specific and are not a standard MQM scoring configuration.

Scores were mapped to four broad quality bands: Excellent for 90–100, Good for 75–89, Fair for 60–74, and Poor below 60. Severe failures could also receive a Reject label.

LLMs sometimes produce inconsistent arithmetic when converting a list of deductions into a total score. To avoid this failure mode, the evaluation script recomputed each dimension and the total score from the listed defect penalties:

$$s_d = \max \left(0, M_d - \sum_{e \in E_d} |p_e| \right), \quad (1)$$

$$s_{\text{total}} = \sum_{d=1}^4 s_d. \quad (2)$$

Here M_d is the maximum for dimension d , E_d its listed defects, and p_e the penalty assigned to defect e .

This makes the numeric result auditable against the model-generated evidence.

4.3 Human verification

The human stage was designed as verification and correction rather than an independent second annotation campaign. This differs from expert-led MQM studies with full document context, such

as Freitag et al. (2021); we do not claim to reproduce that annotation protocol. Reviewers checked whether the reported errors were supported by the source and translation, whether cited cue numbers matched the actual subtitle lines, whether error categories and penalties were appropriate, and whether the total score was consistent with the corrected defect list.

When the automatic assessment was accurate, it was retained. When the assessment contained false positives, missed or duplicated penalties, cue-number mismatches, or inappropriate severity, reviewers corrected the structured assessment. The corrected `total_score` was then used as the final score. If no corrected score could be extracted, the automatic median score was retained.

The evaluation workbook contained 2,198 unique units. Human-corrected scores could be extracted for 2,052 units, corresponding to 93.36% of the evaluation units. This is corrected-score field coverage, not the proportion of fully valid JSON assessments or evidence of independent human annotation. Among these units, 314 scores changed, or 15.30%. The median score change was zero and the mean was -0.313 points, but individual corrections ranged from -55 to $+90$ points. These outliers show that aggregate stability does not eliminate the need to inspect severe judge failures.

4.4 Aggregation and ranking

Final scores were aggregated in three steps:

1. **Video score:** the human-corrected score when available; otherwise the automatic median score.
2. **Team-language score:** the arithmetic mean over all evaluated videos for that team and language.
3. **Overall score:** the macro-average over the language scores submitted by the team, with each submitted language receiving equal weight.

The macro-average prevents languages with more evaluated videos from receiving greater weight. However, it does not reward broader language coverage: a team submitting one language can rank above a multilingual system if its single-language average is higher. For this reason, Table 4 reports both the overall score and the number of submitted languages.

Dimension	Maximum	Examples of deductions
Accuracy and fidelity	40	Critical mistranslation, terminology error, omission, inconsistency, semantic deviation
Fluency and readability	35	Grammar or structural error, translationese, spelling and punctuation
Subtitle text conventions	5	Excessive line length, formatting violations
Cultural and stylistic adaptation	20	Cultural failure, cultural distance, inappropriate tone or register

Table 3: The deduction-based LQA rubric. Dimension scores are floored at zero before summation.

One duplicated Indonesian evaluation unit for rupiah was removed using the unique key (team, language, video ID). The duplicate carried the same score in both rows.

5 Results

This section reports overall and language-specific results and the effect of human verification.

5.1 Overall and language-specific ranking

Table 4 presents the final results. Each language cell contains the mean score followed by the rank within that language in parentheses. A dash indicates that the team did not submit that direction.

ReopenAI ranked first in all five submitted directions. Its macro-average exceeded lanigo by 11.092 points. lanigo placed second in Indonesian, Malay, Thai, and Traditional Chinese, but fourth in English. English had the broadest participation, with all eight teams. The top three English systems were ReopenAI, LensLinguals, and AuditTranslate.

Participation was more limited in the other directions: four teams submitted Indonesian, three submitted Malay, four submitted Thai, and three submitted Traditional Chinese. Rankings in these directions should therefore be interpreted as comparisons among smaller system sets.

5.2 Effect of human verification

Human verification had different effects across target languages. English scores were generally stable: for most teams, the difference between the automatic and final mean was within 0.3 points, although SubVision increased by 1.290 points. Traditional Chinese scores also changed little.

Indonesian showed a consistent downward correction. Final means decreased by 3.900 points for lanigo, 4.345 for rupiah, 5.330 for BVSLP, and 1.780 for ReopenAI. In contrast, Malay scores moved upward by 1.150 points for lanigo, 2.525 for

BVSLP, and 1.300 for ReopenAI. Thai corrections were smaller and mostly positive.

These language-specific directions suggest that the judge was not equally calibrated across target languages. The Indonesian assessments appear to have been comparatively lenient before human review, whereas Malay assessments were comparatively strict. Human verification did not change the overall rank order, but it materially changed absolute scores for several language-system pairs. Such changes can affect threshold-based interpretation and close language-specific comparisons even when the global leaderboard remains stable.

6 Analysis and Discussion

This section relates system design choices to performance and identifies remaining challenges.

6.1 Context is a core requirement, not an optional feature

The strongest systems explicitly modeled context. ReopenAI provided the complete video subtitle sequence to both generators and the selector. LensLinguals used a local window and metadata. rupiah combined a synopsis, preceding cues, and previously translated examples. AuditTranslate provided neighboring cues as read-only context, and lanigo processed multiple cues jointly in its final fusion step.

These approaches address a fundamental property of Chinese subtitles: subjects, referents, and interpersonal relations are frequently omitted when recoverable from discourse. A cue-level system may produce a locally plausible sentence while choosing the wrong person, gender, title, or discourse relation; related context-dependent translation problems are documented by Voita et al. (2019) and Wang et al. (2019). The results suggest that future subtitle systems should use hierarchical context: local windows for syntax and continuation,

Rank	Team	en	id	ms	th	zh-TW	Macro avg.	N_L
1	ReopenAI	90.195 (1)	87.845 (1)	82.160 (1)	89.015 (1)	97.571 (1)	89.357	5
2	laniqo	74.155 (4)	74.270 (2)	69.120 (2)	79.895 (2)	93.885 (2)	78.265	5
3	AuditTranslate	74.705 (3)	—	—	—	—	74.705	1
4	LensLinguals	75.255 (2)	—	—	72.657 (3)	—	73.956	2
5	rupiah	70.895 (5)	67.600 (3)	—	—	—	69.248	2
6	BVSLP	60.480 (6)	62.420 (4)	62.000 (3)	56.025 (4)	51.370 (3)	58.459	5
7	paramanandaAI	55.645 (7)	—	—	—	—	55.645	1
8	SubVision	49.775 (8)	—	—	—	—	49.775	1

Table 4: Final mean scores with within-language ranks in parentheses. Macro-averages weight each submitted language equally; N_L is the number of submitted languages. Higher is better. LensLinguals/Thai and ReopenAI/zh-TW each have 99 scored videos; all other submitted combinations have 100. Overall scores compare different language subsets and do not measure multilingual breadth.

together with video-level memory for characters, terminology, and narrative state.

6.2 Candidate diversity and constrained selection are effective

Four of the five highest-ranked systems used explicit candidate search or selection. ReopenAI generated candidates from two models and two seeds. laniqo used 25 candidates and multi-objective reranking. AuditTranslate used best-of-three generation with a critic and targeted retranslation. rupiah used multi-temperature generation and QE reranking.

Candidate generation reduces dependence on a single stochastic output, but only when candidates are meaningfully different and the selector is reliable. Quality-aware decoding provides a related framework for using learned evaluation signals in candidate reranking (Fernandes et al., 2022); this does not imply that the submitted systems used minimum Bayes risk decoding. Two design choices appear particularly useful. First, selection can be constrained so that it cannot invent an unseen translation, as in ReopenAI. Second, a critic can request new translations from a translation-specialized model rather than directly rewriting Chinese content, as in AuditTranslate. Both approaches reduce the risk that the quality-control stage introduces a new error.

6.3 Translation and criticism benefit from role specialization

Translation-specialized models and general instruction models are trained for different objectives. HyMT2 targets multilingual translation and translation instruction following (Zheng et al., 2026), whereas

Gemma 4 includes general instruction-following and reasoning capabilities (Gemma Team, 2026). These reports motivate, but do not experimentally establish, the relative advantage of either model for the translator or critic role in this task. AuditTranslate explicitly separated these roles, while ReopenAI combined both models for generation but restricted Gemma’s final decision to candidate selection.

The broader design hypothesis is that model composition should consider functional specialization rather than model count alone. Whether a compact translator–critic pipeline is more reliable than an unconstrained single-model pipeline requires a controlled comparison; the submitted systems do not isolate this effect.

6.4 Deterministic control is important for subtitle production

Names, titles, and terminology must remain stable across an episode. LensLinguals and rupiah generated glossaries; AuditTranslate used persistent name memory and a file-level normalization pass; BVSLP maintained a dynamic terminology memory. These components target errors that are individually small but accumulate across a video and directly affect coherence.

Likewise, cue IDs, timestamps, empty outputs, target-language leakage, line length, and SRT serialization are deterministic properties. The strongest pipelines generally delegated ambiguous semantic decisions to models and deterministic constraints to code. This division improves reproducibility and prevents a language model from damaging already correct structure.

6.5 Hy-MT is a strong base, but orchestration determines the ceiling

Six teams used Hy-MT. ReopenAI, LensLinguals, and AuditTranslate occupied the top three positions in English and all used Hy-MT2-7B. rupiah also achieved competitive results in Indonesian. These outcomes support Hy-MT2-7B as a strong compact translation base.

However, using the same model family did not guarantee high performance. SubVision and paramanandaAI used Hy-MT2-1.8B with QLoRA but ranked below systems that combined stronger base models with contextual inference and quality control. The comparison suggests that the base model determines a substantial part of translation quality, while context construction, candidate search, selection, terminology consistency, and post-processing determine the system ceiling.

6.6 Remaining challenges

Cross-cue reference resolution. Chinese frequently omits subjects and uses relationship-dependent forms of address. Local windows help but cannot always recover long-distance narrative information. Full-video context increases computational cost and may dilute attention. Structured character memory and dialogue-state tracking remain promising directions.

Cultural and register adaptation. Historical titles, kinship terms, online slang, humor, and politeness systems require target-language-specific decisions. Literal fidelity is insufficient when the target language expresses social hierarchy or gender differently. These phenomena also require language-specific evaluation guidelines.

Brevity versus semantic completeness. Subtitle output must fit time and screen constraints, but aggressive compression can remove plot-critical information. CPS, CPL, and character limits are useful operational checks, yet they should be optimized jointly with meaning preservation rather than applied as blind truncation rules; earlier subtitle MT work likewise models duration and length constraints jointly with translation and segmentation (Matusov et al., 2019).

Multilingual evaluation calibration. The systematic correction patterns for Indonesian and Malay show that one judge and one general rubric may not be calibrated uniformly across languages. Target-language experts, language-specific prompts, multiple independent judges, and explicit

calibration sets should be considered for future editions.

Judge reliability and evidence alignment. The automatic assessments sometimes contained mismatched cue numbers, duplicate deductions, incorrect severity, and inconsistent arithmetic. Repeated scoring, code-based score recomputation, and human correction reduced these errors. Future evaluations could add judge agreement statistics, blinded duplicate review, and targeted comparison with reference-based metrics.

7 Conclusions

The WMT26 Video Subtitle Translation shared task evaluated eight teams across five directions from Simplified Chinese. The evaluation combined repeated LLM-based LQA assessment with human verification, producing final scores for 2,198 unique team–language–video units. ReopenAI achieved the highest macro-average and ranked first in all five language directions, while lanigo ranked second overall.

The system comparison highlights several promising design patterns associated with stronger results: context-aware prompting, diverse candidate generation, constrained candidate selection, translator–critic role separation, terminology and character memory, and deterministic subtitle validation. These observational comparisons motivate studying fine-tuning together with inference-time context and quality control, but do not isolate their contributions from differences in base models or model sizes.

The evaluation process was broadly stable at the aggregate level, but human reviewers changed 15.30% of the verified scores and exposed language-specific calibration effects. Human verification therefore remains important when LLM-as-a-Judge scores are used for fine-grained system comparison. Future task iterations should strengthen target-language calibration, collect independent human judgments on selected difficult cases, and release structured error annotations to support research on discourse, style, terminology, and subtitle-specific constraints.

Limitations

This analysis has several limitations. First, the automatic evaluation was reference-free and therefore depended on the capabilities and biases of a single judge model and its prompts. Second, the human

stage verified the automatic assessments rather than independently re-annotating every translation; it may therefore inherit the framing of the automatic rubric. Third, 146 unique units did not contain an extractable human-corrected score and retained their automatic scores.

Fourth, language coverage was uneven. Some overall scores were based on one language, whereas others represented five-language systems. Macro-averaging treats submitted languages equally but does not measure breadth. Fifth, system descriptions varied in detail. SubVision did not provide a standalone system paper in the available directory, so its description was reconstructed from its training notebook. Finally, the current analysis reports means and rank order but does not include confidence intervals or paired significance tests. Bootstrap resampling is an established approach to MT evaluation (Koehn, 2004). Future work should apply paired resampling or non-parametric tests when sufficiently aligned human judgments are available, following the statistical ranking practice used in WMT findings reports (Kocmi et al., 2025).

Ethical Considerations

The shared-task data and outputs may contain fictional violence, insults, sensitive dialogue, or culturally specific language. Automatic content filters can fail on such material and may introduce systematic missingness. Evaluators should be informed about potentially sensitive content and should not be required to review harmful material without appropriate safeguards.

The evaluation records should avoid storing personal information about reviewers. System outputs and judgments should be released only under the applicable data and task licenses. LLM-based assessment can encode language and cultural biases. Separately, presentation-related biases have been documented in LLM judges (Zheng et al., 2023). Consequently, final interpretations should not treat automatic scores as an infallible measure of linguistic quality, particularly for lower-resource languages and culture-specific expressions.

References

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [QLoRA: Efficient fine-tuning of quantized LLMs](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Longxu Dou, Qian Liu, Fan Zhou, Changyu Chen, Zili Wang, Ziqi Jin, Zichen Liu, Tongyao Zhu, Cunxiao Du, Penghui Yang, Haonan Wang, Jiaheng Liu, Yongchi Zhao, Xiachong Feng, Xin Mao, Man Tsung Yeung, Kunat Pipatanakul, Fajri Koto, Min Si Thu, and 22 others. 2025. [Sailor2: Sailing in south-east asia with inclusive multilingual LLMs](#). *arXiv preprint arXiv:2502.12982*.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre 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*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. [Experts, errors, and context: A large-scale study of human evaluation for machine translation](#). *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Gemma Team. 2026. [Gemma 4 technical report](#). *arXiv preprint arXiv:2607.02770*.

Google. 2026. [Gemini 3.1 Pro Preview](#). Gemini API documentation. Accessed September 11, 2026.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. [xCOMET: Transparent machine translation evaluation through fine-grained error detection](#). *Transactions of the Association for Computational Linguistics*, 12:979–995.

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 *International Conference on Learning Representations*.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. [MetricX-24: The Google submission to the WMT 2024 metrics shared task](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Drach, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakouga, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. [Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.

- Tom Kocmi and Christian Federmann. 2023a. [GEMBA-MQM: Detecting translation quality error spans with GPT-4](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.
- Tom Kocmi and Christian Federmann. 2023b. [Large language models are state-of-the-art evaluators of translation quality](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.
- Philipp Koehn. 2004. [Statistical significance tests for machine translation evaluation](#). In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph 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, SOSP ’23*, pages 611–626. ACM.
- Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. [Multidimensional quality metrics \(MQM\): A framework for declaring and describing translation quality metrics](#). *Tradumática*, pages 455–463. Issue 12.
- 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*, 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: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- Qwen Team. 2026. [Qwen3.5: Towards native multi-modal agents](#). Official project blog.
- 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 (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.
- 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.
- Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. [A study of translation edit rate with targeted human annotation](#). In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.
- Elena Voita, Rico Sennrich, and Ivan Titov. 2019. [When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. 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](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).
- Longyue Wang, Zhaopeng Tu, Xing Wang, and Shuming Shi. 2019. [One model to learn both: Zero pronoun prediction and translation](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 921–930, Hong Kong, China. Association for Computational Linguistics.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-Bench and Chatbot Arena](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.
- 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*.