

QwenSub-MT Framework for WMT 2026 Video Subtitle Translation

Palak Arora^{1,2}, Mansi Jangid¹, Bharti Nathani¹, Nisheeth Joshi¹

¹Speech and Language Processing Lab, Centre for Artificial Intelligence, Banasthali Vidyapith, Rajasthan, India

²School of Computing, DIT University, Dehradun, Uttarakhand, India

palak.arora.pa55@gmail.com, mansijangid@banasthali.in, nbharti@banasthali.in, nisheeth.joshi@rediffmail.com

Abstract

The paper introduces QwenSub-MT, a system that is context-aware and constraint-controlled and has been developed for the WMT 2026 Video Subtitle Translation shared task. It carried out the translation of Simplified Chinese subtitles into 6 different languages. QwenSub-MT used Qwen3.5-Instruct together with a multistage 4-bit QLoRA adaptation and treated adjacent subtitle cues as contextual blocks. In order to enhance consistency and disambiguation, the system incorporated video synopsis information, previous translations, terminology memory and selective visual grounding. A revision stage was carried out to correct semantic, terminology and formatting errors, while duration-aware length control, line-break optimization and deterministic SRT validation were used to make sure that the outputs were readable and structurally valid. 500 subtitle files were submitted in five language directions and obtained a macro-average score of 58.459, placing sixth in the overall ranking. The results show that contextual modelling and structural control are useful, but candidate diversity, independent quality estimation and stronger adaptation to the target language still needs improvement.

1 Introduction

The rapid expansion of online video, streaming platforms, digital education and social-media content has increased the demand for high-quality multilingual subtitles. Subtitles support accessibility, international distribution and cross-cultural communication, but video subtitle translation is more demanding than conventional sentence-level machine translation. A successful subtitle must preserve meaning while remaining synchronized with the video, respecting limited

display time and maintaining readability. It must also retain dialogue style, emotional tone, speaker intention and should maintain narration continuity. Therefore, even a linguistically correct translation may be unsuitable when it is too long, poorly segmented, inconsistent with surrounding dialogue or disconnected from the visual scene. Subtitle translation is further constrained by temporal and spatial limitations. Subtitle cues are short, time-aligned units that do not always correspond to complete grammatical sentences. A single sentence may span several cues, while individual segments may contain ellipsis, pronouns, discourse markers or short responses whose interpretation depends on contextual information. Consequently, subtitle quality involves not only translation accuracy but also segmentation, timing and readability. This has motivated subtitle-specific evaluation measures such as SubER, which jointly evaluates textual errors, subtitle segmentation and timing rather than considering translated text alone (Wilken et al., 2022).

Context is particularly important in Chinese subtitle translation because subjects and other sentence components are frequently omitted when they can be inferred from discourse. Pronouns, forms of address, character relationships, lexical choices and politeness levels may therefore remain ambiguous when cues are translated independently. Context-insensitive systems can produce translations that appear acceptable in isolation but become inconsistent within the larger dialogue. Voita et al. (2019) identified deixis, ellipsis and lexical cohesion as major sources of contextual errors, while Vincent et al. (2024) showed that extra-textual information such as film metadata can reduce context-related errors during professional subtitle post-editing.

Video subtitle translation is also inherently multimodal. Text alone may not clearly indicate speaker identity, referents, objects, emotions or actions. Visual information can provide evidence

about characters, locations, objects and scene context, thereby resolving ambiguities. However, irrelevant or temporally misaligned frames may introduce noise. This motivates selective visual grounding, where visual evidence is used primarily for subtitle segments containing contextual or referential ambiguity.

The WMT 2026 Video Subtitle Translation task requires translation of Simplified Chinese subtitles into English, Thai, Indonesian, Malay and Traditional Chinese, while considering contextual ambiguity, timing, readability, terminology consistency and SRT preservation. To address these challenges, we propose QwenSub-MT, a context-aware and multimodal framework based on Qwen3.5-Instruct. The system performs block-level subtitle translation using neighbouring dialogue, synopsis information, previous translations, dynamic terminology memory and selective visual keyframes. Duration and readability constraints are incorporated during generation, followed by revision and SRT validation. Multistage QLoRA fine-tuning enables multilingual, subtitle-domain and metadata-enriched adaptation within a unified framework.

2 Literature Review

Machine translation for subtitles has evolved from rule-based and statistical approaches to neural, contextual, multimodal and large-language-model-based systems. Early work by Turcato et al. (2000) showed that subtitle translation requires careful preprocessing because closed captions contain colloquial expressions, incomplete sentences, irregular segmentation and named entities. Volk et al. (2010) demonstrated the practical use of statistical MT for television subtitles and emphasized domain-specific training and professional post-editing. With neural MT, Matusov et al. (2019) integrated domain adaptation, inter-sentence context and subtitle segmentation, showing that translation and presentation constraints should be addressed jointly.

Research has also focused on subtitle-specific corpora and production constraints. Karakanta et al. (2020a) introduced MuST-Cinema, which preserves subtitle boundaries and supports joint modelling of translation, punctuation and segmentation. Karakanta et al. (2020b) further demonstrated that subtitle quality cannot be determined by character limits alone, since timing,

pauses and semantic structure influence segmentation. From a professional perspective, Koponen et al. (2020) found that translators evaluate MT subtitles not only for linguistic accuracy but also for timing, segmentation, contextual consistency and editing effort. For evaluation, Karakanta et al. (2022) proposed Sigma for measuring subtitle segmentation, while Wilken et al. (2022) introduced SubER, which jointly considers translation errors, segmentation and timing.

Context-aware translation has gained importance in subtitle translation, as dialogue often relies on pronouns, ellipsis, and discourse-level information whose interpretation depends on surrounding context. Wang et al. (2023) showed that large language models can exploit document-level context to improve translation coherence, although excessive context may introduce irrelevant information. Karpinska and Iyyer (2023) similarly reported that paragraph-level LLM translation improved grammaticality and consistency but could still produce omissions. Cui et al. (2024) demonstrated that selectively retrieving and summarizing relevant context can improve document-level translation, supporting the use of compressed and focused context rather than unrestricted document input.

Multimodal research has investigated whether visual information can resolve textual ambiguity. Caglayan et al. (2019) found that images are particularly useful when textual information is incomplete, whereas Wu et al. (2021) cautioned that improvements under complete-text conditions may sometimes reflect regularization rather than genuine visual grounding. Ji et al. (2022) proposed information-theoretic mechanisms to increase visual awareness. Kang et al. (2023) introduced BigVideo, containing approximately 4.5 million subtitle pairs and 9,981 hours of video, while Li et al. (2023) developed the EVA dataset to evaluate genuinely video-helpful translation cases. More recent approaches employ selective visual processing: Liu et al. (2025) proposed DeDiT, which activates visual reasoning only for ambiguous inputs, and Guan et al. (2025) introduced SHIFT, which selects useful video frames rather than uniformly processing all frames.

Large language models further enable metadata-guided subtitle translation. Pramodya et al. (2025) showed that movie titles, summaries and genre information can improve subtitle translation with

GPT and LLaMA models. The Yang et al. (2025) introduced the multilingual Qwen3 family with thinking and non-thinking modes, providing a suitable foundation for controlled instruction-based translation.

3 Methodology

This section describes the proposed QwenSub-MT framework developed for the WMT 2026 Video Subtitle Translation task. It translated Simplified Chinese subtitles into English, Thai, Indonesian, Malay, and Traditional Chinese for Taiwan. Unlike conventional sentence-level MT, the system was designed to preserve meaning while also satisfying subtitle-specific requirements.

The complete workflow consisted of SRT parsing, subtitle-block construction, metadata processing, Qwen3.5-based translation, revision, length control, line-break optimization, SRT reconstruction, and final validation. A single Qwen3.5-Instruct model was used for translation, terminology processing, and revision, thereby reducing system complexity and keeping the overall model size within the shared-task restriction.

Hyperparameter/Setting	Initial Configuration
Quantisation method	NF4, 4-bit
Compute datatype	BF16
LoRA rank	32
LoRA alpha	64
LoRA dropout	0.05
Per-device batch size	1
Gradient accumulation steps	16
Effective batch size	16
Learning rate	$1/10^{-4}$
Learning-rate scheduler	Cosine
Gradient checkpointing	Enabled
Training epochs	2–3 per training stage

Table 1: Hyperparameters Used

3.1 Model Selection and Adaptation

The Qwen model operated in non-thinking instruction mode. Although it supports very long context windows, training sequences were restricted to approximately 2,048–4,096 tokens to make fine-tuning practical on a 24 GB GPU. Thinking mode was disabled to reduce unnecessary generation, avoid reasoning traces in subtitle output, and ensure deterministic structured responses. Qwen3.5-4B was retained as a fallback

configuration if the 9B model exceeds available memory during fine-tuning.

The model was adapted through 4-bit QLoRA. The initial configuration used NF4 quantisation. The parameters of the model are shown in table 1. LoRA adapters are applied mainly to the language-model layers. The data for training the model was provided by the organizers of the shared task.

3.2 SRT Processing

Each SRT file was parsed into subtitle events containing sequence number, start time, end time, duration, text, and formatting information. Unicode is normalized, punctuation was standardized. Music markers, speaker labels, numbers, dates, and special symbols were preserved.

Instead of translating every cue independently, adjacent subtitle events were grouped into short translation blocks. A new block was started when sentence-final punctuation was detected, a substantial time gap occurred, the speaker changed, the combined source became too long, or the block reached a maximum of four cues. Qwen translated all cues in the block jointly while returning one translation for each original subtitle identifier. This improved handling of sentences that extended across multiple SRT events while preserving the original subtitle structure.

3.3 Metadata, Context and Terminology Memory

Before translation, the available video synopsis and metadata were compressed into a compact context record describing the domain, setting, tone, characters, and terminology. A dynamic video-level glossary stored character names, organizations, locations, forms of address, technical expressions, and recurrent phrases. After each translated block, new terminology was added to the memory and later translations were checked for consistency.

For each translation block, the model received a package containing three preceding subtitle events, their accepted target translations, the current block, two following source events, the compressed synopsis, relevant glossary entries, and duration or length constraints. Previous target translations helped maintain consistent spelling, pronoun use, terminology, and dialogue style.

3.4 Multistage Finetuning

Fine-tuning was performed in three stages. Stage 1 strengthened multilingual translation using general parallel corpora across all target languages. Stage 2 performed subtitle-domain adaptation using dialogue, slang, short utterances, speaker changes, duration information, and segment-level structured outputs. Stage 3 introduced metadata-enriched and multimodal examples containing synopsis information, character glossaries, previous translations, and subtitle-length constraints.

This progressive training strategy preserved the multilingual knowledge of the pretrained model while specializing it for contextual and constrained subtitle generation.

3.5 SRT Reconstruction

After translation, each target segment was mapped back to its original SRT event. Sequence numbers, timestamps, event order, formatting tags, and encoding were preserved. A deterministic validator checked cue counts, timestamp syntax, empty segments, language correctness.

4 Evaluation and Discussion

The QwenSub-MT system that has been proposed was tested as part of the WMT 2026 Video Subtitle Translation shared task, this task involving the translation from Simplified Chinese into 5 languages. BVSLP provided translations for all five destination languages, including 100 SRT files for each translation direction. Thus, 500 subtitle files in total were submitted. BVSLP was one of only three teams, along with ReopenAI and lanigo, to take part in all five language directions, which meant that the system could be evaluated in a wide multilingual environment rather than just for one target language.

A human-verified LLM-as-a-Judge evaluation method was used in the shared task. For each combination involving a team, a language, and a video, the source and translated SRT files were analyzed and aligned on the basis of the subtitle numbers. The multi-line subtitle entries were combined together and the resulting source-translation pairs were then evaluated at the level of the entire video. Each video was assessed three times using a Language Quality Assessment prompt that was specific to the language in question, and the median of the valid scores was

taken in order to reduce the variability among the automatic judgments. The automatic evaluations were then examined by human assessors, who verified that the errors identified were actually present in both the source and the translation, that the cue numbers were correct, and that the penalties and severity labels were suitable. When needed, the human reviewers made the necessary corrections to the structured assessments and to the final scores.

The assessment was carried out using a rubric based on a 100-point scale divided into four different areas. In the case of Accuracy and Fidelity, 40 points were possible and this area looked at mistranslation, omissions, errors regarding terminology, inconsistencies, and semantic deviation. For Fluency and Readability, 35 points were allocated and took into account grammaticality, structural problems, translationese, spelling, and punctuation. The aspect of Subtitle Text Conventions was worth 5 points and dealt with matters such as excessively long lines and violations of formatting rules. Lastly, Cultural and Stylistic Adaptation accounted for 20 points and examined cultural appropriateness, tone, register, and stylistic distance. Scores between 90 and 100 were considered to be Excellent, those between 75 and 89 were classified as Good, scores between 60 and 74 were regarded as Fair, and anything below 60 was deemed Poor.

4.1 BVSLP Performance

The final language-wise results for BVSLP are presented table 2.

Target Language	Final Score	Rank	Quality Band
English	60.48	6th	Fair
Indonesian	62.42	4th	Fair
Malay	62	3rd	Fair
Thai	56.025	4th	Poor
Traditional Chinese (Taiwan)	51.37	3rd	Poor
Macro-average	58.459	6th overall	Poor

Table 2: BVSLP Results

The system achieved a total macro-average score of 58.459 and thus came in sixth place out of the eight teams that were assessed. The best score was achieved for Indonesian, with a score of 62.420, then Malay at 62.000 and English at 60.480. These three language versions were in the

360 Fair quality range of the shared-task rubric. Thai 410 adaptation using enriched metadata. At inference
361 obtained a score of 56.025, and Traditional Chinese 411 time, adjacent subtitle cues were grouped into short
362 for Taiwan had the lowest BVSLP score, which 412 translation blocks rather than being treated
363 was 51.370, with both of them falling into the Poor 413 separately. The framework included several factors
364 band. 414 viz a dynamic terminology memory, line-break

365 Although it should be considered in light of the 415 optimization, deterministic SRT reconstruction etc.
366 number of systems that took part, English had the 416 BVSLP was the only entry that was explicitly
367 widest participation since all eight teams submitted 417 documented as making use of selective visual
368 results, whereas both Indonesian and Thai each had 418 grounding.

369 four participating systems and Malay and 419 The design decisions made, took into account a
370 Traditional Chinese each had only three. It is 420 number of important aspects of subtitle translation,
371 therefore the case that BVSLP's third-place 421 including cross-cue dependencies, the consistency
372 position for Malay and Traditional Chinese is based 422 of characters and terminology, the preservation of
373 on a much narrower competitive field than that of 423 structure, and readability. However, the results of
374 English. The macro-average assigns equal weight 424 the shared task indicate that these features on their
375 to each of the languages submitted and so shows 425 own were not sufficient to achieve the level of
376 multilingual breadth, but it does not directly reward 426 performance attained by the best-ranked systems.

378 4.2 Effect of Human Verification

379 Human verification had a clear impact on the 430 A major outcome of the shared task was the
380 final BVSLP scores, especially in the cases of 431 importance of having a varied range of candidate
381 Indonesian and Malay. For Indonesian, the final 432 translations and of restricting the number of
382 score was reduced by 5.330 points as compared to 433 candidate translations. Of the five systems that
383 the automatic evaluation, while the Malay score 434 performed best, four produced more than one
384 increased by 2.525 points after human verification. 435 candidate translation and used some form of
385 English showed only a slight change, Thai stayed 436 selection or reranking procedure. ReopenAI
386 the same, and Traditional Chinese was adjusted 437 developed four candidates and applied an
387 only slightly downwards. These trends were not 438 agreement-based selection method. Ianiqo
388 specific to BVSLP; the shared-task results also 439 generated 25 candidates and then carried out Pareto
389 showed a general tendency that the automatic 440 reranking before proceeding to the fusion stage.
390 assessments of Indonesian had been relatively 441 AuditTranslate employed best-of-three generation
391 lenient and those of Malay had been relatively strict 442 together with a critic and targeted retranslation, and
392 before human correction. 443 rupiah generated the candidates using temperature

393 The fact shown here highlights the need for 444 diversity and selected the outputs by reference-free
394 human verification in the evaluation of 445 quality estimation.
395 multilingual text. The automatic judge was not 446 On the other hand, QwenSub-MT consisted of
396 equally calibrated for all the target languages, and 447 producing one translation with Qwen3.5 followed
397 the corrections had an effect on the absolute scores 448 by a revision stage. Although this revision process
398 even though the general rank order stayed the 449 enabled the identification of semantic,
399 same. In the case of BVSLP, it therefore follows 450 terminology, and formatting problems, the
400 that the final results should be regarded as human- 451 approach still relied on the quality of the initial
401 verified estimates rather than as direct outputs from 452 translation. The results obtained in the shared task
402 the automatic judge alone. 453 therefore suggest that selecting among a number of
454 alternative translations rather than repeatedly
455 editing a single output could provide a more
456 effective way of ensuring quality.

403 4.3 Analysis of the QwenSub-MT System

404 BVSLP was very different from most of the 456 Another important point is concerned with
405 other systems in terms of its architectural design. It 457 functional specialization. In several instances,
406 employed Qwen3.5-Instruct together with a 458 more powerful systems have split off the task of
407 multistage 4-bit QLoRA adaptation. The training 459 translation from that of criticism or selection.
408 process involved adaptation for multilingual 460 Specialized models for translation have been used
409 translation, adaptation for the subtitle domain, and 461 to generate the outputs, while general-purpose
462 instruction models or quality-estimation tools have
463 been used to carry out the evaluation or selection of

the candidates. This kind of division reduced the likelihood that a single model would make an error when performing the translation and then reinforce that error during self-revision. In the case of BVSLP, employing Qwen3.5 for both translation and revision simplified the architecture, but it also have decreased the independence of the quality-control phase.

Yet the results do show that QwenSub-MT was capable of generating multilingual and structurally valid subtitle outputs in all five WMT26 directions. The incorporation of subtitle blocks, terminology memory, metadata, previous translations, and deterministic SRT validation is in keeping with the general finding of the shared task that context and structural control are essential for the translation of video subtitles.

The scores achieved by BVSLP indicate that semantic fidelity and target-language adaptation remain the main limitations; the worse performance in the cases of Thai and Traditional Chinese also suggests that multilingual fine-tuning alone may not be sufficient to obtain adequate language-specific adaptation, especially when issues such as register, cultural conventions, lexical preferences, or localization decisions are taken into account.

5 Conclusion

The paper introduced QwenSub-MT, a framework that is context-aware and constraint-controlled and was developed for the WMT 2026 Video Subtitle Translation shared task. The system carried out the translation of Simplified Chinese subtitles into 6 different languages by using Qwen3.5-Instruct which had been adapted through a multistage 4-bit QLoRA approach. This framework incorporated subtitle-block translation, video metadata, previous translations, terminology memory, revision, line-break optimization, and deterministic SRT reconstruction in order to meet both the linguistic and structural demands of subtitle translation.

A total of 500 SRT files for all five of the target languages and secured a macro-average score of 58.459, placing sixth overall. The best results were achieved in Indonesian, then in Malay and English, whereas Thai and Traditional Chinese proved to be more difficult. Additionally, human verification revealed that the automatic evaluation had not been equally calibrated for all languages, especially for Indonesian and Malay, thus highlighting the need

for human review in the assessment of multilingual subtitles.

The results from the shared task also showed that although contextual information and structural control were helpful they were not on their own enough to achieve the best performance. Systems which used multiple candidate generations, reranking, independent criticism, and had a specialization of translator and critic generally performed better. This indicates that future versions of QwenSub-MT should include diverse candidate generation and independent quality estimation methods rather than simply relying on single-output revision. Altogether, the study shows that it is feasible to use a unified Qwen-based approach for multilingual subtitle translation and identifies semantic fidelity, language-specific adaptation, and candidate selection as the key areas needing further improvement.

References

- Caglayan, O., Madhyastha, P., Specia, L., & Barrault, L. (2019). Probing the need for visual context in multimodal machine translation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (pp. 4159–4170). Association for Computational Linguistics. doi:10.18653/v1/N19-1422.
- Cui, M., Du, J., Zhu, S., & Xiong, D. (2024). Efficiently exploring large language models for document-level machine translation with in-context learning. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 10885–10897). Association for Computational Linguistics. doi:10.18653/v1/2024.findings-acl.646.
- Guan, B., Han, C., Zhang, Y., Liang, Y., Zhang, Z., Zhao, Y., & Zong, C. (2025). SHIFT: Selected helpful informative frame for video-guided machine translation. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (pp. 3249–3267). Association for Computational Linguistics. doi:10.18653/v1/2025.emnlp-main.161.
- Ji, B., Zhang, T., Zou, Y., Hu, B., & Shen, S. (2022). Increasing visual awareness in multimodal neural machine translation from an information theoretic perspective. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 6755–6764). Association for Computational Linguistics. doi:10.18653/v1/2022.emnlp-main.453.
- Kang, L., Huang, L., Peng, N., Zhu, P., Sun, Z., Cheng, S., Wang, M., Huang, D., & Su, J. (2023). BigVideo:

- A large-scale video subtitle translation dataset for multimodal machine translation. In Findings of the Association for Computational Linguistics: ACL 2023 (pp. 8456–8473). Association for Computational Linguistics. doi:10.18653/v1/2023.findings-acl.535.
- Karakanta, A., Negri, M., & Turchi, M. (2020a). MuST-Cinema: A speech-to-subtitles corpus. In Proceedings of the Twelfth Language Resources and Evaluation Conference (pp. 3727–3734). European Language Resources Association.
- Karakanta, A., Negri, M., & Turchi, M. (2020b). Is the answer to everything in subtitling-oriented speech translation? In Proceedings of the 17th International Conference on Spoken Language Translation (pp. 209–219). Association for Computational Linguistics. doi:10.18653/v1/2020.iwslt-1.26.
- Karakanta, A., Buet, F., Cettolo, M., & Yvon, F. (2022). Evaluating subtitle segmentation for end-to-end generation systems. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 3069–3078). European Language Resources Association.
- Karpinska, M., & Iyyer, M. (2023). Large language models effectively leverage document-level context for literary translation, but critical errors persist. In Proceedings of the Eighth Conference on Machine Translation (pp. 419–451). Association for Computational Linguistics. doi:10.18653/v1/2023.wmt-1.41.
- Koponen, M., Sulubacak, U., Vitikainen, K., & Tiedemann, J. (2020). MT for subtitling: Investigating professional translators’ user experience and feedback. In Proceedings of the 1st Workshop on Post-Editing in Modern-Day Translation (pp. 79–92). Association for Machine Translation in the Americas.
- Li, Y., Shimizu, S., Chu, C., Kurohashi, S., & Li, W. (2023). Video-helpful multimodal machine translation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 4281–4299). Association for Computational Linguistics. doi:10.18653/v1/2023.emnlp-main.260.
- Liu, D., Kong, F., Sun, X., Patil, D., Vajpayee, A., Liu, Z., Bhat, V., & Sadoughi, N. (2025). Detect, disambiguate, and translate: On-demand visual reasoning for multimodal machine translation with large vision-language models. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (pp. 1559–1570). Association for Computational Linguistics. doi:10.18653/v1/2025.naacl-long.74.
- Matusov, E., Wilken, P., & Georgakopoulou, Y. (2019). Customizing neural machine translation for subtitling. In Proceedings of the Fourth Conference on Machine Translation, Volume 1: Research Papers (pp. 82–93). Association for Computational Linguistics. doi:10.18653/v1/W19-5209.
- Pramodya, A., Sakai, Y., Vasselli, J., Kamigaito, H., & Watanabe, T. (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 (pp. 315–330). Association for Computational Linguistics. doi:10.18653/v1/2025.acl-srw.22.
- Turcato, D., Popowich, F., McFetridge, P., Nicholson, D., & Toole, J. (2000). Pre-processing closed captions for machine translation. In ANLP-NAACL 2000 Workshop: Embedded Machine Translation Systems.
- Vincent, S., Prescott, C., Bayliss, C., Oakley, C., & Scarton, C. (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 (pp. 561–572). European Association for Machine Translation.
- Voita, E., Sennrich, R., & Titov, I. (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 (pp. 1198–1212). Association for Computational Linguistics. doi:10.18653/v1/P19-1116.
- Volk, M., Sennrich, R., Hardmeier, C., & Tidström, F. (2010). Machine translation of TV subtitles for large scale production. In Proceedings of the Second Joint EM+/CNGL Workshop: Bringing MT to the User: Research on Integrating MT in the Translation Industry (pp. 53–62). Association for Machine Translation in the Americas.
- Wang, L., Lyu, C., Ji, T., Zhang, Z., Yu, D., Shi, S., & Tu, Z. (2023). Document-level machine translation with large language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 16646–16661). Association for Computational Linguistics. doi:10.18653/v1/2023.emnlp-main.1036.
- Wilken, P., Georgakopoulou, P., & Matusov, E. (2022). SubER: A metric for automatic evaluation of subtitle quality. In Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022) (pp. 1–10). Association for Computational Linguistics. doi:10.18653/v1/2022.iwslt-1.1.

677 Wu, Z., Kong, L., Bi, W., Li, X., & Kao, B. (2021). 687 Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng,
678 Good for misconceived reasons: An empirical 688 B., Yu, B., Gao, C., Huang, C., Lv, C., et al. (2025).
679 revisiting on the need for visual context in 689 Qwen3 technical report. arXiv preprint
680 multimodal machine translation. In Proceedings of 690 arXiv:2505.09388.
681 the 59th Annual Meeting of the Association for 691 WMT26 Video Subtitle Translation Shared Task.
682 Computational Linguistics and the 11th 692 (2026). Preliminary findings of the WMT26 Video
683 International Joint Conference on Natural Language 693 Subtitle Translation shared task: Context and
684 Processing, Volume 1 (pp. 6153–6166). Association 694 candidate selection matter. Shared-task overview
685 for Computational Linguistics. 695 and findings paper, August 2026.
686 doi:10.18653/v1/2021.acl-long.480.

696