

SCR-Eval: Evaluating Video Subtitle Translation with a Script-like Context Representation

Jongin Lee Juyi Hwang Sehyun Kim

Samsung Electronics

{jongin.g.lee, juyi.hwang, sehyun216.kim}@samsung.com

Abstract

The quality of video subtitle translation cannot be assessed from sentence-level semantic equivalence alone. Subtitles are governed by audiovisual cues, speaker relationships, and narrative context that accumulates across scenes, yet existing automatic metrics and text-centric LLM-based evaluation do not reflect these factors.

We propose SCR-Eval, a context-aware evaluation framework for video subtitle translation, built on a Script-like Context Representation (SCR) reconstructed from the video. The evaluator aligns the translated subtitles to the SCR’s semantic scenes, then performs MQM error analysis with an LLM judge.

On three open movies, we benchmark five English-to-Korean subtitle providers against blind, randomized annotations by seven professional evaluators. The framework’s provider ranking agrees strongly with the experts (system-level Spearman $\rho = 0.967$). In ablations, agreement weakens stepwise from full context through scene-bounded text to isolated source–translation pairs. Automatic baselines agree less: MetricX and xCOMET correlate negatively with the expert ranking, while RUBRIC-MQM correlates only weakly.

With full context, SCR-Eval shows substantially stronger agreement with experts than these baselines.

1 Introduction

Assessing the quality of video subtitle translation is not a sentence-level problem. Adequacy is a matter of judgment, shaped by the scene’s visual setting, the speaker–addressee relationship, the manner of delivery, and the narrative accumulated from earlier scenes. Constrained by display time and per-line character limits, a translated subtitle condenses, omits, and reformulates the source rather than transferring it verbatim, so no single rendering is uniquely correct, and overlap with one reference cannot capture quality.

Existing automatic metrics (BLEU (Papineni et al., 2002), chrF (Popović, 2015), the COMET family (Rei et al., 2020)) and short-context LLM judges (Kocmi and Federmann, 2023a; Kim, 2025) do not explicitly model scene context or speaker relations, limiting their ability to reflect constraints specific to subtitling (Díaz Cintas and Remael, 2007). Document-level evaluation (Voita et al., 2019; Jiang et al., 2022) likewise stops short of the audiovisual and cumulative narrative context of video. Recent paragraph-level LLM evaluators use broader textual context (Fernandes et al., 2025; Zhang et al., 2025), but, as published, neither returns per-source-line judgments that can be compared directly with our expert annotations (Appendix F). Expert MQM evaluation (Lommel et al., 2014; Freitag et al., 2021; Kocmi et al., 2024) is fine-grained but time- and cost-intensive to repeat across titles and providers¹. This gap calls for automatic evaluation that is both context-aware and quantitatively analyzable—a need especially acute in real-world settings that translate and quality-check VoD subtitles at scale across languages (§5.1).

We propose an automatic evaluation framework for video subtitle translation centered on a Script-like Context Representation (SCR) that structures the video’s non-verbal audiovisual context beyond the subtitle text. Using the SCR, we align translated cues to the source semantic scope to produce Scope-Aligned Translated Subtitles (SATS, §2.2). A Translated Subtitle Evaluation Report (TSER, §2.3) then anchors MQM errors to source lines and translated cues, enabling system-, scene-, and line-level analysis (§2).

On three open movies and five providers,

¹*Provider* denotes the source that produced each candidate subtitle—a human translation, LLM configurations, a commercial solution, and generic NMT—i.e., the entities we compare; *system* is reserved for the aggregation level (system/scene/line), not for an entity.

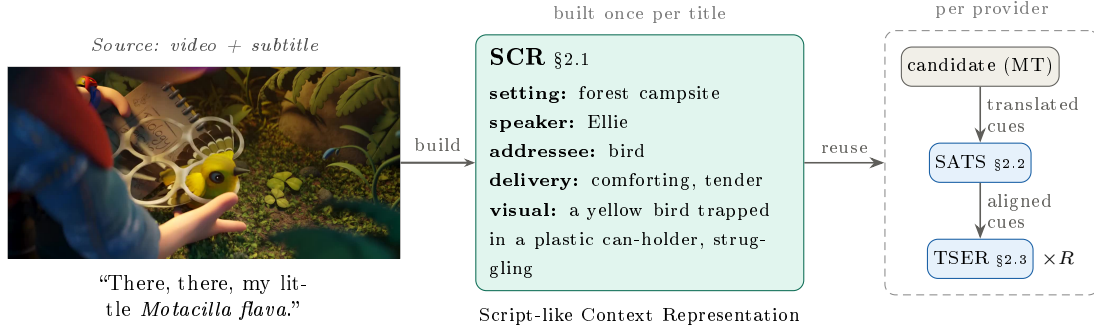


Figure 1: SCR turns raw video and source subtitles into reusable, structured text fields—scene, speaker, delivery, and on-screen visual (selected; full schema in Appendix A)—built once per title. The reused SCR then drives a lightweight per-provider evaluation: SATS aligns each candidate’s cues to the source, and TSER produces line-anchored error reports—repeated R times and aggregated into system/scene/line scores.

we repeat the LLM evaluation $R=10$ times per movie–provider pair to reduce variance. Seven professional subtitle evaluators, blind to provider identity, independently score the candidate translations in randomized order, assigning per-line category and severity. Using the simple mean of the seven as the expert reference, the proposed framework reaches a strong system-level correlation with expert judgment (Spearman $\rho = 0.967$, Kendall $\tau = 0.933$), whereas two learned QE metrics built for sentence-level MT (MetricX, xCOMET) correlate negatively (-0.60 , -0.50) and a published LLM-as-judge MQM metric (RUBRIC-MQM) reaches only 0.267 . We further analyze the stability–cost trade-off across repetition counts and use context ablations to examine how input context contributes to expert agreement.

Our contributions are fourfold:

1. We propose a context-aware automatic evaluation framework for video subtitle translation, centered on the SCR and producing anchored MQM-style reports (TSER) that enable system-, scene-, line-, and category-level analysis (§2).
2. To our knowledge, no public benchmark or protocol exists for context-aware subtitle-translation evaluation. We therefore construct an expert benchmark annotated by seven professional audiovisual translators under a blind, randomized protocol.
3. Using this benchmark, we show that the framework agrees with experts at the system, scene, and line levels. Ablations further show agreement increasing stepwise from line-level evaluation to scene-bounded text and then to the full SCR-Eval context.

4. We find that line-level automatic metrics—two learned QE metrics (MetricX, xCOMET) and an LLM-as-judge metric (RUBRIC-MQM)—do not track the expert ranking, and we analyze the stability–cost trade-off of the repetition count.

2 SCR-Based Evaluation Framework

SCR-Eval first fixes the source video and subtitles into a reviewable context representation (SCR, §2.1), built once per title and reused across providers (Figure 1). Each provider’s translated cues are aligned to the SCR’s scene and source-line scope to form SATS (§2.2). The SCR and SATS are then used to generate an MQM-style error report scene by scene (TSER, §2.3); aggregating repeated runs on the same input reduces evaluator stochasticity (§2.4).

2.1 Script-like Context Representation

The SCR (Script-like Context Representation) makes the video’s non-verbal audiovisual context, which subtitle text alone does not convey, a fixed evaluation input. Concretely, an SCR groups all source-subtitle cues into screenplay-like semantic and narrative scenes. Each cue belongs to exactly one scene, and scene boundaries are placed only between cues. Movie-level fields define the cast. Scene-level fields describe what happens where and carry narrative memory forward. Line-level fields identify who speaks to whom, how each utterance is delivered, and the visual or auditory cues relevant to interpreting it (Figures 1 and 2; Appendix A gives the exact fields and construction procedure). The subsequent SATS alignment (§2.2) and TSER evaluation (§2.3) both use the SCR as

their reference context.

Each SCR is drafted by a VLM-assisted agent and finalized by human review. The draft is built in two stages: the source subtitles are first segmented into scenes, and the VLM then analyzes the corresponding video clip for each fixed scene (Appendix A). This ordering lets the VLM process bounded clips rather than the whole title, while subtitle-based segmentation preserves cue integrity and centers scenes on conversational context relevant to subtitle translation. Construction and review use only the source video and source subtitles, never candidate translations or their scores. Consequently, an SCR cannot be tuned toward any provider; as a source-side representation, it is target-language-agnostic by construction. A single SCR per title is therefore reused without modification across all providers, repeated runs, and ablation conditions. For this benchmark, the bilingual reviewer changed speaker/addressee labels on 12/243 source lines (4.9%); no scene boundary, summary, delivery, or audiovisual field was edited (Appendix A).

One natural alternative—feeding the raw video directly—would require the evaluator to re-infer visual context on every pass, destabilizing evaluation conditions and obscuring whether a low score reflects failed translation judgment or failed context reconstruction. The SCR fixes video understanding before evaluation, so every comparison shares the same context. The SCR is not merely a video summary but an evaluation interface. A human can inspect the exact context given to the evaluator or compare SCR-present and SCR-absent conditions in a controlled way (§4.3). Its scenes and source lines provide common anchors for SATS and TSER, linking automatic results directly to experts’ line-level annotations.

2.2 Scope-Aligned Translated Subtitles

Translated subtitle cues do not align one-to-one with source lines: several source lines may be merged into a single cue (many-to-one), one line may be split across cues (one-to-many), material absent from the source may be added (addition), or a source line may be dropped (missing). For each movie–provider pair, Scope-Aligned Translated Subtitles (SATS) fix this alignment before evaluation. Each translated cue is assigned to an SCR scene and anchored to one or more corresponding source lines. Cues with no source counterpart are marked as additions, while source lines

SCR — <i>Sprite Fright</i> , scene 2	
MOVIE	
characters	Ellie, Victor, Roxy, Spike, Phil, Elder Sprite ...
SCENE 2 00:01:59-00:03:56	
settings	forest campsite, night
summary	while Phil grills, Ellie discovers mushroom-like creatures ("sprites"); her friends see a way to make money ...
visual_contexts	sprites carry a blood-stained sneaker out of the burrow while the Elder Sprite greets the group ...
memory	cumulative_summary Ellie observes the forest quietly while her four friends trash the woods and mock her as a "tree hugger" ...
	previous_scene_detail daytime at the campsite; the conflict surfaces for the first time ...
LINE 2:6 00:02:59-00:03:02	
preceding line 2:5 — Ellie, warm:	"They're absolutely adorable."
line_original	"Yeah, absolutely bloody adorable."
line_speaker	Victor
line_addressee	Ellie, the friends
line_delivery	sarcastic, cynical
visual_context	Victor puts a hand on Ellie's shoulder with a sarcastic expression; he walks past her, still smirking ...
auditory_context	Victor's voice; echoes Ellie's line mockingly, in a cynical tone ...

Figure 2: An SCR excerpt (*Sprite Fright*, scene 2, abridged) at the movie, scene, and line levels, with the memory carried over from scene 1; line 2:6 sarcastically echoes line 2:5. In our deployment the SCR is authored and reviewed in Korean—the operators’ working language—and is shown here in English translation; the authoring language is immaterial as long as the evaluator LLM can read it.

with no aligned cue are identified as missing. For example, when one cue condenses source lines 1:7–1:8, SATS anchors it to both lines; if a TSER error spans both anchors, line-level aggregation splits its severity equally between them (Appendix C).

Alignment is automatic: a lightweight LLM selector, separate from the evaluator, decides for each candidate cue within a scene’s time window which scene it belongs to, which source lines it matches, and whether it is an addition (handling of doubly assigned and unassigned cues is detailed in Appendix C). For this benchmark, a bilingual reviewer checked all alignments; no cue required reassignment. The resulting SATS is a human-inspectable intermediate product that was then reused unchanged across all runs and ablations.

2.3 Translated Subtitle Evaluation Report

For each scene, the evaluator queries an LLM with that scene’s SCR fields (Figure 2) and the source–translation cues aligned by SATS. In addition to the SCR’s cumulative narrative context, the input includes translation-side memory: translated cues from preceding scenes whose speaker–addressee pair also occurs in the current scene (Appendix J). The output is not a single score but a structured Translated Subtitle Evaluation Report (TSER), one report per movie–provider–run. Each error carries an MQM-style category (e.g.,

mistranslation, omission), an integer severity from 1 to 100 (RUBRIC-MQM’s 100-point rubric (Kim, 2025)), and a short rationale. It is anchored to a scene, source lines, and translated cues.

A movie–provider score is the sum of error severities, with lower scores indicating better quality. Because every error is anchored to source lines—with severity split evenly across them when an error spans several—the resulting severity values can be aggregated under the same schema at the system, scene, line, and category levels. This makes a TSER inspectable case by case and directly comparable with experts’ line-level annotations, without further conversion.

2.4 Repeated Evaluation and Aggregation

An LLM evaluator returns slightly different TSERs on identical input, so the score from a single report is unstable. The framework’s final step therefore generates R TSERs from the same SCR and SATS and averages their error severities at the system, scene, and line levels (system: mean overall_score; scene: mean length-normalized severity; line: mean distributed severity). We use R=10; the choice of R and the stability–cost trade-off are analyzed in §4.4.

3 Expert Benchmark and Setup

The task takes the source subtitles, the video, and a candidate translation as input and outputs quality assessments at the system, scene, and line levels; at the system level, it compares the five providers within each movie. Since no public benchmark is known for context-aware subtitle-translation evaluation (§1), we construct an expert benchmark in which professional translators from an audiovisual-translation vendor annotate candidate translations of open movies under a blind, randomized protocol (§3.1 and §3.2), with metrics and evaluator setup in §3.3 and §3.4. What is released and what is withheld are described under Ethics.

3.1 Benchmark Data and Providers

We use three freely licensed English open movies (*Elephants Dream*, *Sprite Fright*, *Tears of Steel*) to build an expert-annotated benchmark of English-to-Korean subtitle translation. Rights to use the material and release derived results—not representativeness—determined the pool (Limitations); within it we selected three titles with substantial dialogue and distinct speaker structures—a two-

Movie	Scenes	Source lines
<i>Elephants Dream</i>	6	85
<i>Sprite Fright</i>	5	82
<i>Tears of Steel</i>	6	76
Total	17	243

Table 1: Benchmark statistics for the three open movies. Each movie has five providers, giving 15 movie–provider, 85 scene–provider, and 1,215 source-line–provider units in total. Every movie–provider is evaluated at R=10, and the same seven experts (P1–P7) assess all candidate translations.

handler, an ensemble horror-comedy, and a science-fiction short. Sizes and units are summarized in Table 1, whose Scenes follow the SCR’s segmentation (§2.1) applied identically to each movie. For each movie we compare five translation providers, referred to by labels: a human-finalized translation in which two bilingual reviewers cross-reviewed and refined a machine-translation draft not drawn from any evaluated provider (Human); an LLM translation of each line in isolation (LLM-Line); an LLM translation of each scene’s lines together, with no further context (LLM-Scene); a global commercial AI subtitle solution under an anonymized label (Comm-AI); and a general-purpose neural machine translation that is not an LLM (Google NMT). The two LLM providers are produced with gemini-2.5-pro. The five range from human-finalized translation to non-LLM machine translation.

3.2 Expert Annotation Protocol

The professional evaluation was carried out by seven translators (P1–P7) employed at a vendor specializing in audiovisual content. None of the seven helped produce any compared translation. They independently assessed all movie–provider pairs—blind to provider identity and in randomized order—recording the single most impactful error, if any, on each line with its MQM category and 1–100 severity (identifiers were mapped only after scoring; Appendix I gives the guide). Their annotations use the same source-line anchors, categories, and 1–100 severity scale as TSER (§2.3), allowing direct line-level comparison with the automatic results. We compared alternative aggregates but use the unweighted mean of all seven (mean₇) to avoid selection, exclusion, or weighting; a standard one-up/one-down trimmed mean serves as a robustness check.

3.3 Metrics

We measure agreement at three levels. Because our framework, experts, and baselines use different score scales, the primary metric is Spearman rank correlation, which depends only on ranking. At the system level we compute Spearman ρ and Kendall τ over the five providers’ ranking within each movie and macro-average over the three movies. At the scene level we use (scene, provider) units and length-normalize the scene score (mean severity per line) to control for the wide variation in scene length (3–27 lines per scene). At the line level we report (line, provider) severity Spearman and, because most lines are error-free and concentrate at zero, additionally report binary error-detection F1 and a rubric-aligned weighted Cohen κ over severity levels (Appendix G). As automatic baselines we use two learned reference-free QE metrics—MetricX-24-XL-QE and xCOMET-XXL—and one published LLM-as-judge MQM metric, RUBRIC-MQM (setup in Appendix D), aggregated to the same units and sign-aligned (scales and directions in the Figure 3 caption). Category agreement and R=1,...,10 convergence are reported in §4.

3.4 Evaluator Setup

We generated each TSER scene by scene (§2.3) using gemini-3.1-pro-preview, the strongest Gemini model available at the time of the experiments; for SATS alignment, we used the lighter gemini-3-flash-preview (models and generation settings per role in Appendix E, prompts in Appendix J). Repeated evaluation and aggregation follow §2.4 (R=10; the same evaluator and R are used across our main and context-ablation experiments).

4 Results and Analysis

4.1 Does It Score Like the Experts?

We first verify that our framework ranks the translation providers similarly to the experts. For each movie we compute the Spearman ρ over the five providers’ ranking and macro-average over the three movies; against the unweighted seven-expert mean (mean₇; §3.2), SCR-Eval reaches strong system-level agreement (Spearman ρ = 0.967, Table 2; only in *Elephants Dream* do the adjacent LLM-Scene and Comm-AI swap by one rank). Across the same five-provider rankings, leave-one-out expert agreement averages ρ = 0.951 over the 21

Level	Spearman ρ	95% CI
System	0.967	[0.87, 1.00]
Scene	0.849	[0.75, 0.87]
Line	0.666	[0.63, 0.70]

Table 2: Framework agreement with the expert mean (mean₇): system is the provider ranking (Kendall τ = 0.933), scene severities are length-normalized, and line is the severity ρ ; 95% CIs are line-level bootstrap intervals (Appendix B). Per-level Kendall τ_b and the baselines appear in Table H1.

evaluator–movie cases; SCR-Eval’s 0.967 is therefore comparable to inter-expert agreement (Appendix B). A one-sided exact shared-label permutation test of the three-movie macro-average gives p = .008, the smallest attainable value with five providers (Appendix B). A standard trimmed mean gives system/scene/line ρ = 0.967/0.871/0.660; excluding both same-family LLM providers leaves the conclusion unchanged (see Limitations).

Even when the system ranking agrees, whether the framework identifies the same problematic scenes and lines must be checked separately (units in §3.3). Agreement weakens from the scene to the line level but stays positive, and a line-level bootstrap over within-movie source lines (n=2,000; Appendix B) keeps the 95% interval above zero at all three levels (Table 2). The line level is also reported as binary error-detection F1 = 0.601 (rationale in §3.3; severity-level agreement in Appendix G); with many error-free items and larger inter-evaluator variance it stays conservative, so our central claims rest on provider ranking and scene-level diagnostic agreement. Category-level agreement (κ = 0.43) lies within the inter-expert range (0.34–0.56), as does severity-bin agreement (κ = 0.34; inter-expert range 0.21–0.56; Appendix G).

4.2 How Do Baselines Behave Here?

We compare our framework with three automatic baselines (MetricX, xCOMET, and RUBRIC-MQM; §3.3). All produce source-line-level scores in a reference-free/QE setting, and we aggregate them to the same units and compare with mean₇ via Spearman. These baselines read the source and candidate text alone, whereas our framework also receives the SCR’s scene structure, speaker relations, and audiovisual context. The comparison therefore shows how each metric behaves in this setting rather than a matched-input contest; §4.3 instead removes that context from our own evalua-

tor to show how much of the agreement it accounts for. The results show a clear contrast (Figure 3). Whereas our framework reaches system $\rho = 0.967$, the two learned metrics correlate negatively (MetricX -0.600 , xCOMET -0.500): they move opposite to the expert ranking, underrating the human-finalized translation (Human) and overrating the non-LLM machine translation (Google NMT). The human subtitles condense and rephrase the source while preserving meaning; the learned metrics tend to rank them below the more literal machine output, whereas the experts—and our framework—do not. The two learned metrics show only weak positive line-level correlations (0.12 to 0.18), versus 0.666 for our framework.

RUBRIC-MQM (Appendix D), itself an LLM-as-judge, avoids the sign reversal but reaches only system $\rho = 0.267$ and scene $\rho = 0.029$. At the scene level, none of the baselines agrees positively with experts beyond $\rho = 0.029$ (MetricX -0.146 , xCOMET -0.372), whereas our framework reaches 0.849. The contrast persists when RUBRIC-MQM uses our evaluator backbone (R=5): system/scene/line $\rho = 0.300/0.175/0.306$ (Appendix D). The baselines receive isolated source-translation lines, whereas our framework receives scene-grouped text and SCR context. Thus, this comparison shows how well each method aligns with experts when applied with its respective inputs in this subtitle setting; it does not establish which metric is better under matched inputs. Built for sentence-level MT without scene or video context, these metrics appear to favor more literal renderings over the context-appropriate paraphrases that subtitles require.

4.3 How Does Agreement Change with Context?

Our framework evaluates each scene with its SCR fields and translation-side memory (§2.3). To test agreement under progressively richer inputs, we compare three conditions on the same movies and providers, using the same evaluator and R=10 as in §4.1: Full (the complete SCR-Eval input; the same runs as the main result), Scene-only (SATS-aligned source and translated cues supplied together within SCR scene boundaries, with other SCR fields and memory removed), and Line-only (each aligned source-translation pair evaluated individually, with no scene grouping or memory).

Agreement increases monotonically at all three levels: Line-only < Scene-only < Full (Figure 4).

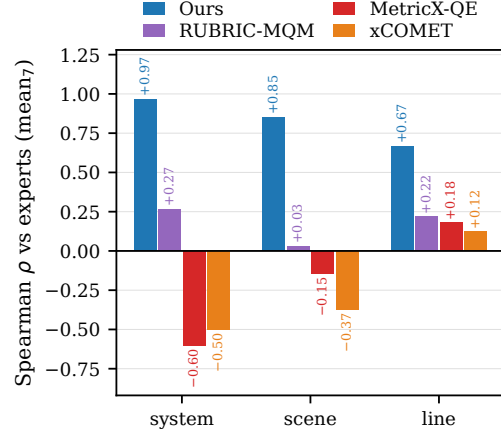


Figure 3: Agreement with the expert mean (mean_7) as Spearman ρ at the three levels; scene severities are length-normalized. The baselines—MetricX-24-XL-QE (0–25, lower-better), xCOMET-XXL (0–1, higher-better), and the LLM-as-judge RUBRIC-MQM (span severity 1–4, lower-better)—are sign-aligned; the learned metrics have negative system-level correlations. Exact values incl. per-movie system ρ in Appendix H.

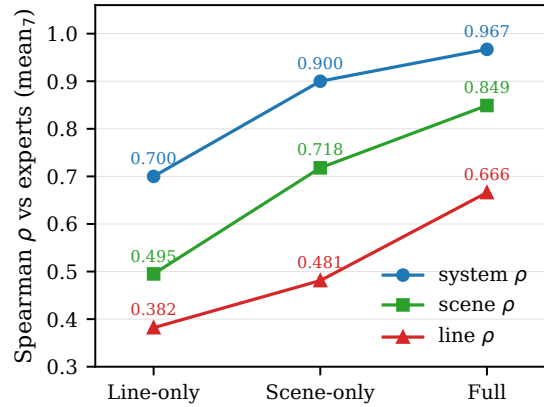


Figure 4: Agreement with the experts (mean_7) by evaluation condition. All conditions use the same evaluator at R=10 (Full = the main result’s runs); scene severities are length-normalized. Line-only evaluates each source line individually but retains anchors and the framework’s scoring protocol. Per-movie values in Appendix H.

(1) Scene-bounded text: moving from Line-only to Scene-only raises system ρ from 0.700 to 0.900; this bundled condition evaluates each SCR-defined scene as a unit with surrounding source-translation cues. (2) Full context input: adding the remaining SCR fields and translation-side memory raises system ρ to 0.967 and improves scene and line agreement by 0.131 and 0.185. The progression thus shows stepwise gains from isolated pairs to scene-bounded text and then full context (see Limitations for field-level analysis).

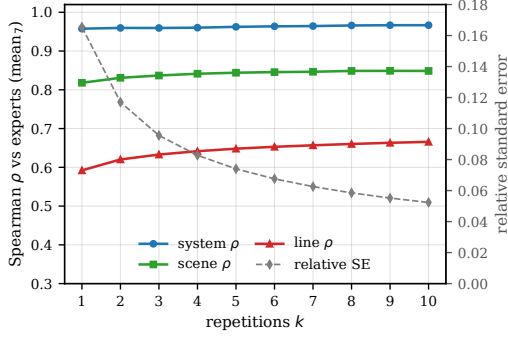


Figure 5: Repeated-run stability–cost over $k=1,\dots,10$: agreement with mean_7 (left axis) and relative standard error (right axis). Exact values appear in Appendix B, Table B1.

4.4 How Many Runs Are Worth the Cost?

Increasing k from 1 to 10, we measure agreement with the experts (mean_7), reproducibility of the $R=10$ ranking, and the relative standard error (Figure 5; averaging over randomized run orderings; full table in Appendix B).

The provider ranking is already highly reproducible from a single run (expected system $\rho = 0.958$; agreement with the $R=10$ ranking ≈ 0.98), whereas repetition mainly benefits line-level estimates: line agreement rises from 0.592 to 0.666, and the relative standard error falls from 0.166 to 0.052 following $1/\sqrt{k}$. A single run may therefore be adequate when only provider ranking is needed on this benchmark. We use $R=10$ for the most stable line estimates and identify $k \approx 5$ as a cost-aware operating point.

4.5 Qualitative Examples

Consider two context-dependent errors that the source text alone cannot adjudicate: Google NMT renders the idiom "I'll give you that one" literally (Figure 6), and the comforting "There, there" as 저기 ... 있습니다 *jeogi ... itseumnida* ('over there ... is'; the SCR context in Figure 1). In both, the learned QE metrics rate the literal version above the human subtitle, yet our framework and all seven experts flag it as an error.

5 Discussion

5.1 Practical Motivation and Use

This framework grew out of a practical need to automate commercial VoD subtitle translation at scale: no benchmark existed for judging improvement, and manual expert review cannot cover every

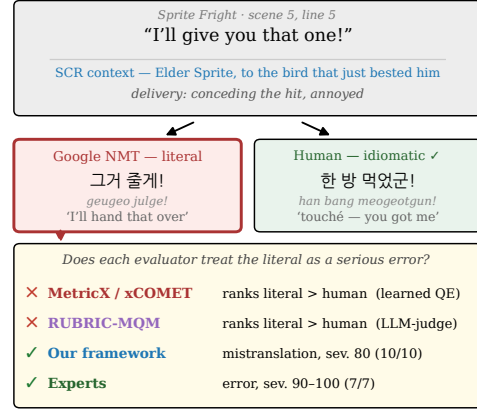


Figure 6: A context-dependent error (*Sprite Fright*, line 5:5): all three baselines—learned QE (MetricX, xCOMET) and LLM-as-judge RUBRIC-MQM—rate Google NMT’s literal rendering of "I'll give you that one" above the human subtitle, whereas our framework (mistranslation, severity 80, 10/10 runs) and all seven experts (severities 90–100) flag it.

line and revision across titles and languages. Reviewing an SCR is a one-time, source-side cost per title, amortized across languages, providers, and runs; automated candidate-translation evaluation recurs for each provider, language, and revision. On this benchmark, $R=10$ automatic evaluation cost $\approx$ US\$3.3 per movie–provider (under US\$50 for all 15), versus $\approx$ US\$2,800 and about two calendar weeks for seven-expert annotation of the same pairs; human SCR review took about one hour per 30 minutes of video (Appendix E). We plan to extend the benchmark to internal content with secured usage rights, quantify build-or-adopt decisions, and pilot line-level evaluation as a triage stage that prioritizes high-severity items for operator review while retaining human oversight.

5.2 Conclusion

We proposed SCR-Eval, a context-aware evaluation framework for video subtitle translation built on the SCR→SATS→TSER pipeline (§2). On our three-movie English-to-Korean benchmark, it agrees strongly with experts on provider ranking (§4.1); agreement decreases stepwise from Full to Scene-only to Line-only (§4.3), while learned metrics correlate negatively with the expert ranking (§4.2). Together, these results show that, on this benchmark, SCR-Eval tracks expert judgments most closely with its full context input while retaining inspectable line-level diagnostics.

Limitations

Our results should be interpreted within the following limitations.

Sample size. The benchmark contains three five-provider rankings (15 movie-provider scores), so the system-level correlation is coarse and sensitive to a single adjacent swap. We therefore do not use it to claim that one provider is statistically better than another. The per-movie correlations (0.90 / 1.00 / 1.00) and line-bootstrap intervals (§4.1; Appendix B) quantify agreement with experts within these films, not representativeness or between-provider significance.

Content scope and representativeness. Expert evaluation required films whose video and source subtitles could be legally reused, whose derived artifacts could be released, and whose dialogue was sufficient for line-level analysis. Very few publicly available films met all of these conditions, leaving a practical pool largely limited to freely licensed shorts; we obtained expert line annotations for three of them. This rights-constrained pool is not representative of film content. Long-form commercial titles could not be included because their proprietary video and subtitles would require rights clearance. Our expert-agreement claims are therefore limited to these three shorts and five translation providers.

Language and domain generalization. Expert validation covers only English-to-Korean movie subtitles. Although the SCR fields are designed to be language-independent, we have no evidence that the same expert agreement holds for other language directions or subtitle domains such as series, documentaries, or news. We selected English-to-Korean because it is the operational setting from which the framework arose and because professional evaluators were available for this pair. Validation across languages, genres, and domains remains future work, and the value of cumulative memory may itself vary by domain.

Unresolved effects of individual context components. The ablation varies context in bundles: Line-only→Scene-only jointly adds SCR-defined scene grouping and surrounding source-translation cues, while Scene-only→Full jointly adds the remaining SCR fields and cumulative memory. It cannot isolate any field group’s or cumulative narrative memory’s effect. Few scenes in these shorts also limit tests of cross-scene accumulation. Direct testing requires field-wise and memory-only

ablations on longer titles; until then, the observed stepwise trend should not be assumed to generalize beyond these films.

Dependence on one LLM evaluator. We use only gemini-3.1-pro-preview as the evaluator. Averaging R=10 runs reduces stochastic variation but does not remove systematic bias (e.g., underuse of the punctuation category; Appendix G). LLM-Line and LLM-Scene also come from an earlier generation of the Gemini family, creating a possible same-family preference. Excluding both providers strengthens agreement at every level, so the result does not appear to be driven by that preference (Appendix H); nevertheless, reproducibility with evaluators from other model families remains untested.

Shared MQM paradigm. The experts and SCR-Eval use the same MQM-style categories and 1–100 severity rubric, so some of the observed agreement may arise from the shared evaluation paradigm. However, RUBRIC-MQM also uses an MQM category-and-severity structure but reaches only system $\rho = 0.267$ and 0.300 with our evaluator backbone; sharing MQM is therefore not sufficient by itself to account for SCR-Eval’s 0.967 (Appendix D). Isolating this contribution would require independently collected non-MQM gold, which we do not have.

No independent validation of intermediate artifacts. The SCR and SATS files underwent code-based structural checks and exhaustive review by one bilingual reviewer, but neither was compared against independently authored professional gold, and inter-reviewer agreement was not measured. We therefore validate not the accuracy of scene boundaries, SCR fields, or SATS alignments themselves, but the agreement of final evaluation results with experts when the reviewed artifacts are fixed. Reusing each title’s SCR across providers and a fixed SATS within each movie-provider pair across runs and ablations controls input variation within comparisons; it does not measure construction variability or eliminate error propagation. In production, human review of the SCR against the video and of SATS anchors’ semantic validity may be warranted depending on the intended use and required level of assurance. Separately, SATS coverage, duplicate assignments, and omissions can be checked structurally in code (Appendices A and C).

Ethics Statement

This study builds its evaluation benchmark on the English source subtitles of three Blender Foundation open movies (*Elephants Dream*, *Tears of Steel*, *Sprite Fright*), all under CC BY licenses². We will release eligible derived artifacts subject to the source licenses and the applicable terms of the generating services. For reproducibility we provide the SCR/SATS/TSER schema specifications, the expert annotation protocol (rubric), and derived results (scores and correlations) as appendices.

Upon publication we will additionally release the source-side SCR for these open movies and, for the four providers other than Comm-AI, the SATS representations, per-run TSER outputs, line-level expert annotations under anonymized evaluator identifiers (P1–P7), and translations—our human-finalized translation together with the LLM-Line, LLM-Scene, and Google NMT outputs we generated through public APIs—to the extent permitted by the applicable terms of the generating services. The released artifacts will be available on the Hugging Face Hub.³ The commercial solution’s translation and its provider-specific artifacts are withheld under its license, and the evaluation pipeline code under internal policy, so exact re-execution of the complete pipeline is not possible. The released artifacts do keep every error anchored to its scene, source lines, and translated cues, so the individual judgments behind the reported agreement can be inspected case by case, and a different evaluator can be run on the same inputs and scored against the same expert gold. The baseline metrics’ outputs are not included, but all three are public implementations (§3.3) that can be run on the released subtitles. Appendix K maps the paper’s provider labels to the identifiers used in these files, and Appendix E records the disclosed model and generation settings for each applicable artifact.

We report Google NMT by name as a public, reproducible baseline, but the commercial solution only under the anonymized label Comm-AI, as it is a vendor’s proprietary product. Expert results are reported under anonymized evaluator identifiers (P1–P7) and contain no personally identifying information.

The proposed framework is not a replacement

for human expert evaluation but an auxiliary, inspectable tool within the cost and scalability limits of expert evaluation. Its anchor structure lets people review automatic results case by case, and we recommend using it to assist human reviewers rather than to replace decisive judgments.

Our results do not disparage the baselines: MetricX, xCOMET, and RUBRIC-MQM remain valid in the sentence-level MT setting they were built for, and their low subtitle agreement reflects a task–domain mismatch rather than a defect (§4.2).

Acknowledgments

We thank the seven professional subtitle evaluators whose annotations form our expert benchmark, and the Blender Foundation for making the three open movies used in this work available under open licenses. We also thank the anonymous reviewers and the area chair for comments that improved this paper.

References

- Aljoscha Burchardt, Arle Lommel, Lindsay Bywood, Kim Harris, and Maja Popović. 2016. [Machine translation quality in an audiovisual context](#). *Target*, 28(2):206–221.
- Jorge Díaz Cintas and Aline Remael. 2007. *Audiovisual Translation: Subtitling*. Translation Practices Explained. St. Jerome Publishing, Manchester.
- Patrick Fernandes, Sweta Agrawal, Emmanouil Zaranis, André F. T. Martins, and Graham Neubig. 2025. [Do LLMs understand your translations? evaluating paragraph-level MT with question answering](#). In *Second Conference on Language Modeling (COLM)*.
- Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André F. T. Martins, Graham Neubig, Ankush Garg, Jonathan H. Clark, Markus Freitag, and Orhan Firat. 2023. [The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation](#). In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, pages 1066–1083.
- 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.
- Markus Freitag, Nitika Mathur, Daniel Deutsch, Chiklu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. [Are LLMs](#)

²Open movies (Blender Foundation), reusable with attribution (also commercially): *Elephants Dream* (CC BY 2.5), *Tears of Steel* (CC BY 3.0), *Sprite Fright* (CC BY 1.0).

³<https://huggingface.co/datasets/Geneen/SCR-Eval>

- breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation (WMT)*, pages 47–81.
- 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.
- Yuchen Eleanor Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. **BlonDe: An automatic evaluation metric for document-level machine translation**. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 1550–1565.
- 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 (WMT)*, pages 492–504.
- Ahrii Kim. 2025. **RUBRIC-MQM: Span-level LLM-as-judge in machine translation for high-end models**. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 147–165.
- 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 (WMT)*, pages 768–775.
- 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 (EAMT)*, pages 193–203.
- Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. **Error span annotation: A balanced approach for human evaluation of machine translation**. In *Proceedings of the Ninth Conference on Machine Translation (WMT)*, pages 1440–1453.
- Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhuja, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. **Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help**. In *Proceedings of the Tenth Conference on Machine Translation (WMT)*, pages 436–483.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. **G-Eval: NLG evaluation using GPT-4 with better human alignment**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2511–2522.
- Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. **Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics**. *Tradumàtica: tecnologies de la traducció*, (12):455–463.
- Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2021. **A survey on document-level neural machine translation: Methods and evaluation**. *ACM Computing Surveys*, 54(2):Article 45.
- Wafaa Mohammed and Vlad Niculae. 2025. **Context-aware or context-insensitive? assessing LLMs’ performance in document-level translation**. In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 126–137, Geneva, Switzerland. European Association for Machine Translation.
- 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.
- 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.
- 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.
- 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.
- Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. **BLEURT: Learning robust metrics for text generation**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.
- Umut Sulubacak, Ozan Caglayan, Stig-Arne Grönroos, Aku Rouhe, Desmond Elliott, Lucia Specia, and Jörg Tiedemann. 2020. **Multimodal machine translation through visuals and speech**. *Machine Translation*, 34(2–3):97–147.
- 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 (ACL)*, pages 1198–1212.

- Patrick Wilken, Panayota Georgakopoulou, and Evgeny Matusov. 2022. [SubER - a metric for automatic evaluation of subtitle quality](#). In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT)*, pages 1–10.
- Ran Zhang, Wei Zhao, Lieve Macken, and Steffen Eger. 2025. [LiTransProQA: An LLM-based literary translation evaluation metric with professional question answering](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29099–29121, Suzhou, 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 P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena](#). In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Datasets and Benchmarks Track*.

A Schemas and SCR Construction

Released schemas are versioned; the artifacts in this paper follow `scr@1`, `sats@1`, and `tser@1`. Tables A1–A3 summarize the key fields of each schema.

Level	Key fields	Role in evaluation
Movie	<code>characters</code> , <code>time</code>	Cast glossary (names, aliases, descriptions) for resolving speakers/addressees, and the time base
Scene	<code>scene_id</code> , <code>summary</code> , <code>settings</code> , <code>visual_contexts</code> , <code>memory</code>	Semantic scope and scene/narrative context
Line	<code>line_uid</code> , <code>line_original</code> , <code>line_speaker</code> , <code>line_addressee</code> , <code>line_delivery</code> , <code>visual_context</code> , <code>auditory_context</code>	Basis for line-level translation judgments and error anchors

Table A1. SCR fields (one record per movie)

Level	Key fields	Role in evaluation
File	<code>schema_version</code> , <code>scr_ref</code> , <code>ts_ref</code> , <code>stats</code>	Provenance (SCR/subtitle references) and coverage stats
Scene	<code>scene_id</code> , <code>start_time/end_time</code> , <code>source_lines</code> , <code>ts_cues</code>	Scene scope shared with the SCR
Cue	<code>ts_index</code> , <code>start_time/end_time</code> , <code>text</code> , <code>matched_source_uids</code> , <code>is_addition</code> , <code>confidence</code>	Per-cue anchor to source lines, addition flag, selector confidence

Table A2. SATS fields (one file per movie–provider pair)

Level	Key fields	Role in evaluation
File	<code>schema_version</code> , <code>movie</code> , <code>provider</code> , <code>overall_score</code> , <code>scene_scores</code> , <code>scenes</code>	Report identity and aggregate scores
Error	<code>span</code> , <code>original</code> , <code>translated</code> , <code>category</code> , <code>severity</code> , <code>reason</code>	MQM-style error record (category from the 11 error categories in Appendix J ; the 12th, No-Error, marks an error-free span and is excluded from error records; severity 1–100)
Anchor	<code>scene_id</code> , <code>source_line_uids</code> , <code>ts_indices</code>	Joins each error to source lines and translated cues

Table A3. TSER fields (one report per movie–provider–run)

SCR construction procedure

The SCR is drafted automatically in two stages and then corrected by human reviewers. (i) *Scene segmentation*: an LLM splits the source subtitles—without the video—into scenes that cover the full timeline without gaps or overlaps, with boundaries snapped to subtitle-cue edges; each scene is capped at three minutes (a heuristic for the length the VLM handles reliably), and within that cap the model cuts at semantic-unit boundaries. (ii) *Scene analysis*: proceeding scene by scene, a VLM receives the scene video and its source lines together with a rolling memory—the cumulative plot summary, a character lexicon, and the detail of the previous scene—and emits the scene-level fields (summary, settings, visual contexts) and line-level fields (speaker, addressee, delivery, visual and auditory context) of Table A1 as schema-constrained structured output; the memory is then updated by restating facts from the analyzed scene before the next scene is processed. Operators then review and correct the draft—in Korean, their working language—before it is used for evaluation. In this review the corrections were confined to speaker or addressee labels on 12 of the 243 source lines (4.9%); no scene boundary, summary, delivery, or audiovisual field required correction.

B Supplementary Figures and Tables

Figure B1 relates the expert and framework scores behind [§4.1](#), Figure B2 gives the bootstrap distributions behind the intervals in [Table 2](#), and Table B1 lists the full convergence values behind [Figure 5](#).

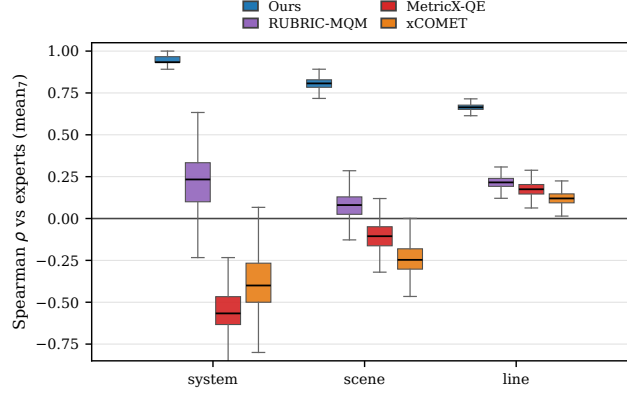


Figure B2. Line-level bootstrap distribution ($n=2,000$) of the agreement (Spearman ρ vs the expert mean_7) for the framework and the three reference metrics, at the system, scene, and line levels; box = interquartile range, line = median, whiskers = $1.5 \times \text{IQR}$ (outliers omitted). The intervals quantify line-sampling uncertainty only — the reference metrics are single-run (MetricX, xCOMET) or R=5-mean (RUBRIC-MQM) point scores, so no run-to-run variance is included. At the system level the framework’s distribution lies entirely above zero and MetricX’s entirely below it; xCOMET’s upper whisker reaches zero and RUBRIC-MQM’s straddles zero, so the sign reversal is robust for MetricX while the five-provider base leaves the other baselines’ system-level intervals wide.

k	system ρ	scene ρ	line ρ	rank stab.	relative SE
1	0.958	0.818	0.592	0.978	0.166
2	0.960	0.831	0.620	0.986	0.117
3	0.959	0.837	0.633	0.990	0.096
4	0.960	0.841	0.642	0.993	0.083
5	0.962	0.844	0.648	0.996	0.074
6	0.964	0.846	0.653	0.997	0.068
7	0.964	0.847	0.657	0.998	0.063
8	0.966	0.849	0.660	0.999	0.059
9	0.967	0.849	0.663	1.000	0.055
10	0.967	0.849	0.666	1.000	0.052

Table B1. Convergence over repetition count k (full $k=1, \dots, 10$). Values are means over random run-order permutations; ρ is Spearman agreement with the experts (mean_7 , scene length-normalized), rank stab. is reproducibility of the R=10 provider ranking, and relative SE is the cumulative-mean relative standard error ($\text{single-run CV} / \sqrt{k}$).

C SATS Alignment Details

SATS alignment uses a local rather than global search. For each scene, a candidate window extends several seconds before its start and after its end; if a neighboring scene boundary lies farther out, the window extends to that boundary. Only cues within this window are retained as candidates; in this benchmark, we used 5 s before the start and 10 s after the end. A lightweight LLM selector decides, for each candidate cue, which scene it belongs to, which source lines it matches, and whether it is an addition, together with a confidence. A cue claimed by two scenes is reassigned, and a cue left unassigned is classified into an adjacent scene. For example, in *Tears of Steel* the source is split across two cues—1:7 “I’m freaked out! I have nightmares that I’m being chased...” and 1:8 “...by these giant robotic claws of death...” —which a provider may condense into a single Korean cue such as 정말 무서웠어! 거대한 죽음의 로봇 발톱에 쫓기는 악몽을 꾀어 *jeongmal museowosseo! geodaehan jugeumui robot baltobe jjotgineun angmongeul kkwasseo* (‘I’m freaked out! I had a nightmare of being chased by giant robotic claws of death’). SATS records this as one many-to-one cue anchored to both 1:7 and 1:8 (addition = false, no missing line).

SATS explicitly records each cue’s anchors and addition status, as well as unmatched source lines, so coverage, duplicate assignment, and omission are structurally checkable in code. These checks do not establish that an anchor is semantically correct; depending on the intended use and required level of assurance, separate bilingual review may be warranted. In the finalized benchmark SATS, all cues were assigned, and the exhaustive review found no cue requiring further reassignment. This is an internal-review result for one language pair and five providers, not a general guarantee of alignment accuracy.

D RUBRIC-MQM Baseline Setup

RUBRIC-MQM (Kim, 2025) is a published span-level LLM-as-judge MQM metric. We ran the latest released implementation (v2.0) unmodified, with its default configuration: gpt-4.1-mini-2025-04-14 as the evaluator, temperature 0.7, max tokens 2048, seed 42, and a single pass. RUBRIC-MQM is reference-free and emits per-span categories with severity 1–4 (the released implementation’s default; RUBRIC-MQM also defines 1–8 and 1–100 severity rubrics, the last of which our own evaluator adopts, §2.3); we map each span to source-line and cue units through the same SATS alignment used elsewhere (the original `data.text` suggestion field is not used as an anchor).

Across five runs, per-provider scores varied with CV 1.3–6.1% (mean 3.3%), and adjacent provider pairs—including top and bottom pairs—swapped rank between runs in every movie. We therefore likewise average repeated runs (R=5, versus R=10 for our framework and ablation). RUBRIC-MQM’s agreement with mean_7 is essentially stable across R=1/3/5 (system $\rho = 0.28 / 0.29 / 0.27$, scene $\rho = 0.04 / 0.04 / 0.03$; line $\rho = 0.19 / 0.21 / 0.22$, a small and diminishing gain); even its line agreement stays far below our framework’s 0.666, so R=5 does not understate the baseline relative to our R=10.

Per-movie, per-level agreement with mean_7 appears in Appendix H (Table H1). To examine the effect of the evaluator backbone, we additionally ran RUBRIC-MQM with our evaluator (gemini-3.1-pro-preview, R=5) in place of its default gpt-4.1-mini-2025-04-14. Agreement with mean_7 barely changes and stays far below our framework—system ρ 0.300 (vs. 0.967), scene 0.175, and line 0.306. Thus, the observed gap is not explained by the evaluator backbone alone and is associated with the methods’ remaining design differences, including SCR-Eval’s structured-context input. The backbone control therefore does not turn RUBRIC-MQM into our method without context; that condition is Line-only (§4.3), which keeps our evaluation prompt and severity rubric and reaches system $\rho = 0.700$.

E Models per Role and Experiment

As the evaluator we used the strongest Gemini model available at the time of the experiments (SATS alignment is an exception: a lightweight flash-tier model sufficed). The LLM-based translation providers were generated with the strongest model available at generation time (gemini-2.5-pro); because the benchmark concerns agreement with experts rather than the providers’ absolute translation quality, they were not regenerated with a newer model. The main results (§4.1) and the context ablation (§4.3) share the same single evaluator, gemini-3.1-pro-preview, and the same R=10 runs, so these internal comparisons are not confounded by an evaluator difference. For the RUBRIC-MQM baseline we did not select a model—we ran the released v2.0 implementation with its default model; the framework-vs-RUBRIC-MQM comparison therefore also differs in evaluator model, which we control for in Appendix D.

Licenses of external artifacts: we use MetricX-24 (Apache-2.0) and RUBRIC-MQM (MIT) without usage restrictions, and xCOMET-XXL (CC BY-NC-SA-4.0) solely for non-commercial research comparison—not as part of any deployed or commercial pipeline; Korean is rendered with the NanumGothic font (SIL OFL 1.1), and the Gemini and GPT models are accessed via their providers’ APIs under the applicable terms.

One evaluation averages $\approx 50\text{k}$ tokens and US\$0.33, so evaluating one provider on one film costs $\approx \text{US\$}3.3$ at R=10, and all 15 movie-provider pairs under US\$50; this is the recurring per-provider cost. For comparison, the complete annotation campaign—in which seven experts evaluated all 15 movie-provider pairs across three films—cost approximately US\$2,800 and took about two calendar weeks. Elapsed time is reported for the campaign rather than divided across concurrently evaluated units. SCR construction is instead a one-time cost amortized over all providers, runs, and ablations: $\approx \text{US\$}1$ per short-film VLM draft, plus one bilingual review pass per title, which took about one hour per 30 minutes of video on this benchmark. That review rate reflects these titles and our current workflow rather than a fixed rate; we retain human review because VLM drafts can still contain errors, and expect the burden to fall as VLMs improve.

Models and generation settings per role. The role-specific settings below state the values actually used, including model defaults.

SCR construction (§2.1). `gemini-3.1-pro-preview`. Scene segmentation, from the subtitles alone, runs at temp 0.1 and top-p 0.7; scene analysis and field generation run at temp 1.0 and top-p 0.95. Both use fixed top-k 64, candidate count 1, thinking level high, and max output 65536, the model cap.

SATS alignment (§2.2). `gemini-3-flash-preview` at temp 0.0, top-p 1.0, top-k 64, candidate count 1, thinking level low, and max output 16384.

TSER evaluation (§4.1, §4.3). `gemini-3.1-pro-preview` at temp 1.0, top-p 0.95, fixed top-k 64, candidate count 1, thinking level high, and max output 65536; the main results and the context ablation share this evaluator and the same R=10 runs.

Translation providers (§3.1). LLM-Line and LLM-Scene use `gemini-2.5-pro` on Vertex AI. No generation config was specified, yielding temp 1.0, top-p 0.95, fixed top-k 64, candidate count 1, and an automatic thinking budget of up to 8192 tokens. Google NMT uses Cloud Translation - Basic (v2)’s default `nmt` model (`general/nmt`), without a glossary or customization.

RUBRIC-MQM baseline (§4.2). `gpt-4.1-mini-2025-04-14` with the released v2.0 defaults — temp 0.7, max tokens 2048, seed 42, a single pass per run — averaged over R=5.

F Extended Related Work

Automatic and Human MT Evaluation

Automatic translation evaluation has progressed from string-based metrics (BLEU (Papineni et al., 2002), chrF (Popović, 2015)) to neural metrics (BLEURT (Sellam et al., 2020), COMET (Rei et al., 2020), MetricX (Juraska et al., 2024), xCOMET (Guerreiro et al., 2024)), which correlate well with human judgment in general translation. Recent WMT shared tasks show that metric performance varies across evaluation granularity and language coverage, while fine-tuned neural metrics remain strong (Freitag et al., 2024; Lavie et al., 2025). Yet subtitle quality depends on non-verbal context, including on-screen characters and relationships, and subtitles condense or reformulate the source for readability; text-only metrics and source fidelity alone therefore cannot determine subtitle quality (Díaz Cintas and Remael, 2007; Burchardt et al., 2016). MQM (Lommel et al., 2014; Freitag et al., 2021) supports fine-grained error-type and severity annotation but requires substantial expert time and cost. To reduce this burden, simplified human-evaluation protocols such as ESA (Kocmi et al., 2024) mark error spans only as minor or major and then assign an overall segment score. We adopt MQM’s category–severity structure as our output schema but anchor every error to source lines and subtitle cues (TSER), joining automatic and expert line-level annotations at the same unit, and use the simple mean of seven evaluators as gold.

LLM-as-a-Judge for Translation Evaluation

Building on general LLM-as-a-judge methods (Zheng et al., 2023; Liu et al., 2023), translation evaluation has adopted this paradigm through GEMBA (Kocmi and Federmann, 2023b), GEMBA-MQM (Kocmi and Federmann, 2023a), and AutoMQM (Fernandes et al., 2023). RUBRIC-MQM (Kim, 2025) uses span-level rubrics to improve performance and NO-ERROR recognition, but only partly addresses label bias and leaves category hallucination unresolved. More recent evaluators use context beyond the sentence: TREQA (Fernandes et al., 2025) scores paragraph-level translations through reading-comprehension questions, whereas LiTransProQA (Zhang et al., 2025) applies a fixed professional questionnaire and returns a paragraph-level score. Neither yields per-source-line judgments, so applying either here would alter its published input, output, and aggregation. We therefore isolate text-only context with the matched Scene-only condition (§4.3). LLM-based evaluation also raises concerns about output stochasticity and possible same-family preference between evaluator and translation models. We reduce the former by averaging R=10 evaluations and probe the latter by excluding the two Gemini-generated providers (Appendix H); Appendix D separately tests whether the baseline gap can be explained by the evaluator backbone alone. Anchored TSERs over fixed SCR/SATS inputs make individual judgments inspectable but do not themselves eliminate model-family effects.

Subtitle, Multimodal, and Document-Level Evaluation

Subtitle translation carries distinctive constraints—cue segmentation, condensation and reformulation for readability, and dependence on on-screen action, expression, and audio (Díaz Cintas and Remael, 2007). Prior work addresses related challenges at different stages. Multimodal MT (Sulubacak et al., 2020) and movie-meta-augmented LLM subtitling (Pramodya et al., 2025) study context use in translation generation. In contrast, audiovisual MT quality research (Burchardt et al., 2016) and SubER (Wilken et al., 2022) concern evaluation, with SubER explicitly incorporating subtitle segmentation and timing. Separately, the limits of sentence-level metrics were noted early in document- and discourse-level evaluation (Voita et al., 2019); discourse phenomena such as pronouns and deixis, tense, and lexical or named-entity consistency are handled by document-level methodology (Maruf et al., 2021), the discourse-aware metric BlonDe (Jiang et al., 2022), and recent work asks whether LLMs exploit document context (Mohammed and Niculae, 2025). Video subtitling is where this long-range dependence meets a movie’s cumulative narrative; rather than feeding video to the evaluator, we fix scene summaries, speaker/addressee, visual cues, delivery, and cumulative narrative (`cumulative_summary`, `previous_scene_detail`) into an explicit, reviewable SCR reused across all providers, repetitions, and ablations.

Position of This Work

In sum, our work quantifies how sentence-level MT metrics behave in this subtitle setting and combines LLM-as-judge with (a) explicit structuring of the video’s non-verbal audiovisual context (SCR), (b) anchored evaluation output (TSER), and (c) repeated-evaluation averaging, yielding evaluations directly joinable with experts’ line-level annotations.

G Line-Level Category and Severity Agreement

The framework’s TSER and the expert annotations share the MQM-style category vocabulary, so labels compare directly without a mapping (lowercase normalization only). The unit is the (movie, provider, source line) where both sides flag an error—detection agreement is reported separately as the line-level F1 in the body, so here we ask only whether the category also matches when detection agrees. Each side’s label is the plurality: for the framework, the most frequent category among errors anchored to the line across the R=10 runs; for the experts, the most frequent among the seven annotators’ labels; ties are excluded (16 framework, 56 expert).

Unlike the category analysis, binary error-detection F1 in the body uses all (movie, provider, source line) units. Expert mean₇ and framework line severities are each binarized as error-positive when greater than zero, and F1 is the harmonic mean of precision and recall; the reported 0.601 is the macro-average of the three per-movie F1 values.

Plurality categories agree 56.1% of the time (chance 23.7%, $\kappa = 0.43$), rising monotonically to 64.6% ($\kappa = 0.52$) on detection-confident subsets—much of the disagreement comes from borderline cases (Table G1). Measured under the same protocol, inter-expert agreement (one annotator vs. the plurality of the remaining six, leave-one-out) averages $\kappa = 0.50$ (per-annotator 0.34–0.56), so the framework’s category agreement lies within the range that individual experts show against their peers’ consensus. Note the structural asymmetry of this comparison: the framework side is a plurality over 10 runs (aggregation reduces noise), whereas each expert LOO label is a single annotator’s.

Confusions concentrate among semantically adjacent categories (Figure G1): style $\leftrightarrow$ mistranslation (31+14 cases) and terminology $\rightarrow$ inconsistency (11). One systematic tendency: the framework rarely uses the punctuation category that experts use frequently, labeling those errors otherwise (recall 0.11).

Severity-level agreement. Line scores are binned into no error plus four contiguous severity bands—minor (1–40), moderate (41–60), major (61–80), and critical (81–100), with cuts at 40/60/80—and compared using linear-weighted Cohen κ . The framework’s severity κ with mean₇ is 0.341. Leave-one-out expert agreement, comparing one annotator’s bins with the mean of the other six, averages $\kappa = 0.408$ (per annotator 0.208–0.564), placing the framework value within the observed expert range. Because the framework side aggregates R=10 runs whereas each leave-one-out target is one expert annotation, we do not treat the expert mean as a hard ceiling or attach a qualitative agreement label.

Comparison	n	agree	chance	κ
Framework vs expert plurality — any (both flag ≥ 1)	369	0.561	0.237	0.425
— confident (framework $\geq 5/10$ runs, experts $\geq 2/7$)	261	0.613	0.251	0.483
— strict (framework $\geq 5/10$ runs, experts $\geq 4/7$)	240	0.646	0.262	0.520
Expert leave-one-out mean (1 vs plurality of 6)	3,341	0.626	—	0.497
Expert pairwise mean (21 pairs)	—	0.525	—	0.363

Table G1. Category agreement by subset, with expert–expert baselines

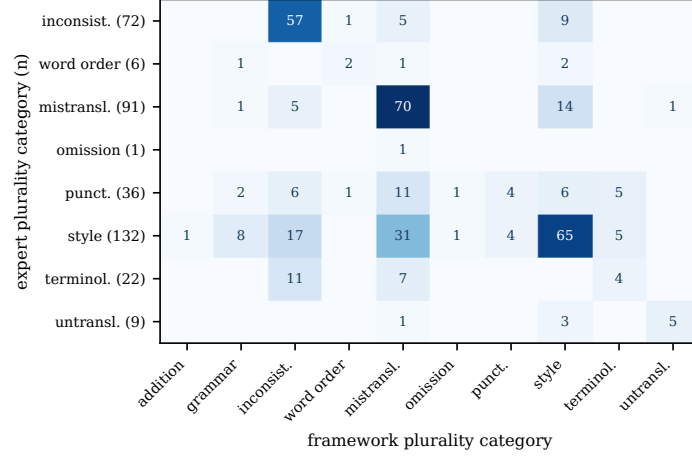


Figure G1. Category confusion between expert and framework plurality labels (primary subset; expert rows with $n \geq 1$; cell counts annotated). Confusions concentrate among adjacent categories (style $\leftrightarrow$ mistranslation, terminology $\rightarrow$ inconsistency), and the framework under-uses punctuation.

H Detailed Result Tables

Tables H1 and H2 give the exact values plotted in Figures 3 and 4, and Table H3 lists the per-movie provider rankings behind the sign reversal described in §4.2.

Target	sys ρ (τ)	scene ρ (τ)	line ρ (τ)	ED	SF	ToS
SCR-Eval	+0.967 (+0.933)	+0.849 (+0.690)	+0.666 (+0.541)	0.90	1.00	1.00
RUBRIC-MQM	+0.267 (+0.200)	+0.029 (+0.006)	+0.218 (+0.157)	0.40	0.30	0.10
MetricX	−0.600 (−0.467)	−0.146 (−0.088)	+0.179 (+0.126)	−0.60	−0.70	−0.50
xCOMET	−0.500 (−0.400)	−0.372 (−0.245)	+0.121 (+0.086)	−0.60	−0.70	−0.20

Table H1. Baseline agreement with the expert mean (mean_7): Spearman ρ , with Kendall τ_b in parentheses; scene-level length-normalized; sign-aligned (exact values for Figure 3). Per-movie columns are system-level ρ . τ_b tracks ρ at every level—same sign and ordering—so the gap reflects the two coefficients’ differing scales, not weaker agreement. ED = *Elephants Dream*, SF = *Sprite Fright*, ToS = *Tears of Steel*

Condition	system ρ	scene ρ	line ρ	ED (sys)	SF (sys)	ToS (sys)
Full (= main runs)	0.967	0.849	0.666	0.90	1.00	1.00
Scene-only	0.900	0.718	0.481	0.90	0.90	0.90
Line-only (per-line)	0.700	0.495	0.382	0.60	0.70	0.80

Table H2. Context-ablation agreement (mean_7), same evaluator at matched R=10 (exact values for Figure 4)

Provider	Expert	SCR-Eval	MetricX	xCOMET	RUBRIC-MQM
Human	1, 1, 1	1, 1, 1	3, 5, 4	3, 4, 4	2, 3, 3
LLM-Scene	3, 2, 2	2, 2, 2	4, 3, 2	4, 3, 1	4, 2, 2
Comm-AI	2, 3, 3	3, 3, 3	5, 2, 5	5, 5, 5	3, 4, 5
LLM-Line	4, 4, 4	4, 4, 4	1, 4, 3	1, 2, 3	1, 1, 1
Google NMT	5, 5, 5	5, 5, 5	2, 1, 1	2, 1, 2	5, 5, 4

Table H3. Provider rankings (1 = best) by each scorer; sign-aligned. Each cell lists the rank in the three movies, in the order *Elephants Dream*, *Sprite Fright*, *Tears of Steel*. Across all three movies, the learned QE metrics rank Human below and Google NMT above their expert positions, whereas SCR-Eval tracks the expert rankings much more closely.

Same-family robustness of the main agreement. Our LLM evaluator (gemini-3.1-pro-preview) and the LLM-based providers LLM-Line and LLM-Scene (gemini-2.5-pro) share the Gemini product family but come from different model generations,

raising a same-family circularity concern. Excluding the two Gemini-based providers and recomputing agreement over Human, Comm-AI, and Google NMT does not reduce agreement; instead, it strengthens at every level: system-level $\rho = 0.967 \rightarrow 1.000$, scene-level $0.849 \rightarrow 0.854$, and line-level $0.666 \rightarrow 0.686$. Thus, within this limited provider subset, same-family providers do not appear to account for the high agreement.

I Expert Annotation Protocol (Rubric)

The guide distributed to the seven evaluators (originally in Korean; faithfully summarized):

Purpose. Compare the source utterances and the translated subtitles within each movie scene and identify and assess translation errors under MQM. Evaluators watch the video and judge, for each source line, whether the subtitle conveys the source meaning and the speaker’s intent appropriately and reads naturally as a Korean subtitle.

Method. Evaluation is per line, taking the scene and the surrounding utterances into account; in line with subtitling norms, the criteria are meaning transfer in scene context and naturalness in the target language rather than word-for-word correspondence.

Categories. Eleven MQM categories (addition, mistranslation, style, omission, untranslated text, grammar, inconsistency, punctuation, source issue, incorrect word order, terminology), with definitions matching in substance those given to the framework’s evaluator ([Appendix J](#)). The Style category follows RUBRIC-MQM v2 in flagging renderings that are unnatural or overly literal even when the meaning is preserved, which matters for subtitle fluency. The framework’s evaluator uses these same eleven error categories plus an explicit No-Error label, so it can affirmatively mark an error-free span instead of being forced to choose an error category; experts need no such label because they leave error-free lines blank—hence eleven categories for experts and twelve for the evaluator.

Severity. A 1–100 scale, judged by the impact of the error on meaning transfer and understanding: 1–40 minor (little impact on meaning; awkward phrasing or minor mechanical issues), 41–60 moderate (the message is mostly understood, with some loss of meaning or nuance, reduced clarity, or partial misunderstanding possible), 61–80 major (part of the core meaning is distorted, or information important to understanding the scene is conveyed incorrectly), 81–100 critical (the source’s core intent or facts are barely conveyed, or could be understood in reverse).

Recording. Per line: if a line contains multiple errors, only the single error with the greatest impact on meaning transfer is recorded, with its category, severity, and a brief comment explaining the decision; error-free lines are left blank.

J Prompts

Except for auxiliary prompts used only to handle exceptional alignment cases, all prompts used by the pipeline are reproduced below: the TSER evaluation prompt (Figure J1) with the expansions of its category and severity placeholders (Figures J2–J3), and the SATS selector prompt (Figure J4). An error-free span may be returned either as an empty array or as a No-Error record (Figure J2’s twelfth category); both carry no severity and are excluded from error counts and aggregation. Within [Previous Scene Memory], the `prev_cum_block` and `prev_detail_block` placeholders carry the cumulative plot summary and the previous scene’s detail, while `speaker_memory_block` carries the already-translated cues from all preceding scenes whose source line shares a speaker-addressee pair with the current scene; additions, cues aligned to no source line, and lines with multiple speakers or addressees are excluded.

SYSTEM INSTRUCTION

persona + Evaluation Guide; identical across scenes

You are an expert MQM evaluator specializing in film subtitle translation, fluent in both `{{source_lang}}` and `{{target_lang}}`.

Read this [Evaluation Guide] first. The scene context and the source/translated lines to evaluate are provided in the user message; evaluate the translation strictly according to this guide.

[Evaluation Guide]

– Evaluation scope:

Treat the entire scene as one semantic context, but identify errors, assign categories, and score severity at the corresponding line level.

- Principles for film subtitle translation:
 - Subtitle translation should prioritize meaning in scene context and naturalness in the target language, rather than word-for-word correspondence.
 - Do not mark condensation, merging, or restructuring as errors by default if they improve subtitle readability or naturalness.
 - Even if some source elements are not explicitly translated, do not mark them as errors if the core meaning, speaker intent, and scene understanding are preserved.
 - Do not mark pronoun-to-noun shifts, added relational cues, or explicit wording of visually obvious information as errors by default when they make the subtitle more natural in the target language.
 - However, allowed condensation, merging, omission, or explicitness does not excuse unnatural wording.
- If the meaning is preserved but the subtitle sounds unnatural, overly literal, translation-like, or explanatory in the target language, identify it as an error.
- How to read the inputs (context only; none of these need to appear in the translation):
 - [Previous Scene Memory]: narrative context. Check that the translation fits the story flow.
 - Visual/Audio Context (under each source line): how the meaning and intent are expressed in the scene.
 - [speaker → addressee][delivery]: speaker-addressee relationship, intent, and delivery style. Flag the error if the relationship, tone, or delivery shifts when read with the scene.
 - SATS terms and matched_source_uids: reference and alignment hints only - verify alignment from the text and scene, and treat SATS wording as non-binding. For merged lines (e.g., [11 - 2:8,2:9]), judge how the source meanings are integrated.
- How to judge:
 - Evaluate by meaning transfer, not word-for-word matching: is the core meaning, speaker intent, and target-language naturalness preserved?
 - For each error, assign the single most appropriate MQM category and a severity score (integer, 1-100).
- MQM error category definitions:
{mqm_definitions}
- Severity scoring guidelines (RubricMQM 100 Scale):
{severity_guidelines}
- For reason, briefly explain which phrase, word, or sentence is problematic and why.

USER MESSAGE

scene data + task; varies per scene

```
[Previous Scene Memory]
{prev_cum_block}{prev_detail_block}{speaker_memory_block}

[Current Scene]
- scene_id: {scene_id}
- Summary: {scene_summary}
{settings_block}{visual_block}

[Lines]
<Source: {source_lang} Lines>
- Format: [{{line_uid}}] {{text}} [{{speaker}} → {{addressee}}][{{delivery}}]
- Example: [2:1] Get down! [SWAT officer → suspects][shouting, commanding]
- Data:
{source_block}

<Translation: {target_lang} Lines>
- Format: [{{ts_index}}-{{matched_uid}}] {{text}}
- Example: [11-2:8,2:9] translated subtitle text
- Data:
{ts_block}

[Task]
Based on the source dialogue and translated subtitles above, evaluate the
translation and identify all MQM errors, following the [Evaluation Guide] provided
in the system instruction.

[Output]
```

- Return only a JSON array. If there are no errors, return an empty array: [].
- anchors.scene_id must be set to {scene_id}.
- anchors.source_line_uids must include the relevant source line_uid whenever possible (one or more). For addition errors unrelated to any source line, an empty array is allowed.
- anchors.ts_indices must include the relevant translated subtitle ts_index whenever possible (one or more). For omission errors, specify nearby ts_indices where the omission appears to occur. If no relevant ts_index can be identified, an empty array is allowed.
- category must exactly match one of the 12 MQM categories defined in the guide.
- reason must be concise.

Figure J1. The TSER scene-level evaluation prompt (deployed in English), shown as the system instruction followed by the per-scene user message, with input placeholders in braces. The JSON output structure—the keys `span`, `original`, `translated`, `category`, `severity`, `reason`, and `anchors`—is enforced by an API response schema rather than enumerated in the prompt. The Line-only ablation (§4.3) uses a pairwise variant with the same principles, category definitions, and severity scale, but with all scene and memory context removed.

* Addition: This error occurs when extra content not in the original text leads to repetition, unnecessary details, or redundancy, distorting the message and potentially confusing readers or diverging from the original intent.

* Mistranslation: This error involves inaccurate translation or interpretation, often due to poor word choice, leading to a message that strays from the original content’s meaning and intent.

* Style: This error occurs when the translation sounds unnatural, awkward, or overly literal in the target language, reducing fluency and readability even when the meaning is generally accurate.

* Omission: This error occurs when essential elements from the original text are missing in the translation, resulting in incomplete meaning and loss of critical information or nuances needed for full understanding.

* Untranslated Text: This error refers to parts of the source language that remain in the translation without being converted, resulting in an incomplete or inaccurate translation. Source-language items such as proper nouns or brand names that are naturally retained in the target language are not considered errors.

* Grammar: This error involves incorrect grammar, such as tense, verb form, pronouns, agreement, articles, or gender, disrupting fluency and coherence and risking misunderstandings or credibility loss.

* Inconsistency: It refers to variations in style, structure, wording, or expression within the same context that undermine the fluency and readability of the translated text.

* Punctuation: This error stems from incorrect punctuation, quotation marks, spacing, or hyphenation, disrupting clarity and reading flow, and potentially causing misunderstandings.

* Source Issue: It refers to any problematic elements originating from the source text (i.e., ambiguities, grammatical errors, or unclear phrasing) that hinder accurate translation and lead to misunderstandings.

* Incorrect Word Order: This error occurs when the translation uses unnatural word order or phrasing in the target language, which can alter the meaning, clarity, or emphasis, leading to awkward or confusing text.

* Terminology: This error occurs when a term or word choice is contextually inappropriate or inconsistent, leading to misaligned meaning or intent and potentially causing confusion or lack of clarity, especially with technical or specialized terms.

* No-Error: This category denotes a flawless translation, accurately conveying the source text’s meaning, tone, nuances, consistency, and style with clarity, cultural appropriateness, and grammatical accuracy in the target language.

Figure J2. Expansion of the `mqm_definitions` placeholder in Figure J1: the twelve MQM category definitions.

-
- * 10: Negligible impact; intent remains fully intact.
 - * 20: Minor wording issue; message preserved.

- * 30: Slight clarity issue; intent is clear.
- * 40: Slight misunderstanding possible, but message understood.
- * 50: Affects clarity; some interpretation needed.
- * 60: Distorts part of the message; ambiguous intent.
- * 70: Alters message significantly; misunderstanding likely.
- * 80: Core meaning is misinterpreted.
- * 90: Serious miscommunication; original intent lost.
- * 100: Translation is misleading and completely unfaithful.

Figure J3. Expansion of the severity_guidelines placeholder in Figure J1: the 1–100 severity scale.

```

You are a subtitle aligner.
Goal:
- Select only the translated subtitle lines that correspond to the provided "source utterance list".
- Judge solely by source correspondence. Do not use scene summaries, key phrases, or contextual inference.
Rules:
- Selection criterion: select a translated subtitle line only when it explicitly contains a translation of one or more source utterances.
- No context/inference: do not supplement or guess meaning that is not present in the source.
- Order: the ts_index values of the selected subtitles must be in ascending (monotonically increasing) order.
- Additions: on-screen text, sound effects, signs, etc. that are absent from the source may be marked is_addition=true.
- Uncertainty: if confidence is low, exclude the line or mark it with a low confidence.
- Return matched_source_indices as 0-based indices (by the order of the source utterance list).
  If nothing matches, leave it as an empty array ([]); do not force a match.
- Do not include text (subtitle strings) or matched_source_text in the output JSON; return only ts_index, matched_source_indices, and is_addition.
- Output only the specified JSON. No explanations or comments.
Input:
- Scene time: [{scene_start_ms}, {scene_end_ms_plus10}]
- Source utterances (FULL, in order):
{source_block}
- TS candidates (in the time window):
{ts_block}
Output (JSON schema): output JSON containing only the following keys.
{schema}

```

Figure J4. The SATS cue-to-scene selector prompt used in §2.2 (Appendix C); an English rendering is shown.

K Provider Label Mapping

The paper itself refers to providers only by the labels in Table K1. In the released data files (scope under [Ethics](#)), providers appear under the listed identifiers; the commercial solution is not part of the released data.

Paper label	Released data id	Description
Human	hand_written	Human-finalized translation (§3.1)
LLM-Line	gemini_line	Line-level LLM translation: each source line translated independently (gemini-2.5-pro)
LLM-Scene	gemini_scene	Scene-level LLM translation: each scene’s source lines translated together, with no context beyond the scene’s own lines (gemini-2.5-pro)
Comm-AI	— (not released)	Global commercial AI subtitle solution, withheld per Ethics
Google NMT	google_translate	General-purpose NMT (non-LLM)

Table K1. Provider labels used in the paper and the corresponding identifiers in the released data files

See [Appendix E](#) for models and generation settings, where applicable.