

Dubformer at WMT26 General MT: Translating the Video, Not the Transcript

Sergey Dukanov
Dubformer
sergey@dubformer.ai

Maria Tsyurupa
Dubformer
mashats@dubformer.ai

Anton Dvorkovich
Dubformer
dvorkanton@dubformer.ai

Abstract

We describe the system Dubformer entered in the WMT26 General MT shared task, an unconstrained system built on our video-dubbing pipeline. It rests on two commitments. A document is never translated as raw text: it is first decomposed into a structure chosen for its domain — scenes carrying speaker and audio-visual context for spoken dialogue, markup-masked segments for web documents, value-only segments for localization resources — so that non-linguistic material is protected by construction and everything needed for consistency is in front of the model at once. And no translation is taken on trust: every segment is checked by a second model from a different family for hallucination and for drift from the meaning of the source, and a segment it rejects is re-translated with the reason attached rather than filled in from a fallback.

1 Introduction

The system has two paths. Spoken dialogue is translated by our video-dubbing pipeline, driven directly; its central commitment is that a spoken-dialogue document is translated *from the video* rather than from a transcript of it. Web documents, social posts and localization resources need none of that machinery, so for them we built a thin harness of our own, shaped by what running that pipeline had taught us: protect structure by construction, and check every translation before it ships.

The task setup follows the previous edition of the shared task (Kocmi et al., 2025). §2 states what the two paths share and where they diverge. This report describes the design and the reasoning behind it. It reports no results on the shared-task test set; where it appeals to evidence, that evidence is internal and is marked as such.

2 What the domains share, and where they differ

Two things are common to both paths (Figure 1). No part of a document is translated in isolation: a short document goes to the model whole, a scene of a longer one carries the document’s context and its neighbours. Translating a document as a whole rather than sentence by sentence is known to help, and to leave critical errors that context alone does not remove (Karpinska and Iyyer, 2023) — which is what the second property is for. And every line is written by the model from the source, never edited from a machine-translated draft, then checked by an independent auditor whose rejections are re-translated rather than replaced from a fallback (§5).

What differs is how a document is prepared for that call: how it is cut into segments that can be re-assembled afterwards, what context accompanies them, and which instruction the model is given. Spoken dialogue is cut into scenes over a transcript we produce ourselves — a clip is usually short enough to be one scene — carrying speaker identity, per-speaker gender and a note taken from the video and its audio; the prompt is the dubbing engine’s own, plus a per-target directive fixing language variety, script and region (§3). In the other three domains the instruction shipped with each blindset row is authoritative and reaches the model verbatim on top of a fixed base translator prompt; web documents and social posts are cut into text segments with inline markup masked, social posts also carrying screenshots of the page, and localization resources into translatable values only, their keys exposed as context but never translated (§4). Domains are not labeled in the blindset, so the path is recovered from marker phrases in each row’s instruction. Table 1 lists the model used at each stage.

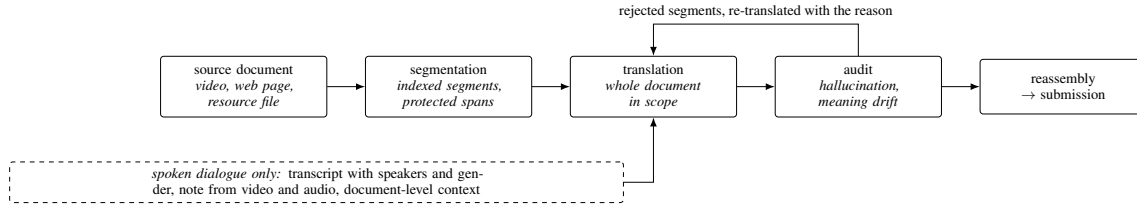


Figure 1: The pipeline is the same for every domain; what changes is how a document is cut into segments and, for spoken dialogue only, the extra context assembled before translation. A rejected segment goes back through the same translation call rather than to any fallback.

Stage	Model
transcription	Scribe v2
speakers	ECAPA-TDNN clustering
gender	SVM over the same embeddings
note from video and audio	Gemini 3.1 Pro
document context	Claude Opus 4.8
scene translation	Claude Sonnet 4.6
document translation	Claude Sonnet 4.6
audit	Gemini 3.1 Pro

Table 1: Models by stage: everything above *document translation* is the spoken-dialogue path. The audit runs the same model on both paths, configured more strictly for speech. The strongest model goes on document-level context, which every later stage inherits. Speaker embeddings are ECAPA-TDNN (Desplanques et al., 2020) as packaged by SpeechBrain (Ravanelli et al., 2021); the gender head is an SVM we trained on Vox-Celeb1 (Nagrani et al., 2017).

3 Spoken dialogue

The video is the input. We transcribe the video ourselves rather than take the transcript shipped with the blindset. Recognition stays with a dedicated ASR system: the multimodal model that later reads the scene is used to catch where a transcript is wrong, never to produce it. The transcription is then ours to control, which matters most for timing: speaker segmentation and the per-speaker gender attribution derived from the audio are computed against those timings, and several of our target languages need gender to get agreement right — a known difficulty when the source language does not mark it, and one where the audio carries information the transcript does not (Bentivogli et al., 2020).

Disfluencies are removed at translation time, not at transcription time. Transcription runs in verbatim mode — our ASR has a flag that strips filled pauses, repetitions and false starts, and we deliberately leave it off. The disfluency policy is expressed instead as instructions to the translator,

which drops them while keeping genuine interjections, since those carry attitude the dialogue would lose. The principle is to separate *what was heard* from *what gets written*: the transcript remains a faithful record of the audio, and the cleanup becomes a translation decision made with the surrounding context available, rather than a transcription setting applied blindly at the word level.

What the picture and the audio add. Before translation, a multimodal model both watches and listens to each scene and produces a note about it: the setting, who is on screen and their apparent gender, the tone, register cues, and perception notes wherever the recording resolves something the text leaves open — deixis, an addressee, whether a line is sarcastic. Hearing the audio and not only the transcript, it also flags where the two disagree, so a recognition error can be corrected rather than translated. The translator is told to prefer these notes over the text wherever they disagree about gender, address or deixis, which is what makes the video the source rather than an illustration of it. Video-guided translation is an established line of work (Sulubacak et al., 2020), and the visual signal has been shown to pay off precisely where the textual input is impoverished (Caglayan et al., 2019) — the condition an automatic transcript creates. In our internal evaluations on comparable material, adding the audio-visual note gave a statistically significant improvement.

Document context before scenes. A document-level pass first extracts what a translator needs to stay consistent across scenes: a summary, the cast with gender and forms of address, and terminology, so a character is named and characterized the same way throughout. The pass earns its place on long material; here the clips run under a minute, about one scene for our splitter, so it mostly acts as an explicit reasoning step before any line is translated.

Written from scratch, not post-edited. Every line is generated by the model from the source itself — for spoken dialogue, from the source text, the video and the language-variety directive, and nothing else. The engine can instead show the model a machine-translated draft to post-edit, a mode effective in its own right (Raunak et al., 2023); that path is disabled for every target, including the lower-resource ones where it is normally offered. In our experience a baseline anchors the model to someone else’s phrasing and narrows the range of renderings it will consider. We observed this on high-resource pairs only; the path is disabled everywhere for consistency, not because we measured its effect on the lower-resource ones.

4 Web, social and software documents

A web page or a resource file is a handful of paragraphs or interface strings, and the three domains share one thin harness. Its per-domain parameters are the decomposition and, for two of them, one extra input each: a framing line for software, the page screenshots for social.

One call per document. The whole document is translated in a single call whose input and output are index-keyed lists of segments. That call is the document-context pass: the model sees every sibling segment at once, so it can hold terminology, register and named entities together, and indexing keeps the output re-insertable into the original structure. The decomposition is supplied as a pair: a gate deciding whether a candidate preserved the document’s structure, and a renderer producing the plain text the auditor compares. A domain is a decomposition, not a fork of the pipeline.

Mask markup; expose placeholders. The two decompositions make opposite choices, and the contrast is the point. For marked-up documents, inline tags are replaced by opaque markers before the model sees the text: it places a marker and never sees a URL or an attribute value in the text it is given (a screenshot, where one is attached, is a separate channel), and the gate requires the marker multiset to survive and to remain well nested. For localization resource files, placeholders such as {{count}} are left *literal*: their names are exactly the context the model needs for the text around them — whether a value is numeric, whether two placeholders are a pair wrapping a link, what an ambiguous short label refers to. The

test is whether the hidden material carries information the surrounding translation depends on: a URL does not, a placeholder name does. The guarantee is the same one obtained differently — the gate requires the placeholder multiset to survive, and a document-level check re-verifies it after re-insertion. The principle we would carry forward: *mask what the model must not read, expose what it must reason about, and in both cases enforce structure by conservation rather than by trust.*

5 Gating: audit, corrective retry, no fallback

An independent auditor. Every translation is compared to its source by a second model, in the manner of reference-free LLM error detection (Kocmi and Federmann, 2023), but used as a gate rather than as a metric. It flags material factual divergence — flipped polarity, altered numbers or entities, dropped qualifiers, invented claims, swapped roles — while leaving idiomatic rephrasing and register shifts alone. The auditor is deliberately from a different provider than the translator. In our internal evaluations of automatic judging, models scored generations from their own family above others’, so an auditor drawn from the translator’s family risks certifying its blind spots; the same self-preference effect is reported in the literature (Panickssery et al., 2024). That independence has one hole we did not close: the note taken from the video comes from the auditor’s own model, and the auditor is told to treat it as ground truth about the scene, so where a note is wrong the translation follows it and the audit is the last place that would catch it.

Audit reliability depends on how much it judges at once. The same auditor model misses errors in a long list of segment pairs that it catches in a short one. We chose the batch size empirically on our internal test sets, as the setting with the best recall; the explanation we assume — that attention dilutes across the batch — is an interpretation. We accept the cost in latency and calls.

The retry is surgical and corrective. A rejected segment is re-translated on its own, through the same call and under the same instruction, with accepted segments left untouched; only the refilled ones are re-audited. In the speech path the whole scene is re-sent with it, so its context is unchanged. Crucially, the retry is not blind: the auditor’s stated

reason travels back into the prompt for that segment, together with an instruction that for it, fidelity to the source outranks the document’s own terminology list. In one case, automatic terminology extraction had normalized a character’s name to another used elsewhere in the document; the auditor flagged it, a blind retry reproduced the error exactly, and only the corrective retry fixed it.

No MT fallback. The engine can fill an unresolved segment from a generic third-party MT system; here it is disabled everywhere. The reason is fluency. In our experience generic MT output is markedly less fluent than the model-generated text around it, and in dialogue the damage is worse than an average would suggest: a single flat, literal line between lines that read like speech is more conspicuous than the same line would be in a uniformly machine-translated document. We would rather degrade to a worse translation than to a different kind of text. What replaces the fallback differs between the paths: an unresolved segment fails the document in spoken dialogue, which is then re-run — nothing reached that point in this run — while in the document domains the last structurally valid output ships, flagged for a later re-attempt. Both keep every document in the submission, and because the audit was configured more strictly for speech, its rejection rate is not comparable across paths.

References

- Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. [Gender in danger? evaluating speech translation technology on the MuST-SHE corpus](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6923–6933, Online. Association for Computational Linguistics.
- Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 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 (Long and Short Papers)*, pages 4159–4170, Minneapolis, Minnesota. Association for Computational Linguistics.
- Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. [ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification](#). In *Proc. Interspeech 2020*, pages 3830–3834.
- Marzena Karpinska and Mohit Iyyer. 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*, pages 419–451, Singapore. Association for Computational Linguistics.
- Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakounga, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. [Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.
- Tom Kocmi and Christian Federmann. 2023. [GEMBA-MQM: Detecting translation quality error spans with GPT-4](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.
- Arsha Nagrani, Joon Son Chung, and Andrew Senior. 2017. [VoxCeleb: A large-scale speaker identification dataset](#). In *Proc. Interspeech 2017*.
- Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. [LLM evaluators recognize and favor their own generations](#). In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*.
- Vikas Raunak, Amr Sharaf, Yiren Wang, Hany Awadalla, and Arul Menezes. 2023. [Leveraging GPT-4 for automatic translation post-editing](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12009–12024, Singapore. Association for Computational Linguistics.
- Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, and 2 others. 2021. [SpeechBrain: A general-purpose speech toolkit](#). *arXiv preprint arXiv:2106.04624*.
- 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.