LEARNING TIMELINE-PRESERVING SPEECH-TO-SPEECH TRANSLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Speech-to-speech translation (S2ST) increasingly preserves *what* is said and *how* it sounds, but not necessarily *when* speech and silence occur. For dubbing and other temporally grounded applications, matching total duration is not enough: speech and pauses must also occur at corresponding moments. We formulate *timeline-preserving S2ST*, which aims to preserve the source endpoint and salient internal pauses on a shared absolute timeline. We present TRANSDUBBER, an end-to-end autoregressive system that learns this temporal structure directly from aligned target waveforms, without inference-time frame-level timing control or post-hoc temporal correction. To provide this supervision, we construct 642K exactly duration-matched English–Chinese speech pairs by translating and synthesizing target speech on source-defined temporal scaffolds. Training combines a progressive supervised curriculum with verifier-guided reinforcement learning to balance translation quality and temporal alignment. On a 4,353-utterance test set spanning read and spontaneous speech, TRANSDUBBER increases the proportion of translations within 2% of the source duration from 25.6% to 61.5% and pause-alignment F1 from 0.118 to 0.550 compared with UniSS, while maintaining competitive translation quality. Our analysis shows that aligned supervision establishes timeline following, while reinforcement learning refines temporal precision and partially offsets the translation-quality cost of strict synchronization. Together, these results demonstrate the effectiveness of our data and training framework for timeline-preserving S2ST. We will release the training data, model checkpoints, and code upon publication. An anonymous demo is available at https://anonymous.4open.science/w/transdubber-demo-38E6.

1 INTRODUCTION

In a dubbed scene, silence can be part of the dialogue. A character may pause to breathe, hesitate, or wait for a reaction; a translation that speaks through that interval can disrupt the timing of the scene even when its words are accurate. Psycholinguistic research shows that prosodic boundaries guide listeners' online interpretation of sentence structure (Steinhauer et al., 1999), motivating attention to how speech is organized as well as what it conveys. In audiovisual translation, phrase and pause alignment is also tied to the visual context, with different synchronization requirements for on-screen and off-screen speech (Virkar et al., 2022). We therefore focus on settings in which translated speech must fit an existing scene or event sequence.

Modern speech-to-speech translation (S2ST) systems increasingly preserve semantic content, speaker identity, and expressive characteristics (Jia et al., 2019; Lee et al., 2022; Barrault et al., 2023a;b; Cheng et al., 2026). Temporal constraints have also been studied at both the translation and speech-generation levels: isometric MT encourages source–target text-length agreement (Lakew et al., 2022), while duration-based translation uses source-speech duration to guide translated text (Won et al., 2025). Prior work addresses internal temporal structure: isochrony-aware MT jointly translates and projects pause markers (Tam et al., 2022), and TransVIP conditions speech generation on duration and frame-level voice activity through an explicit isochrony-control module (Le et al., 2024). Matching the source endpoint alone does not determine this internal timeline: two translations can have exactly the same duration while placing pauses and speaking regions at entirely different moments. Building on these efforts, we study *timeline-preserving S2ST*, in which target speech aims to preserve the source utterance endpoint and salient internal speech–silence events on a shared absolute timeline,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

as illustrated in Fig. 1. Our focus is on how an autoregressive S2ST model can learn this behavior directly from target waveforms aligned to the source timeline, without an inference-time frame-level isochrony controller or post-hoc padding, truncation, or global time stretching. This is an application-specific constraint, rather than an assumption that source-like timing is universally optimal: studies of professional dubbing emphasize translation quality and vocal naturalness and suggest a more conditional role for timing constraints (Brannon et al., 2023).

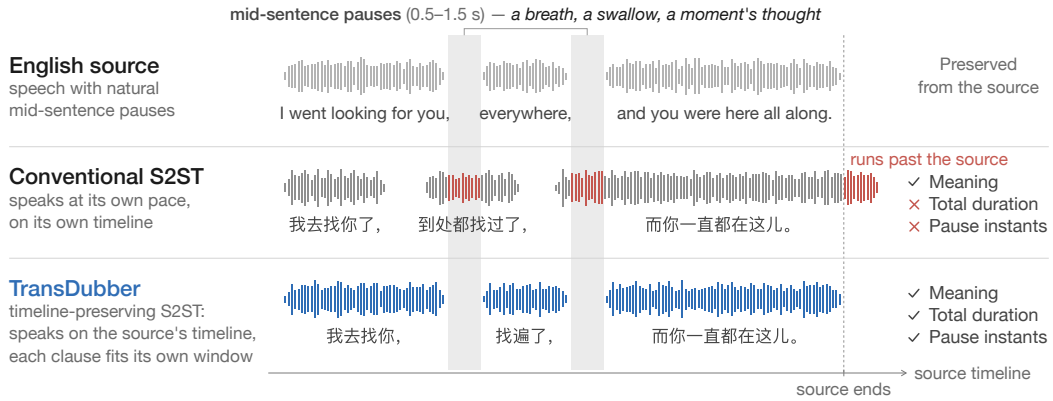


Figure 1: Illustration of timeline-preserving speech translation. Conventional S2ST may shift pauses or overrun the source endpoint; TRANSDUBBER instead preserves meaning while aligning translated speech with the source speaking windows, salient pauses, and utterance endpoint.

Learning this behavior requires supervision that standard parallel speech corpora do not provide. Independently recorded translations naturally follow the phrasing and rhythm of the target language rather than the speech-activity pattern of the source. To create the missing supervision, we develop a scalable source-canvas construction pipeline. We extract a speech-and-pause scaffold from each source utterance, translate individual speech regions under local duration budgets while retaining utterance-level semantic context, synthesize speaker-conditioned target speech, and place the accepted segments back into their corresponding source windows. The resulting targets preserve the source endpoint and selected internal pauses by construction, providing explicit supervision for both pause-bearing and pause-free timelines.

On this data, we train TRANSDUBBER with a progressive multi-stage curriculum. The model first acquires pause-aware recognition and translation, then learns natural target-speech generation, and finally learns to generate translated speech on the source timeline. We subsequently apply verifier-guided policy optimization with separate objectives for translation quality, global duration agreement, and local pause alignment. Rather than collapsing these signals into a single reward before optimization, we normalize them independently and assign semantic feedback to translated-text tokens while assigning temporal feedback to speech-codec tokens. On our English-to-Chinese test set, TRANSDUBBER improves Dur@2% from 25.6% to 61.5% and pause F1 from 0.118 to 0.550 relative to UniSS, with a similar COMET score.

Our contributions are as follows:

- **Timeline preservation as a learned generation behavior.** We formulate timeline-preserving S2ST and train TRANSDUBBER to reproduce the source endpoint and salient internal pauses directly through autoregressive generation, without an inference-time frame-level timing controller or post-hoc temporal correction.
- **A 642K-pair timeline-preserving corpus.** Our scalable pipeline combines duration-budgeted translation, constrained speaker-conditioned synthesis, and source-canvas rendering to produce exactly duration-matched English–Chinese speech pairs with 298K supervised pause spans.
- **Progressive training with fine-grained RL.** We progress from pause-aware text grounding to natural speech generation and timeline-preserving fine-tuning, followed by verifier-guided RL. Quality, duration, and pause rewards are normalized independently, with semantic and temporal feedback assigned to text and speech tokens, respectively.

- **State-of-the-art timeline-preserving S2ST.** TRANSDUBBER substantially outperforms strong S2ST baselines in strict duration fidelity and local pause alignment while maintaining competitive translation quality.

2 CONSTRUCTING TIMELINE-PRESERVING SUPERVISION

We construct the supervision data in three stages. Stage 1 derives a source-side speech-and-pause scaffold from each English utterance, transcribes the resulting speech regions, and separately produces a natural Chinese translation. Stage 2 uses the natural translation to construct complete Chinese speech at its natural speaking rate. Stage 3 reuses the Stage 1 temporal scaffold and region-level English transcripts, but translates these transcripts again under local duration constraints before synthesizing and rendering the target speech on the source timeline.

Stages 1 and 2 are constructed from the 8–25 s source pool drawn from Emilia-EN (He et al., 2024) and GigaSpeech (Chen et al., 2021). Stage 3 uses eligible utterances from this pool and additionally introduces 2–8 s utterances to improve short-form coverage. Detailed construction and filtering settings are provided in Appendix B.

2.1 STAGE 1: TEXT GROUNDING

Source-derived temporal scaffold. For each source waveform, we first apply source-side VAD (Silero Team, 2021) to identify speech spans and the internal non-speech gaps between adjacent spans. Internal gaps of at least 0.5 s are selected as pause boundaries, and the utterance is partitioned into speech regions only at these selected gaps. The pause locations are therefore derived directly from the source waveform rather than from punctuation or translated text.

Segment-level ASR (Shi et al., 2026) is applied to the resulting speech regions, producing English transcripts $e_{1:N}$. For utterances with selected internal pauses, the region transcripts are joined using a dedicated textual marker. Each marker corresponds to a selected source-side pause interval.

$$e^{\text{pause}} = e_1 \texttt{ <pause> } e_2 \texttt{ <pause> } \cdots \texttt{ <pause> } e_N$$

Utterances without a selected internal pause contain no <pause> marker.

Separately, Hy-MT2-30B-A3B (Zheng et al., 2026) translates the complete English utterance into a natural Chinese sentence z^{nat}. This translation is generated at the utterance level without a duration budget, character-count constraint, region-level alignment requirement, or pause-placement constraint. Unlike the English ASR representation, z^{nat} contains no <pause> token.

Stage 1 therefore yields ASR pairs $(x, e^{\text{pause}}) \in \mathcal{D}_{\text{ASR}}$ and AST pairs $(x, z^{\text{nat}}) \in \mathcal{D}_{\text{AST}}$. The temporal scaffold and $e_{1:N}$ are retained for Stage 3, while z^{nat} also serves as the synthesis text for Stage 2.

2.2 STAGE 2: NATURAL-PACE SPEECH SUPERVISION

Stage 2 uses the natural Chinese translation z^{nat} from Stage 1 as synthesis text. For each source utterance, VoxCPM2 (Zhou et al., 2026) synthesizes a complete Chinese waveform $y^{\text{nat}} = \text{VoxCPM2}(z^{\text{nat}}, r_x)$, conditioned on source-speaker reference audio r_x.

The synthesized waveform follows the natural phrasing, rhythm, and duration of the Chinese translation. No source-duration constraint, segment-level duration budget, time-scale modification, or source-canvas rendering is applied.

2.3 STAGE 3: TIMELINE-PRESERVING SPEECH SUPERVISION

Stage 3 reuses the Stage 1 temporal scaffold and English region transcripts, but generates new duration-constrained translations rather than using z^{nat} (Fig. 2). It includes a *no-pause* subset without selected internal pauses and a *pause-aligned* subset preserving selected pauses of approximately 0.5–1.5 s.

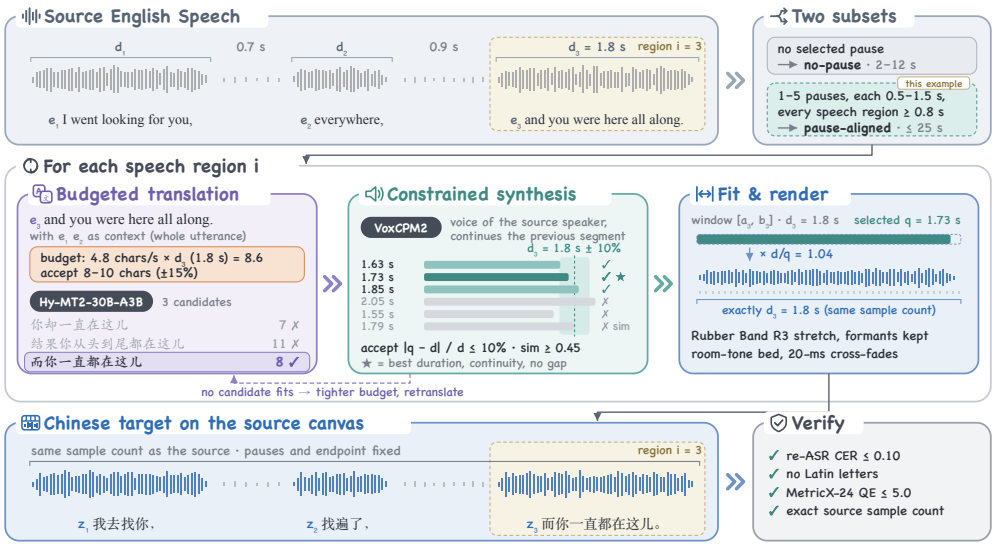


Figure 2: Stage 3 construction pipeline for timeline-preserving speech supervision.

Duration-budgeted translation. Each region transcript e_i is translated under a duration-dependent character budget, with the full transcript $e_{1:N}$ as semantic context. When synthesis cannot satisfy the duration constraint, we refine the budget and retranslate the region.

Constrained synthesis and source-canvas rendering. VoxCPM2 generates multiple source-speaker-conditioned candidates, selected for duration compatibility, speaker similarity, and cross-segment continuity. Only near-duration candidates are eligible for bounded time-scale correction; an utterance is discarded if any region lacks an acceptable candidate. Accepted segments are fitted to their assigned source windows, while audio outside these windows is retained. The resulting waveform preserves the selected source pause intervals and utterance endpoint and has exactly the same sample count as the source. These corrections are used only to construct training targets.

Verification and retained data. Quality and timing checks yield 642K speech pairs with 298K supervised pause spans (Table 1).

Table 1: Retained training data by stage and supervision type. Source hours are summed across tasks and stages, not unique corpus hours. Stage 1 pause markers annotate source transcripts; only Stage 3 provides target-waveform timing supervision.

Stage	Data type	# Samples	Src. h	Tgt. h	Target timing	Pause supervision
1	ASR text	928,661	3,014	–	Not applicable	Source-side markers only
1	AST text	883,563	2,854	–	Not applicable	None in Chinese target
2	Natural-pace S2ST	806,407	2,610	2,431	Unconstrained natural timing	No target-time labels
3	Timeline no-pause	412,945	687	687	Exact source duration	No selected pause spans
3	Timeline pause-aligned	229,331	682	682	Exact source duration	All selected pauses retained
Stage-summed total		3,260,907	9,847	3,800	–	–

3 TIMELINE-PRESERVING TRAINING

3.1 MODEL AND SEQUENCE REPRESENTATION

TRANSDUBBER is initialized from Qwen3-ASR-1.7B (Shi et al., 2026), a speech recognition model that maps input speech to text but does not natively support translation or speech generation. We extend its vocabulary with 6,561 discrete speech-codec symbols (Du et al., 2025), representing speech

at a frame rate of 25 Hz, and train the model to generate translated text and target speech in addition to the source-language ASR transcript.

Given a source utterance, TRANSDUBBER first autoregressively generates the ASR transcript, followed by the translated text and target-speech codec tokens in an interleaved fashion, as illustrated in Fig. 3. Following Wu et al. (2025), we adopt a fixed 5:10 schedule, where every five text tokens are followed by ten speech tokens. Once the translated text is complete, no additional text padding tokens are inserted; the model continues generating speech tokens until the end-of-sequence token is produced. A flow-matching-based speech decoder (Du et al., 2025) then maps the generated discrete speech-token sequence to the corresponding raw waveform.

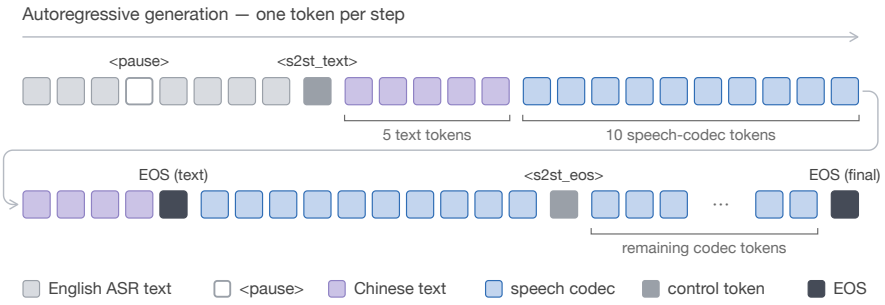


Figure 3: Autoregressive decoding in TRANSDUBBER. A pause-marked English transcript precedes Chinese text and speech-codec tokens, interleaved in blocks of five text and ten speech tokens. After text completion, speech generation continues until the final end-of-sequence token.

3.2 STAGE 1: TEXT GROUNDING

The first stage trains pause-aware ASR and English-to-Chinese text translation. Source transcripts include explicit pause markers when valid pauses are present. It does not yet require speech generation or source-timeline following.

3.3 STAGE 2: NATURAL-PACE SPEECH-GENERATION WARM-UP

Stage 2 trains the model to generate translated text and target-speech codec tokens using naturally paced target speech. The output follows the interleaved text–speech sequence format described above, but the target waveforms are not constrained to match the source duration or its internal pause locations. This stage therefore introduces target-speech generation and sequence termination without explicitly imposing source-timeline constraints.

3.4 STAGE 3: TIMELINE-PRESERVING SUPERVISED FINE-TUNING

Stage 3 fine-tunes the model on the timeline-aligned target waveforms constructed in Section 2. Each target matches the source utterance duration and retains the selected internal pause intervals, when present. Maximum-likelihood training on these targets provides supervision for placing translated speech within the source speaking windows and terminating near the source endpoint.

3.5 STAGE 4: VERIFIER-GUIDED POLICY OPTIMIZATION

After Stage 3, we further optimize the policy with GSPO (Zheng et al., 2025) using verifier-based rewards for translation quality and temporal alignment. We sample RL prompts from the same source-utterance pool used for Stage 3, including both no-pause and pause-aligned examples. For each source prompt x, the policy samples a group of G candidates $\{y_i\}_{i=1}^{G}$. Each candidate is evaluated by three sequence-level rewards: translation quality R_{qual}, global duration agreement R_{dur}, and pause alignment R_{pause}.

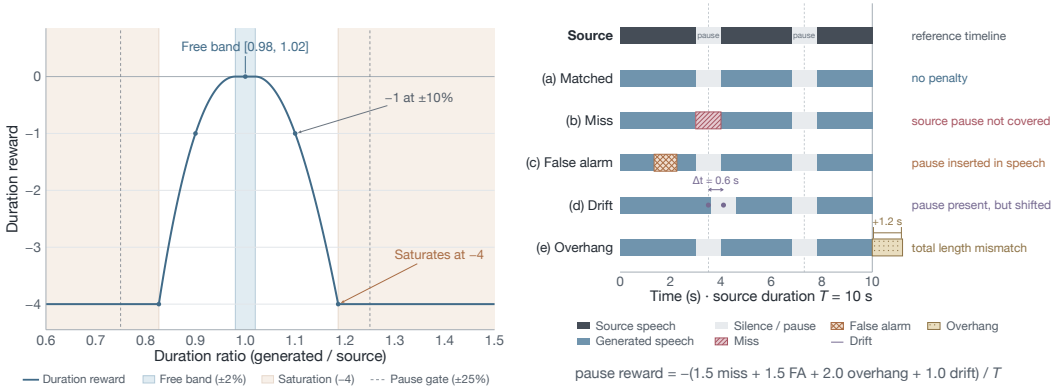
Importantly, we do not combine these rewards into a scalar before advantage estimation. Instead, each reward is normalized independently within the rollout group and optimized using the token subset most directly associated with that reward.

Verifier rewards. We measure translation quality using MetricX-QE (Juraska et al., 2024) in its reference-free mode. Given the dataset-provided source transcript s_x and the generated translation $\hat{t}_y$, MetricX-QE assigns a score in $[0, 25]$, with lower values indicating better translation quality. We convert this score into a higher-is-better reward in $[0, 1]$ via $R_{\text{qual}}(x, y) = 1 - \text{MetricX-QE}(s_x, \hat{t}_y)/25$.

Using the dataset transcript, rather than the model's own ASR output, prevents recognition errors from directly affecting the quality verifier.

For duration alignment, let T_x and T_y denote the source and generated speech durations, respectively. The duration reward is defined from the ratio $\rho = T_y/T_x$. As illustrated in Fig. 4a, it assigns no penalty within $\pm 2\%$ of the source duration, decreases quadratically with the normalized log-distance outside this region, reaches -1 at $\pm 10\%$, and saturates at -4 for large mismatches.

The pause reward compares source and generated speech–silence occupancy on a shared timeline. It penalizes four types of temporal error: uncovered source pauses, pauses inserted during source speech, pause-center displacement, and endpoint overhang. A small temporal collar allows minor boundary differences between corresponding pause intervals. To prevent a temporally mis-scaled candidate from receiving a favorable local-alignment reward, we additionally enforce $|\rho - 1| > 0.25 \Longrightarrow R_{\text{pause}}(x, y) \leq -1$. The exact duration function, pause-interval construction, and penalty weights are provided in Appendix C.



(a) Global duration reward.

(b) Timeline-based pause reward.

Figure 4: Temporal rewards for Stage 4 reinforcement learning. (a) The duration reward is zero within ±2% of source duration and saturates at -4. (b) The pause reward penalizes missed or spurious silence, shifted pause centers, and endpoint mismatch.

Component-specific credit assignment. The quality reward is assigned only to the tokens of the generated text translation, excluding speech-codec and special tokens. In contrast, the duration and pause rewards are assigned only to generated speech-codec tokens.

Let M_{text} and M_{speech} denote the corresponding token masks, and let A_c be the independently group-normalized advantage for component c. We optimize

$$\mathcal{L} = \frac{1}{\lambda_q + \lambda_p + \lambda_d} \Big(\lambda_q \, \mathcal{L}_{\text{GSPO}}(A_{\text{qual}}, M_{\text{text}}) + \lambda_p \, \mathcal{L}_{\text{GSPO}}(A_{\text{pause}}, M_{\text{speech}}) + \lambda_d \, \mathcal{L}_{\text{GSPO}}(A_{\text{dur}}, M_{\text{speech}}) \Big) + \beta \, D_{\text{KL}}^{M_{\text{text}} \cup M_{\text{speech}}} (\pi_\theta \,\|\, \pi_{\text{SFT}}) \,. \quad (1)$$

Each GSPO term uses a sequence-level importance ratio computed over its own token mask. Details of reward normalization, advantage clipping, importance-ratio clipping, and rollout filtering are reported in Appendix C.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Evaluation data. We construct an English-to-Chinese evaluation set by selecting 4,353 real speech utterances from the CoVoST2 test set (Wang et al., 2021) and GigaSpeech test set (Chen et al., 2021). The resulting set contains 1,439 CoVoST2 utterances, which are predominantly read speech and come with human Chinese references, and 2,914 GigaSpeech utterances, which consist of spontaneous speech and are paired with Chinese references generated by Gemini 3.1 Pro and subsequently manually reviewed by experts for correctness. The source utterances have an average duration of 8.9 seconds and a maximum duration of 19.4 seconds.

Systems. We compare **TRANSDUBBER** against five representative baselines spanning cascaded pipelines, specialized end-to-end S2ST models, expressive S2ST systems, and general-purpose multimodal models. **Cascaded** is a strong three-stage pipeline consisting of Qwen3-ASR-1.7B (Shi et al., 2026) for English speech recognition, Hy-MT2-30B-A3B (Zheng et al., 2026) for English-to-Chinese text translation, and VoxCPM2 (Zhou et al., 2026) for Chinese speech synthesis. **Seamless-L** (Barrault et al., 2023b) is a large multilingual end-to-end speech translation model from the SeamlessM4T family, while **Seamless-Ex** (Barrault et al., 2023a) refers to SeamlessExpressive, an end-to-end S2ST model designed to preserve expressive and paralinguistic information. **Qwen2.5-Omni-7B** (Xu et al., 2025) is a general-purpose multimodal model supporting direct speech input and speech output. **UniSS** (Cheng et al., 2026) is a unified expressive S2ST baseline. TransVIP (Le et al., 2024) is closely related but is not included because its published system targets French–English rather than our English–Chinese setting.

Evaluation metrics and protocol. We evaluate the proposed system as follows:

- **Translation quality and speaker preservation.** We report MetricX-QE, MetricX-ref (Juraska et al., 2024) and COMET score (Rei et al., 2020). We also report speaker similarity (SIM; Chen et al., 2022) to evaluate speaker preservation.
- **Isochrony (global duration fidelity).** We report Dur@2%, Dur@5%, and Dur@10%, which measure the percentage of utterances whose relative duration error falls within the corresponding threshold.
- **Pause alignment (local timeline fidelity).** We report pause F1@0.4 and onsetMedAE@0.4 for event-level pause alignment, and pause-span tIoU and SER for span-level temporal alignment.
- **Human evaluation.** We report 1–5 ratings for translation quality (Qual.) and speech naturalness (Nat.).

See Appendix E for detailed definitions of the isochrony and pause-alignment metrics, and Appendix F for the human-evaluation protocol. Unless otherwise stated, all automatic translation-quality scores, including COMET and MetricX-QE, are computed on the intended-text track, using each system's target-language translation text rather than an ASR transcription of its generated speech. MetricX-QE is used for RL optimization, whereas COMET with the `wmt22-comet-da` checkpoint is held out from training and used as an independent audit metric.

Decoding. Unless otherwise stated, TRANSDUBBER decoding uses nucleus sampling with $p = 0.85$, temperature 0.85, no top-k cutoff, a maximum of 896 new tokens, and seed 42.

4.2 MAIN RESULTS

Table 2 demonstrates the benefit of explicitly learning source-timeline preservation. TRANSDUBBER achieves Dur@2% and Dur@5% of 61.5% and 92.2%, respectively, compared with 25.6% and 56.1% for UniSS, the strongest baseline on these metrics. Across Cascaded, Seamless-L, Seamless-Ex, and Qwen2.5-Omni-7B, Dur@2% remains between 3.7% and 5.4%. The improvements extend beyond the utterance endpoint to internal pause placement: TRANSDUBBER reaches 0.550 pause F1, compared with 0.118 for UniSS and at most 0.080 for the other baselines. It also obtains the best tIoU and SER, at 0.322 and 0.128, respectively.

These temporal gains are accompanied by competitive translation quality. TRANSDUBBER achieves 0.788 COMET and 4.12 MetricX-QE, outperforming both Seamless variants; relative to UniSS,

Table 2: English-to-Chinese speech-to-speech translation on 4,353 CoVoST2 and GigaSpeech test utterances. Dur@ϵ denotes the percentage within ϵ of source duration. onMedAE denotes onsetMedAE@0.4. Human ratings are on a 1–5 scale.

System	SIM↑	Translation Quality		Isochrony			Pause Alignment				Human Evaluation	
		COMET↑	MetricX-QE↓	Dur@2%↑	Dur@5%↑	Dur@10%↑	F1↑	tIoU↑	SER↓	onMedAE↓	Qual.↑	Nat.↑
Cascaded	0.451	0.854	2.67	5.4	13.6	27.3	0.076	0.071	0.402	0.213	3.97	3.93
Seamless-L	0.029	0.705	6.79	3.7	9.5	19.5	0.002	0.008	0.407	0.533	1.83	1.10
Seamless-Ex	0.261	0.718	5.99	4.0	10.7	21.9	0.062	0.059	0.395	0.230	2.20	1.53
Qwen2.5-Omni-7B	0.020	**0.809**	**3.88**	4.0	11.1	21.5	0.080	0.060	0.446	0.211	3.63	**3.97**
UniSS	0.293	0.784	5.65	25.6	56.1	83.3	0.118	0.108	0.288	0.266	3.13	2.87
TRANSDUBBER	**0.464**	0.788	4.12	**61.5**	**92.2**	**98.7**	**0.550**	**0.322**	**0.128**	**0.162**	**3.63**	3.47

COMET is comparable (0.788 vs. 0.784), while MetricX-QE improves from 5.65 to 4.12. Human evaluation provides complementary evidence: on the 1–5 scales, TRANSDUBBER receives 3.63 for translation quality and 3.47 for speech naturalness, exceeding UniSS (3.13 and 2.87) and both Seamless variants on both dimensions. Its translation-quality rating also matches Qwen2.5-Omni-7B at 3.63, despite the latter's higher COMET score (0.809). Cascaded still achieves higher automatic and human-rated translation quality (0.854 COMET and 3.97 Qual.), and both Cascaded and Qwen2.5-Omni-7B receive higher naturalness ratings (3.93 and 3.97, respectively).

Table 3 further shows how these capabilities emerge through the training curriculum. Stage 1 establishes text translation, while Stage 2 introduces natural-pace speech generation but provides limited temporal fidelity, with only 5.2% Dur@2% and 0.099 pause F1. Its COMET score of 0.823 is computed only on the 73.7% of samples that successfully complete generation. Timeline-preserving SFT in Stage 3 produces the main gains in source-timeline following, raising Dur@2% to 48.6% and pause F1 to 0.518. Stage 4 then refines this behavior: Dur@2% and Dur@5% increase to 61.5% and 92.2%, respectively, pause F1 reaches 0.550, and SER decreases from 0.198 to 0.128. These temporal improvements are accompanied by an increase in COMET from 0.772 to 0.788.

Table 3: Capability acquisition across TRANSDUBBER's training stages. Stage 1 generates text only.

Stage	Capability	COMET ↑	Dur@2% ↑	Dur@5% ↑	Pause F1 ↑	tIoU ↑	SER ↓
Stage 1	ASR and text translation	**0.823**	–	–	–	–	–
Stage 2	Natural-pace speech translation	0.823†	5.2	14.2	0.099	0.086	0.420
Stage 3	Timeline-preserving translation	0.772	48.6	83.3	0.518	0.293	0.198
Stage 4	Quality and timing refinement	0.788	**61.5**	**92.2**	**0.550**	**0.322**	**0.128**

† Computed only on the 73.7% of samples that successfully completed generation.

4.3 ABLATION STUDIES AND ANALYSIS

Role of pause markers. We intervene on the <pause> markers in the ASR prefix to examine whether the pause-aware representation guides target-timeline generation (Table 4). Shifting the markers forward by three tokens reduces pause F1 from 0.550 to 0.476 while leaving the number of generated pauses almost unchanged (0.89 vs. 0.88 per utterance). Shuffling their positions further reduces F1 to 0.369. Thus, the marker locations affect pause alignment beyond simply controlling how often the model pauses. Removing the markers produces the largest degradation: pause F1 falls to 0.231, tIoU decreases from 0.322 to 0.156, and the generated pause count approximately halves to 0.44 per utterance. Nevertheless, Dur@2% remains 58.3%, and COMET improves slightly to 0.795. These interventions show that both the presence and placement of pause markers contribute to local timeline fidelity, which can deteriorate substantially even when translation quality and utterance-level duration agreement remain comparatively stable.

Natural-pace speech-generation warm-up. Table 5 compares curricula with and without Stage 2 before timeline-preserving SFT. Including the natural-pace warm-up modestly improves translation quality, increasing COMET from 0.768 to 0.772 and lowering both MetricX scores. Its main benefit is stronger text–speech consistency: average CER between generated text and re-transcribed speech decreases from 5.14% to 4.24%, while the fraction of outputs with CER at least 30% falls from 3.81%

Table 4: Pause-marker interventions in TRANSDUBBER. ASR inputs contain an average of 0.89 <pause> markers per utterance.

	COMET↑	MetricX-QE↓	Dur@2% ↑	Dur@5% ↑	Pause F1↑	tIoU↑	onMedAE↓	**Pauses / utt.**
No intervention	0.788	4.12	**61.5**	**92.2**	**0.550**	**0.322**	**0.162**	0.89
Shift +3 tokens	0.784	4.24	60.7	91.5	0.476	0.278	0.174	0.88
Shuffle	0.786	4.17	58.8	90.2	0.369	0.227	0.174	0.81
Remove	**0.795**	**3.91**	58.3	89.1	0.231	0.156	0.192	0.44

to 2.92%. Temporal performance changes little, with nearly identical Dur@10% (94.9% vs. 94.8%) and a slight increase in pause F1 (0.511 to 0.518). These results suggest that the warm-up primarily strengthens speech realization rather than timeline alignment.

Table 5: Ablation on progressive capability acquisition. The intermediate natural-pace speech translation stage (S_2) improves both translation quality and speech intelligibility.

Training Curriculum	Translation Quality			Speech Intelligibility†		Dur@10% ↑	Pause F1 ↑
	MetricX-ref ↓	MetricX-QE ↓	COMET ↑	Avg. CER (%) ↓	CER ≥ 30% (%) ↓		
$S_1 \rightarrow S_3$	4.35	4.49	0.768	5.14	3.81	94.9	0.511
$\mathbf{S_1 \rightarrow S_2 \rightarrow S_3}$	**4.21**	**4.38**	**0.772**	**4.24**	**2.92**	94.8	**0.518**

† CER is computed between the Paraformer transcription of the generated speech and the corresponding generated text.

Contributions of the reward components. Table 6 shows the trade-off among the reward components. Quality-only optimization gives the highest translation scores but substantially weakens temporal alignment. The pause reward yields the highest pause F1 and tIoU, whereas combining the temporal rewards improves strict duration matching. Adding the quality reward to this combination raises COMET from 0.772 to 0.788 while retaining strong temporal performance. The full objective achieves the best duration accuracy and SER, although its pause F1 remains below that of pause-only optimization.

Table 6: Reward-component ablation under a common initialization and optimization budget.

Reward	COMET↑	MetricX-QE↓	Dur@2% ↑	Dur@5% ↑	Pause F1↑	tIoU↑	SER↓
Quality only	**0.795**	**3.90**	10.2	38.2	0.354	0.206	0.134
Duration only	0.771	4.40	59.0	89.2	0.556	0.316	0.175
Pause only	0.769	4.47	54.4	87.6	**0.567**	**0.334**	0.134
Duration + Pause	0.772	4.37	61.1	90.8	0.544	0.315	0.131
Full (Quality + Duration + Pause)	0.788	4.12	**61.5**	**92.2**	0.550	0.322	**0.128**

5 CONCLUSION

We presented TRANSDUBBER, an end-to-end autoregressive S2ST system that learns to translate on the source timeline, rather than merely match utterance duration. A 642K-pair source-aligned corpus and progressive training enable the model to learn utterance-endpoint fidelity and salient pause alignment without an inference-time frame-level timing controller or post-hoc temporal correction. On English-to-Chinese translation, TRANSDUBBER raises Dur@2% from 25.6% to 61.5% and pause F1 from 0.118 to 0.550 over UniSS, with comparable COMET. Our analyses show that aligned SFT establishes timeline following, while verifier-guided RL refines timing and improves translation quality. Closing the translation-quality and naturalness gap to the cascaded baseline remains an important next step. Together, these findings support treating source timing as part of the translation target itself: learning not only what to say, but when to speak and when to remain silent.

Reproducibility Statement

Data construction, reward definitions, training settings, and evaluation protocols are described in Appendices B–F. We will release the training data, model checkpoints, and code upon publication, subject to applicable data licenses.

Ethics Statement

Our work targets speech translation and dubbing. Speaker-preserving generation could be misused for impersonation, while translation errors could misrepresent speakers. Downstream use should involve appropriate speaker authorization and disclosure of synthetic speech, and data redistribution should respect source licenses and privacy constraints.

AI Use Statement

We used Qwen3-ASR-1.7B, Hy-MT2-30B-A3B, and VoxCPM2 for transcription, translation, and speech synthesis during data construction. Gemini 3.1 Pro generated Chinese references for the GigaSpeech evaluation subset, which were manually reviewed. ChatGPT assisted with drafting parts of the manuscript. The authors take full responsibility for the final paper and all AI-assisted content.

References

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, et al. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*, 2023a.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. SeamlessM4T: Massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*, 2023b.

William Brannon, Yogesh Virkar, and Brian Thompson. Dubbing in practice: A large scale study of human localization with insights for automatic dubbing. *Transactions of the Association for Computational Linguistics*, 11:419–435, 2023.

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, et al. GigaSpeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio. In *Proceedings of Interspeech 2021*, pp. 3670–3674. ISCA, 2021.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6): 1505–1518, 2022.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6255–6271, 2025.

Sitong Cheng, Weizhen Bian, Xinsheng Wang, Ruibin Yuan, Jianyi Chen, Shunshun Yin, Yike Guo, and Wei Xue. UniSS: Unified expressive speech-to-speech translation with your voice. In *International Conference on Learning Representations*, 2026.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, et al. CosyVoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*, 2025.

Zhifu Gao, Shiliang Zhang, Ian McLoughlin, and Zhijie Yan. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. In *Proceedings of Interspeech 2022*, pp. 2063–2067. ISCA, 2022.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, et al. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, pp. 885–890. IEEE, 2024.

Ye Jia, Ron J. Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. Direct speech-to-speech translation with a sequence-to-sequence model. In *Proceedings of Interspeech 2019*, pp. 1123–1127. ISCA, 2019.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Ninth Conference on Machine Translation*, November 2024.

Surafel M. Lakew, Yogesh Virkar, Prashant Mathur, and Marcello Federico. Isometric MT: Neural machine translation for automatic dubbing. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6242–6246. IEEE, 2022.

Chenyang Le, Yao Qian, Dongmei Wang, Long Zhou, Shujie Liu, Xiaofei Wang, Midia Yousefi, Yanmin Qian, Jinyu Li, Sheng Zhao, et al. TransVIP: Speech to speech translation system with voice and isochrony preservation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 89682–89705, 2024.

Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Sravya Popuri, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, et al. Direct speech-to-speech translation with discrete units. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3327–3339. Association for Computational Linguistics, 2022.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2685–2702. Association for Computational Linguistics, 2020.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, et al. Qwen3-ASR technical report. *arXiv preprint arXiv:2601.21337*, 2026.

Silero Team. Silero VAD: Pre-trained enterprise-grade voice activity detector (VAD), number detector and language classifier. GitHub repository, 2021.

Karsten Steinhauer, Kai Alter, and Angela D. Friederici. Brain potentials indicate immediate use of prosodic cues in natural speech processing. *Nature Neuroscience*, 2(2):191–196, 1999.

Derek Tam, Surafel M. Lakew, Yogesh Virkar, Prashant Mathur, and Marcello Federico. Isochrony-aware neural machine translation for automatic dubbing. In *Proceedings of Interspeech 2022*, pp. 1776–1780. ISCA, 2022.

Yogesh Virkar, Marcello Federico, Robert Enyedi, and Roberto Barra-Chicote. Prosodic alignment for off-screen automatic dubbing. In *Proceedings of Interspeech 2022*, pp. 496–500. ISCA, 2022.

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. CoVoST 2 and massively multilingual speech translation. In *Proceedings of Interspeech 2021*, pp. 2247–2251. ISCA, 2021.

Hyun-Sik Won, DongJin Jeong, Hyunkyu Choi, and Jinwon Kim. End-to-end multilingual automatic dubbing via duration-based translation with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 515–521. Association for Computational Linguistics, 2025.

Haibin Wu, Yuxuan Hu, Ruchao Fan, Xiaofei Wang, Kenichi Kumatani, Bo Ren, Jianwei Yu, Heng Lu, Lijuan Wang, Yao Qian, et al. Towards efficient speech-text jointly decoding within one speech language model. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–7. IEEE, 2025.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

Mao Zheng, Zheng Li, Tao Chen, Bo Lv, Mingrui Sun, Mingyang Song, Jinlong Song, Hong Huang, Decheng Wu, Hai Wang, et al. Hy-MT2: A family of fast, efficient and powerful multilingual translation models in the wild. *arXiv preprint arXiv:2605.22064*, 2026.

Yixuan Zhou, Guoyang Zeng, Xin Liu, Xiang Li, Renjie Yu, Jiancheng Gui, Jiaheng Wu, Ziyang Wang, Xudong Shen, Runchuan Ye, et al. VoxCPM2 technical report. *arXiv preprint arXiv:2606.06928*, 2026.

A Extended Related Work

Direct and expressive speech-to-speech translation. Early direct S2ST systems demonstrated speech translation through spectrogram prediction (Jia et al., 2019) and discrete-unit generation (Lee et al., 2022), while SeamlessM4T extended unified speech and text translation to a massively multilingual setting (Barrault et al., 2023b). Beyond semantic accuracy, SeamlessExpressive models vocal style and prosody, including speech rate and pauses (Barrault et al., 2023a), and UniSS combines speech semantics and style within a unified text–speech language model to preserve speaker identity, emotion, and duration consistency (Cheng et al., 2026). These developments establish expressive transfer as an important dimension of S2ST. Our work focuses on a complementary requirement: preserving the source's speech–silence structure on a shared absolute timeline. Preserving voice or speaking rate does not guarantee duration agreement. Even matching total duration does not ensure that internal pauses occur at corresponding moments.

Isochrony and automatic dubbing. Temporal constraints have been studied at both the translation and speech-generation levels. Isometric MT encourages translations to match the source text length (Lakew et al., 2022), while duration-based translation uses source-speech duration to determine phoneme budgets and iteratively adjust translated text (Won et al., 2025). Importantly, prior work also addresses internal structure: isochrony-aware MT jointly translates and projects pause markers (Tam et al., 2022), and TransVIP conditions speech generation on duration and frame-level voice activity through an explicit isochrony-control module (Le et al., 2024). Whereas TransVIP derives its isochrony conditioning from target speech during training and source speech at inference, we construct target waveforms that already follow the source timeline. This supervision enables TransDubber to learn endpoint fidelity, salient pause alignment, and autonomous termination jointly, without post-hoc padding, truncation, or global time stretching at inference.

Speech timing and context-dependent synchronization. Psycholinguistic evidence indicates that intonational phrasing contributes to the initial analysis of sentence structure and that misleading prosodic boundaries can disrupt interpretation (Steinhauer et al., 1999). These findings concern prosodic structure, rather than exact cross-lingual timestamp correspondence. Dubbing research further suggests that synchronization should be considered together with translation quality, vocal naturalness, and visual context. Brannon et al. (2023) analyze professional human dubs and argue for a qualified view of timing constraints. Virkar et al. (2022) distinguish on-screen from off-screen dubbing and report improved subjective viewing experience when the alignment strategy accounts for speaker visibility. Against this background, we study how to learn source-timeline following when that constraint is required, rather than prescribing a single synchronization policy for all translation scenarios.

Joint text–speech modeling and reward-based post-training. Joint speech language models provide a common sequence space for text and speech generation, with interleaved decoding coordinating the two modalities (Wu et al., 2025). Reward-based post-training has also been explored in speech synthesis, with CosyVoice 3 improving content consistency beyond standard likelihood training (Du et al., 2025). For language-model optimization, GSPO uses sequence-level importance ratios and clipping to align policy updates with sequence-level rewards (Zheng et al., 2025). We adapt these ingredients to timeline-preserving S2ST by separating semantic and temporal feedback within the interleaved output. Translation quality, global duration agreement, and local pause alignment are normalized independently, with semantic feedback assigned to translation-text tokens and temporal feedback to speech-codec tokens.

B Data Construction and Filtering Details

This appendix complements Section 2 with the source-selection criteria, construction settings, and quality checks. We follow the same three-stage organization: Stage 2 synthesizes the natural Chinese translation z^{nat}, whereas Stage 3 generates new translations and speech from the Stage 1 temporal scaffold and English region transcripts $e_{1:N}$. Stage 3 does not reuse z^{nat} or the Stage 2 target waveform. Figure 2 illustrates the Stage 3 construction.

B.1 SOURCE SELECTION

We draw source utterances from the training partitions of Emilia-EN (He et al., 2024) and GigaSpeech (Chen et al., 2021). Stages 1 and 2 use utterances of 8–25 s with at least eight words. Stage 3 additionally includes a short-form pool of 2–8 s with at least three words. Short-form sources undergo the same source-side processing but contribute only to Stage 3.

Both pools require an estimated speaking rate of 3.0–5.5 syllables/s and leading and trailing non-speech intervals no longer than 1.5 s each. Within Stage 3, no-pause examples are restricted to 2–12 s, whereas pause-aligned examples may extend to 25 s.

B.2 STAGE 1: TEMPORAL SCAFFOLD AND TEXT GROUNDING

Source-derived speech regions and pauses. We use Silero VAD (Silero Team, 2021) to identify source speech spans and the internal gaps between them. Gaps of at least 0.5 s define the selected pause boundaries. Speech regions are split only at these gaps, with the following region anchored to the next strict-VAD speech onset. Leading and trailing non-speech remain part of the source canvas but are not internal pause labels. The strict pass uses a speech-posterior threshold of 0.8, minimum speech and silence durations of 250 and 100 ms, and 30-ms boundary padding. A separate 0.4-s threshold is used for diagnostic gap detection, not for selecting supervised pauses.

Source-speech leakage screening. Because source-canvas rendering retains audio outside the speech windows, we screen these intervals for source speech missed by the strict VAD pass. A second, more permissive pass with a speech-posterior threshold of 0.35 flags speech spans with less than 90% coverage by strict-VAD spans expanded by 0.15 s on each side. Utterances with a suspected leakage region of at least 0.6 s are discarded. Remaining suspect regions of at least 0.1 s are checked with Qwen3-ASR-1.7B (Shi et al., 2026); any nonempty hypothesis triggers rejection.

Region transcription. The same ASR model transcribes the source speech regions to obtain $e_{1:N}$. All required transcripts must be nonempty, and their number-normalized word error rate against the corpus transcript must not exceed 0.15. We join the region transcripts with `<pause>` markers only at the selected source boundaries, yielding the Stage 1 ASR target described in Section 2.

Natural translation. Hy-MT2-30B-A3B (Zheng et al., 2026) translates the complete utterance into z^{nat}. The prompt requests faithful, colloquial Chinese without unnecessary fillers and imposes no duration, character-count, or pause-placement constraint. If the output contains Latin letters, one additional request explicitly prohibits them; the retry replaces the original only when it is entirely Latin-free. This translation provides the Stage 1 AST target and the Stage 2 synthesis text.

B.3 STAGE 2: NATURAL-PACE SPEECH SUPERVISION

VoxCPM2 (Zhou et al., 2026) synthesizes z^{nat} in a single whole-utterance call conditioned on source-speaker reference audio. The target follows the natural phrasing and duration of the Chinese translation, without segment chaining, synthesis-guided retranslation, time-scale modification, or source-canvas rendering. The acceptance criteria in Table 7 remove unreliable synthesis without imposing source-timeline alignment. In particular, the broad target-to-source duration-ratio range is a sanity check rather than an isochrony constraint.

B.4 STAGE 3: TIMELINE-PRESERVING SPEECH SUPERVISION

We form the no-pause and pause-aligned subsets from the shared source scaffold. No-pause examples contain no selected internal pause, although shorter hesitations may remain within a speech region. Pause-aligned examples contain one to five selected pauses, each approximately 0.5–1.5 s long, with every speech region lasting at least 0.8 s. All selected pauses are retained.

Let $[a_i, b_i]$ denote source speech region i, with available speaking time $d_i = b_i - a_i$, measured in seconds. Stage 3 generates a new Chinese translation and speech segment for each region while keeping these windows fixed.

Duration-budgeted translation. We initialize the character budget from the source-region duration using a reference rate of 4.8 Chinese characters per second:

$$m_i = 4.8 d_i, \qquad L_i = \max(2, \lceil 0.85 m_i \rceil), \qquad U_i = \max(3, \lceil 1.15 m_i \rceil). \tag{2}$$

Only Chinese ideographs count toward the budget; punctuation and whitespace are excluded. The translation prompt provides the complete transcript $e_{1:N}$ as semantic context, the current region index and transcript e_i, and the requested character range $[L_i, U_i]$.

We generate up to three fixed-seed translation candidates. Candidates within the requested range are preferred to those outside it. Within either group, candidates closer to the interval center are preferred. The selected text undergoes the same Latin-removal retry as the natural translation.

If the selected translation cannot be synthesized within the permitted temporal range, the speaking rate of the closest synthesis candidate is used to tighten the character budget and retranslate the same English region. This feedback changes the translated text, not the source speech windows or pause boundaries.

Speaker-conditioned synthesis. VoxCPM2 generates candidates for each translated region using the longest detected source speech span, capped at 10 s, as a common speaker reference. For each region after the first, synthesis additionally conditions on the preceding selected target segment and its text. This preceding segment is used before duration correction to encourage local continuity. Each region receives at most two synthesis rounds, with six fixed-seed candidates per round.

Candidate acceptance and selection. After VAD-based edge trimming, let $q_{i,j}$ and $s_{i,j}$ denote the duration and speaker similarity of candidate j for region i. Ordinary candidates are eligible for duration correction only when

$$\frac{|q_{i,j} - d_i|}{d_i} \le 0.10 \qquad \text{and} \qquad s_{i,j} \ge 0.45.$$

A separate short-tail option accepts candidates satisfying $0 \le d_i - q_{i,j} \le 0.15\,\text{s}$ under the same speaker-similarity threshold. These candidates retain their natural duration rather than being stretched.

Among eligible candidates, selection jointly favors duration agreement, a duration ratio $q_{i,j}/d_i$ close to that of the preceding selected segment, higher speaker similarity, and fewer generated internal gaps of at least 0.4 s. Internal gaps incur a soft penalty rather than causing automatic rejection.

Generation may stop early when a candidate achieves at most 2% duration error, speaker similarity of at least 0.60, and no detected internal gap. These are early-stopping conditions, not requirements for every accepted candidate. If even one region has no acceptable candidate, the complete utterance is discarded.

Duration correction and source-canvas rendering. We fit each accepted target segment to its assigned source window. For an ordinary selected candidate of duration q_i, Rubber Band R3 applies the factor d_i/q_i using the finer engine with formant preservation. The corrected segment is then cropped or padded to the exact window sample count. Short-tail candidates are not stretched; source-derived room tone fills the remaining window. Candidates outside the acceptance ranges are rejected rather than subjected to larger corrections.

We construct the target waveform by replacing the speech in each source window $[a_i, b_i]$ with the corresponding target segment while retaining the source audio outside these windows. Source-derived room tone provides background ambience beneath the replacement speech. We match the source-region RMS level with gain limited to ± 6 dB and apply 20-ms equal-power cross-fades at replacement boundaries.

The completed waveform is stored as 16-kHz PCM16 with exactly the same sample count as the source, keeping the selected source pause intervals and the utterance endpoint fixed on the source canvas. These operations construct training targets and are not applied during model inference.

B.5 Quality control and retained data

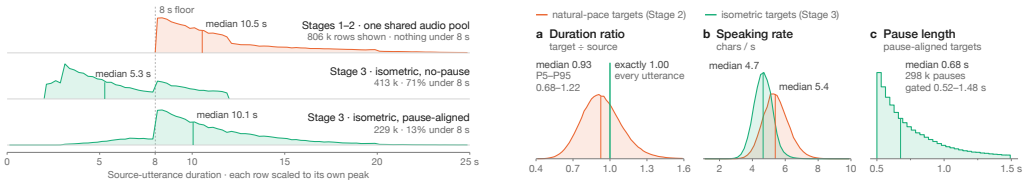
Acceptance criteria. Table 7 summarizes the target-speech acceptance criteria. Chinese-character error rate (CER) compares Paraformer-zh re-ASR (Gao et al., 2022) with the intended synthesis text

Table 7: Acceptance criteria for synthesized training targets. Stage 3 speaker checks are applied during candidate selection.

Criterion	Stage 2: natural pace	Stage 3: timeline preserving
Content realization	Chinese CER ≤ 0.10	Chinese CER ≤ 0.10
Script purity	No Latin letters in re-ASR	No Latin letters in intended text or re-ASR
Speaker consistency	Similarity ≥ 0.45	Similarity ≥ 0.45 per selected segment
Translation adequacy	No QE gate	MetricX-24 QE ≤ 5.0
Duration	Target/source ratio in $[0.35, 2.0]$	Exact source sample count

after retaining only Chinese ideographs. Stage 3 translation quality is evaluated with MetricX-24 QE (XXL), using the concatenated English region transcripts and intended Chinese translations. Weak-speech screening takes place before translation, and Stage 3 speaker similarity is enforced during candidate selection rather than checked again at final verification.

Timing audits. Pause tIoU and source-pause matching with a 0.4-s tolerance are computed for auditing, not used as hard acceptance gates. Endpoint equality follows from exact sample-count matching, while the selected source pause intervals are retained through source-canvas rendering.



(a) Source data duration distributions.

(b) Target timing statistics.

Figure 5: Training-data timing statistics. (a) Source-duration distributions, including short utterances added in Stage 3. (b) Target/source duration ratios, speaking rates, and supervised pause lengths. Stage 3 targets exactly match source duration, unlike Stage 2 natural-pace targets.

Data export and retained corpus. Verified waveforms are encoded with s3tokenizer-v3 at 25 tokens/s using its 6,561-entry codebook. Training records retain all supervised pause spans and waveform timing metadata. These annotations are consumed directly without rerunning VAD or deriving pauses from punctuation.

Table 1 and Figs. 5a and 5b summarize the retained data. Stage 3 contains 642,276 pairs and 1,368.8 hours per side, including 320,421 short-form examples below 8 s and 321,855 examples of at least 8 s. Reported hours count training exposure across tasks and stages, not unique source audio.

C Reward Definitions and Policy-Optimization Details

This appendix details the verifier rewards and policy update used in Stage 4 (Section 3.5). We first specify the reward functions, then describe component-wise advantage estimation and masked policy optimization. The temporal rewards are illustrated in Figure 4. All temporal quantities are measured in seconds.

C.1 Translation-Quality and Global-Duration Rewards

Translation quality. We use the reference-free MetricX-QE reward,

$$R_{\text{qual}}(x, y) = 1 - \frac{\text{MetricX-QE}(s_x, \hat{t}_y)}{25}. \tag{3}$$

The verifier compares the generated translation text $\hat{t}_y$ with the dataset-provided source transcript s_x, rather than the model's own ASR output. The linear transformation maps the lower-is-better MetricX-QE score in $[0, 25]$ to a higher-is-better reward in $[0, 1]$.

Global duration. Let T_x and T_y denote the source and generated speech durations, respectively, and let $\rho = T_y/T_x$. We allow a free region of $\epsilon_{\text{free}} = 0.02$ and use $\epsilon_{\text{ref}} = 0.10$ to calibrate the penalty outside it. The normalized log-distance is

$$u(\rho) = \begin{cases} 0, & |\rho - 1| \le \epsilon_{\text{free}}, \\ \dfrac{\log(\rho/(1+\epsilon_{\text{free}}))}{\log((1+\epsilon_{\text{ref}})/(1+\epsilon_{\text{free}}))}, & \rho > 1 + \epsilon_{\text{free}}, \\ \dfrac{\log((1-\epsilon_{\text{free}})/\rho)}{\log((1-\epsilon_{\text{free}})/(1-\epsilon_{\text{ref}}))}, & \rho < 1 - \epsilon_{\text{free}}. \end{cases} \tag{4}$$

The corresponding reward is

$$R_{\text{dur}}(x, y) = -\min\big(u(\rho)^2,\, 4\big)\,. \tag{5}$$

Thus, the reward is zero within $\pm 2\%$ of the source duration, equals -1 at $\rho = 0.9$ and $\rho = 1.1$, and saturates at -4 for large mismatches.

C.2 Timeline-based pause reward

The pause reward compares internal silence on a shared absolute timeline. It combines missed and spurious silence, pause-center displacement, and utterance-end mismatch.

Pause preprocessing. The temporal origin is the beginning of each utterance. Leading and trailing silence is excluded from the pause-event sets, without shifting the timestamps of internal pauses. Silence intervals shorter than $\tau_{\text{sil}} = 0.12\,\text{s}$ are discarded. For generated speech, neighboring silence intervals separated by at most two codec tokens are merged; at 25 codec tokens per second, this gives $\tau_{\text{merge}} = 2/25 = 0.08\,\text{s}$. After preprocessing, the source and generated pause sets are

$$\mathcal{I}_x = \{[a_k^x, b_k^x]\}_{k=1}^{K_x}, \qquad \mathcal{I}_y = \{[a_j^y, b_j^y]\}_{j=1}^{K_y},$$

where K_x and K_y are their respective event counts. The generated-silence occupancy is $\mathcal{S}_y = \bigcup_{j=1}^{K_y} [a_j^y, b_j^y]$.

Collar-tolerant occupancy errors. A boundary collar of $\delta = 0.1\,\text{s}$ tolerates small differences in pause boundaries. We erode each source pause to obtain a silence core and dilate it to obtain a tolerated silence region:

$$\mathcal{S}_x^{\text{core}} = \bigcup_{\substack{1 \le k \le K_x \\ b_k^x - a_k^x > 2\delta}} [a_k^x + \delta,\, b_k^x - \delta], \tag{6}$$

$$\mathcal{S}_x^{\text{tol}} = \bigcup_{k=1}^{K_x} \left([a_k^x - \delta,\, b_k^x + \delta] \cap [0, T_x]\right). \tag{7}$$

Intervals that become empty after erosion contribute no silence core. Let $\mu(\cdot)$ denote total interval duration. We define

$$m(x, y) = \mu(\mathcal{S}_x^{\text{core}} \setminus \mathcal{S}_y)\,, \tag{8}$$

$$f(x, y) = \mu\big(\mathcal{S}_y \setminus \mathcal{S}_x^{\text{tol}}\big)\,, \tag{9}$$

$$o(x, y) = |T_y - T_x|. \tag{10}$$

Here, m measures uncovered source-silence cores, f measures generated silence outside the tolerated source regions, and o measures endpoint mismatch, including both early and late termination.

Pause-center displacement. For pause centers $c_k^x = (a_k^x + b_k^x)/2$ and $c_j^y = (a_j^y + b_j^y)/2$, each source event is compared with its nearest generated event. The total displacement error is

$$d(x,y) = \sum_{k=1}^{K_x} \Delta_k,$$
$$\Delta_k = \begin{cases} \min\left(\min\limits_{1 \le j \le K_y} |c_k^x - c_j^y|,\, \tau_{\text{drift}} \right), & K_y > 0, \\ \tau_{\text{drift}}, & K_y = 0, \end{cases} \tag{11}$$

with $\tau_{\text{drift}} = 2.0\,\text{s}$. A source event receives the maximum displacement penalty when no generated pause is present or its nearest one is more than 2.0 s away. These nearest-event comparisons do not impose one-to-one matching. When $K_x = 0$, the sum is zero.

Combined penalty and duration gate. The complete pause penalty is

$$P_{\text{pause}}(x,y) = \frac{1.5\, m(x,y) + 1.5\, f(x,y) + 2.0\, o(x,y) + \mathbf{1}[K_x > 0]\, d(x,y)}{\max(T_x, \epsilon)} + 0.3\, \mathbf{1}[K_x > 0 \wedge K_y = 0], \tag{12}$$

where $\epsilon > 0$ safeguards duration normalization. The time-based errors are normalized by the source duration; an additional penalty of 0.3 applies when the source contains pauses but none is generated. We first clip the negative penalty from below,

$$R^0_{\text{pause}}(x,y) = \max(-P_{\text{pause}}(x,y),\, r_{\text{floor}}), \tag{13}$$

where $r_{\text{floor}} = -4$ denotes the pre-gate reward-clipping floor used during training. The final reward then applies the global-duration gate:

$$R_{\text{pause}}(x,y) = \begin{cases} \min\big(R^0_{\text{pause}}(x,y), -1\big), & |\rho - 1| > 0.25, \\ R^0_{\text{pause}}(x,y), & \text{otherwise.} \end{cases} \tag{14}$$

This prevents a candidate with a large duration mismatch from receiving a favorable pause reward solely through local overlap.

C.3 Component-wise advantage estimation

For each source prompt, we sample $G = 16$ candidates. Let $R_{i,c}$ be the reward of candidate i for component $c \in \{\text{qual}, \text{pause}, \text{dur}\}$. Each component is standardized independently within the rollout group:

$$\overline{R}_c = \frac{1}{G} \sum_{i=1}^{G} R_{i,c}, \qquad Z_{i,c} = \frac{R_{i,c} - \overline{R}_c}{\max\Big(\text{Std}_{j=1}^{G}[R_{j,c}],\, 0.05\Big)}. \tag{15}$$

We clip the standardized values and then recenter them:

$$\widetilde{A}_{i,c} = \text{clip}(Z_{i,c}, -3, 3), \qquad A_{i,c} = \widetilde{A}_{i,c} - \frac{1}{G} \sum_{j=1}^{G} \widetilde{A}_{j,c}. \tag{16}$$

Clipping therefore precedes recentering, and the resulting advantages have zero mean within each component. Rewards are not combined before this calculation. Extreme-rollout filtering is disabled: all G candidates are retained, and each contributes to a component when its corresponding token mask is nonempty.

C.4 Masked GSPO and KL regularization

Reward-specific token masks. The translation-text mask M_{text} selects only generated translation tokens after `<s2st_text>` and before `<s2st_eos>`. It excludes the control markers, speech-codec tokens, and sequence-level end-of-sequence tokens. The speech mask M_{speech} selects only generated speech-codec tokens. Thus,

$$M_{\text{qual}} = M_{\text{text}}, \qquad M_{\text{pause}} = M_{\text{dur}} = M_{\text{speech}}. \tag{17}$$

Masked sequence-level policy loss. For candidate y_i and component c, let $M_{i,c,t} \in \{0,1\}$ indicate whether token t is selected, with mask length $L_{i,c} = \sum_t M_{i,c,t}$. For a nonempty mask, the length-normalized importance ratio of GSPO (Zheng et al., 2025) becomes

$$s_{i,c}(\theta) = \exp\left(\frac{1}{L_{i,c}} \sum_t M_{i,c,t} \log \frac{\pi_\theta(y_{i,t} \mid x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t} \mid x, y_{i,<t})}\right), \tag{18}$$

where $\pi_{\theta_{\text{old}}}$ is the rollout policy. The mask selects the probabilities entering the ratio; each probability remains conditioned on the full preceding sequence. We use asymmetric clipping with $\epsilon_{\text{low}} = 0.2$ and $\epsilon_{\text{high}} = 0.28$:

$$\bar{s}_{i,c}(\theta) = \text{clip}(s_{i,c}(\theta),\, 1 - \epsilon_{\text{low}},\, 1 + \epsilon_{\text{high}})\,.$$

Under the loss-minimization convention, the component surrogate is

$$\mathcal{L}^{(c)}_{\text{GSPO}} = -\frac{1}{G} \sum_{i=1}^{G} \min(s_{i,c}(\theta) A_{i,c},\, \bar{s}_{i,c}(\theta) A_{i,c})\,. \tag{19}$$

Loss aggregation and regularization. The component losses are combined only after their advantages and masked importance ratios have been computed:

$$\mathcal{L}_{\text{policy}} = \frac{\lambda_q \mathcal{L}^{(\text{qual})}_{\text{GSPO}} + \lambda_p \mathcal{L}^{(\text{pause})}_{\text{GSPO}} + \lambda_d \mathcal{L}^{(\text{dur})}_{\text{GSPO}}}{\lambda_q + \lambda_p + \lambda_d}, \tag{20}$$

with $\lambda_q = 0.2$ and $\lambda_p = \lambda_d = 1$. The fixed Stage 3 checkpoint serves as the reference policy π_{SFT}. We apply a single KL term over $M_{\text{KL}} = M_{\text{text}} \cup M_{\text{speech}}$, rather than a separate term for each reward component:

$$\mathcal{L} = \mathcal{L}_{\text{policy}} + \beta\, D^{M_{\text{KL}}}_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{SFT}})\,, \qquad \beta = 0.01. \tag{21}$$

Both the reference policy and the KL coefficient remain fixed throughout Stage 4. With these component weights, this recovers Eq. 1 in the main text.

D Training Hyperparameters

Table 8 summarizes Stages 2–4, each initialized from the final checkpoint of the preceding stage. Stage 4 samples its source prompts from the Stage 3 training set. All three stages use eight GPUs, random seed 42, and AdamW with $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-6}$, and gradient clipping at 1.0. Learning rates warm up over the first 5% of updates and then decay to zero.

Table 8: Training hyperparameters for natural-pace SFT (Stage 2), timeline-preserving SFT (Stage 3), and GSPO (Stage 4). Dashes indicate inapplicable settings.

Hyperparameter	Stage 2	Stage 3	Stage 4
Peak learning rate	10^{-4}	10^{-4}	3×10^{-6}
Weight decay	0.02	0.02	0.01
LR decay schedule	Linear	Linear	Cosine
Optimization steps	20,000	12,000	100
Global batch per update	256 utterances	256 utterances	28 prompts
Codec-parameter LR multiplier	10×	10×	–
Pause-codec loss weight	1.0	2.0	–
End-of-sequence loss weight	4.0	2.0	–
Rollouts per prompt	–	–	16
Training microbatch size	–	–	4
Rollout temperature / top-p	–	–	1.0 / 1.0
Maximum new tokens	–	–	896
GSPO clipping $(\epsilon_{\text{low}}, \epsilon_{\text{high}})$	–	–	(0.2, 0.28)
KL coefficient	–	–	0.01
Component weights $(\lambda_q, \lambda_p, \lambda_d)$	–	–	(0.2, 1, 1)

For SFT, the effective batch size is $8 \times 8 \times 4 = 256$ (GPUs $\times$ per-GPU batch $\times$ gradient accumulation), with BF16 mixed precision and sequence-normalized loss. Stage 4 uses seven training GPUs and one reward-server GPU; four generation batches per update yield $7 \times 4 = 28$ prompts and 448 rollouts before zero-signal groups are discarded. The policy weights are FP32, while reference weights and computation use BF16. We use the checkpoint after 100 updates as the final model.

E ISOCHRONY AND PAUSE-ALIGNMENT METRICS

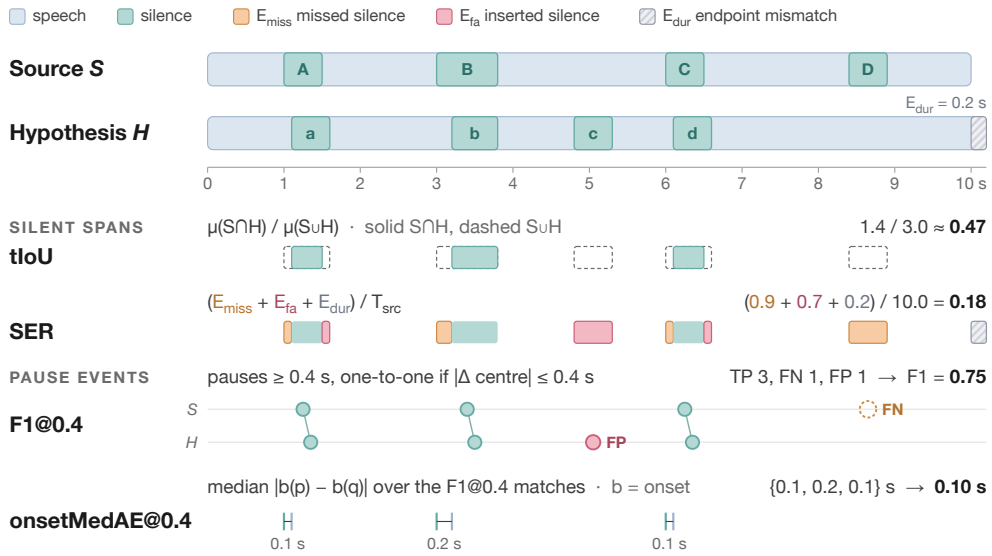


Figure 6: Pause-alignment metrics on one shared timeline. Source S and hypothesis H pauses (all $\geq \tau_{\min} = 0.4\,\text{s}$) are compared at the same temporal origin. tIoU is the overlap of the silent spans; SER sums the missed silence E_{miss}, the inserted silence E_{fa}, and the endpoint mismatch E_{dur}, normalized by the source duration. F1@0.4 matches pause events one-to-one when their centers differ by at most 0.4 s (D is a false negative, c a false positive); onsetMedAE@0.4 is the median absolute onset difference over the matched pairs.

Isochrony. For utterance i, let T_i^{src} and T_i^{hyp} denote the durations of the source and translated speech, respectively. We define the relative duration error as

$$e_i^{\text{dur}} = \frac{\left| T_i^{\text{hyp}} - T_i^{\text{src}} \right|}{T_i^{\text{src}}}. \tag{22}$$

The metric $\text{Dur@}\epsilon$ is the percentage of utterances satisfying $e_i^{\text{dur}} \leq \epsilon$:

$$\text{Dur@}\epsilon = \frac{100}{N} \sum_{i=1}^{N} \mathbb{I}\left[e_i^{\text{dur}} \leq \epsilon \right]. \tag{23}$$

We report $\epsilon \in \{2\%, 5\%, 10\%\}$; higher values indicate better utterance-level duration matching.

Pause alignment. As illustrated in Fig. 6, we anchor the source and translated utterances at the same temporal origin and compare their pauses on a shared absolute time axis. Let S and H denote the unions of the source and hypothesis silent intervals, respectively, and let $\mu(\cdot)$ denote total interval duration. The metrics provide two complementary views: tIoU and SER operate on *silent spans*, whereas F1 and onset error operate on discrete *pause events*.

Pause F1. For the event-based view, we retain pauses of duration at least $\tau_{\min} = 0.4\,\text{s}$. Given a matching tolerance δ, a source pause and a hypothesis pause may be matched one-to-one when the

difference between their center times is at most δ. Matched pairs are true positives, unmatched source pauses are false negatives, and unmatched hypothesis pauses are false positives:

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \qquad R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \qquad \mathrm{F1} = \frac{2PR}{P + R}. \tag{24}$$

We report F1@0.4 using $\delta = 0.4\,\mathrm{s}$. Higher is better.

Pause-span tIoU. Temporal intersection over union measures the overlap between the source and hypothesis silent regions:

$$\mathrm{tIoU}(S, H) = \frac{\mu(S \cap H)}{\mu(S \cup H)}. \tag{25}$$

It jointly penalizes missed pauses, additional pauses, temporal shifts, and mismatched pause durations. Higher is better.

Structure Error Rate (SER). SER decomposes pause-timeline disagreement into three components:

$$E_{\mathrm{miss}} = \mu(S \setminus H), \qquad E_{\mathrm{fa}} = \mu(H \setminus S), \qquad E_{\mathrm{dur}} = \left| T^{\mathrm{hyp}} - T^{\mathrm{src}} \right|. \tag{26}$$

Here, E_{miss} is the duration for which the source is silent but the translation is speaking, E_{fa} is the duration for which the translation is silent but the source is speaking, and E_{dur} measures utterance-end mismatch. We normalize their sum by the source duration:

$$\mathrm{SER} = \frac{E_{\mathrm{miss}} + E_{\mathrm{fa}} + E_{\mathrm{dur}}}{T^{\mathrm{src}}}. \tag{27}$$

Lower is better. Unlike tIoU, SER also exposes the contributions of the different error types.

Pause-onset error. Let $\mathcal{M}_\delta$ be the pause pairs matched under tolerance δ, and let $b(p)$ denote the onset time of pause p. We compute the median absolute onset error over the matched pairs:

$$\mathrm{onsetMedAE@}\delta = \underset{(p,q)\in\mathcal{M}_\delta}{\mathrm{median}} \left| b(p) - b(q) \right|. \tag{28}$$

We report onsetMedAE@0.4 in seconds using the same matches as F1@0.4, and abbreviate it as onMedAE in the tables. Lower is better. Because this metric is conditioned on successfully matched pauses, it must be interpreted jointly with pause F1.

F Human Evaluation

Participants and setup. Participants were recruited through Prolific and were adults aged 22–58 years residing in the United Kingdom and United States. A total of 25 participants took part in the study. They were required to self-confirm fluency in English and Mandarin Chinese before participation. System identities were concealed, and output presentation order was randomized within each source utterance. Each output was evaluated on two separate five-point scales.

Audio naturalness (Nat.). Participants assessed how natural the Mandarin speech sounded, considering pronunciation, rhythm, phrasing, intonation, fluency, pauses, and noticeable speech artifacts. They judged this dimension independently of translation accuracy and similarity to the original speaker. Scores ranged from 1 (*very unnatural*) to 5 (*completely natural*).

Translation quality (Qual.). Participants assessed how accurately and completely the Mandarin speech conveyed the English source's meaning and important information. Natural rephrasing was permitted provided that the original meaning was preserved. Scores ranged from 1 (*meaning not conveyed*) to 5 (*fully accurate and complete*).

Rating procedure. As illustrated in Fig. 7, participants first listened to the English source with its transcript. They then listened to each Mandarin output in full and rated audio naturalness before its transcript was revealed. Subsequently, they assessed translation quality with the Mandarin transcript available. Whenever the transcript differed from the spoken output, participants were instructed to judge the speech rather than the displayed text.

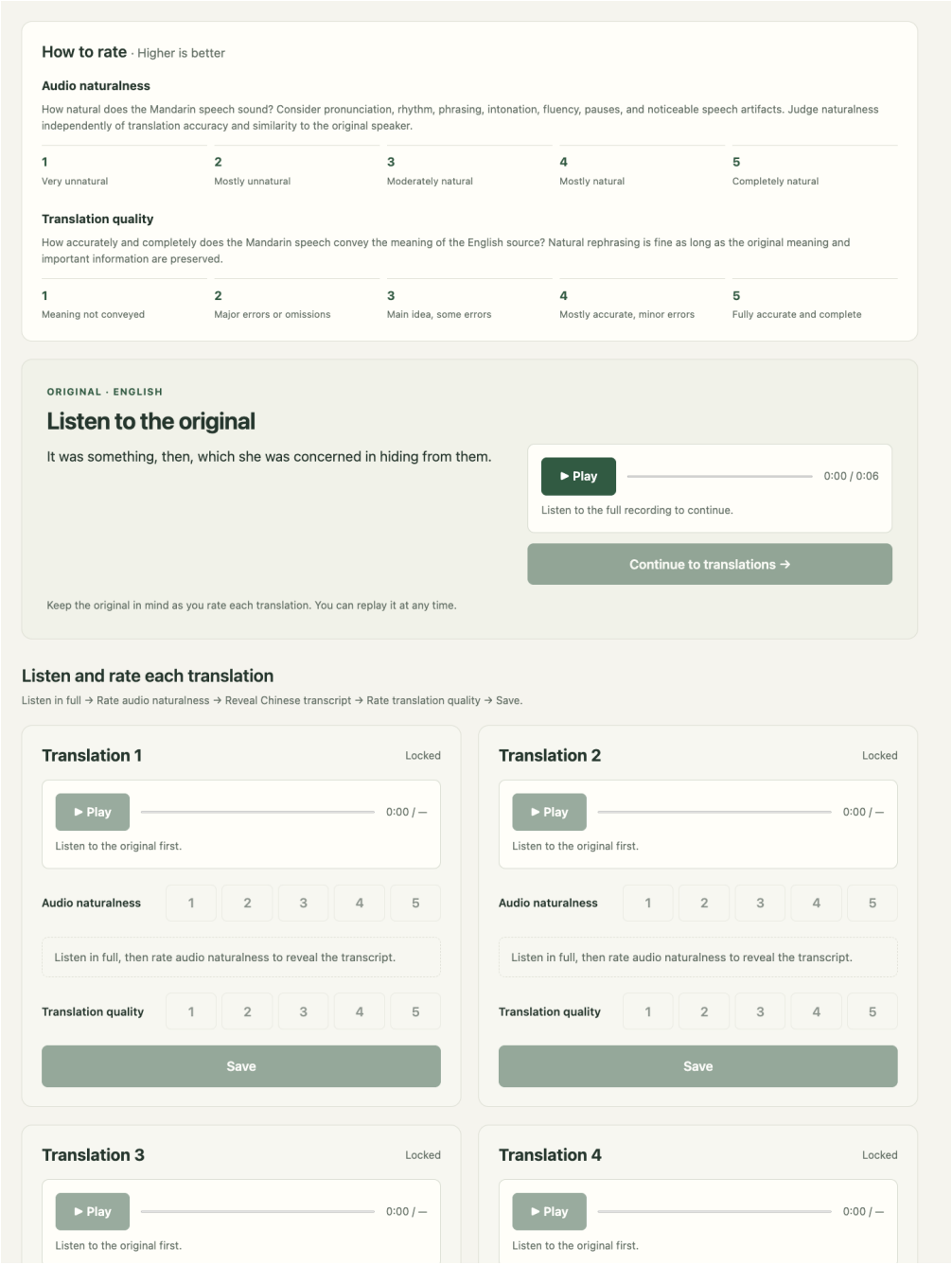


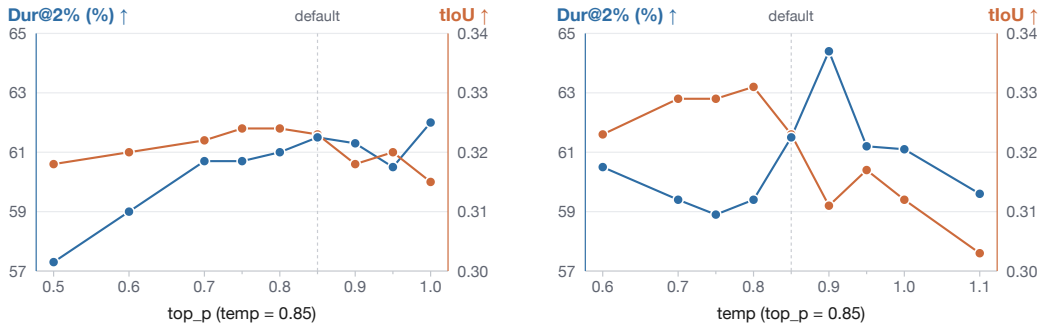
Figure 7: Human evaluation interface for audio naturalness and translation quality. Participants rate naturalness before viewing the Mandarin transcript, which is subsequently revealed for translation-quality assessment.

G Additional Experimental Results and Ablations

This appendix complements Section 4 with analyses of decoding sensitivity, domain-specific translation quality, Stage 3 supervision composition, and alternative approaches to temporal control.

Sensitivity to decoding parameters. Figure 8 varies the nucleus-sampling threshold p and temperature separately, keeping the other parameter fixed at its default value of 0.85. The two temporal

metrics do not improve monotonically together. In the top-p sweep, restricting sampling to $p = 0.5$ reduces strict duration agreement, while intermediate values provide better pause overlap than $p = 1.0$. The temperature sweep exhibits a clearer trade-off: increasing the temperature from 0.85 to 0.9 improves Dur@2% but reduces tIoU, and further increasing it to 1.1 degrades both metrics relative to the default. Thus, the default setting of $p = 0.85$ and temperature 0.85 provides a compromise between endpoint fidelity and local pause alignment rather than maximizing either metric in isolation.



(a) Top-p sweep

(b) Temperature sweep

Figure 8: Top-p and temperature decoding settings. Left axes show duration accuracy (Dur@2%); right axes show pause overlap (tIoU). Dashed lines mark the default settings.

Domain-specific translation quality. Table 9 reports translation quality separately for CoVoST2 and GigaSpeech on the ASR-mediated track. On CoVoST2, TRANSDUBBER achieves 0.745 COMET and 5.18 MetricX-QE, compared with 0.738 and 5.68 for UniSS. The difference is larger on GigaSpeech, where COMET increases from 0.732 to 0.809 and MetricX-QE decreases from 5.64 to 3.61. The reported advantage therefore appears in both the predominantly read-speech and spontaneous-speech subsets, although its magnitude differs substantially. These are within-domain comparisons; differences in absolute scores across the two subsets should not be attributed solely to speaking style, since their reference construction also differs, as described in Section 4.1.

Table 9: Domain-specific translation quality on the ASR-mediated track. Unlike Table 2, these scores are computed from ASR transcriptions of the generated speech and therefore also reflect speech realization errors.

Domain	System	COMET ↑	MetricX-QE ↓
CoVoST2	UniSS	0.738	5.68
	TRANSDUBBER	**0.745**	**5.18**
GigaSpeech	UniSS	0.732	5.64
	TRANSDUBBER	**0.809**	**3.61**

Composition of timeline-preserving supervision. Table 10 compares Stage 3 training with only no-pause examples, only pause-aligned examples, and their union. No-pause supervision yields the strongest translation scores in this comparison, with 0.793 COMET and 3.77 MetricX-QE, but provides limited local alignment: pause F1 and tIoU reach only 0.180 and 0.134. Using only pause-aligned examples raises pause F1 to 0.523, but achieves lower strict duration agreement, with Dur@2% of 28.1% compared with 33.4% for no-pause supervision. Combining both subsets yields the best Dur@2% and Dur@5%, at 48.6% and 83.3%, together with the highest tIoU of 0.293 and lowest SER of 0.198. Its pause F1 of 0.518 remains close to that of pause-aligned-only training, although its translation scores do not match those of no-pause-only training. Combining the two subsets improves duration accuracy and pause-span overlap. No-pause-only training gives better translation scores, while pause-aligned-only training gives a slightly higher pause F1.

Initialization for policy optimization. Table 11 examines RL initialization from different training stages. The Stage 1 checkpoint yields no valid rollouts, while Stage 2 achieves a 93.8% valid-rollout

Table 10: Ablation of the Stage-3 supervision composition.

Stage-3 supervision	COMET ↑	MetricX-QE ↓	Dur@2% ↑	Dur@5% ↑	Pause F1 ↑	tIoU ↑	SER ↓
$\mathcal{D}_{no\text{-}pause}$	**0.793**	**3.77**	33.4	63.9	0.180	0.134	0.258
$\mathcal{D}_{\text{pause-aligned}}$	0.767	4.52	28.1	59.7	**0.523**	0.284	0.230
Full ($\mathcal{D}_{no\text{-}pause} \cup \mathcal{D}_{\text{pause-aligned}}$)	0.772	4.38	**48.6**	**83.3**	0.518	**0.293**	**0.198**

rate. However, RL from Stage 2 leaves Dur@2% unchanged at 5.2% and improves tIoU only from 0.086 to 0.097. In contrast, Stage 3 provides a 100% valid-rollout rate and enables RL to raise Dur@2% from 48.6% to 61.5% and tIoU from 0.293 to 0.322. Thus, valid speech generation alone is insufficient for substantial temporal gains in this setting. RL is more effective at refining timeline-following behavior established through supervised training than at acquiring it from natural-pace speech generation alone.

Table 11: Reinforcement learning initialized from text grounding (Stage 1), natural-pace speech training (Stage 2), or timeline-preserving supervision (Stage 3).

RL initialization	COMET ↑		Dur@2% (%) ↑		tIoU ↑		Valid rollout rate
	Pre-RL	Post-RL	Pre-RL	Post-RL	Pre-RL	Post-RL	
Stage 1 checkpoint	0.823	N/A	–	N/A	–	N/A	0%
Stage 2 checkpoint	0.823	**0.829**	5.2	5.2	0.086	0.097	93.8%
Stage 3 checkpoint	0.772	0.788	48.6	**61.5**	0.293	**0.322**	100%

Global duration correction versus local timeline preservation. Table 12 examines whether global time-scale stretching after generation can recover source-timeline alignment. All listed systems reach 100% Dur@2% after stretching, but substantial differences in local alignment remain. The stretched baselines obtain pause F1 scores of 0.026–0.159 and tIoU scores of 0.034–0.116, still well below the 0.550 pause F1 and 0.322 tIoU achieved by TRANSDUBBER without stretching. For UniSS, correcting global duration raises pause F1 only from 0.118 to 0.131. Stretching TRANSDUBBER produces comparatively small additional gains, increasing pause F1 from 0.550 to 0.561 and tIoU from 0.322 to 0.334. The reductions in SER must also be interpreted in light of its definition: correcting duration directly reduces its endpoint-mismatch component and therefore does not, by itself, establish accurate internal pause placement. Uniformly rescaling an utterance can compensate for a global rate mismatch, but cannot independently adjust individual speaking regions and pauses. This experiment supports the distinction between matching the utterance endpoint and preserving its internal timeline. Stretching is used only for this diagnostic comparison; the main TRANSDUBBER results do not use post-hoc temporal correction.

Table 12: Impact of global time-scale stretching on temporal alignment. Each entry reports the performance before and after stretching (*before* → *after*).

System	Dur@2% ↑	Pause F1 ↑	tIoU ↑	SER ↓	onMedAE ↓
Seamless-L	3.7 → 100	0.002 → 0.026	0.008 → 0.034	0.407 → 0.143	0.533 → 0.252
Seamless-Ex	4.0 → 100	0.062 → 0.124	0.059 → 0.101	0.395 → 0.177	0.230 → 0.243
Qwen2.5-Omni-7B	4.0 → 100	0.080 → 0.159	0.060 → 0.114	0.446 → 0.177	0.211 → 0.227
UniSS	25.6 → 100	0.118 → 0.131	0.108 → 0.116	0.288 → 0.201	0.266 → 0.262
TRANSDUBBER	61.5 → 100	0.550 → **0.561**	0.322 → **0.334**	0.128 → **0.103**	0.162 → **0.156**

Comparison with synthesis-based alternatives. Table 13 compares TRANSDUBBER with F5-TTS (Chen et al., 2025) resynthesis of its translated text and the Stage 3 construction pipeline. Under the evaluated configuration, F5-TTS achieves 100% Dur@10% but yields lower pause F1 (0.263 vs. 0.550) and higher CER (8.14% vs. 2.68%), despite using the same translated text. The Stage 3 pipeline obtains better pause-alignment scores but succeeds on only 47.5% of inputs, compared with 99.4% for TRANSDUBBER; its temporal scores should therefore be interpreted alongside this substantially lower output coverage.

Table 13: Comparison with F5-TTS resynthesis of TRANSDUBBER translations and the Stage 3 construction pipeline.

System	Success (%) ↑	COMET ↑	MetricX-QE ↓	CER (%) ↓	Dur@10% ↑	Pause F1 ↑	tIoU ↑	SER ↓	onMedAE ↓
TRANSDUBBER	99.4	0.788	4.12	2.68	98.7	0.550	0.322	0.128	0.162
TRANSDUBBER translated text + F5-TTS	100.0	0.788	4.12	8.14	100.0	0.263	0.172	0.216	0.277
Stage-3 construction pipeline	47.5	0.724	4.88	5.07	100.0	**0.642**	**0.481**	**0.081**	**0.067**

H LIMITATIONS

Our experiments focus on English-to-Chinese translation of short utterances, leaving generalization to other language pairs and long-form speech unverified. Moreover, our timeline-preserving supervision relies on automatically translated and synthesized targets. Despite quality filtering, these targets may inherit errors or biases from the construction pipeline, while rejecting utterances that cannot satisfy local duration constraints may favor examples that are easier to synchronize.

Source-timeline preservation is an application-specific objective that may constrain translation flexibility and natural target-language phrasing. Our formulation captures only a subset of temporal structure, focusing on utterance-level duration and selected speech–silence events, while not explicitly modeling finer-grained prosody, word-level timing, semantic alignment, or audiovisual lip synchronization. Temporal control also remains imperfect.