Temporal-Linguistic Adaptive Streaming for Continuous Sign Language Translation

Arshia Kermani, Habib Irani, Deautaun Ross, Vangelis Metsis
Department of Computer Science
Texas State University
San Marcos, TX 78666, USA
{arshia.kermani, habibirani, drr175, vmetsis}@txstate.edu

Abstract

Real-time sign language translation must generate text incrementally as signs arrive, yet existing streaming policies treat glosses as a flat token sequence and discard the temporal rhythm of signing. Inter-gloss pauses reliably mark sentence boundaries in continuous discourse, but policies such as Wait-k cause arbitrary cross-boundary fragmentation. We propose Temporal-Linguistic Adaptive Streaming (TLAS), which fuses a Temporal Pause Detector (TPD, tracking inter-gloss interval statistics via an exponential moving average) and a Linguistic Readiness Estimator (LRE, a trained neural head on a frozen T5 encoder) through an Adaptive Fusion Gate (AFG). A proactive timeout fires *before* the next gloss arrives when the inter-gloss gap exceeds a threshold, producing clean sentence segmentation without oracle boundary information. We also contribute a synthetic discourse dataset of 1,400 ASL discourse groups with LLM-generated per-gloss timestamps and introduce a continuous-stream evaluation paradigm requiring autonomous boundary detection from an unbroken gloss stream. Under such conditions, TLAS significantly outperforms current heuristic baselines, such as Wait-k, and methods relying solely on linguistic content.

1 Introduction

Approximately 70 million people worldwide rely on sign languages as their primary mode of communication (World Health Organization, 2021). In healthcare triage, legal proceedings, and emergency response, the absence of a real-time interpreter can delay diagnoses, produce inadmissible testimony, or prevent timely access to services. Automated sign language translation systems capable of operating in real time would substantially reduce these barriers; yet the large majority of existing systems require a complete, pre-recorded input before producing any output, making them unsuitable for interactive deployment.

A central challenge in building such systems is the structural mismatch between visual sign production and spoken-language text. American Sign Language (ASL) employs spatial grammar, topic-comment ordering, and non-manual markers that do not correspond one-to-one with English syntax. A direct video-to-text approach must simultaneously solve recognition (mapping continuous visual motion to discrete linguistic units) and translation (mapping one language to another), two tasks whose error rates compound. The standard solution is a two-stage pipeline: a vision module converts video frames into a sequence of *glosses* (discrete lexical labels for individual signs), and a translation model maps the resulting gloss sequence to fluent target text. The monotonic gloss-to-token correspondence makes streaming translation tractable: as each gloss arrives, the system can decide, based on accumulated context, whether to wait for additional signs or commit to generating output. This paper addresses the gloss-to-text stage under real-time, continuous-stream conditions.

Existing streaming policies borrowed from spoken-language simultaneous translation fail to exploit a modality-specific signal that is readily available in the gloss stream: the temporal rhythm of signing. Within a sentence, inter-gloss intervals range from approximately 300 to 650 ms; between sentences, signers pause for 2 to 7 seconds. Wait-k (Ma et al., 2019) and TransLLaMa (Koshkin et al., 2024) treat the gloss stream as a flat token sequence and discard this temporal information entirely, fragmenting multi-sentence discourse arbitrarily across boundaries and forcing the translation backend to complete incoherent partial sentences or produce hallucinated continuations. A policy that monitors inter-gloss gap statistics could instead detect sentence boundaries before the next sentence begins, enabling clean segmentation and preserving discourse coherence across translations.

A further obstacle to progress is the absence of

publicly available datasets that pair continuous discourse gloss streams with per-gloss arrival timestamps. Without such a benchmark, it is impossible to measure continuous-stream segmentation quality or to study the effect of timestamp degradation (e.g., jitter introduced by a real-time vision module) on translation policies. Existing resources such as ASLG-PC12 (Othman and Jemni, 2012) provide isolated sentence pairs without temporal annotations, and sentence-level evaluation conceals the boundary-detection failures that dominate real-world deployment.

We propose **Temporal-Linguistic Adaptive Streaming** (TLAS), a streaming policy that fuses temporal pause detection and neural linguistic readiness estimation through an adaptive gate, with a proactive timeout mechanism for clean inter-sentence segmentation in continuous discourse. We further contribute (a) a synthetic discourse dataset of 1,400 ASL discourse groups with per-gloss timestamps spanning three conversation types, and (b) a continuous-stream evaluation paradigm in which policies must segment and translate an unbroken multi-sentence gloss stream without oracle boundaries. Timestamp robustness experiments demonstrate that the temporal and linguistic signals are complementary: TLAS-temporal maximizes quality under reliable timing while TLAS-linguistic provides a timestamp-invariant floor. The main contributions of this work are:

- **TLAS architecture**: a streaming policy that fuses temporal pause detection and neural linguistic readiness estimation through an adaptive gate, with a proactive timeout mechanism for clean inter-sentence segmentation in continuous discourse.
- **Continuous-stream evaluation paradigm (E2)**: a discourse-level benchmark protocol in which policies must segment and translate an unbroken multi-sentence gloss stream, together with quantitative analysis of segmentation quality across five streaming policies.
- **Synthetic discourse dataset**: 1,400 ASL discourse groups with per-gloss timestamps spanning three conversation types, released to support future research in continuous-stream sign language translation.
- **Timestamp robustness analysis**: a systematic study of translation quality under three timestamp conditions (realistic, uniform, and noisy), demonstrating that TLAS-linguistic is effectively timestamp-invariant while TLAS-temporal degrades gracefully, with implications for deployment in the presence of vision-module latency.

2 Related Work

2.1 Sign Language Translation

Sign language translation decomposes into two stages: visual recognition, which maps continuous video to a discrete gloss sequence, and gloss-to-text translation, which produces fluent target language output (Kermani et al., 2025). Early recognition systems combined hand-crafted visual features with HMM-based sequence models to handle multiple signers across large vocabularies (Koller et al., 2015). Camgöz et al. (Camgöz et al., 2018) introduced the first neural gloss-to-text system, adapting attention-based sequence-to-sequence architectures originally developed for spoken language translation and establishing BLEU as the field's standard metric. Transformer-based systems that jointly optimize recognition and translation end-to-end further improved performance (Camgöz et al., 2020). Data augmentation via sign back-translation (Zhou et al., 2021) and pretraining with multilingual models such as mBART (Liu et al., 2020) and T5 (Raffel et al., 2020) have subsequently pushed translation quality on standard benchmarks. Continuous-discourse resources such as How2Sign (Duarte et al., 2021) for American Sign Language and the TVB-HKSL-News corpus (Niu et al., 2024) for Hong Kong Sign Language provide multi-sentence content beyond isolated sentence pairs, but they do not pair gloss streams with the per-gloss arrival timestamps required to evaluate streaming policies under realistic temporal conditions. All of these systems, however, assume batch access to a complete gloss sequence before producing any output, rendering them unsuitable for the interactive scenarios in which real-time translation is most urgently needed.

2.2 Simultaneous Translation

Simultaneous (streaming) translation generates target tokens incrementally as source tokens arrive. Ma et al. (Ma et al., 2019) proposed the Wait-k policy, which delays output by a fixed k source tokens and then alternates read and write steps; its deterministic simplicity makes it a durable baseline

despite the inability to adapt to content boundaries. Arivazhagan et al. (Arivazhagan et al., 2019) introduced Monotonic Infinite Lookback (MILk) attention, which replaces the hard lag with a learned, monotonically constrained attention distribution; Monotonic Multihead Attention (MMA) (Ma et al., 2020) extended this mechanism to multi-head settings with independent per-head step decisions, enabling end-to-end training of the read/write policy. More recently, Koshkin et al. (Koshkin et al., 2024) demonstrated that large language models can be fine-tuned to emit a special <WAIT> token when context is insufficient for translation, integrating the policy decision into the model itself.

These approaches share a critical limitation when applied to sign language: read/write decisions are conditioned exclusively on linguistic content, and the temporal dimension of the gloss stream is discarded entirely. In continuous signing, inter-gloss intervals within a sentence (300–650 ms) differ by an order of magnitude from the pauses between sentences (2–7 s). To our knowledge, no prior streaming translation policy explicitly models inter-gloss timing as a boundary signal, and applying existing policies to a continuous discourse stream causes arbitrary cross-boundary fragmentation rather than coherent sentence-by-sentence translation.

3 Methodology

3.1 Problem Formulation

We formalize streaming sign language translation as a read/write decision problem over a time-stamped token stream. Each incoming gloss is a pair (g_t, τ_t), where $g_t \in \mathcal{V}$ is a discrete gloss token and $\tau_t \in \mathbb{R}^+$ is the wall-clock arrival time in seconds since stream onset. The system maintains a gloss buffer $\mathcal{B}_t = \{(g_1, \tau_1), \ldots, (g_t, \tau_t)\}$ and at each step issues either **READ** (wait for the next gloss) or **WRITE** (flush $\mathcal{B}_t$ to the translation backend and reset). In the continuous-stream setting, the input is a multi-sentence discourse group delivered as a single unbroken sequence with no oracle boundary markers; a WRITE at a sentence boundary yields a coherent translation segment, while a premature or delayed WRITE produces a fragment or cross-boundary concatenation that no translation model can recover from. The objective is to maximize translation quality while maintaining discourse coherence across consecutive translations.

3.2 TLAS Architecture

The TLAS policy integrates two complementary readiness signals through a learned decision gate. A Temporal Pause Detector (TPD) monitors inter-gloss arrival times and emits a high score during inter-sentence pauses. A Linguistic Readiness Estimator (LRE) scores the grammatical completeness of the accumulated buffer using a neural head on a frozen T5 encoder (Raffel et al., 2020). An Adaptive Fusion Gate (AFG) combines both scores to issue READ or WRITE decisions. A proactive timeout mechanism fires between glosses when the current inter-gloss gap exceeds M times the signer's running average, producing clean sentence boundaries before the first gloss of the next sentence arrives. Figure 1 illustrates the complete data flow.

3.2.1 Temporal Pause Detector (TPD):

Upon receiving gloss g_t at time τ_t, the TPD computes the interval $\Delta_t = \tau_t - \tau_{t-1}$ and updates an exponential moving average:

$$\hat{\mu}_t = \alpha \cdot \Delta_t + (1 - \alpha) \cdot \hat{\mu}_{t-1} \tag{1}$$

where $\alpha = 0.3$ controls the adaptation rate. The EMA is initialized at $\hat{\mu}_0 = 450$ ms, set to the midpoint of the within-sentence inter-gloss range (300–650 ms), to suppress cold-start fluctuations during the first few glosses. The pause score is a linear ramp between normal pace and the pause multiplier $M = 2.5$:

$$s_t^{\text{TPD}} = \text{clamp}\left(\frac{\Delta_t/\hat{\mu}_t - 1}{M - 1},\ 0,\ 1\right) \tag{2}$$

When $\Delta_t = \hat{\mu}_t$ (ratio 1.0), the score is 0; when $\Delta_t \geq M \cdot \hat{\mu}_t$, the score saturates at 1. Because the score is defined relative to $\hat{\mu}_t$ rather than a fixed threshold, the TPD adapts automatically to each signer's natural pace without recalibration.

3.2.2 Linguistic Readiness Estimator (LRE):

The LRE estimates the semantic completeness of the accumulated gloss buffer using a lightweight neural head on the T5 encoder. Given encoder hidden states $H \in \mathbb{R}^{B \times L \times d}$, the head computes an attention-masked mean pool $\bar{h} \in \mathbb{R}^d$ over non-padding positions and maps it to a scalar readiness score:

$$l_t = \sigma\big(W_2\,\text{ReLU}(W_1\bar{h} + b_1) + b_2\big) \tag{3}$$

where $W_1 \in \mathbb{R}^{256 \times 768}$ and $W_2 \in \mathbb{R}^{1 \times 256}$, with dropout 0.1 between layers ($\approx$197K parameters).

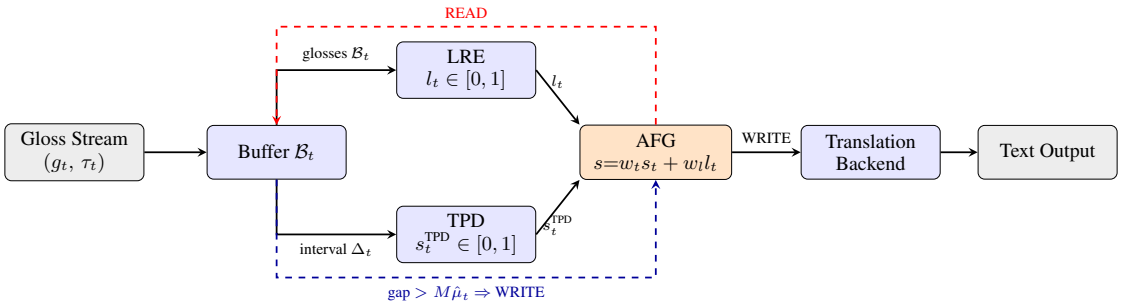

Figure 1: TLAS architecture. The LRE encodes accumulated glosses with a frozen T5 encoder and maps the pooled representation to a completeness score l_t; the TPD maps the inter-gloss interval Δ_t to a pause score s_t^{TPD}. The AFG combines both signals and issues WRITE or READ (dashed red). A proactive timeout (dashed blue) issues WRITE directly when the elapsed gap exceeds $M \cdot \hat{\mu}_t$.

The head is trained separately after T5 fine-tuning. Oracle readiness labels are generated by translating every gloss prefix $G_{1:t}$ with the frozen T5 model and computing ROUGE-L (Lin, 2004) against the reference:

$$r_t^* = \text{ROUGE-L}(T5(G_{1:t}),\ y^*) \tag{4}$$

Monotonicity is enforced by replacing each label with the running maximum: $r_t^* \leftarrow \max(r_t^*,\ r_{t-1}^*)$. The head is then trained with MSE loss on these oracle scores. For non-T5 backends, the LRE operates as a standalone scorer: the fine-tuned T5 encoder and LRE head are loaded locally and shared across all TLAS instances, incurring no per-gloss API calls.

3.2.3 Adaptive Fusion Gate (AFG):

The AFG combines both signals through a weighted sum:

$$s = w_t \cdot s_t^{\text{TPD}} + w_l \cdot l_t \tag{5}$$

with $w_t = 0.4$ and $w_l = 0.6$, giving linguistic completeness a slightly higher weight because it directly predicts translation quality. The weights w_t and w_l are empirically tuned hyperparameters. The gate evaluates three ordered conditions and issues WRITE on the first satisfied:

1. **Safety valve**: buffer length $|\mathcal{B}_t| \geq L_{\max} = 6$. Prevents unbounded accumulation when both signals are weak.

2. **Joint threshold**: $s \geq \theta = 0.40$. With $w_l = 0.6$, this requires $l_t \geq 0.67$ in the absence of any temporal signal, i.e., the LRE is confident that the buffer forms a complete translatable unit.

3. **Strong pause override**: $s_t^{\text{TPD}} \geq 0.8$ and $l_t \geq 0.3$. Trusts an unambiguous signer pause even when the LRE is not yet fully confident, reflecting the high correlation between long inter-sentence pauses and syntactic completion in ASL.

If none is satisfied, the AFG returns READ. We note that the fusion weights w_t and w_l and the AFG thresholds are tuned on a held-out validation set rather than learned end-to-end (see Section 4.4); treating the gate as a differentiable policy and optimizing it directly under a quality–latency reward is a natural direction for future work (Section 6).

3.2.4 Proactive Timeout:

The most discriminative boundary signal in continuous discourse arrives *between* glosses. The proactive timeout operates outside the per-gloss TPD→LRE→AFG pipeline: when the elapsed time since the last gloss exceeds $M \cdot \hat{\mu}_t$ and at least one gloss is buffered, the system issues a WRITE directly, before the next token arrives. Given typical within-sentence averages of $\hat{\mu}_t \approx 450$ ms and $M = 2.5$, this timeout fires approximately 1.1 s into an inter-sentence pause, cleanly separating two sentences without waiting for the first gloss of the following sentence to confound segmentation. The proactive timeout is disabled for the TLAS-linguistic ablation, which must operate purely on linguistic content.

3.3 Ablations and Baselines

We evaluate three TLAS configurations to isolate each component's contribution:

- **TLAS** (full): w_t=0.4, w_l=0.6; both TPD and LRE active; proactive timeout enabled.

- **TLAS-temporal**: w_t=1.0, w_l=0.0; the LRE returns a constant 0 but the AFG structure is otherwise unchanged; proactive timeout enabled. Isolates the temporal signal.
- **TLAS-linguistic**: w_t=0.0, w_l=1.0; the TPD EMA is updated normally but its score is clamped to 0 before reaching the AFG; proactive timeout disabled. Isolates the linguistic signal.

All three configurations use identical AFG thresholds, ensuring that performance differences reflect signal quality rather than threshold tuning. We compare against four baselines: **Batch (oracle)**, which splits at ground-truth boundaries (upper bound); **Batch (non-oracle)**, which accumulates every gloss in the discourse group and submits them as a single concatenated string to the translator, producing one output for what may be five or six sentences of content—because the single output aligns only partially with the first reference sentence and has no correspondence to later ones, BLEU collapses at positions 1 and beyond, and the gap between Batch (non-oracle) and Batch (oracle) quantifies the full value of correct sentence boundary detection; **Wait-k** (k=3) (Ma et al., 2019), which alternates read/write with fixed lag; and **TransLLaMa** (Koshkin et al., 2024), which emits `<WAIT>` when context is insufficient. All baseline policies are evaluated with a discourse context window of zero prior translations, while TLAS variants use a sliding window of three. This asymmetric configuration reflects each policy's natural operating regime rather than disadvantaging the baselines: when a baseline fragments a discourse stream across sentence boundaries, its prior outputs are cross-boundary fragments rather than coherent sentence translations, so populating its context buffer with such outputs would propagate segmentation errors into subsequent prompts and further degrade quality. Baselines therefore operate under the more favorable zero-context setting. An evaluation under matched zero-context conditions for all policies, which would further isolate the segmentation contribution from any residual benefit of context, is left to future work.

3.4 Translation Backends

TLAS is backend-agnostic: the policy layer calls `translate(buffer, context)` and receives a text string, with no assumptions about the underlying model. We evaluate three backends.

T5 (local, fine-tuned): We fine-tune T5-base (Raffel et al., 2020) on a multi-source training mixture comprising 50K ASLG-PC12 gloss–English pairs, SIGNUM German Sign Language pairs for cross-domain vocabulary exposure, and TransLLaMa-style streaming examples in which early prefixes (first third of glosses) target the special `<WAIT>` token and mid-prefixes (first half) target partial translations. Discourse context is incorporated by prepending a sliding window of the three most recent prior translations as a prefix, separated by a context delimiter. Training uses AdamW with effective batch size 32, learning rate 5×10^{-5}, warmup ratio 0.1, weight decay 0.01, label smoothing 0.1, and FP16 mixed precision for 5 epochs on a single GPU.

Gemini (cloud API): We use `gemini-3.1-flash-lite-preview` via the Google AI API. Each WRITE invocation sends a structured prompt containing the discourse context window and the buffered glosses; the model returns a single-sentence translation. No fine-tuning is applied.

Ollama (local API): We deploy `gpt-oss:120b` through a local Ollama server, providing an on-premise alternative that avoids data egress. The same prompt format as Gemini is used. For both non-T5 backends, the LRE standalone scorer loads the fine-tuned T5 encoder locally and computes readiness scores without invoking the external API.

4 Experimental Setup

In this section, we describe our experiments and evaluation methodology. All hyperparameters, training scripts, evaluation code, and the synthetic discourse dataset are publicly available.[1]

4.1 Datasets

ASLG-PC12 (Othman and Jemni, 2012) is a parallel corpus of approximately 87,000 ASL gloss–English pairs. We use a shuffled split of 50,000 training pairs, 10,000 validation pairs, and 10,000 test pairs (seed 42); sentence-level evaluation (E1) draws 100 test examples with synthetic uniform timestamps (450 ms per gloss).

Synthetic Discourse Dataset (contributed in this work) comprises 1,400 multi-sentence discourse groups generated via Gemini LLM, spanning monologue (≈40%), deaf-deaf dialog

[1] https://github.com/imics-lab/tlas-gloss2text

(≈30%), and deaf-hearing dialog (≈30%). Per-gloss timestamps are calibrated to real ASL timing: 300–650 ms within sentences, 1.5–7 s between sentences. The first 200 groups (888 deaf sentences) constitute the test split; the remaining 1,200 groups (≈5,234 context pairs) form the training split.

SIGNUM consists of 779 German Sign Language sentence pairs from everyday conversational topics; we include it during T5 fine-tuning for cross-domain vocabulary exposure and reserve a held-out subset for cross-dataset generalization evaluation.

4.2 Evaluation Paradigms

E1 (sentence-level) evaluates 100 ASLG-PC12 test sentences with synthetic uniform timestamps (450 ms per gloss). Since all policies operate within known sentence boundaries, the temporal signal is neutralized and TLAS degenerates to max-lag triggering. E1 isolates pure linguistic translation quality but cannot assess boundary detection.

E2 (continuous-stream, primary experiment) feeds 200 test discourse groups as unbroken gloss streams using LLM-generated timestamps, with no oracle sentence boundaries provided. Inter-sentence pauses of 1.5–7 s give TLAS a strong temporal signal via the proactive timeout, while baselines fragment the stream arbitrarily. We report corpus-level BLEU and break results down by discourse position to measure quality drift within a group.

4.3 Evaluation Metrics

We report four metrics: **BLEU** (Papineni et al., 2002) (corpus-level n-gram precision, computed with SacreBLEU); **ROUGE-L** (Lin, 2004) (longest common subsequence recall); **SBERT cosine similarity** (semantic similarity via Sentence-BERT, robust to paraphrase); and **chrF++** (character n-gram F-score with word unigram recall, informative for morphologically varied output). For E2 we additionally report **retention rate** (policy BLEU divided by oracle BLEU) and **position-stratified BLEU** (computed per discourse position to reveal whether segmentation errors accumulate across a group).

4.4 Hyperparameters

Table 1 summarizes all TLAS hyperparameters. The TPD smoothing factor α and pause multiplier M were set from the inter-gloss timing statistics of the training data; the AFG weights and thresholds were validated on 20 held-out discourse groups not included in the test split. None of these values are learned end-to-end; the gate is a tuned policy rather than a trained one.

Table 1: TLAS hyperparameter settings.

Component	Parameter	Value
TPD	EMA smoothing α	0.3
	Pause multiplier M	2.5
	EMA prior $\hat{\mu}_0$	450 ms
LRE	Hidden dim	256
	Dropout	0.1
AFG	w_t (temporal weight)	0.4
	w_l (linguistic weight)	0.6
	Joint threshold θ	0.40
	Strong pause threshold	0.80
	Min. readiness for pause	0.30
	Max lag $L_{\max}$	6
Policy	Discourse context window	3
	Wait-k lag k	3

Table 2: E2 continuous-stream results, T5 backend (200 groups, 888 sentences). Ret. = BLEU / oracle BLEU. Best streaming value in **bold**.

Policy	BLEU	ROUGE-L	Ret.
Batch (oracle) ↑	24.49	.593	—
Batch	1.74	.068	7.1%
Wait-k (k=3)	1.86	.286	7.6%
TransLLaMa	4.76	.385	19.4%
TLAS-linguistic	17.18	.474	70.1%
TLAS	21.51	.536	87.9%
TLAS-temporal	**23.70**	**.564**	**96.8%**

5 Results and Discussion

5.1 E2: Continuous-Stream Discourse Evaluation

Table 2 presents T5 results on 200 discourse groups (888 sentences). Two distinct performance tiers are immediately apparent. The three TLAS variants cluster between 17 and 24 BLEU (70–97% retention), while all baselines fall below 5 BLEU (under 20% retention). This separation holds across every metric (BLEU, ROUGE-L, SBERT, and chrF++), confirming that the gap is not an artifact of the BLEU brevity penalty.

5.1.1 Cross-Backend Results:

Table 3 extends the comparison across all three backends. TLAS-temporal and full TLAS consistently achieve retention above 80%, while traditional baselines peak at 56.8%. The higher absolute retention of the Gemini and Ollama baselines relative to T5 is consistent with the larger LLMs partially masking segmentation failures by com-

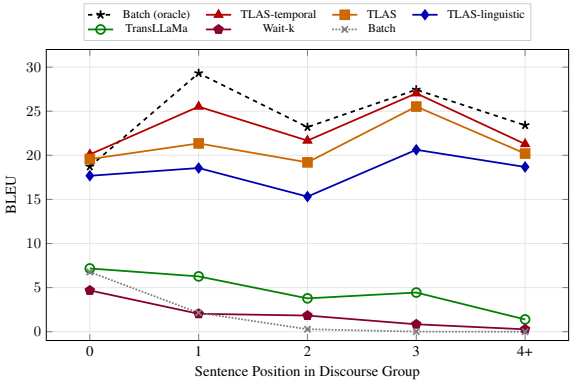

Figure 2: Position-stratified BLEU across discourse positions in E2 (T5 backend, 200 groups). TLAS variants track the oracle trajectory at all positions, while baselines collapse by position 2 due to cross-boundary fragmentation. The gap between tiers widens at later positions, confirming that segmentation errors accumulate monotonically for policies lacking temporal boundary detection.

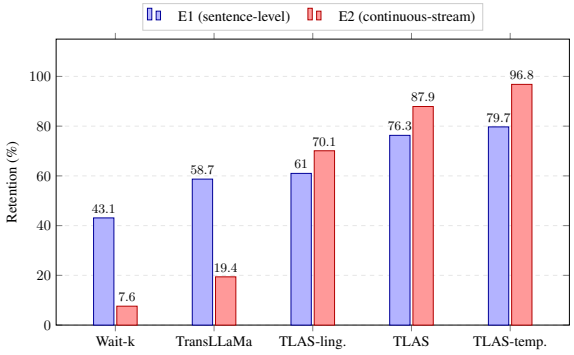

Figure 3: Retention rates under sentence-level (E1) vs. continuous-stream (E2) evaluation (T5 backend). Wait-k and TransLLaMa collapse from E1 to E2 (43.1%→7.6% and 58.7%→19.4%, respectively), while TLAS variants maintain or improve retention under discourse conditions, exposing a critical failure mode that sentence-level evaluation conceals.

pleting incoherent fragments into plausible-looking text rather than reflecting genuinely better segmentation; in either case, TLAS variants still lead every traditional baseline by at least 24 percentage points on every backend. Figure 2 confirms this: TLAS variants track the oracle trajectory across all five discourse positions, while Wait-k and Batch decline monotonically to near-zero BLEU by position 3, indicating cumulative segmentation failure rather than translation model weakness.

5.1.2 Position-Stratified Analysis:

Table 4 reports BLEU stratified by sentence position within the discourse group (T5 backend). This analysis reveals the mechanism underlying the two-tier separation. Wait-k collapses from 4.68 BLEU at position 0 to 0.27 at position 4+; Batch (non-oracle) reaches 0.00 by position 3. The three TLAS variants maintain stable quality across all positions, with TLAS-temporal staying within 5 BLEU points of the oracle at every depth. The root cause is boundary detection, not translation capability per se: baselines that fragment across sentence boundaries submit incoherent gloss subsets to the translator, producing outputs with near-zero overlap with any individual reference sentence.

5.1.3 Timestamp Robustness:

Table 5 reports BLEU under three timestamp conditions for TLAS variants on E2 (T5 backend). TLAS-linguistic varies by less than 0.1 BLEU points across all three conditions, confirming zero sensitivity to timestamp quality. TLAS-temporal degrades by 57% from realistic to noisy timestamps (23.70 $\rightarrow$ 10.07): Gaussian jitter of σ=500 ms corrupts the EMA, causing premature or delayed proactive timeouts. A crossing point occurs at uniform timestamps: TLAS-linguistic (17.09) overtakes TLAS-temporal (15.61), because uniform gaps suppress the temporal signal to exactly zero while the LRE continues to operate normally. Full TLAS degrades more gracefully than temporal-only (11.53 vs. 10.07 under noise), since the linguistic weight provides a stabilizing floor; however, the remaining 46% decline from realistic to noisy conditions confirms that the temporal component remains the dominant vulnerability when timestamp quality is poor.

5.2 E1: Sentence-Level Results

Table 6 presents sentence-level results on 100 ASLG-PC12 test examples with synthetic uniform timestamps (T5 backend). Because all policies receive isolated sentences with no inter-sentence pauses, the TPD emits near-zero scores throughout and TLAS degenerates to max-lag ($L_{\max}$=6) triggering. Under these conditions TLAS-temporal (79.7% retention) and full TLAS (76.3%) lead all streaming methods, demonstrating that the max-lag safety valve provides adequate within-sentence segmentation quality even when the temporal cue is suppressed.

Table 3: E2 results across all three backends. Ret. = BLEU / backend oracle BLEU.

	T5 (oracle: 24.49)		**Gemini** (oracle: 24.39)		**Ollama** (oracle: 12.53)	
Policy	BLEU	Ret.	BLEU	Ret.	BLEU	Ret.
Batch	1.74	7.1%	5.71	23.4%	1.90	15.2%
Wait-k	1.86	7.6%	11.49	47.1%	7.12	56.8%
TransLLaMa	4.76	19.4%	12.54	51.4%	5.64	45.0%
TLAS-linguistic	17.18	70.1%	15.79	64.7%	7.53	60.1%
TLAS	21.51	87.9%	21.46	88.0%	10.23	81.6%
TLAS-temporal	**23.70**	**96.8%**	**22.13**	**90.7%**	**11.56**	**92.3%**

Table 4: Position-stratified BLEU in E2 (T5 backend). Position 0 = first sentence in group; 4+ = fifth or later.

Policy	**Pos 0**	**Pos 1**	**Pos 2**	**Pos 3**	**Pos 4+**
Batch (oracle) ↑	18.73	29.31	23.20	27.44	23.41
Batch	6.77	2.14	0.28	0.01	0.00
Wait-k	4.68	2.03	1.83	0.84	0.27
TransLLaMa	7.17	6.27	3.78	4.44	1.39
TLAS-linguistic	17.68	18.56	15.32	20.63	18.68
TLAS	19.57	21.35	19.20	25.53	20.20
TLAS-temporal	**20.10**	**25.52**	**21.67**	**27.03**	**21.28**

Table 5: Timestamp robustness: BLEU under three conditions (T5, E2). Oracle: 24.49.

Policy	**Realistic**	**Uniform**	**Noisy**
TLAS-temporal	**23.70**	15.61	10.07
TLAS	21.51	16.68	11.53
TLAS-linguistic	17.18	**17.09**	**17.18**

Table 6: E1 sentence-level results, T5 backend (n=100, synthetic timestamps). Ret. = BLEU / oracle BLEU.

Policy	**BLEU**	**ROUGE-L**	**Ret.**
Batch ↑	73.84	.919	—
TLAS-temporal	58.88	.869	79.7%
TLAS	56.36	.867	76.3%
TLAS-linguistic	45.04	.824	61.0%
TransLLaMa	43.35	.821	58.7%
Wait-k (k=3)	31.84	.757	43.1%

5.3 Discussion

The E2 results establish that correct sentence boundary detection is the primary determinant of translation quality in continuous discourse. TLAS-temporal's 96.8% retention demonstrates that the temporal signal alone is sufficient for near-perfect segmentation when timestamps are reliable; the EMA $\hat{\mu}_t$ tracks within-sentence timing and cleanly detects inter-sentence pauses via the proactive timeout. The fusion design is not a simple average: the strong-pause override ensures unambiguous temporal boundaries are respected even when the LRE is uncertain, while the joint threshold allows the LRE to accelerate translation for linguistically complete prefixes when temporal cues are weak. Figure 3 quantifies this paradigm gap: baselines lose 35–39 percentage points of retention moving from E1 to E2, while TLAS-temporal gains 17 points (79.7%→96.8%), demonstrating that continuous-stream evaluation is essential for any policy intended for real-world deployment.

The contrast between E1 and E2 underscores the necessity of continuous-stream evaluation. Wait-k achieves 43.1% retention on E1 but only 7.6% on E2; TransLLaMa achieves 58.7% on E1 but only 19.4% on E2. These collapses do not appear in E1 because oracle boundaries are implicit in the single-sentence structure, eliminating cross-boundary fragmentation entirely. Prior streaming translation benchmarks operate exclusively in E1-like settings, which cannot detect this failure mode. Under E2, TLAS-temporal's retention advantage over Wait-k grows from 36.6 percentage points (E1) to 89.2 percentage points (E2), a three-fold amplification that sentence-level evaluation conceals. Any streaming policy intended for deployment in continuous signing should be evaluated under an E2-like paradigm; an E1-only evaluation cannot assess the core challenge of real-time boundary detection.

The timestamp robustness experiments (Table 5) carry direct deployment implications. In systems with reliable sub-100 ms timestamp precision (purpose-built sign recognition pipelines), TLAS-temporal maximizes quality. In systems with known timing variance (video codecs, high-latency GPU inference), TLAS-linguistic provides a robust fallback. The full fusion is recommended as the default when timestamp reliability is unknown or variable, as it outperforms linguistic-only under reliable timing and outperforms temporal-only under noise.

Several limitations bound the current study. First, the discourse dataset is synthetically generated; while per-gloss timestamps are validated against

empirical signing rates, the conversation topics, turn structures, and inter-sentence pauses may not fully capture the rhythm of authentic Deaf community interactions, which include hesitations, self-corrections, and more variable pacing. Second, the end-to-end system from raw video to text has not been integrated, and recognition errors from an upstream vision module would compound with the TPD and LRE in ways not measured here: gloss insertions and deletions would corrupt the EMA, perturbing both the inter-gloss interval statistic and the proactive timeout, while noisy gloss content would degrade the LRE's readiness estimates. Third, wall-clock inference latency under real-time constraints, where GPU memory bandwidth, tokenizer overhead, and API round-trip times all compound, has not been characterized; a full latency analysis is necessary before deployment claims can be substantiated. Finally, the AFG fusion weights and thresholds are tuned on a held-out validation set rather than learned end-to-end, and an evaluation under matched zero-context conditions for all policies, which would most cleanly isolate the segmentation contribution from any residual context benefit, is left for future work.

6 Conclusion

We presented TLAS, a streaming policy that fuses temporal pause detection and learned linguistic readiness as complementary signals for continuous sign language segmentation and translation. On the primary E2 evaluation (200 discourse groups, 888 sentences), TLAS-temporal achieves 96.8% retention of oracle quality while Wait-k reaches only 7.6%. Position-stratified BLEU confirms that TLAS sustains near-uniform quality across all discourse positions, whereas baselines collapse by position 2 due to cross-boundary fragmentation. Timestamp robustness analysis shows the signals are complementary: TLAS-temporal maximizes performance under reliable timing while TLAS-linguistic provides a stable 70.1% floor invariant to timestamp noise. Future work will explore replacing the T5 decoder with a state-space model for $O(1)$ per-step inference and implicit cross-sentence context propagation, and optimize the AFG weights via reinforcement learning to replace the currently tuned gate with a policy adapted directly to the deployment environment.

References

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, Chung-Cheng Chiu, Semih Yavuz, Ruoming Pang, Wei Li, and Colin Raffel. 2019. Monotonic infinite lookback attention for simultaneous machine translation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28–August 2, 2019, Volume 1: Long Papers*, pages 1313–1323. Association for Computational Linguistics.

Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018*, pages 7784–7793. IEEE.

Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13–19, 2020*, pages 10020–10030. Computer Vision Foundation / IEEE.

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. 2021. How2Sign: A large-scale multimodal dataset for continuous American sign language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2735–2744.

Arshia Kermani, Habib Irani, and Vangelis Metsis. 2025. Finetuning pre-trained language models for bidirectional sign language gloss to text translation. In *Proceedings of the Workshop on Sign Language Processing (WSLP)*, pages 73–81.

Oscar Koller, Jens Forster, and Hermann Ney. 2015. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125.

Roman Koshkin, Katsuhito Sudoh, and Satoshi Nakamura. 2024. TransLLaMa: LLM-based simultaneous translation system. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12–16, 2024*, pages 461–476. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: a package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. 2019. STACL: simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28–August 2, 2019, Volume 1: Long Papers*, pages 3025–3036. Association for Computational Linguistics.

Xutai Ma, Juan Miguel Pino, James Cross, Liezl Puzon, and Jiatao Gu. 2020. Monotonic multihead attention. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020*. OpenReview.net.

Zhe Niu, Ronglai Zuo, Brian Mak, and Fangyun Wei. 2024. A Hong Kong sign language corpus collected from sign-interpreted TV news. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 636–646, Torino, Italy. ELRA Language Resources Association (ELRA) and the International Committee on Computational Linguistics (ICCL).

Achraf Othman and Mohamed Jemni. 2012. English-ASL gloss parallel corpus 2012: ASLG-PC12. In *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, LREC 2012*, pages 151–154, Istanbul, Turkey.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6–12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

World Health Organization. 2021. Deafness and hearing loss. `https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss`. Accessed: 2026-03-18.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving sign language translation with monolingual data by sign back-translation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19–25, 2021*, pages 1316–1325. IEEE.