HOMURA: Taming the Sand-Glass for Time-Constrained LLM Translation via Reinforcement Learning

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have achieved remarkable strides in multilingual translation but are hindered by a systemic cross-lingual verbosity bias, rendering them unsuitable for strict time-constrained tasks like subtitling and dubbing. Current prompt-engineering approaches struggle to resolve this conflict between semantic fidelity and rigid temporal feasibility. To bridge this gap, we first introduce Sand-Glass, a benchmark specifically designed to evaluate translation under syllable-level duration constraints. Furthermore, we propose HOMURA, a reinforcement learning framework that explicitly optimizes the trade-off between semantic preservation and temporal compliance. By employing a constrained reinforcement learning objective featuring a novel dynamic syllable-ratio reward, HOMURA effectively "tames" the output length. Experimental results demonstrate that HOMURA significantly outperforms strong baselines, achieving precise length control that respects linguistic density hierarchies without compromising semantic adequacy.

1 Introduction

With the rapid advancement of Large Language Models (LLMs), their multilingual capabilities have achieved unprecedented semantic adequacy and fluency in Neural Machine Translation (NMT) (Hendy et al., 2023; Jiao et al., 2023; Zhu et al., 2024). However, this linguistic prowess is accompanied by a systemic cross-lingual verbosity bias (Briakou et al., 2024; Manakhimova et al., 2024). As illustrated in Figure 1, LLM-generated translations tend to be significantly longer than source utterances, a "chattiness" that persists even when semantic content remains constant. While acceptable in general text generation, this expansion fails in time-constrained multimedia localization (e.g., subtitles, dubbing, interpretation), where translation length constitutes a rigid temporal budget that strictly limits output duration (Karakanta et al., 2020b; Jørgensen and Mengshoel, 2025; Gaido et al., 2024). To operationalize the temporal budget, we approximate the available speaking time with a temporally grounded syllable budget scaled to each target language.

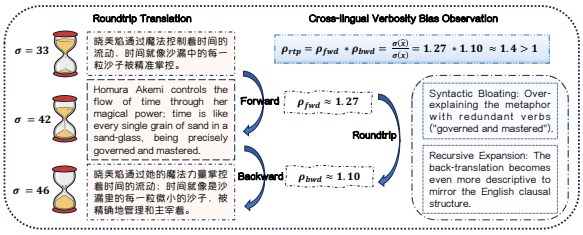


Figure 1: Illustration of cross-lingual verbosity bias. Unlike forward metrics, the **Roundtrip Expansion Ratio** ρ_{rtp} isolates model-induced redundancy from linguistic density shifts. A value of $\rho_{\text{rtp}} > 1.0$ indicates systemic inflation despite constant semantic content. Formal definition and diagnostic study are provided in Section 2.

Viewed through an information-theoretic lens (Shannon, 1948; Davisson, 1972), these temporal budgets act as a strict rate constraint: with a limited number of target syllables available, the model must pack maximal meaning into minimal length, operating near a rate-distortion limit on the achievable fidelity compression trade-off. In our setting, the rate is the syllable budget, and distortion can be understood as the loss in meaning.

Unfortunately, off-the-shelf LLMs struggle to respect these limits. This failure is rooted in the post-training alignment process, where learned reward models frequently exhibit a "length bias" that conflates verbosity with quality (Saito et al., 2023; Singhal et al., 2023). Moreover, cross-lingual typological differences systematically skew length ratios, transforming length control from a formatting task into a first-class modeling challenge. [1]

[2] Our code and data are available at: https://anonymous.4open.science/r/HOMURA-4A79/.

[1] Detailed related works can be found in appendix A.

To address this fundamental mismatch between LLM capabilities and temporal constraints, we argue that length control should be treated as an optimization objective rather than a post-hoc adjustment. To this end, we first operationalize the problem by constructing *Sand-Glass*, a dedicated benchmark for time-constrained media localization. The name reflects the rigid temporal budget that "drains" as the model produces redundant tokens. Unlike generic translation datasets, Sand-Glass incorporates syllable-based duration proxies calibrated by Information Density (ID), derived from real-world colloquial subtitles enabling precise evaluation of length compliance.

Building on this testbed, we propose HOMURA (**H**ard **O**ptimization for **M**ultilingual **U**tterance **R**eduction and **A**lignment), a reinforcement learning framework designed to "tame the sandglass" by explicitly regulating translation length without requiring supervised compression datasets. Instead of relying on brittle prompts, HOMURA optimizes the quality compression trade-off directly. We design a novel reference-free reward function that combines a dynamic syllable ratio penalty with a semantic fidelity reward, encouraging the model to condense information density without sacrificing core meaning. Our core contributions are as follows:

- **Diagnostic Analysis:** We systematically quantify the verbosity bias in LLM-based translation under temporal constraints, providing cross-lingual diagnostics (e.g., the ρ_{rtp} metric) that reveal the limitations of unconstrained models as visualized in Figure 1.
- **Benchmark Construction:** We introduce SAND-GLASS, a specialized dataset tailored for time-constrained translation, featuring syllable-level duration budgets to simulate real-world dubbing and subtitling scenarios.
- **Methodological Innovation:** We introduce HOMURA, a constrained RL framework for temporally grounded translation, which substantially improves budget compliance while maintaining semantic fidelity over strong LLM baselines.

2 Pilot Study: Benchmark and Verbosity Bias

2.1 The Sand-Glass Benchmark

To diagnose cross-lingual verbosity bias under temporal constraints, we introduce SAND-GLASS, a benchmark built from colloquial video transcripts. We collect real-world videos from five domains—*Gaming*, *Film & Television*, *Travel & Tourism*, *ACGN*, and *General Knowledge*—and segment ASR transcripts into subtitle-style units with explicit duration budgets.

Construction Procedure.

(1) Data Collection and Segmentation. We transcribe real-world videos with ASR and segment the transcripts according to pauses and semantic boundaries. Each segment inherits its original speech duration as the temporal budget for translation.

(2) Preprocessing and Packaging. We remove non-speech artifacts such as fillers and music markers, and package segments into ten-unit sliding windows to preserve local context. All instances are formatted into a fixed JSON schema for model interaction.

(3) Filtering and Sampling. We filter out overly short or extreme-CPS segments to ensure stable syllable-ratio estimation, and further score samples using heuristic signals such as perplexity, repetition, and script consistency. After LSH-based de-duplication (Jha et al., 2023), we perform domain-balanced quota sampling to obtain *1,000 golden instances*.

(4) Core Event Annotation. We extract golden core events from each source Mandarin segment by identifying content-bearing predicate–argument structures (nouns, proper nouns, numerals) via Stanza.[2] To evaluate semantic preservation under length compression, candidate translations are first back-translated into Mandarin. We then extract candidate event realizations using the same pipeline. Source and back-translated event representations are compared via contextual embedding similarity; unmatched events are flagged as semantic violations to compute *BT-CERR*.

Following this pipeline, the finalized SAND-GLASS benchmark contains 1,000 high-quality golden instances balanced across five distinct media domains. Detailed descriptive statistics are deferred to Appendix N.

2.2 Syllable as Duration Proxy

Our use of syllables as duration proxies is grounded in the Iso-Information Principle. Pellegrino et al.

[2] https://stanfordnlp.github.io/stanza/

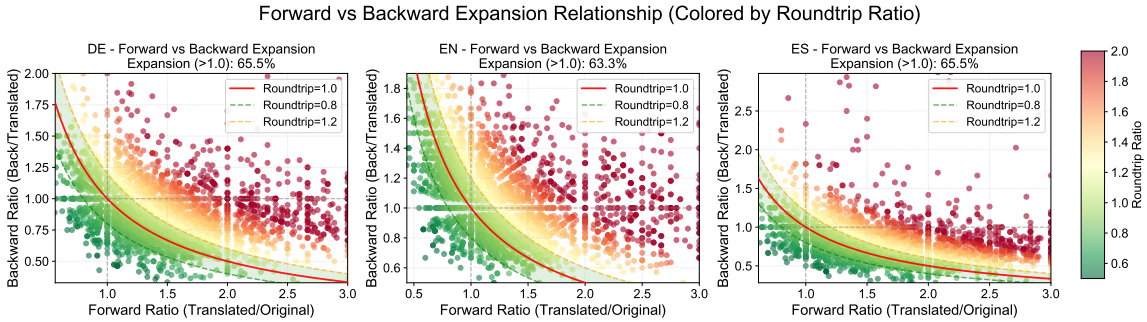


Figure 2: Forward vs. backward expansion relationship (colored by round-trip ratio ρ_{rtp}). The red curve represents the unity baseline $\rho_{\text{rtp}} = 1.0$ ($y = 1/x$). Points above this curve indicate systemic model-induced length inflation. Unconstrained models show a median $\rho_{\text{rtp}} > 1.10$.

Language	**Syllable Rate** (σ/s)	**Info. Density** (Normalized)	**Expansion** (from Zh)
Mandarin (Zh)	5.18 ± 0.15	0.94	$1.00\times$
English (En)	6.19 ± 0.16	0.91	$\approx 1.03\times$
German (De)	5.97 ± 0.19	0.79	$\approx 1.19\times$
Spanish (Es)	7.82 ± 0.16	0.63	$\approx 1.49\times$

Table 1: Cross-linguistic statistics adapted from Pellegrino et al. (2011). Lower density necessitates higher expansion factors when translating from Mandarin.

(2011) demonstrated that while Syllable Rate (SR) vary drastically across languages (e.g., Spanish $7.82\sigma/s$ [syllables/s] vs. Mandarin $5.18\sigma/s$), the universal Information Rate remains nearly constant at ≈ 39 bits/s (Coupé et al., 2019).Recent findings further corroborate this by quantifying a universal surprisal–duration trade-off (Pimentel et al., 2021), showing that languages systematically adjust segment duration to smooth information transmission.

As shown in Table 1, this trade-off implies that low-density languages must utilize higher syllable counts to convey equivalent content. Translating from a high-density source (Mandarin, ID=0.94) to a lower-density target (e.g., Spanish, ID=0.63) necessitates a *linguistically required expansion*. A direct 1:1 mapping would result in $\approx 33\%$ information loss (derived from the density ratio $\frac{0.63}{0.94}$). Thus, our constraints are non-uniform and calibrated to these density ratios to distinguish legitimate expansion from model verbosity.

2.3 Quantifying Cross-Lingual Verbosity Bias

Standard metrics like the Forward Expansion Ratio defined as $\rho_{fwd} = \sigma(y)/\sigma(x)$ for a source x and its translation y often confound *linguistic necessity* with *model hallucinations*. For instance, a high ρ_{fwd} in Spanish might simply reflect its naturally lower information density rather than model-induced redundancy.

To isolate systemic bias, we introduce the *Roundtrip Expansion Ratio* (ρ_{rtp}). By translating the target output y back into the source language ($\hat{x}$) and comparing its syllable count to the original input x, we define:

$$\rho_{rtp} = \rho_{fwd} \cdot \rho_{bwd} = \frac{\sigma(y)}{\sigma(x)} \cdot \frac{\sigma(\hat{x})}{\sigma(y)} = \frac{\sigma(\hat{x})}{\sigma(x)}, \quad (1)$$

where $\sigma(\cdot)$ denotes the syllable count of a sequence. Theoretically, since this comparison is performed within the same source language, ρ_{rtp} should gravitate towards 1.0. Any value where $\rho_{rtp} > 1.0$ serves as a density-invariant indicator of model-induced inflation.

Empirical analysis on Sand-Glass (Figure 2) confirms that this verbosity is systemic rather than anecdotal. By plotting the joint distribution of expansion ratios, we observe a pervasive upward shift from the unity baseline ($\rho_{rtp} = 1.0$), with the vast majority of segments clustered in the inflation zone. This concentration demonstrates a consistent model-induced redundancy that persists across various frontier models and language pairs (see Appendix J for statistical breakdowns).

Calibrated Target Bounds. To bridge this gap, we define Target Syllable Bounds ($\mathcal{B}_L$) by contrasting model inflation (μ_{LLM}) against Theoretical Expansion (μ_{theo}). As shown in Table 2, LLMs produce significant surplus bias (Δ_{bias}). We establish $\mathcal{B}_L$ to enforce a strict condensation regime. Notably, we lower these targets beyond theoretical baselines to address real-world dubbing and subtitling complexities, specifically to buffer against rapid speech and synchronization constraints. Accordingly, we configure the bounds as follows: English ($\mathcal{B}_{En} \in [0.8, 0.9]$) is set below parity to force active summarization; German ($\mathcal{B}_{De} \in [0.9, 1.0]$) maintains relative pressure despite lower density;

Target Lang	Theoretical (μ_{theo})	LLM Baseline (μ_{LLM})	Bias (Δ)	Ours (Target) ($\mathcal{B}_L$)
En	1.03×	1.35×	+0.32	$\mathbf{0.8 \sim 0.9}$
De	1.19×	1.59×	+0.40	$\mathbf{0.9 \sim 1.0}$
Es	1.49×	1.84×	+0.35	$\mathbf{1.0 \sim 1.1}$

Table 2: Comparison of expansion ratios. Target Bounds ($\mathcal{B}_L$) are set to remove model bias (Δ) and enforce semantic condensation.

and Spanish ($\mathcal{B}_{Es} \in [1.0, 1.1]$) significantly reduces the unconstrained baseline of 1.84× to ensure intelligibility under strict time limits while respecting its information density.

3 Method

We formulate syllable-level control as a standard controllable text generation task(Li et al., 2024; Gu et al., 2025), where our objective is to maximize translation quality while satisfying the prescribed syllable-ratio constraint. As illustrated in Figure 3, we optimize the model with GRPO (Shao et al., 2024), a lightweight PPO-style algorithm designed for sequence-level rewards.

3.1 Reward Design

In our setting, the control condition c corresponds to satisfying a strict length budget while maintaining semantic consistency with the source. This requires a reward that simultaneously encourages length compliance and discourages meaning drift. We instantiate this task reward as:

$$R(x, c, y) = \lambda_{\text{len}}\, R_{\text{len}}(x, y) + \lambda_{\text{qual}}\, R_{\text{qual}}(x, y), \tag{2}$$

where x is the source utterance, y is the generated translation, and $\lambda_{\text{len}}, \lambda_{\text{qual}} \geq 0$ control the trade-off between temporal alignment and semantic fidelity.

3.1.1 Dynamic Syllable-ratio Reward

To ensure the translation length aligns naturally with the source speech across varying scales, we propose a dynamic syllable-ratio reward. The syllable ratio ρ for a source-target pair (x, y) is defined as:

$$\rho(x, y) = \frac{\sigma(y)}{\max(\sigma(x), 1)}, \tag{3}$$

where $\sigma(\cdot)$ is the syllable count of a sequence.

Recognizing that strict length constraints must be sensitive to both linguistic density and utterance length, we implement a dynamic acceptance interval $[L(x), R(x)]$. We initialize the baseline bounds $[L_0, R_0]$ using the language-specific Target Bounds $\mathcal{B}_L$ set in Section 2.3. This ensures the optimization target respects the cross-lingual density hierarchy established in our theoretical analysis.

However, for shorter utterances, the "granularity" of syllable counting introduces high variance (discretization noise), making rigid bounds brittle. To address this, for utterances where the source length $\sigma(x)$ is below the corpus mean μ, we define a scaling factor $\gamma(x)$ to adaptively relax the lower threshold:

$$\gamma(x) = \alpha_1 + \alpha_2 \left(\frac{\sigma(x)}{\mu}\right)^{1/2}, \tag{4}$$

where α_1, α_2 are hyperparameters governing the degree of relaxation. The final dynamic bounds are formulated as $[L(x), R(x)] = [L_0 \cdot \gamma(x),\ R_0]$.

To facilitate stable policy optimization, we avoid sparse or binary signals in favor of a smoother reward landscape. The reward function $R_{\text{len}}(x, y)$ evaluates compliance with these dynamic boundaries via a squared exponential decay:

$$R_\rho = \begin{cases} 1 & \text{if } \rho \in [L(x), R(x)] \\ \exp(-k\delta^2) & \text{otherwise} \end{cases}, \tag{5}$$

where $\delta(x, y) = \max\big(|\rho(x, y) - L(x)|, |\rho(x, y) - R(x)|\big)$ denotes the deviation. This non-linear formulation provides a soft penalty that effectively guides the model toward the density-calibrated region while preventing training instability associated with hard-truncation rewards. In Appendix E.1, we further investigate the effectiveness of a static syllable-ratio reward.

3.1.2 Rubric-based Quality Reward

We adopt a two-part quality rubric(Hashemi et al., 2024; Liu et al., 2023), covering (i) semantic fidelity and (ii) linguistic well-formedness, and implement it as the following reward components.

Semantic Fidelity Reward Inspired by the probabilistic duality in (He et al., 2016), we enforce content preservation via a reconstruction-based reward. We employ a frozen `DeepSeek-V3` (DeepSeek-AI, 2024) to back-translate the hypothesis y into $\hat{x} = f_{bt}(y)$. To measure semantic retention, we compute the cosine similarity $S(x, \hat{x})$ between embeddings of the source x and reconstruction $\hat{x}$ extracted by `Qwen-Embedding-0.6B` (Zhang et al., 2025):

$$S(x, \hat{x}) = \frac{\mathbf{v}_x \cdot \mathbf{v}_{\hat{x}}}{\|\mathbf{v}_x\| \|\mathbf{v}_{\hat{x}}\|}. \tag{6}$$

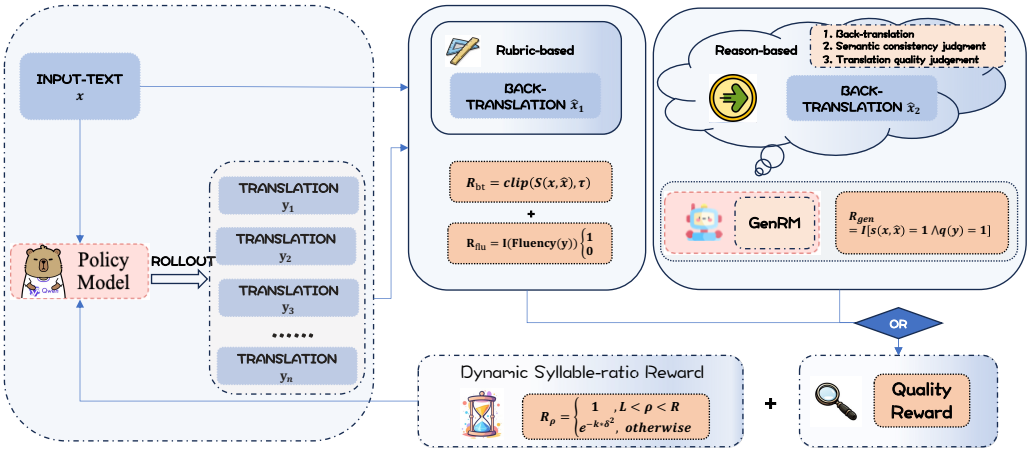


Figure 3: Overview of our HOMURA framework. Given a source utterance x and a syllable-ratio budget c, we optimize the policy $\pi_\theta(y \mid x, c)$ with GRPO. The reward combines a length-compliance term R_{len} and a quality term R_{qual} (rubric-based or reason-based), encouraging concise yet faithful translations.

The final reward is clipped to stabilize optimization:

$$R_{bt}(x, y) = \text{clip}(S(x, \hat{x}), \tau_{min}, \tau_{max}). \quad (7)$$

Linguistic Well-formedness Reward To prevent the telegraphic outputs often associated with aggressive compression, we explicitly penalize linguistic degradation. We utilize a frozen `DeepSeek-V3` as a judge to evaluate translations against a rubric covering: (i) *grammaticality*, (ii) *readability*, and (iii) *language consistency* (e.g., no code-switching). The model outputs a binary decision $s_{flu} \in \{0, 1\}$, serving directly as the reward:

$$R_{flu}(x, y) = s_{flu}. \quad (8)$$

Full prompting details are in Appendix H.

3.1.3 Reason-based Quality Reward

Beyond the simple compositional rewards, we explore an end-to-end generative reward model (GenRM)(Mahan et al., 2024; Liang et al., 2025) that evaluates translation quality via explicit reasoning. GenRM is trained using structured Chain-of-Thought (CoT) annotations synthesized by `Gemini-2.5-Pro` (Comanici et al., 2025). To ensure high-quality reasoning, we filter the raw CoT data using the aforementioned rubric-based method. The model is instantiated from `Qwen3-32B` (Yang et al., 2025) and optimized via supervised fine-tuning (SFT) over the full CoT sequence and the final decision token.

Given a source sentence x and a translation hypothesis y, GenRM performs a single long CoT reasoning process that can be abstracted as

$$\hat{x} = f_{\text{bt}}(y), \; s(x, \hat{x}), q(y) \in \{0, 1\}, \quad (9)$$

where f_{bt} denotes implicit back-translation, $s(x, \hat{x})$ indicates whether the core information in x is preserved, and $q(y)$ evaluates the intrinsic quality of y in terms of adequacy and fluency. Based on these decisions, GenRM outputs a binary quality reward

$$R_{gen}(x, y) = \mathbb{I}\big[s(x, \hat{x}) = 1 \wedge q(y) = 1\big]. \quad (10)$$

This formulation yields a reference-free, interpretable quality signal and enforces an explicit conjunction between information preservation and translation quality. Concrete examples of the CoT reasoning are provided in the Appendix G. Ablation studies on the reward components and design choices are provided in Appendix E.

4 Experiments and Results

4.1 Experimental Setup

We evaluate SOTA LLMs under five strategy categories that span the length-control spectrum:

- **Category 1: Unconstrained LLMs.** Frontier models (`GPT-5` (OpenAI, 2025), `Claude-4.1-Opus` (Anthropic, 2025), `Gemini-2.5-Pro`, `DeepSeek-V3`) with standard decoding. For `GPT-5` and `Gemini-2.5-Pro`, we set low reasoning effort to approximate direct translation. These serve as semantic upper bounds.
- **Category 2: Prompt-based Compression.** Category 2 models with brevity-oriented system prompts enforcing syllable-level conciseness.
- **Category 3: Best of N Refinement.** Generating N candidates with varying lengths,

Table 3: Performance Comparison on Sand-Glass Across All Language Pairs (BLEU-ρ, Cometkiwi, BT-CERR). All results are averaged by 3 times.

Model	Evaluation Metrics on Sand-Glass																	
	Zh → En						Zh → De						Zh → Es					
	IB	ρ	CK	BC	B-ρ	Tok.	IB	ρ	CK	BC	B-ρ	Tok.	IB	ρ	CK	BC	B-ρ	Tok.
Category 1: Unconstrained LLMs																		
`claude-4.1-opus`	22.4%	1.361	0.741	0.943	0.313	17.1	12.1%	1.582	0.676	0.882	0.247	24.2	8.2%	1.812	0.668	0.900	0.220	21.3
`deepseek-v3`	22.6%	1.413	0.748	0.941	0.287	46.3	12.4%	1.674	0.676	0.889	0.222	62.3	9.1%	1.990	0.671	0.884	0.191	99.6
`gemini-2.5-pro`*	24.9%	1.289	0.739	0.941	0.307	12.1 (916.9)	15.5%	1.511	0.666	0.877	0.238	14.2 (125.8)	9.9%	1.718	0.668	0.883	0.212	14.0 (1532.7)
`gpt-5`*	23.2%	1.330	0.746	0.947	0.300	22.4 (412.8)	11.0%	1.633	0.681	0.894	0.220	23.2 (605.6)	8.3%	1.862	0.671	0.899	0.206	24.0 (676.8)
Category 2: Prompt-based Compression																		
`claude-4.1-opus`	46.1%	0.852	0.691	0.907	0.376	12.2	31.5%	0.951	0.535	0.838	0.293	17.2	26.6%	1.106	0.631	0.849	0.261	13.5
`deepseek-v3`	38.4%	0.960	0.700	0.924	0.307	12.4	20.3%	1.196	0.557	0.866	0.245	15.8	16.1%	1.355	0.556	0.868	0.225	16.3
`gemini-2.5-pro`*	49.7%	0.920	0.698	0.927	0.322	10.7 (1231.3)	52.3%	1.016	0.517	0.853	0.266	12.3 (1450.1)	46.7%	0.974	0.586	0.827	0.252	10.7 (554.5)
`gpt-5`*	53.6%	0.937	0.705	0.928	0.345	21.7 (2024.9)	51.5%	1.025	0.534	0.874	0.283	23.1 (1962.3)	62.6%	0.982	0.586	0.833	0.270	21.0 (3105.8)
Category 3: Best-of-N Refinement																		
`claude-4.1-opus`	67.6%	0.845	0.691	0.903	0.367	109.7	50.9%	0.946	0.611	0.848	0.324	109.2	45.5%	1.039	0.621	0.857	0.304	110.0
`deepseek-v3`	63.6%	0.852	0.691	0.902	0.332	122.6	44.7%	0.959	0.527	0.863	0.286	160.5	44.4%	1.042	0.522	0.851	0.275	113.6
`gemini-2.5-pro`*	61.9%	0.900	0.673	0.904	0.276	133.2 (1089.2)	44.1%	0.992	0.517	0.854	0.241	136.0 (972.7)	42.3%	1.087	0.554	0.858	0.235	133.8 (611.1)
`gpt-5`*	67.2%	0.888	0.673	0.906	0.296	114.2 (1810.2)	47.4%	0.994	0.571	0.861	0.275	111.2 (2229.6)	51.5%	1.076	0.529	0.855	0.256	110.8 (2158.8)
Category 4: Pipeline-based Post-editing (2 run)																		
`claude-4.1-opus`	38.3%	0.918	0.694	0.904	0.327	30.2	24.1%	1.089	0.653	0.868	0.276	42.7	18.7%	1.284	0.643	0.876	0.246	37.6
`deepseek-v3`	38.4%	0.967	0.703	0.934	0.322	88.1	23.3%	1.154	0.645	0.874	0.239	89.4	17.8%	1.301	0.543	0.873	0.221	102.1
`gemini-2.5-pro`*	50.0%	0.837	0.691	0.902	0.348	22.3 (2517.7)	41.9%	1.001	0.627	0.830	0.267	27.5 (1043.1)	35.2%	1.042	0.602	0.840	0.250	22.4 (2975.2)
`gpt-5`*	78.6%	0.833	0.690	0.917	0.321	42.2 (2666.2)	75.5%	0.934	0.514	0.842	0.275	46.6 (2773.0)	66.3%	1.018	0.557	0.820	0.234	44.9 (4012.9)
Category 5: HOMURA																		
Ours (HOMURA$_{\text{Rubric}}$)	86.3%	0.809	0.700	0.914	0.378	9.2	84.3%	0.901	0.591	0.853	0.317	16.3	89.0%	1.067	0.620	0.863	0.260	13.1
Ours (HOMURA$_{\text{Reason}}$)	89.6%	0.846	0.701	0.925	0.376	10.1	84.6%	0.918	0.603	0.861	0.321	14.7	87.3%	1.057	0.605	0.858	0.309	13.2

Note: **IB**: In-Bounds Rate (%). ρ: syllable ratio. **CK**: CometKiwi. **BC**: BT-CERR. **B-ρ**: BLEU-ρ. **Tok.**: avg output tokens; for reasoning models (*), parenthesized values include internal thinking tokens.

followed by a greedy selection of the highest-fidelity output that fulfills the syllable budget.

- **Category 4: Pipeline-based Post-editing.** Two-stage translation → length-constrained rewriting.
- **Category 5: Our Method (HOMURA).** RL fine-tuning on `Qwen3-8B` (Yang et al., 2025) to enforce temporal budgets as hard constraints.

More experimental details are provided in Appendix D.

4.2 Target Syllable Bounds Configuration

Following Section 2.3, we set density-calibrated target bounds $\mathcal{B}_L$ to impose comparable compression pressure across languages: En $\in [0.8, 0.9]$, De $\in [0.9, 1.0]$, and Es $\in [1.0, 1.1]$. These intervals are used as hard constraints for HOMURA.

4.2.1 Evaluation Metrics

We report complementary metrics for efficiency, semantic preservation, and linguistic quality:

IB: In-Bounds Rate (% ↑) We define the **In-Bounds Rate** (**IB**) as the percentage of translations that satisfy the target interval bounds:

$$\text{IB} = \frac{1}{|D|} \sum_{(x,\mathcal{B}_L)\in D} \mathbb{I}\left(\frac{\sigma(y)}{\sigma(x)} \in \mathcal{B}_L\right), \quad (11)$$

where D is the evaluation dataset, and $\mathbb{I}(\cdot)$ is the indicator function.

Translation Density (BLEU-ρ ↑) We define **BLEU-ρ** as semantic fidelity per temporal unit:

$$\text{BLEU-}\rho = \frac{\text{BLEU}(x, \hat{x})}{\rho(x, y)}, \quad (12)$$

where $\text{BLEU}(x, \hat{x})$ is the back-translation BLEU (Papineni et al., 2002). Since BLEU alone may not fully capture fluency in back-translation paradigms (Edunov et al., 2020), we use BLEU-ρ specifically to measure **information density**, while linguistic quality assessed via Cometkiwi.

Semantic Integrity (BT-CERR ↑) BT-CERR is a binary indicator that evaluates whether all core events extracted from x are retained in the back-translation $\hat{x}$.

Linguistic Quality (Cometkiwi ↑) We use reference-free Cometkiwi(Rei et al., 2022) to assess fluency and adequacy under compression.

4.3 Main Results

4.3.1 Cross-Lingual Performance and Temporal Compliance

Table 3 summarizes the main results on SAND-GLASS, with additional zero-shot generalization

results on WMT24pp(Deutsch et al., 2025) and MuST-Cinema(Karakanta et al., 2020c) reported in Appendix L. We report human evaluation in Appendix C and Statistical Significance Testing in Appendix B. Overall, the results highlight the trade-off between temporal compliance and semantic density.

Budget Violation in Frontier Models. As shown by the ρ column, unconstrained frontier models(Cat 1) universally fail to meet temporal budgets, particularly in low-density languages like Spanish. While specialized prompting (Cat 2) shifts the distribution toward brevity, it remains unreliable, with models like `GPT-5` still exhibiting "verbosity drift" that exceeds target bounds. This confirms that pure instruction-following is insufficient for rigorous syllable-level constraints.

The Search-Fidelity Trade-off. Search-based (Cat 3) and pipeline (Cat 4) strategies achieve higher In-Bounds rates (IB) but at a significant semantic cost. For instance, Cat 4 variants often show a drop in BT-CERR, suggesting that the disjointed post-editing process leads to "hallucinated omissions" of critical source predicates. Moreover, the increased inference latency of Best-of-N makes it less practical in real-time settings.

Information Density and Superior Compliance of Homura. Concurrently, Homura achieves up to an 89.6% temporal compliance rate while achieving the highest BLEU-ρ across all language pairs, demonstrating active *semantic packing* that maximizes the semantic payload per syllable. We further validate the practical deliverability and scalability of our framework via two key analyses detailed in the Appendix: (1) an evaluation of Syllables Per Second (SPS) and speech-rate thresholds, which confirms that Homura achieves the lowest average SPS and highest pass rate for comfortable video articulation (Appendix M); and (2) a scaling analysis demonstrating consistent gains with a 32B model (Appendix F).

Inference efficiency. Homura is markedly more inference-efficient. Unlike Best-of-N, which multiplies decoding cost by generating N candidates, and unlike translate-then-modify pipelines that require an extra rewriting pass, Homura satisfies the syllable/temporal budget in a single decoding. Consequently, it attains high temporal compliance with substantially fewer tokens (AVG-Tokens in Table 3), making it suitable for real-time and high-throughput deployment.

Robustness of Reason-based GenRM. Within Category 5, $\text{Homura}_{\text{Reason}}$ achieves higher BT-CERR than $\text{Homura}_{\text{Rubric}}$ on two of the three target languages, with the largest gain observed for English. This suggests the reason-based reward better prunes redundancy while preserving the source's logical structure. We further validate this by replacing the self-trained GenRM in $\text{Homura}_{\text{Reason}}$ with an LLM-as-RM baseline (`DeepSeek-V3`); results are in Appendix E.2.

4.3.2 Analysis of Quality–Compression Trade-off

Figure 4 illustrates the trade-off between translation quality and output length by plotting semantic metrics against the average syllable ratio ρ.

As shown, unconstrained baselines cluster in the high-quality, high-ρ region. Although standard LLMs can be prompted to shorten outputs, their quality drops sharply under aggressive compression. This reflects a low rate–distortion efficiency: when forced to reduce the phonetic "rate" (syllables), these models suffer from severe semantic "distortion" due to suboptimal content selection.

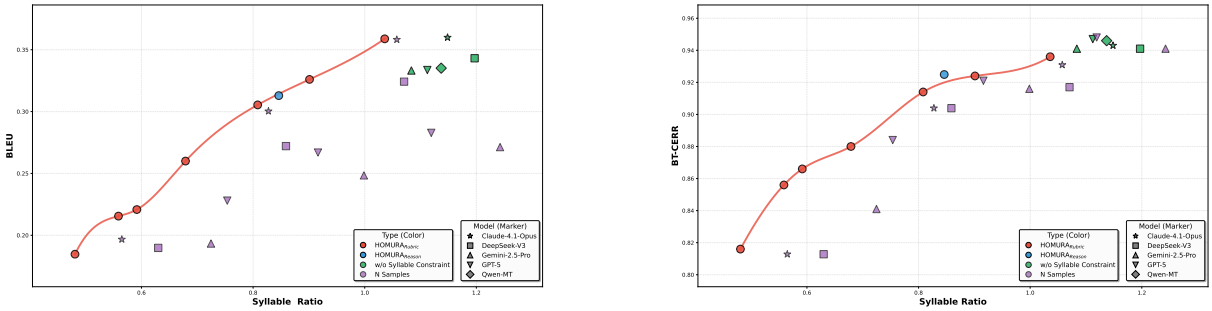
In contrast, Homura demonstrates superior rate–distortion efficiency, maintaining higher semantic fidelity at any given syllable budget. Rather than executing naive truncation, Homura actively optimizes the generation trajectory to operate at a more efficient trade-off point, packing a higher semantic payload into the limited phonetic constraints required by media localization.

4.3.3 Qualitative Case Studies

Table 10 qualitatively benchmarks Homura against `Claude-4.1-Opus` baselines. Unlike baseline strategies that suffer from over-translation, syntactic fragmentation, or a coarse *cliff pattern*, Homura reliably hits the target window in a single pass while preserving pragmatic integrity (full analyses are deferred to Appendix I).

4.4 Autonomous Compression: Navigating the Compression Limit

The preceding experiments show that Homura can reliably translate under a user-specified ratio interval. We further ask whether, without a preset lower bound, the model can autonomously discover the practical compression limit: the point where additional shortening triggers a sharp loss in meaning.



(a) BLEU vs. Avg. Syllable Ratio

(b) BT-CERR vs. Avg. Syllable Ratio

Figure 4: Quality–compression trade-off on Zh→En. We plot BLEU (left) and BT-CERR (right) versus the average syllable ratio ρ. Each point is a model setting (marker: backbone; color: variant: Unconstrained, Rubric, GenRM, Multi-length Prompting). The red curve shows the empirical trend: quality degrades as compression strengthens.

Table 4: The Compression Wall. The autonomous model reliably reaches the feasible limit; stricter manual targets miss ρ and produce worse compression.

Model & Configuration	ρ	BLEU-ρ	BT-CERR
Auto HOMURA	**0.485**	0.404	0.811
Manually Constrained HOMURA			
$R_{\text{target}} \in [0.3, 0.4]$	0.558	0.392	0.819
$R_{\text{target}} \in [0.4, 0.5]$	**0.481**	0.384	0.816
$R_{\text{target}} \in [0.5, 0.6]$	0.559	0.386	0.856
$R_{\text{target}} \in [0.8, 0.9]$	0.808	0.378	0.914

To this end, we formulate an autonomous compression task by replacing the hard-interval length reward with a continuous incentive:

$$R_{\text{len}}^{(\text{auto})}(x, y) = \min\left(\cos\left(\theta \cdot \rho(x, y)\right),\ 0\right), \quad (13)$$

where θ scales the incentive gradient. The policy is encouraged to minimize ρ while still being constrained by the quality reward R_{qual}.

The Empirical Compression Wall. We observe a stable *compression wall* when pushing the model to its structural limit. As detailed in Table 4, autonomous optimization consistently converges to a narrow operating regime around $\rho \in [0.46, 0.49]$, regardless of the reward configuration. Crucially, when we enforce an ultra-aggressive manual target of $[0.3, 0.4]$, the model does *not* become shorter; instead, it rebounds to $\rho = 0.558$ while semantic coherence collapses. This non-linear failure indicates a rigid *feasibility threshold* around $\rho \approx 0.49$ for Zh→En translation, beyond which the model cannot maintain basic syntactic well-formedness.

Rate-Distortion Efficiency and Feasibility Boundary. To quantify this behavior, Auto HOMURA attains the optimal BLEU-ρ (0.404) while operating at a genuinely shorter length. Qualitative analysis reveals that autonomous optimization navigates this quality–compression frontier via active *semantic condensation* (e.g., stripping non-essential modifiers and deploying dense lexical substitutions). Conversely, forcing the model into an unnatural manual interval triggers unstable truncation, reflecting the severe tension between mechanical length reduction and core meaning preservation. This empirical wall effectively represents the *minimal information carrier* required to express core predicate–argument content, closely aligning with the theoretical density floor predicted by the Iso-Information Principle (Section 2.3).

5 Conclusion

We address systemic cross-lingual verbosity bias in LLM translation, a key obstacle for time-constrained scenarios such as subtitling and dubbing. To study this problem, we introduce SANDGLASS, a syllable-budgeted benchmark grounded in the Iso-Information Principle, and propose HOMURA, a constrained RL framework that optimizes the trade-off between semantic fidelity and temporal feasibility.

HOMURA uses a dynamic syllable-ratio reward to encourage translations that remain within the prescribed budget while preserving meaning, without relying on reference translations. Experiments show that HOMURA substantially improves length compliance and information density, as measured by BLEU-ρ. Future work will examine whether the observed compression feasibility boundary, e.g., $\rho \approx 0.49$ for Zh→En, generalizes across language pairs, and extend HOMURA to end-to-end speech translation with joint semantic and duration optimization.

Limitations

First, while we adopt syllable count as a duration proxy grounded in the Iso-Information Principle, this metric remains a textual approximation of acoustic reality. In real-world dubbing and subtitling, physical duration is heavily influenced by prosodic features and speech tempo, and our current text-only framework does not account for multimodal constraints like lip-synchronization or isochrony. Consequently, generated outputs may still require minor timing adjustments during recording.

Second, our experimental validation is currently restricted to Chinese-to-English, German, and Spanish translation tasks. The universality of the "compression feasibility boundary" observed in our study remains to be verified across linguistically distant pairs or low-resource languages. Future work is needed to determine if such limits are language-specific or intrinsic to the translation task itself.

Acknowledgments

References

Anthropic. 2025. Claude 4.1 Opus System Card. Technical report, Anthropic.

Eleftheria Briakou, Zhongtao Liu, Colin Cherry, and Markus Freitag. 2024. On the implications of verbose llm outputs: A case study in translation evaluation. *arXiv preprint arXiv:2410.00863*.

Harveen Singh Chadha, Aswin Shanmugam Subramanian, Vikas Joshi, Shubham Bansal, Jian Xue, Rupeshkumar Mehta, and Jinyu Li. 2025. Length aware speech translation for video dubbing. *Preprint*, arXiv:2506.00740.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Christophe Coupé, Yoon Mi Oh, Dan Dediu, and François Pellegrino. 2019. Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances*, 5(9):eaaw2594.

L Davisson. 1972. Rate distortion theory: A mathematical basis for data compression. *IEEE Transactions on Communications*, 20(6):1202–1202.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Daniel Deutsch, Eleftheria Briakou, Isaac Rayburn Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, and 1 others. 2025. Wmt24++: Expanding the language coverage of wmt24 to 55 languages & dialects. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12257–12284.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2020. On the evaluation of machine translation systems trained with back-translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2836–2846, Online. Association for Computational Linguistics.

Marco Gaido, Sara Papi, Mauro Cettolo, Roldano Cattoni, Andrea Piergentili, Matteo Negri, and Luisa Bentivogli. 2024. Automatic subtitling and subtitle compression: Fbk at the iwslt 2024 subtitling track. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 86–96.

Yuxuan Gu, Wenjie Wang, Xiaocheng Feng, Weihong Zhong, Kun Zhu, Lei Huang, Ting Liu, Bing Qin, and Tat-Seng Chua. 2025. Length controlled generation for black-box llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16878–16895.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. Llm-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. *arXiv preprint arXiv:2501.00274*.

Di He, Yingce Xia, Tao Qin, Laiwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are GPT models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Dávid Javorský, Ondřej Bojar, and François Yvon. 2025. Prompting LLMs: Length control for isometric machine translation. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 119–137, Vienna, Austria (in-person and online). Association for Computational Linguistics.

Aditi Jha, Sam Havens, Jeremy Dohmann, Alex Trott, and Jacob Portes. 2023. Limit: Less is more for instruction tuning across evaluation paradigms. *arXiv preprint arXiv:2311.13133*.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Tollef Emil Jørgensen and Ole Jakob Mengshoel. 2025. Cross-lingual sentence compression for length-constrained subtitles in low-resource settings. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*, pages 6447–6458.

Alina Karakanta, Matteo Negri, and Marco Turchi. 2020a. Is 42 the answer to everything in subtitling-oriented speech translation? In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 209–219, Online. Association for Computational Linguistics.

Alina Karakanta, Matteo Negri, and Marco Turchi. 2020b. MuST-cinema: A speech-to-subtitles corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3727–3734.

Alina Karakanta, Matteo Negri, and Marco Turchi. 2020c. Must-cinema: a speech-to-subtitles corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3727–3734.

Wendi Li, Wei Wei, Kaihe Xu, Wenfeng Xie, Dangyang Chen, and Yu Cheng. 2024. Reinforcement learning with token-level feedback for controllable text generation. *arXiv preprint arXiv:2403.11558*.

Xiaobo Liang, Haoke Zhang, Juntao Li, Kehai Chen, Qiaoming Zhu, and Min Zhang. 2025. Generative reward modeling via synthetic criteria preference learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26755–26769.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Dakota Mahan, Duy Van Phung, Rafael Rafailov, Chase Blagden, Nathan Lile, Louis Castricato, Jan-Philipp Fränken, Chelsea Finn, and Alon Albalak. 2024. Generative reward models. *arXiv preprint arXiv:2410.12832*.

Shushen Manakhimova, Vivien Macketanz, Eleftherios Avramidis, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian Möller. 2024. Investigating the linguistic performance of large language models in machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 355–371.

OpenAI. 2025. GPT-5 System Card. Technical report, OpenAI.

Longshen Ou, Xichu Ma, Min-Yen Kan, and Ye Wang. 2023. Songs across borders: Singable and controllable neural lyric translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–467.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

François Pellegrino, Christophe Coupé, and Egidio Marsico. 2011. A cross-language perspective on speech information rate. *Language*, 87:539–558.

Tiago Pimentel, Clara Meister, Elizabeth Salesky, Simone Teufel, Damián Blasi, and Ryan Cotterell. 2021. A surprisal–duration trade-off across and within the world's languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 949–962, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José GC De Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, and 1 others. 2022. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645.

Le Ren, Xiangjian Zeng, Qingqiang Wu, and Ruoxuan Liang. 2025. Lyricar: A difficulty-aware curriculum reinforcement learning framework for controllable lyric translation. *arXiv preprint arXiv:2510.19967*.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. 2023. A long way to go: Investigating length correlations in rlhf. *arXiv preprint arXiv:2310.03716*.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. 2022. Defining and characterizing reward gaming. In *Advances in Neural Information Processing Systems*, volume 35, pages 9460–9471.

Yihan Wu, Junliang Guo, Xu Tan, Chen Zhang, Bohan Li, Ruihua Song, Lei He, Sheng Zhao, Arul Menezes,

and Jiang Bian. 2023. Videodubber: machine translation with speech-aware length control for video dubbing. AAAI'23/IAAI'23/EAAI'23. AAAI Press.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the association for computational linguistics: NAACL 2024*, pages 2765–2781.

A Related Work

A.1 Constrained Translation and Subtitling

Media localization imposes strict spatiotemporal constraints beyond semantic fidelity (Karakanta et al., 2020b). To address these limitations, early research focused on supervised learning paradigms that integrate specialized markers to guide line length and reading speed (Karakanta et al., 2020a). An advancement was introduced by Wu et al. (2023) with *VideoDubber*, which pioneered direct speech-length control by incorporating token-level duration predictions into the decoder stack. However, this approach remains constrained by its dependence on supervised training with non-isochronous datasets and heuristic duration labeling that introduces annotation noise. Building on this, Chadha et al. (2025) demonstrated that even with automatically generated length tags derived from phoneme ratios, supervised training on non-isochronous speech-translation data remains a fundamental bottleneck for achieving precise isochronicity. While architectures like CLSC (Jørgensen and Mengshoel, 2025) further refine this task using control tokens to manage sequence length, they remain fundamentally dependent on supervised training with task-specific compression corpora. In contrast, our work targets general-purpose Large Language Models (LLMs) via a reinforcement learning framework. This approach explicitly optimizes the trade-off between semantic preservation and temporal compliance without requiring ground-truth compression datasets. Recently, the focus has shifted toward leveraging general-purpose Large Language Models (LLMs). For instance, Javorský et al. (2025) investigated prompt-based strategies for machine translation under strict isometric constraints, highlighting the potential of in-context learning for length control. In contrast, our work moves beyond pure prompting paradigms by introducing a reinforcement learning framework. This approach explicitly optimizes the trade-off between semantic preservation and temporal compliance without requiring ground-truth compression datasets or fragile prompt engineering.

A.2 Hard Constraints and Lyric Translation

The imposition of strict feasibility constraints parallels automatic lyric translation, where models must adhere to syllable counts and rhythm (Ou et al., 2023). Building on curriculum RL strategies for discrete constraints (Ren et al., 2025), HOMURA treats temporal duration as a hard constraint. We employ a KL-regularized objective to balance these rigid length requirements with semantic quality, ensuring the model respects the "sand-glass" temporal budget.

A.3 Verbosity Bias in LLMs

A major challenge for time-constrained translation is the systemic "verbosity bias" in LLMs, often exacerbated by RLHF alignment that conflates length with helpfulness (Saito et al., 2023; Skalse et al., 2022; Singhal et al., 2023). Our framework counteracts this by introducing a dynamic syllable-ratio reward. Instead of encouraging length drift, we specifically penalize verbosity, aligning the model's output with the strict density requirements of subtitling and dubbing.

B Statistical Significance Testing

We perform hypothesis testing to examine whether the improvements brought by HOMURA are statistically significant. For a fair comparison, for each baseline we choose its best-performing single-pass configuration and compare it with HOMURA under the same setting. Table 5 reports the resulting p-values.

Across both language pairs and evaluation metrics, HOMURA yields statistically significant improvements over its strongest single-pass baselines, with p-values consistently below the conventional 0.01 threshold (often far smaller, e.g., $< 1e{-}10$ for

Table 5: p-values for testing HOMURA against each variant.

Model	p-value of HOMURA			
	Zh → En		Zh → De	
	B-ρ	CK	B-ρ	CK
Homura$_{\text{Rubric}}$	6.46e-03	8.20e-03	$<1e-10$	$<1e-10$
Homura$_{\text{Reason}}$	-	6.30e-03	$<1e-10$	$<1e-10$

Zh→De). These results confirm that the observed gains are unlikely to be due to random variation. The "–" entry denotes a degenerate case where the two compared results are identical, yielding no measurable difference and thus no meaningful significance test.

C Human Evaluation

We additionally report human evaluation results for the Zh-to-En setting to further validate the effectiveness of our method and the appropriateness of our evaluation metrics.

Table 6: Human evaluation results on Accuracy and Completeness (Zh → En).

Setting	Model	IB	ρ	Acc.	Comp.	Avg.
Unconstrained LLMs	Gemini-2.5-Pro	24.9%	1.289	4.95	5.00	4.97
	Claude-4.1-Opus	22.4%	1.361	4.91	4.98	4.94
	GPT-5	23.2%	1.330	4.96	5.00	4.98
	DeepSeek-V3	22.6%	1.413	4.97	5.00	4.99
Prompt-based Compression	Gemini-2.5-Pro	49.7%	0.920	4.44	4.90	4.67
	Claude-4.1-Opus	46.1%	0.852	4.42	4.93	4.67
	GPT-5	53.6%	0.937	4.42	4.87	4.65
	DeepSeek-V3	38.4%	0.960	4.42	4.87	4.65
Best-of-N Refinement	Gemini-2.5-Pro	61.9%	0.900	4.57	4.57	4.57
	Claude-4.1-Opus	67.6%	0.845	4.51	4.49	4.50
	GPT-5	67.2%	0.888	4.71	4.80	4.75
	DeepSeek-V3	63.6%	0.852	4.56	4.58	4.57
Ours	HOMURA$_{\text{Rubric}}$	86.3%	0.809	4.91	4.83	4.87
	HOMURA$_{\text{Reason}}$	89.6%	0.846	4.85	4.93	4.89

Note: **IB**: In-Bounds Rate (%). ρ: syllable ratio. **Acc.**: Accuracy. **Comp.**: Completeness.

Table 6 shows that (i) without syllable constraints, all frontier LLMs achieve near-ceiling Accuracy/Completeness, serving as an unconstrained upper bound; (ii) enforcing syllable budgets by prompting substantially degrades human-rated quality, especially Accuracy; and (iii) Best-of-N provides limited and unstable gains. In contrast, our HOMURA variants deliver the best overall constrained performance, approaching the unconstrained ceiling, with HOMURA$_{\text{Rubric}}$ slightly favoring Accuracy and HOMURA$_{\text{Reason}}$ favoring Completeness.

Table 7: Key hyperparameters used in training HOMURA.

Hyperparameter	Value
α_1	0.4
α_2	0.5
λ_{len}	0.5
λ_{qual}	0.5
k	300
τ_{min}	0
τ_{max}	0.8

D Implementation Details

Our RM training framework is built on Megatron. We use the Qwen3-32B-Chat model as the initialization. The training is conducted with a batch size of 256, using a cosine learning rate scheduler with an initial learning rate of 5e-6. All models are trained on 64 Huawei's Ascend 910B NPUs.

Our RL training framework is based on the Verl framework. We use the Qwen3-8B-Chat model as the initialization for RL training. During training, we configure a batch size of 16 and perform 16 rollouts per prompt using the GRPO algorithm. The learning rate is initialized at 1e-8, and a cosine scheduler with warm-up is applied toward the final iteration. Sampling is conducted with a temperature of 1.0, and the maximum generation length is limited to 1,024 tokens. The KL penalty coefficient β is set to 0, effectively removing the KL constraint relative to the reference policy. The PPO clipping range ϵ is fixed at 0.2. All models are trained for one epoch using 8 NVIDIA H800 80G GPUs. Other detail training hyperparameters are listed in Table 7.

Our training set consists of 70K instances drawn from the same data sources as the Sand-Glass benchmark, and is constructed following the same procedure.

E Ablation Study on Reward Design

This appendix provides additional ablations on the reward design of HOMURA. We focus on two key components: (i) the dynamic syllable-ratio reward, which provides the temporal compliance signal, and (ii) the GenRM-based quality reward, which guides semantic preservation under compression. These analyses examine whether the proposed reward design improves optimization stability and whether a compression-oriented reward model provides better guidance than a generic external evaluator.

E.1 Effectiveness of Dynamic Syllable-ratio Reward

In Section 3.1.1, we hypothesized that rigid syllable-ratio bounds are susceptible to discretization noise, particularly in shorter utterances. To validate the necessity of our dynamic relaxation mechanism, we conduct an ablation study comparing our *Dynamic Syllable-ratio Reward* against a static baseline on the Zh→En translation task. Both methods share the same target objective $\mathcal{B}_L \in [0.8, 0.9]$, but the static version enforces fixed boundaries regardless of source length.

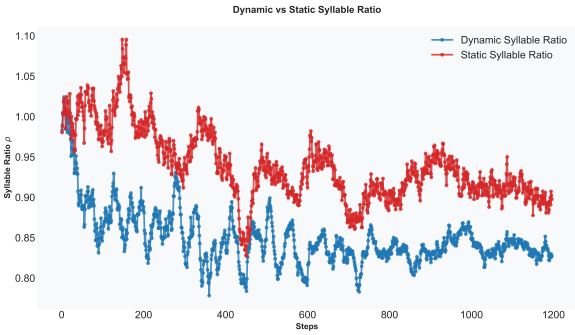


Figure 5: Training dynamics of syllable ratios during policy optimization on Zh→En. The blue curve represents our dynamic bound relaxation, while the red curve represents static fixed bounds. Values indicate the mean syllable ratio ρ across training steps.

As shown in Figure 5, the dynamic reward demonstrates significantly smoother optimization and faster convergence into the target interval $[0.8, 0.9]$. In contrast, the static reward exhibits violent oscillations, especially in early training phases. This instability confirms that hard rewards based on syllable counts are overly sensitive to small variations in short segments: a single-syllable difference in a short utterance can cause a disproportionate jump in the ratio, producing inconsistent learning signals for the RL agent.

By adaptively relaxing the lower bound for shorter sequences via $\gamma(x)$, our method provides a more continuous and achievable reward landscape. This prevents the model from being penalized for unavoidable linguistic granularity and leads to more efficient learning of the desired compression behavior. These findings justify the use of length-aware dynamic bounds in HOMURA to ensure stable policy refinement across diverse utterance lengths.

E.2 Ablation of GenRM Design

We further study the design of the quality reward within the GenRM framework. In particular, we compare our self-trained GenRM with an external LLM-as-RM baseline. In this ablation, the external reward model fully replaces GenRM, while all other components of the compression-oriented reinforcement learning pipeline are kept identical.

External LLM-as-RM Setup. For the external reward model baseline, we adopt an LLM-as-RM setup using `DeepSeek-V3`. Given a source sentence and its translation, the model is prompted to perform a quality assessment. Specifically, it evaluates whether the translation (i) preserves the core semantic content of the source sentence under the target syllable constraint, and (ii) remains fluent and grammatical. The model outputs a binary acceptability judgment, which is directly used as the reward signal during policy optimization. The prompt for the external LLM-as-RM baseline is shown below.

Prompt for External LLM-as-RM

You are a translation quality reward model. Given contextual information, a source sentence, and its translation, you will perform step-by-step reasoning to assess translation quality. Your evaluation process must strictly follow the sequence below.

Step 1: Back-Translation
Translate the given translation back into the source language.

Step 2: Semantic Consistency Assessment
Compare the semantic consistency between the original source text and the back-translated text.

- If the meanings are consistent or highly consistent, proceed to Step 3.
- If they are inconsistent, skip Step 3 and assign a final score of 0.

Step 3: Translation Quality Assessment
Evaluate whether the translation is of high quality based on factors such as fluency, grammatical correctness, and cultural appropriateness.

Your final output must include a clear reasoning process and a reward score:

- If semantic consistency is satisfied and the translation quality is high, set `score = 1`.
- Otherwise, set `score = 0`.

Output Format
The output must strictly follow the format below. Do not add or omit any fields:

```
{
"reasoning":      "<reasoning
process>",
"score": 0 or 1
}
```

Input Format
Context: `{context}`
Source text: `{current_text}`
Translation: `{translated_text}`

Results and Analysis. Table 8 compares compression translation models optimized with different reward model designs. The model trained with the self-trained GenRM consistently outperforms its counterpart using an external LLM-as-RM baseline, demonstrating that a reward model trained specifically for compression-oriented translation provides more effective optimization guidance than a generic prompt-based evaluator.

Table 8: Effect of GenRM choice on compressed translation under HOMURA$_{\text{Reason}}$. Results are averaged across all language pairs.

Model Variant	CometKiwi	BT-CERR	BLEU-ρ
HOMURA$_{\text{Reason}}$	0.636	0.881	0.335
HOMURA$_{\text{Reason}}$ w/ External RM	0.628	0.879	0.330

The self-trained GenRM achieves better CometKiwi, BT-CERR, and BLEU-ρ scores than the external LLM-as-RM variant. This suggests that task-specific reward modeling is beneficial for compression translation, where the evaluator must judge not only general translation quality but also whether the model preserves essential meaning under strict syllable budgets. Moreover, GenRM is only used during training and does not introduce additional inference-time cost, preserving the single-pass decoding advantage of HOMURA.

Table 9: Scaling Analysis of HOMURA (Zh→En). Comparison between 8B and 32B backbones at $\mathcal{B}_L \in [0.8, 0.9]$.

Scale	Model	IB	ρ	CK	BC	B-ρ
8B	HOMURA$_{\text{Rubric}}$	86.3%	0.809	0.700	0.914	0.378
	HOMURA$_{\text{Reason}}$	89.6%	0.846	0.701	0.925	0.376
32B	HOMURA$_{\text{Rubric}}$	90.2%	0.803	0.740	0.917	0.388
	HOMURA$_{\text{Reason}}$	89.9%	0.801	0.740	0.933	0.385

Note: **IB**: In-Bounds Rate (%). ρ: syllable ratio. **CK**: CometKiwi. **BC**: BT-CERR. **B-ρ**: BLEU-ρ.

F Scaling to Larger Model.

We analyze the effect of model scaling by comparing 8B and 32B variants of HOMURA under identical training settings, with results reported in Table 9. Scaling consistently improves translation quality and semantic preservation, as shown by higher BLEU-ρ, Cometkiwi and BT-CERR scores. Moreover, the relative advantage of HOMURA$_{\text{Reason}}$ over HOMURA$_{\text{Rubric}}$ is preserved at larger scale, suggesting that the proposed reward design scales robustly with model capacity.

F.1 Ablation Study on Reward Design

We further analyze the two key components of HOMURA's reward design: the dynamic syllable-ratio reward for temporal compliance and the GenRM-based quality reward for semantic preservation. These ablations aim to answer two questions: (i) whether dynamic length constraints are more effective than static bounds under syllable-level granularity, and (ii) whether the design of the quality reward affects the compression–fidelity trade-off.

Effectiveness of Dynamic Syllable-ratio Reward. We first examine the importance of using dynamic syllable-ratio bounds. A static interval applies the same lower and upper thresholds to all utterances, regardless of source length. However, syllable counts are discrete, and short utterances are particularly sensitive to small counting errors: adding or removing only one syllable can cause a large change in the ratio $\rho(x, y)$. As a result, static bounds may over-penalize otherwise acceptable translations for short inputs and provide unstable learning signals.

To address this issue, HOMURA relaxes the lower bound for shorter utterances while keeping the upper bound fixed, producing an adaptive interval $[L(x), R(x)]$. Compared with the static variant, the dynamic reward yields more stable length con-

trol and better preserves translation quality. This suggests that the gain of HOMURA does not come from merely imposing a rigid length penalty, but from providing a smoother and length-aware optimization signal that accounts for syllable-level discreteness.

Ablation of GenRM Design. We then study the design of the quality reward. In addition to the rubric-based reward, HOMURA explores a reason-based generative reward model, GenRM, which jointly assesses whether the compressed translation preserves the source meaning and remains linguistically well-formed. To isolate the contribution of GenRM, we compare it with alternative quality reward designs, including the rubric-based reward and an LLM-as-RM variant.

The results show that GenRM improves semantic preservation under strict length constraints, especially when compression requires selecting which information should be retained. Unlike simple scalar rubric scores, GenRM explicitly reasons about back-translation consistency and translation quality before producing the final reward decision. This makes the reward less likely to favor overly short but semantically incomplete outputs. Importantly, GenRM is only used during training and does not introduce additional inference-time cost. Therefore, HOMURA retains the single-pass decoding advantage over Best-of-N and translate-then-rewrite pipelines while benefiting from a stronger quality signal during RL training.

Discussion. Taken together, these ablations confirm that both sides of the reward are necessary. The dynamic syllable-ratio reward improves temporal compliance by stabilizing the length-control signal, while GenRM improves semantic robustness by discouraging meaning drift under compression. Their combination enables HOMURA to operate at a better compression–fidelity trade-off: translations remain within the target syllable budget while preserving the core predicate–argument information of the source.

G Examples of Structured CoT Reasoning for GenRM

This appendix provides illustrative examples of the structured CoT reasoning produced by GenRM.

Structured CoT Example #1

Source sentence (x):
我曾经在这看到过几艘陌生的船。

Translation hypothesis (y):
I have seen a few strange ships here before.

COT process:

1. **Back-translation:** The translation "I have seen a few strange ships here before." is back-translated into Chinese. The back-translation result is "我以前曾在这里见过几艘奇怪的船。" (I have previously seen several strange ships here). Alternatively, it can also be "我曾经在这里看到过几艘陌生的船。" (I once saw several unfamiliar ships here).

2. **Semantic consistency judgment:** Compare the source sentence "我曾经在这看到过几艘陌生的船。" (I once saw several unfamiliar ships here) with the back-translated content "我以前曾在这里见过几艘奇怪的船。" (I have previously seen several strange ships here). The two are completely consistent in core semantics, both expressing the fact of "过去在某地看到过不熟悉的船只" (having seen unfamiliar ships at a certain place in the past). "曾经" (once / before) corresponds to "before," "在这" (here) corresponds to "here," "看到过" (have seen) corresponds to "have seen," "几艘" (a few) corresponds to "a few," "陌生的" (unfamiliar / strange) corresponds to "strange," and "船" (ships) corresponds to "ships." The semantic consistency is very high.

3. **Translation quality judgment:** The translation "I have seen a few strange ships here before." is grammatically correct, using the present perfect tense to express a past experience, which conforms to English usage. The expression is fluent and natural, with accurate word choice. Therefore, the translation quality is high.

Based on the above, the semantics are consistent and the translation quality is high, resulting in a final score of 1.

Score:
1

Structured CoT Example #2

Source sentence (x):
所以天命人只有超度了金池长老的冤魂

Translation hypothesis (y):
So the Mandate Person can only rescue the lingering soul of Jinchi Elder

COT process:

1. **Back-translation:** The translation "So the Mandate Person can only rescue the lingering soul of Jinchi Elder" is back-translated into Chinese as "所以天命人只能解救金池长老逗留的灵魂。" (So the Mandate Person can only rescue the lingering soul of Elder Jinchi).

2. **Semantic consistency judgment:** The key terms in the source text are "超度" (perform salvation rites / help a soul attain deliverance) and "冤魂" (wronged or vengeful spirit). "超度" is a term with strong religious and cultural connotations, referring to helping the deceased soul attain liberation and enter reincarnation or the afterlife through rituals. However, the word "rescue" in the translation is overly general and fails to convey the deeper religious meaning of "超度". "冤魂" refers to a ghost that died unjustly and carries strong resentment, whereas "lingering soul" (逗留的灵魂) only expresses the idea of lingering or wandering, losing the core semantic component of "冤" (wronged / resentful). Therefore, the back-translated content is semantically inconsistent with the original text.

3. **Translation quality judgment:** Due to the semantic inconsistency, this step is skipped.

Based on the above analysis, the semantics are inconsistent, and the final score is 0.

Score:
0

H Prompt for Linguistic well-formedness Reward.

This appendix provides the training prompt used for the linguistic well-formedness reward.

Training Prompt
You are an expert in {target_lang} translation. Given context `<context>` for background reference. Here is a text `<text>` from the context and its corresponding {target_lang} translation `<translation>`. You need to determine whether `<translation>` is a qualified {target_lang} translation.

Requirements: Consider only the following criteria; if any are not met, the score is 0:

1. **Grammaticality:** No obvious grammatical or syntactic errors, consistency errors (e.g., subject–verb agreement, tense, case, gender/number), collocation errors, or linguistic defects caused by punctuation or spelling.

2. **Readability & Coherence:** The sentence structure is complete, logically coherent, and naturally connected, making it smooth and easy to understand; there should be no obvious fragmentation, garbled text, repetitive stacking, or malformed sentences.

Input:
`<context>`: {context}
`<text>`: {text}
`<translation>`: {translation}

Output Requirements:

- If `<translation>` satisfies both criteria above, return «`1`».
- Otherwise, return «`0`».
- Please output «`0`» or «`1`» directly; do not output any explanation.

Output:

I Case Study

Table 10 presents two representative Zh→En cases to qualitatively compare HOMURA against three baseline strategies built on top of a frontier LLM (Claude-4.1-Opus). For each case, we report: (1) **Direct**—the unconstrained translation produced by the LLM, which serves as an upper bound on quality but ignores the syllable budget; (2) **Truncated**—the same raw translation hard-cut at the tar-

get syllable count, simulating naïve post-hoc length control; and (3) **Best-of-N Refinement** (k=4)—four translations generated at progressively increasing compression levels via prompting, from which the candidate closest to the target interval $[0.8, 0.9]$ is selected. We present all four candidates (t_1–t_4) to visualize the coverage of the ratio space. Finally, we show the output of HOMURA under the same target ratio.

Observation 1: The Discrete Coverage Gap of Best-of-N Refinement. Across both cases, the four Best-of-N candidates exhibit a characteristic *cliff pattern*: adjacent candidates jump across the target interval $[0.8, 0.9]$ without landing inside it. In Case 1, the ratio drops from t_2=1.45 to t_3=0.93, and further to t_4=0.55; in Case 2, it falls from t_2=1.16 directly to t_3=0.78. This confirms that prompting a frontier LLM for a fixed number of compression variants cannot reliably cover the continuous ratio space—the model's coarse-grained "short vs. long" heuristic produces large gaps precisely in the moderate-compression regime where dubbing applications operate. HOMURA, by contrast, achieves fine-grained syllable awareness during decoding and hits ρ=0.83 and 0.84 in a single pass.

Observation 2: Truncation Destroys Sentence Integrity. When the Direct baseline overtranslates (Case 1: ρ=1.24; Case 2: ρ=1.22), hard truncation to the syllable budget produces syntactically broken output—"*boldly experience this...*" loses its object, and "*continued the...*" leaves a dangling determiner. These fragments may achieve the target ratio numerically but are unusable for dubbing, where every utterance must be a self-contained, prosodically complete phrase. HOMURA avoids this entirely: its outputs are natively generated to the target length, preserving both grammatical completeness and pragmatic tone (e.g., the encouraging "*No worries! Just dive right in*" in Case 1 and the poetic "*the ice tale Chinese write, for a thousand years*" in Case 2).

Observation 3: Register-Aware Compression Strategies. Beyond ratio accuracy, HOMURA adapts its compression strategy to the register of the source. For the colloquial, exhortative Case 1, it employs informal contractions and exclamatives ("*Haven't read...? No worries!*") that faithfully reproduce the casual, inviting tone of the Chinese source. For the literary, parallel-structured Case 2, it opts for noun-phrase compaction ("*hunting on wood across snow peaks*") and appositive restructuring ("*the ice tale Chinese write*"), achieving compression through poetic density rather than information loss. This register sensitivity suggests that the phoneme-indexed reward signal trains the model to internalize not just *how many* syllables to produce, but *which linguistic register* best accommodates the budget.

J Detailed Breakdown of Expansion Metrics

To provide a comprehensive view of the verbosity bias identified in Section 2, we present a fine-grained statistical breakdown of expansion metrics across various frontier LLMs and language pairs. As summarized in Table 11, the *Roundtrip Expansion Ratio* (R_{rtp}) consistently exceeds the unity baseline across all tested models and languages, with a high percentage of segments ($> 60\%$) exhibiting systemic inflation.

The distribution shifts visualized in Figure 6 further confirm that this bias is not merely a result of specific difficult cases but represents a structural characteristic of current LLM-based translation.

K Declaration of AI Assistance

To enhance development efficiency, GitHub Copilot was utilized as a coding assistant in this study. Specifically, it facilitated the implementation of data preprocessing modules, the execution of statistical analyses, and the rendering of complex experimental figures. The core algorithmic design, reinforcement learning frameworks, and the interpretation of results remain the original work of the authors, who maintain full responsibility for the code's accuracy.

L Generalization to General Translation

Generalization to Out-of-Distribution Benchmarks. To examine whether our method generalizes beyond the SAND-GLASS distribution, we conduct zero-shot evaluations on two external benchmarks. First, we evaluate on WMT24pp using models trained directly on the SAND-GLASS training set. We compare HOMURA with strong baseline models under four settings: unconstrained generation, prompting-based constrained generation, Best-of-N selection, and a translate-then-rewrite pipeline.

Table 10: Qualitative case studies on Zh → En translation with syllable-ratio control (target $\rho \in [0.8, 0.9]$). *Direct* translates without length awareness; *Truncated* hard-cuts the raw output to the syllable budget; *Multi* generates four candidates at varying compression levels and selects the closest to the target interval.

Category	Method	Output	Syl.	ρ	Observation
Case 1. 没看过原著的朋友更不用担心，大胆去体验这个瑰奇的打工人故事吧 *(29 syl.)*					
Direct	Raw	Friends who haven't read the original work, don't worry even more. Go ahead and boldly experience this wonderful story of a working man.	36	1.24	Fluent; ignores syllable budget.
	Trunc.	Friends who haven't read the original work, don't worry even more. Go ahead and boldly experience this . . .	26	0.90	In range but object missing; broken.
Multi (k=4)	t_1	For those of you who haven't read the original novel, there's even less to worry about; feel free to boldly experience this wonderfully peculiar story of a working person.	47	1.62	Heavily over-expanded.
	t_2	Friends who haven't read the original novel have even less to worry about; boldly go and experience this fantastic story of a working-class individual.	42	1.45	Still far above target.
	t_3 (sel.)	If you haven't read the book, don't worry; go experience this amazing story of a working person.	27	0.93	Closest candidate; overshoots.
	t_4	Unread the book? No worries, experience this worker's tale.	16	0.55	Over-compressed; target skipped.
HOMURA	ρ≈0.83	Haven't read the book? No worries! Just dive right in and enjoy this fantastic working-man's tale!	24	**0.83**	**On target**; encouraging tone preserved.
Case 2. 从雪山踏木狩猎而行，再到世界舞台上的比拼，冰雪的故事，中国人续写了千年 *(32 syl.)*					
Direct	Raw	From hunting in the snowy mountains with wooden skis to competing on the world stage, the Chinese have continued the story of ice and snow for thousands of years.	39	1.22	Over-translation; verbose.
	Trunc.	From hunting in the snowy mountains with wooden skis to competing on the world stage, the Chinese have continued the . . .	28	0.88	Broken at "continued the".
Multi (k=4)	t_1	From hunting on snow-covered mountains with wooden skis to competing on the global stage, the story of ice and snow has been continued by Chinese people for thousands of years.	44	1.38	Heavily over-expanded.
	t_2	From mountain hunting with wooden skis to global competitions, Chinese people have continued the story of ice and snow for a millennium.	37	1.16	Still above target.
	t_3 (sel.)	From hunting on snowy mountains to world competitions, China's ice and snow story spans millennia.	25	0.78	Below target; interval skipped.
	t_4	China's ice and snow story, from ancient hunts to modern competition, spans millennia.	23	0.72	Further compressed; target missed.
HOMURA	ρ≈0.84	From hunting on wood across snow peaks, to competing on the world stage—the ice tale Chinese write, for a thousand years.	27	**0.84**	**On target**; poetic compression.

Second, we evaluate on MUST-CINEMA, a subtitle-oriented speech translation benchmark. We randomly sample 1,000 examples from the English-to-German subset. To match our high-density-to-low-density setting, we first translate the English source texts into Chinese using DeepSeek-V3.2, and then evaluate models on the resulting Zh→De task. We compare HOMURA with `claude-4.1-opus` prompted with explicit syllable-ratio constraints. Importantly, HOMURA is evaluated zero-shot on this dataset, without any additional training or fine-tuning.

As shown in Table 12, HOMURA achieves stable zero-shot performance on WMT24pp, suggesting that the learned syllable-level control policy can transfer beyond the SAND-GLASS distribution. Although WMT24pp differs substantially from SAND-GLASS and contains relatively simpler translation instances, HOMURA maintains competitive translation quality while preserving effective verbosity control. In particular, HOMURA obtains a BT-CERR score of 0.990, a BLEU-ρ score of 0.605, and a CometKiwi score of 0.803.

Compared with prompting-based constrained

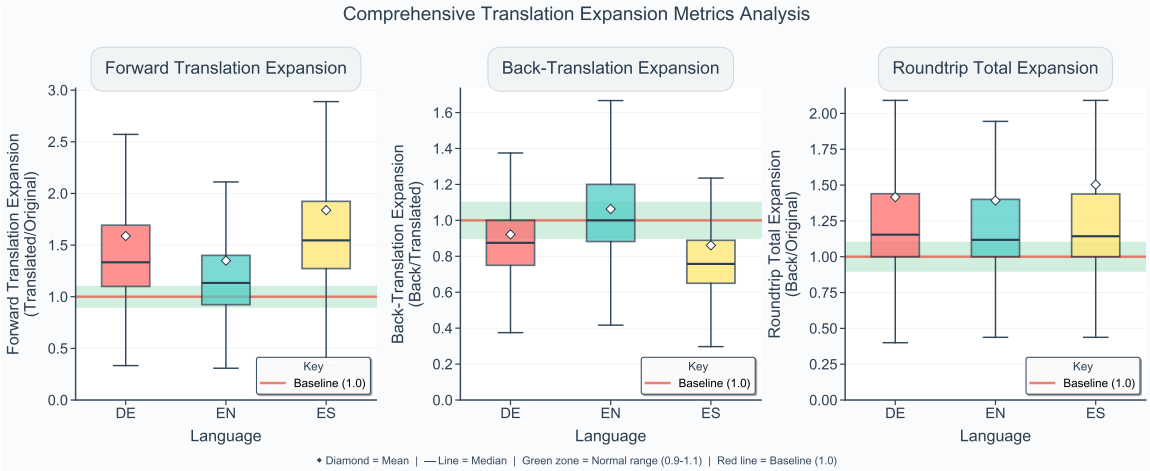


Figure 6: Statistical distribution of translation expansion metrics across DE, EN, and ES. The boxplots illustrate the variations in Forward, Back, and Roundtrip expansion ratios, with diamonds indicating means.

Table 11: Detailed Breakdown of Expansion Metrics across Models and Language Pairs. R_{fwd}, R_{bwd}, and R_{rtp} denote Forward, Backward (Back-translation), and Roundtrip Expansion Ratios, respectively. Values are presented as Mean and Standard Deviation.

Language	Model	R_{fwd} (Mean $\pm$ Std)	R_{bwd} (Mean $\pm$ Std)	R_{rtp} (Mean $\pm$ Std)	$R_{rtp} > 1$ (%)
	`claude-4.1-opus`	1.58 ± 0.96	0.95 ± 1.54	1.39 ± 0.99	66.2
	`deepseek-v3`	1.67 ± 1.37	0.91 ± 0.29	1.54 ± 1.89	63.8
De	`gemini-2.5-pro`	1.51 ± 0.88	0.96 ± 0.44	1.35 ± 0.84	59.0
	`gpt-5`	1.63 ± 0.96	0.90 ± 0.19	1.43 ± 0.87	71.8
	`claude-4.1-opus`	1.36 ± 1.13	1.07 ± 0.55	1.41 ± 1.32	64.8
	`deepseek-v3`	1.41 ± 1.15	1.05 ± 0.48	1.47 ± 1.71	59.1
En	`gemini-2.5-pro`	1.29 ± 0.81	1.07 ± 0.25	1.33 ± 0.80	61.0
	`gpt-5`	1.33 ± 0.81	1.07 ± 0.23	1.36 ± 0.80	65.9
	`claude-4.1-opus`	1.81 ± 1.05	0.90 ± 2.48	1.41 ± 1.19	66.9
	`deepseek-v3`	1.99 ± 1.83	1.01 ± 1.12	1.99 ± 2.90	64.0
Es	`gemini-2.5-pro`	1.72 ± 0.99	0.83 ± 0.36	1.35 ± 0.87	60.4
	`gpt-5`	1.86 ± 1.09	0.78 ± 0.18	1.40 ± 0.83	68.5

generation and Best-of-N selection, HOMURA provides a more direct way to internalize syllable-ratio constraints into the model policy, avoiding repeated sampling or external selection at inference time. Pipeline-based post-editing methods can also achieve strong performance on WMT24pp, but they rely on an additional rewriting stage and may introduce extra latency and error propagation. By contrast, HOMURA performs controlled translation in a single decoding pass.

As shown in Table 13, HOMURA also transfers to MUST-CINEMA without dataset-specific training. It achieves in-bound outputs, whereas the prompted Claude baseline exceeds the target budget. Despite using shorter outputs, HOMURA maintains competitive quality, with a slightly higher CometKiwi score and comparable BT-CERR and BLEU-ρ.

Overall, these results suggest that HOMURA mitigates verbosity bias without substantially compromising general translation performance. The stable transfer from SAND-GLASS to WMT24pp and MUST-CINEMA further indicates that SAND-GLASS provides a challenging and practically relevant benchmark for length- and time-constrained translation.

M Syllables Per Second (SPS) Analysis

Beyond matching the target syllable budget, a practical dubbing or subtitling system must ensure that the translated output is *deliverable*—i.e., a speaker can naturally articulate it within the allotted duration. We therefore introduce **Syllables Per Second (SPS)** as a complementary metric to evaluate information density from a temporal perspective.

Leveraging the precise timestamps provided by our benchmark, we compute SPS as the ratio of the predicted syllable count to the source segment duration. Established psycholinguistic research reports an average English syllabic rate of approximately 6.19 SPS Pellegrino et al. (2011), represent-

Table 12: Zero-shot generalization results on WMT24pp. All systems are evaluated without additional adaptation.

Category	Model	IB	ρ	CK	BC	B-ρ
Unconstrained LLMs	Claude-4.1-Opus	33.3%	1.021	0.842	1.000	0.586
	DeepSeek-V3	30.5%	1.073	0.833	0.969	0.548
	Gemini-2.5-Pro	34.3%	0.998	0.837	0.989	0.599
	GPT-5	30.9%	1.034	0.843	1.000	0.600
Prompt-based Compression	Claude-4.1-Opus	62.6%	0.867	0.734	1.000	0.608
	DeepSeek-V3	46.4%	1.094	0.773	0.979	0.540
	Gemini-2.5-Pro	59.6%	0.803	0.720	0.989	0.606
	GPT-5	84.8%	0.888	0.829	0.989	0.611
Best-of-N Refinement	Claude-4.1-Opus	63.6%	0.852	0.720	1.000	0.619
	DeepSeek-V3	61.2%	0.874	0.730	1.000	0.578
	Gemini-2.5-Pro	77.7%	0.845	0.733	1.000	0.548
	GPT-5	89.6%	0.847	0.797	1.000	0.605
Pipeline-based Post-editing	Claude-4.1-Opus	31.3%	0.707	0.814	0.989	0.616
	DeepSeek-V3	28.4%	0.704	0.804	0.969	0.609
	Gemini-2.5-Pro	54.5%	0.792	0.824	0.989	0.622
	GPT-5	77.3%	0.804	0.816	0.989	0.594
Ours	HOMURA	86.8%	0.803	0.803	0.990	0.605

Note: **IB**: In-Bounds Rate (%). ρ: syllable ratio. **CK**: CometKiwi. **BC**: BT-CERR. **B-ρ**: BLEU-ρ.

Table 13: Zero-shot results on MUST-CINEMA Zh→De.

Model	IB	ρ	CK	BC	B-ρ
`Claude-4.1-Opus`	57.0%	0.999	0.777	**0.976**	**0.498**
HOMURA$_{\text{Rubric}}$	68.5%	0.942	**0.781**	0.966	0.475

Note: **IB**: In-Bounds Rate (%). ρ: syllable ratio. **CK**: CometKiwi. **BC**: BT-CERR. **B-ρ**: BLEU-ρ.

ing the upper bound of comfortable speech delivery. Consistent with this finding, our internal human-annotated studies on live dubbing operations identify **SPS < 6** as the critical comfort threshold for English audiences.

We evaluate all baseline categories defined before on the Zh → En task and report two metrics: (1) Average SPS across all segments, and (2) Pass Rate, the proportion of segments satisfying SPS < 6. Results are summarized in Table 14.

Several key observations emerge. First, unconstrained frontier LLMs (Category 1) consistently exceed the 6 SPS comfort threshold, with average SPS ranging from 6.14 to 6.69 and pass rates barely surpassing 50%. This confirms a systematic *verbosity bias*: when translating from a high-density language (Chinese) to a lower-density one (English), standard models produce outputs that are too long to be naturally spoken within the original duration.

Adding prompt-level syllable constraints (Category 2) substantially reduces SPS, yet pass rates remain between 69–79%, suggesting that most models struggle to reliably internalize explicit length instructions. Best-of-N sampling (Category 3) further improves pass rates to 76–81%, but at the cost of $N\times$ inference overhead. Pipeline post-editing

Table 14: SPS analysis on the Zh → En task. **Avg. SPS** is the mean syllables per second across all segments. **Pass Rate** is the fraction of segments with SPS < 6.

Model	Avg. SPS	Pass Rate @ SPS < 6
Category 1: Unconstrained LLMs		
Claude-4.1-Opus	6.45	54.2%
DeepSeek-V3	6.69	51.7%
Gemini-2.5-Pro	6.14	56.7%
GPT-5	6.30	52.6%
Category 2: Prompt-based Compression		
Claude-4.1-Opus	4.839	79.3%
DeepSeek-V3	5.404	72.3%
Gemini-2.5-Pro	5.247	71.3%
GPT-5	5.346	68.9%
Category 3: Best-of-N Refinement		
Claude-4.1-Opus	4.820	80.6%
DeepSeek-V3	4.849	79.4%
Gemini-2.5-Pro	5.065	77.3%
GPT-5	5.042	76.5%
Category 4: Pipeline Post-editing		
Claude-4.1-Opus	5.178	79.6%
DeepSeek-V3	5.443	70.5%
Gemini-2.5-Pro	4.745	80.3%
GPT-5	4.725	83.0%
Ours		
HOMURA	**4.58**	**86.7%**

(Category 4) yields competitive SPS for Gemini-2.5-Pro and GPT-5; however, the two-stage process introduces its own failure modes—Claude-4.1-Opus and DeepSeek-V3 still exhibit elevated SPS, indicating that post-hoc rewriting does not consistently respect the syllable budget.

In contrast, HOMURA achieves the lowest average SPS (4.58) and the highest pass rate (86.7%), outperforming the best pipeline configuration by 3.7 percentage points. By learning temporal constraints as hard rewards during RL fine-tuning, HOMURA internalizes the information density trade-off directly into the generation process, eliminating the need for external post-processing or repeated sampling.

Domain	Samples	Avg. Chars (ZH)	Total Dur. (Min)	Avg. Dur. (Sec)	Min Dur. (Sec)	Max Dur. (Sec)
ACGN	200	12.02	6.37	1.91	0.56	4.11
Film & Television	200	11.71	7.90	2.37	0.76	6.31
Travel & Tourism	200	11.52	9.40	2.82	0.60	7.52
Gaming	200	12.13	8.93	2.68	0.92	6.61
General Knowledge	200	12.14	9.73	2.92	0.64	9.06
Overall	**1,000**	**11.90**	**42.32**	**2.54**	**0.56**	**9.06**

Table 15: Fine-grained per-domain and global descriptive statistics of the SAND-GLASS source benchmark.

N Domain Distribution and Fine-Grained Textual Attributes

The benchmark comprises exactly 1,000 high-quality golden instances, strictly balanced with 200 instances allocated to each of the five target media domains to prevent domain-specific evaluation bias. Table 15 provides a fine-grained, comprehensive breakdown of the dataset's textual and temporal attributes, capturing both per-domain variations and global corpus metrics.

As indicated by the empirical properties summarized in Table 15, the source segments are derived from colloquial, fast-paced digital platform content.

At the macro level, the complete benchmark encompasses an aggregate audio footprint of **42.32 minutes** of continuous, synchronized speech. The individual segment durations span a highly dynamic and non-trivial range, from an ultra-short minimum of 0.56 seconds (frequently capturing urgent interjections or compact predicate structures in *ACGN*) up to a maximum upper bound of 9.06 seconds (characterizing dense information delivery and explanatory clauses in *General Knowledge*).

The corpus exhibits a stable global mean temporal budget of 2.54 seconds per utterance tightly coupled with an average textual density of 11.90 Chinese characters. This wide yet strictly calibrated variance guarantees that constrained reinforcement learning policies optimized under the *Homura* framework are systematically evaluated against diverse temporal bottlenecks, faithfully simulating the fluid pacing and volatile constraint environments of real-world media localization pipelines.