DocGRPO: Adaptive Adequacy and Fluency Preference Optimization for LLM-Based Document-Level Machine Translation

Anonymous authors
Paper under double-blind review

Abstract

Large language models (LLMs) have shown strong potential for machine translation, yet preserving source information while maintaining coherent discourse remains challenging in document-level machine translation (DocMT). Our preliminary analysis shows that word-alignment-based coverage declines as source documents grow longer, whereas fluency improves and human quality judgments follow a non-monotonic pattern. Coverage alone, however, cannot distinguish genuine omissions from contextually appropriate ellipsis and implicit references, motivating the joint consideration of adequacy and fluency. We propose DocGRPO (**D**ynamic **O**bjective **C**oordination with **G**roup **R**elative **P**olicy **O**ptimization), a reinforcement-learning framework that combines an alignment-based coverage reward for source-content preservation with a discourse-aware fluency reward. To further alleviate the tension between adequacy and fluency, DocGRPO adapts reward weights to training progress and source-document length while combining group-relative and reference-relative advantages to guide policy updates. Experiments across five translation directions and two model scales show improvements over strong document-level baselines. Human evaluation further confirms better overall translation quality, while ablations support the benefits of complementary rewards and adaptive weighting.[1]

1 Introduction

Large language models (LLMs) have substantially improved sentence-level machine translation (SentMT), producing more accurate translations (Zhu et al., 2023; Alves et al., 2024; Wu et al., 2024a). However, their performance remains limited in document-level machine translation (DocMT), which involves longer translation sequences and more complex cross-sentence contextual dependencies (Wang et al., 2023; Lyu et al., 2024). A central challenge is therefore to faithfully preserve source information in longer documents while maintaining fluency and coherence across sentences.

To improve source-content preservation, Wu et al. (2024b) introduced a word-alignment-based preference for sentence-level MT that encourages translations with higher source–target alignment coverage. Although this provides an explicit adequacy signal, word-level coverage alone is insufficient to characterize document-level translation quality. In DocMT, cross-sentence context can support ellipsis and anaphoric references that preserve meaning without explicit word-for-word correspondence. Such contextually appropriate realizations may improve discourse fluency and cohesion yet receive lower coverage scores. Maximizing coverage alone may therefore penalize valid implicit realizations and favor unnecessarily explicit or repetitive translations. Conversely, optimizing fluency alone does not ensure that source information is faithfully preserved. This creates a tension between coverage-based adequacy optimization and discourse fluency, motivating their joint consideration in DocMT.

To examine how this tension varies with document length, we analyze coverage, fluency, and human direct assessment (DA) scores on Zh $\Rightarrow$ En and De $\Rightarrow$ En using Qwen3-8B (Figure 1). As source length increases, coverage scores decline while fluency scores rise. Human DA scores initially decrease but partially recover at longer lengths, despite the continued decline in coverage. This non-

[1]We will release our code on GitHub later.

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

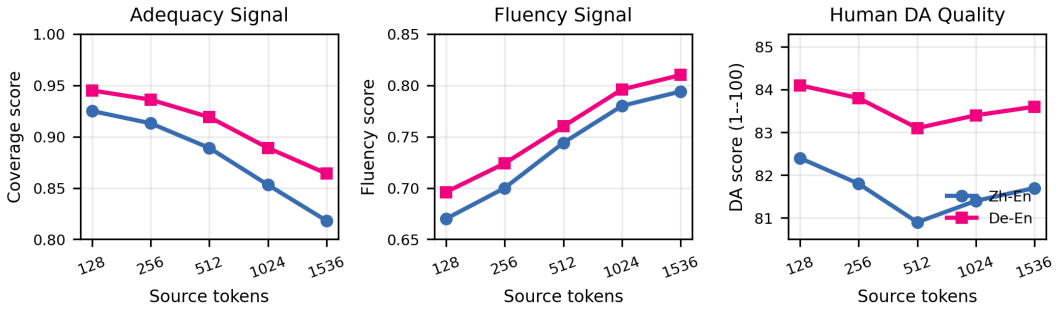


Figure 1: Preliminary analysis on Zh ⇒ En and De ⇒ En with Qwen3-8B. Coverage and fluency follow the definitions in *Fluency and Coverage Reward Estimation*. Refer to Appendix C for human evaluation details.

monotonic pattern suggests that neither coverage nor fluency alone fully captures human judgments of document-level translation quality.

Motivated by these observations, we propose Dynamic Objective Coordination with Group Relative Policy Optimization (DocGRPO), a reinforcement-learning framework that jointly optimizes adequacy and fluency for LLM-based DocMT. DocGRPO combines an alignment-based coverage reward for source-content preservation with a discourse-aware fluency reward for natural and coherent translation. Meanwhile, we introduce a **Stage- and Length-Aware Reward** to adjust their relative weights according to training progress and source-document length. This design prioritizes coverage early in training, progressively increases the fluency weight, and assigns greater fluency emphasis to longer documents. Additionally, to guide policy updates, DocGRPO combines a group-relative advantage that compares policy-generated candidates with a reference-relative advantage that compares each candidate against the human translation. Together, these signals provide within-group discrimination and an external quality anchor, so that optimization does not rely solely on relative differences among potentially low-quality candidates.

We evaluate DocGRPO across five translation directions using two model scales. Automatic and human evaluations show improvements in overall translation quality and discourse fluency, alongside reductions in over- and under-translation errors.

In summary, our main contributions are as follows: 1) We analyze length-dependent trends in coverage, fluency, and human judgments, highlighting the limitations of either signal alone; 2) We propose DocGRPO to jointly optimize coverage and fluency with stage- and length-aware rewards and dual relative advantages; 3) Comparative experiments and ablations demonstrate the benefits of complementary rewards and adaptive weighting.

2 Related Work

Document-Level Machine Translation. Document-level machine translation aims to move beyond sentence-isolated translation by incorporating broader discourse context. Earlier work has shown that document context can help models handle lexical consistency, pronoun resolution, coherence, and other cross-sentence dependencies Jean et al. (2017); Mace & Servan (2019); Xiong et al. (2019); Xu et al. (2021). Recent studies further investigate document-level translation with LLMs, including whole-document prompting and multidimensional evaluation Sun et al. (2025), source-primed multi-turn translation Hu et al. (2025), and multi-granularity knowledge augmentation Hong et al. (2025). More recent work explores filtered synthetic corpora with two-stage LLM adaptation Kim et al. (2026) and graph-guided selection of long-range document context Ji et al. (2026). These approaches show that LLMs can exploit document context, but also report persistent challenges in coherence, omission, and hallucination. Our work instead focuses on dynamically balancing source-content adequacy and discourse fluency during policy optimization.

Hallucination and Adequacy in Machine Translation. Hallucination, omission, and other adequacy errors remain critical reliability issues for machine translation systems. In MT, hallucination is often

reflected as target-side content that is unsupported by the source, while omission or under-translation indicates that source information is missing from the output. Prior work has studied these problems through automatic detection, annotation, and diagnostic measures, including alignment-based signals that estimate whether source content is covered by the translation Dhuliawala et al. (2023); Zhang et al. (2023); Bang et al. (2023); Wu et al. (2024b). Recent work also constructs hallucination-focused preference data to suppress unsupported translations during LLM fine-tuning Tang et al. (2025). These studies motivate explicit adequacy modeling, especially for document-level translation where long contexts make it harder to track all source information. Different from approaches that treat higher coverage as the only goal, we argue that coverage-based adequacy should be considered together with discourse-level fluency, because some implicit or elliptical translations can be legitimate document-level choices.

Preference Optimization for Translation. Preference optimization has become an effective way to align language models with desired output properties. Reinforcement-learning-based alignment and direct preference optimization methods optimize models using pairwise or ranked preferences instead of relying only on maximum-likelihood training Ouyang et al. (2022); Schulman et al. (2017); Rafailov et al. (2024). In machine translation, recent work applies preference-based training to improve overall translation quality Xu et al. (2024) or specifically reduce hallucinated outputs Tang et al. (2025). However, most preference formulations use a single overall quality signal or focus primarily on sentence-level translation. In contrast, our work introduces complementary adequacy and fluency preferences for LLM-based DocMT and adapts their balance to both training stage and document length.

3 Proposed Approach

We introduce Dynamic Objective Coordination with Group Relative Policy Optimization (DocGRPO), a reinforcement-learning framework tailored to LLM-based document-level machine translation. Given a source document $X = \{x_1, \ldots, x_m\}$, DocGRPO optimizes a policy π_θ to generate translations that preserve source information while maintaining coherent target-language discourse. As illustrated in Figure 2, the framework comprises three stages: candidate generation, fluency–coverage reward estimation, and policy optimization with dual relative advantages. The reward formulation incorporates both training progress and source length, enabling the optimization emphasis to shift from adequacy toward fluency as training proceeds and document context increases.

3.1 Candidate Translation Generation

For each source document X, the current policy π_θ independently samples K candidate document translations conditioned on the complete source context:

$$\mathcal{G}(X) = \{Y_1, \ldots, Y_K\}, \qquad Y_i \sim \pi_\theta(\cdot \mid X). \tag{1}$$

The resulting set $\mathcal{G}(X)$ constitutes the on-policy comparison group used for group-relative optimization. In parallel, the corresponding human translation Y_{ref} is retained as an external reference anchor, as illustrated in Figure 2. Because Y_{ref} is not sampled from π_θ, it is excluded from the group statistics and the policy-ratio objective. It is instead evaluated by the same reward functions as the sampled candidates and is used exclusively to construct the reference-relative advantage. This separation preserves the on-policy formulation of GRPO while introducing a stable human-quality calibration signal.

3.2 Fluency and Coverage Reward Estimation

Each sampled translation Y_i receives a fluency score F_i and a coverage score C_i. The same scorers are applied to the human translation, yielding F_{ref} and C_{ref}. Applying identical reward functions to both sources ensures that candidate and reference rewards remain directly comparable.

Coverage Score. Coverage is computed from word alignments between X and Y_i. Let $\mathcal{M}_i$ denote the set of aligned source–target token pairs. A source token is regarded as covered if it is aligned to at least one target token. We define,

$$C_i = \frac{1}{|X|} \sum_{x \in X} \mathbf{1}\big[\exists y \in Y_i : (x, y) \in \mathcal{M}_i\big]. \tag{2}$$

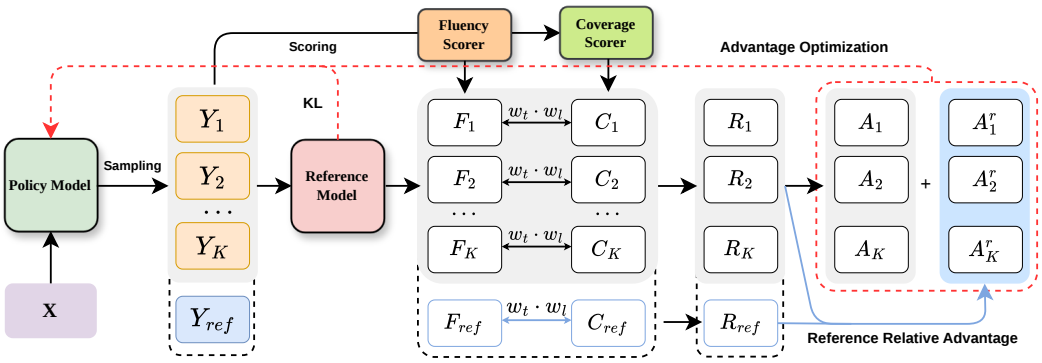


Figure 2: Overview of DocGRPO. Coverage and fluency rewards are adaptively fused by training stage and document length to derive group-relative and reference-relative advantages.

Here, $|X|$ is the number of source tokens and $\mathbf{1}[\cdot]$ is the indicator function. The reference coverage score C_{ref} is computed analogously from alignments between X and Y_{ref}. This reward directly measures source-content realization and therefore penalizes omissions in long-context translation.

Fluency Score. We estimate fluency at both the sentence and token levels. Given a translation Y, we first segment it into a sequence of sentences $Y = \{s_1, \ldots, s_M\}$. To quantify cross-sentence coherence, we apply the Next Sentence Prediction (NSP) head of multilingual BERT (mBERT)[2] to each pair of adjacent sentences. The sentence-level score is the mean probability that s_{j+1} constitutes a coherent continuation of s_j:

$$F_{\text{sent}}(Y) = \frac{1}{M-1} \sum_{j=1}^{M-1} p_{\text{NSP}}\big(\text{IsNext} \mid s_j, s_{j+1}\big). \tag{3}$$

This component measures local discourse continuity across sentence boundaries. For single-sentence translations, for which no adjacent pair is available, we set $F_{\text{sent}}(Y) = 1$ and rely on the token-level component.

We further evaluate token-level well-formedness using the perplexity assigned by the open-source Qwen3.6-27B model[3]. For a translation containing N tokens, its length-normalized perplexity is:

$$\text{PPL}(Y) = \exp\left(-\frac{1}{N} \sum_{n=1}^{N} \log p_{\text{LM}}(y_n \mid y_{<n})\right). \tag{4}$$

Because lower perplexity indicates greater fluency, we convert it into a positively oriented token-level score using its reciprocal,

$$F_{\text{tok}}(Y) = \text{PPL}(Y)^{-1}. \tag{5}$$

The two components are normalized to $[0, 1]$ and combined as:

$$F(Y) = \alpha F_{\text{sent}}(Y) + (1-\alpha) F_{\text{tok}}(Y), \tag{6}$$

where $\alpha \in [0, 1]$ controls the relative contribution of sentence-level coherence and token-level fluency. We set $\alpha = 0.7$, which yields the strongest correlation with human fluency judgments in our preliminary evaluation. We then define $F_i = F(Y_i)$ and $F_{\text{ref}} = F(Y_{\text{ref}})$. Applying the same estimator to sampled and human translations ensures that their fluency rewards remain directly comparable.

3.3 DocGRPO Optimization

Stage- and Length-Aware Reward. As indicated by $w_t \cdot w_l$ in Figure 2, we factor the fluency coefficient into a training-stage term $w_t(t)$ and a source-length term $w_l(X)$. Their product is clipped to obtain:

$$\lambda(X, t) = \min\big(1,\ w_t(t) w_l(X)\big). \tag{7}$$

[2] https://huggingface.co/google-bert/bert-base-multilingual-cased
[3] https://huggingface.co/Qwen/Qwen3.6-27B

The stage term increases monotonically over training, whereas the length term increases with the number of source tokens. One practical instantiation is:

$$w_t(t) = \lambda_0 + (\lambda_1 - \lambda_0)\frac{t}{T}, \quad w_l(X) = 1 + \gamma\frac{|X|}{L}, \tag{8}$$

where T denotes the total number of training steps. The boundary parameters λ_0 and λ_1 specify the initial and final values of the stage-dependent fluency weight, respectively: λ_0 keeps early optimization coverage-oriented, whereas λ_1 determines the eventual emphasis on fluency. Accordingly, $w_t(t)$ controls the temporal transition from adequacy-focused to fluency-focused learning.[4] The length factor $w_l(X)$ adjusts this transition according to source-document length. Here, L is a reference length used to normalize $|X|$, and γ controls the sensitivity of the fluency weight to length variation. The combined rewards for sampled and human translations are:

$$R_i = (1-\lambda)C_i + \lambda F_i, \qquad R_{\text{ref}} = (1-\lambda)C_{\text{ref}} + \lambda F_{\text{ref}}. \tag{9}$$

Consequently, early updates emphasize coverage, while later updates and longer documents assign greater influence to fluency.

Dual Relative Advantages. DocGRPO combines two complementary comparison signals. First, the group-relative advantage measures candidate quality relative to other policy samples:

$$A_i = \frac{R_i - \mu_{\mathcal{G}}}{\sigma_{\mathcal{G}} + \epsilon}, \tag{10}$$

where $\mu_{\mathcal{G}}$ and $\sigma_{\mathcal{G}}$ denote the mean and standard deviation of the rewards assigned to the K policy-generated candidates:

$$\mu_{\mathcal{G}} = \frac{1}{K}\sum_{j=1}^{K} R_j, \qquad \sigma_{\mathcal{G}} = \sqrt{\frac{1}{K}\sum_{j=1}^{K}(R_j - \mu_{\mathcal{G}})^2}. \tag{11}$$

The human-reference reward R_{ref} is excluded from both statistics because it is not sampled from the current policy. Mean centering determines whether a candidate performs above or below the group average, while division by $\sigma_{\mathcal{G}}$ normalizes the reward scale across source documents and candidate groups. The constant ϵ prevents numerical instability when the candidate rewards exhibit little or no variance. Second, the reference-relative advantage compares each candidate with the human anchor:

$$A_i^r = \frac{R_i - R_{\text{ref}}}{\sigma_{\mathcal{G}} + \epsilon}. \tag{12}$$

Following the additive structure in Figure 2, the final advantage is $\widehat{A}_i = A_i + A_i^r$. The first term preserves the variance-reduction property of group-relative optimization, while the second introduces an absolute quality anchor and discourages the group from converging to uniformly poor translations.

Overall Training Objective. The policy ratio can be formalized as follows:

$$\rho_i(\theta) = \frac{\pi_\theta(Y_i \mid X)}{\pi_{\theta_{\text{old}}}(Y_i \mid X)} \tag{13}$$

We first define the clipped surrogate term as follows:

$$\ell_i = \min\big\{\rho_i\widehat{A}_i, \text{clip}(\rho_i, 1-\varepsilon, 1+\varepsilon)\widehat{A}_i\big\}. \tag{14}$$

The DocGRPO objective is then:

$$\mathcal{L}_{\text{DocGRPO}} = -\frac{1}{K}\sum_{i=1}^{K}\ell_i + \beta D_{\text{KL}}, \tag{15}$$

where D_{KL} regularizes π_θ against the frozen reference policy π_{ref}, ε is the clipping threshold, and β controls the regularization strength. This objective jointly exploits within-group ranking and human-reference calibration while preserving conservative policy updates.

[4] Section 5.4 presents ablation results for adaptive reward weighting; Appendix D visualizes the weighting schedule.

4 EXPERIMENTATION

4.1 EXPERIMENTAL SETTING

Datasets. We evaluate DocGRPO on five source-to-English document-level translation tasks including Chinese-to-English (Zh ⇒ En), French-to-English (Fr ⇒ En), German-to-English (De ⇒ En), Spanish-to-English (Es ⇒ En), and Russian-to-English (Ru ⇒ En). To ensure direct comparability with prior work, we use the processed document-level corpora released by Lyu et al. (2024).[5] Detailed dataset statistics and preprocessing information are provided in Appendix A.

Models and Training Settings. We instantiate DocGRPO with Qwen3-4B[6] and Qwen3-8B[7] to examine whether the proposed optimization remains effective across model scales. For each source document, the policy samples $K = 8$ candidate translations with temperature 0.7 and nucleus probability $p = 0.95$. We use AdamW with a learning rate of 1×10^{-6}, an effective batch size of 32 documents, and maximum source and target lengths of 2,048 tokens. The clipping threshold is set to $\varepsilon = 0.2$, the KL coefficient to $\beta = 0.04$, and the numerical-stability constant to $\epsilon = 10^{-6}$. For fluency estimation, we use $\alpha = 0.7$ as determined by its correlation with human judgments. The stage-aware schedule uses $\lambda_0 = 0.2$ and $\lambda_1 = 0.7$, while the length-aware component uses $\gamma = 0.2$ and $L = 256$. All hyperparameters are selected on the validation sets and then held fixed across translation directions and model scales. Following the training recipe of GRPOShao et al. (2024), we introduce a cold-start stage on the document-level training set before DocGRPO optimization, providing the policy with a stable translation initialization. Training uses bfloat16 precision and gradient checkpointing on 8× NVIDIA A100 GPUs. Additional implementation, LoRA, and optimization details are provided in Appendix B.

Baselines. We compare DocGRPO with representative systems covering sentence-level translation, document-level supervised adaptation, preference optimization, and single-reward reinforcement learning. **SentMT-SFT** is trained and evaluated on isolated sentence pairs, providing a context-independent translation baseline. **DocMT-SFT** is supervised on document-level examples and translates each source document jointly, allowing the model to exploit cross-sentence context without reinforcement learning. **WPO** applies word-alignment-based preference optimization Wu et al. (2024b); we include it as an adequacy-oriented preference-learning baseline. **GRPO-Cov** and **GRPO-Flu** retain the same policy initialization, candidate sampling strategy, and GRPO objective as DocGRPO, but optimize exclusively for coverage and fluency, respectively. These two variants isolate the contribution of each reward and reveal the limitations of single-objective optimization. We further include **DeMPT** Lyu et al. (2024), a context-aware prompt-tuning approach for LLM-based document-level MT, as a previously published system on the same translation tasks. All Qwen-based baselines use identical data partitions, decoding configurations, and model backbones to ensure a controlled comparison.[8]

Evaluation. We evaluate translation quality using document-level SacreBLEU Post (2018) and COMET Rei et al. (2020). We employ BlonDe Jiang et al. (2022) to evaluate discourse-sensitive translation phenomena.[9]

4.2 MAIN RESULTS

Table 1 presents the principal evaluation results on five translation directions.

Overall Performance. DocGRPO achieves the strongest average performance under both model scales. With Qwen3-4B, it obtains 40.9 BLEU, 86.1 COMET, and 56.6 BlonDe, outperforming DeMPT by 0.8, 0.9, and 0.6 points, respectively. The improvements remain consistent with Qwen3-8B, where DocGRPO surpasses DeMPT by 0.9 BLEU, 0.8 COMET, and 0.8 BlonDe. Compared with DocMT-SFT, DocGRPO improves the three average metrics by 2.8, 2.7, and 4.0 points with

[5]The datasets are available at `https://github.com/xllyu-nlp/DeMPT`.

[6]`https://huggingface.co/Qwen/Qwen3-4B`

[7]`https://huggingface.co/Qwen/Qwen3-8B`

[8]For fair comparision, we implement the DeMPT and WPO via their relased code with our experimental settings.

[9]As the original COMET and BlonDe implementations are not directly applicable to long-document evaluation, we follow Align-then-Slide Guo et al. (2025) to extend them to the document-level setting.

Table 1: BLEU, COMET (CMT), and BlonDe (BlD) scores on five translation directions. ***Average*** denotes the corresponding metric averaged across all five directions. **Bold** and underlined scores indicate the best and second-best performance within each backbone, respectively. † indicates a statistically significant improvement over DeMPT with $p < 0.01$.

	Methods	Zh ⇒ En			Fr ⇒ En			De ⇒ En			Es ⇒ En			Ru ⇒ En			*Average*		
		BLU	CMT	BlD	BLEU	CMT	BlD	BLEU	CMT	BlD	BLEU	CMT	BlD	BLEU	CMT	BlD	BLEU	CMT	BlD
Qwen3-4B	SentMT-SFT	30.1	83.2	45.2	40.1	82.3	53.1	43.1	86.1	55.1	49.1	85.3	61.2	29.1	81.4	46.1	38.3	83.7	52.1
	DocMT-SFT	29.3	82.8	46.2	40.3	81.3	54.0	42.6	86.2	55.7	48.9	85.8	61.1	29.5	81.0	45.8	38.1	83.4	52.6
	DeMPT	31.9	84.7	50.1	42.4	83.9	56.2	44.7	87.4	59.3	51.0	87.1	**64.3**	30.7	83.1	**50.1**	40.1	85.2	56.0
	WPO	30.2	83.5	46.8	40.8	82.1	55.6	42.9	86.7	56.2	49.5	86.0	60.1	29.7	81.2	44.6	38.6	83.9	52.7
	DocGRPO-Cov	30.9	84.1	47.1	41.6	83.0	54.3	43.4	86.9	55.5	50.3	86.9	60.4	30.1	82.1	44.1	39.3	84.6	52.3
	DocGRPO-Flu	31.2	84.9	49.7	42.3	82.1	56.5	44.3	87.1	58.9	51.2	87.2	63.9	30.4	82.3	49.3	39.9	84.7	55.7
	DocGRPO	**32.4**	**85.7**†	**51.2**†	**43.2**†	**84.3**	**57.8**†	**45.1**	**88.3**†	**60.1**†	**51.7**†	**87.9**†	64.1	**32.1**†	**84.1**†	49.6	**40.9**	**86.1**	**56.6**
Qwen3-8B	SentMT-SFT	32.8	86.6	48.1	42.1	84.8	56.9	44.8	88.7	59.9	52.9	89.1	64.1	32.9	84.5	48.1	41.1	86.7	55.4
	DocMT-SFT	32.2	86.3	49.0	42.3	84.0	57.6	44.4	88.8	60.6	52.6	89.5	64.0	33.2	84.2	47.8	40.9	86.6	55.8
	DeMPT	34.7	88.0	52.9	44.4	86.5	59.3	47.0	90.0	63.8	55.0	90.8	67.2	34.1	86.4	52.3	43.0	88.3	59.1
	WPO	33.0	86.9	49.8	42.8	84.9	58.5	44.9	89.3	60.9	53.5	89.8	63.2	33.3	84.4	46.9	41.5	87.1	55.9
	DocGRPO-Cov	33.6	87.5	50.3	43.5	85.8	57.4	45.6	89.6	60.5	54.2	90.6	63.6	33.8	85.4	46.6	42.1	87.8	55.7
	DocGRPO-Flu	34.0	87.9	52.3	44.2	84.9	59.1	46.7	89.8	63.2	55.3	90.9	66.9	34.0	85.7	51.4	42.8	87.8	58.6
	DocGRPO	**35.2**	**89.1**†	**53.8**†	**45.0**	**86.8**	**60.8**†	**47.6**	**91.0**†	**64.5**†	**55.9**†	**91.6**†	**67.4**	**35.7**†	**87.2**†	**52.9**	**43.9**	**89.1**	**59.9**

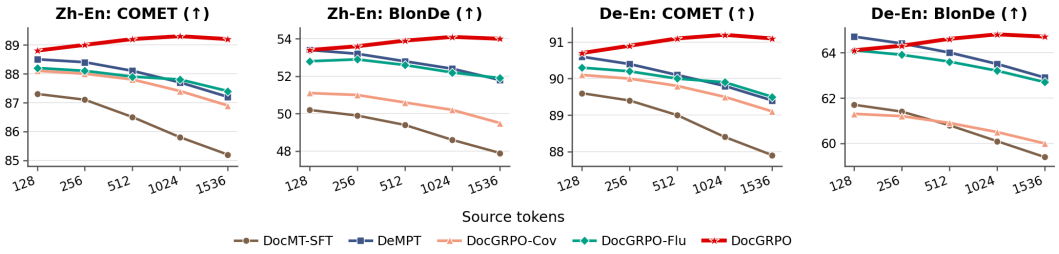


Figure 3: Performance across source-document lengths on Zh ⇒ En and De ⇒ En with Qwen3-8B. Lines show COMET and BlonDe trends as source context increases.

Qwen3-4B, and by 3.0, 2.5, and 4.1 points with Qwen3-8B. Direction-level results exhibit a similar trend, with statistically significant improvements over DeMPT across multiple translation directions and metrics.

Observations. First, the two single-reward variants exhibit complementary behavior. DocGRPO-Cov primarily benefits adequacy-oriented BLEU and COMET, whereas DocGRPO-Flu produces substantially stronger BlonDe scores. Their combination consistently outperforms both variants on all three average metrics, confirming that coverage and fluency provide complementary optimization signals. Second, scaling the backbone from 4B to 8B improves the average DocGRPO scores by 3.0 BLEU, 3.0 COMET, and 3.3 BlonDe while preserving its advantage over the baselines. Third, Qwen3-4B DocGRPO slightly underperforms DeMPT on BlonDe for Es ⇒ En and Ru ⇒ En, although it achieves higher BLEU and COMET. This trade-off disappears with Qwen3-8B, where DocGRPO obtains the best result for every translation direction and metric, suggesting that the larger model more effectively balances source-content preservation and document-level discourse realization.

5 Discussion and Analysis

We conduct further analysis on Zh ⇒ En and De ⇒ En using Qwen3-8B.

5.1 Performance across Document Lengths

We examine whether the benefit of DocGRPO persists as document-level dependencies become longer. The test documents are grouped into five source-length bins ranging from 128 to 1,536 tokens, and Figure 3 reports COMET and BlonDe for each bin. The baseline systems exhibit a steady decline on both metrics as source length increases. In contrast, DocGRPO improves up to 1,024 tokens and

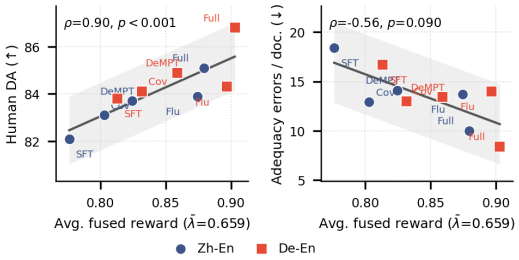


Figure 4: Correlations of the averaged reward proxy $\bar{R} = (1 - \bar{\lambda})C + \bar{\lambda}F$ ($\bar{\lambda} = 0.659$) with human DA and adequacy errors across five systems and two translation directions.

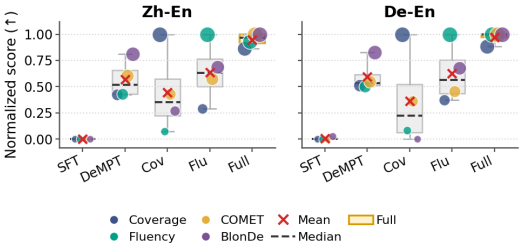


Figure 5: Adequacy–fluency balance with Qwen3-8B on Zh ⇒ En and De ⇒ En. Scores are normalized independently within each metric and language direction.

remains nearly unchanged in the longest bin. Consequently, its advantage becomes larger in more challenging contexts. On Zh ⇒ En, the COMET gain over DeMPT increases from 0.3 points at 128 tokens to 2.0 points at 1,536 tokens, while the BlonDe gain grows from 0.0 to 2.2 points. A similar pattern appears on De ⇒ En: the COMET margin expands from 0.1 to 1.7 points, and DocGRPO changes from trailing DeMPT by 0.6 BlonDe points in the shortest bin to leading it by 1.8 points in the longest bin. These widening margins indicate that the length-aware coordination of adequacy and fluency is particularly beneficial when longer contexts place greater demands on content preservation and discourse realization.

5.2 Reward Validity

We examine whether the fused reward formulation aligns with independent human judgments at the aggregate level. Since the actual coefficient $\lambda(X, t)$ varies for every training stage and source document, it cannot be reconstructed exactly from the reported system-level scores. We therefore uniformly average the coefficient over the full training trajectory and the five source-length bins, yielding $\bar{\lambda} = 0.659$, and construct the proxy $\bar{R} = (1 - \bar{\lambda})C + \bar{\lambda}F$.

Figure 4 shows that this proxy is strongly correlated with human DA ($\rho = 0.90$, $p < 0.001$). Its correlation with the total number of adequacy errors is negative but not statistically significant ($\rho = -0.56$, $p = 0.090$), which is expected because the fused reward also assigns substantial weight to discourse fluency. This analysis validates the aggregate behavior of the reward formulation rather than replacing a document-level analysis based on the original training logs.

5.3 Analysis of Adequacy–Fluency Preferences Tension

Figure 5 illustrates how coverage and fluency preferences affect document-level translation quality. SFT, Cov, Flu, and Full denote DocMT-SFT, DocGRPO-Cov, DocGRPO-Flu, and DocGRPO, respectively. Across both translation directions, DocGRPO-Cov achieves the highest coverage but offers limited fluency gains, whereas DocGRPO-Flu achieves the highest fluency but lower coverage than the coverage-only variant. Neither single-preference model achieves the best overall translation quality. By jointly optimizing the two preferences, DocGRPO attains a more balanced coverage–fluency profile and the highest COMET and BlonDe scores on both Zh ⇒ En and De ⇒ En. These results suggest that maximizing either signal alone is insufficient: coordinating the two preferences better supports source-content preservation and natural discourse realization. Appendix E provides illustrative examples of the discourse phenomena underlying this tension.

5.4 Ablation of Adaptive Reward Weighting

Figure 6 reports ablation results on Zh ⇒ En and De ⇒ En with Qwen3-8B. DocGRPO consistently achieves the best performance across all metrics and both translation directions. Replacing the adaptive coefficient with a fixed value reduces BLEU by 0.8 points and COMET by 0.9 points on both Zh ⇒ En and De ⇒ En, together with clear decreases in BlonDe, coverage, and fluency. Removing either adaptive component also degrades performance, confirming that training stage and source

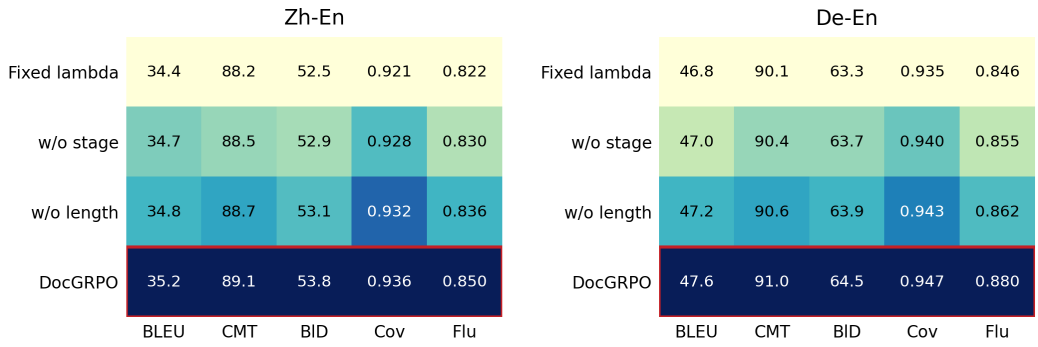


Figure 6: Ablation results for adaptive reward weighting on Zh ⇒ En and De ⇒ En with Qwen3-8B. Colors are normalized independently within each metric, cell annotations show raw values, and the red boxes highlight DocGRPO.

length provide complementary weighting signals. The variant without stage awareness exhibits a larger decline than the variant without length awareness, indicating that progressively shifting the optimization emphasis is particularly important, while length awareness provides additional gains by adapting the reward to document complexity.

5.5 Over- and Under-Translation Analysis

We conduct a human evaluation to directly assess translation adequacy and overall quality. Annotators assign a direct assessment (DA) score (Graham et al., 2017) to each translated document and independently count under-translation and over-translation errors. Under-translation denotes semantically necessary source content that is omitted, while over-translation denotes unsupported, repeated, or spuriously added content. Refer to Appendix C for annotation details.

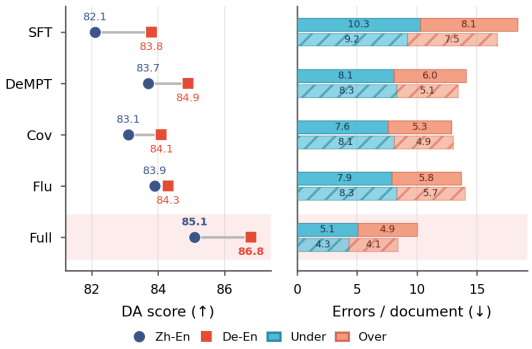


Figure 7: Human evaluation on Zh ⇒ En and De ⇒ En with Qwen3-8B.

Figure 7 reports average DA score and the average number of each error type per document. Human judgment distinguishes legitimate ellipsis and target-language restructuring from genuine adequacy errors, clarifying whether DocGRPO reduces information loss without suppressing valid discourse phenomena.

6 Conclusion

We address the tension between source-content preservation and discourse fluency in LLM-based document-level machine translation. Word-level coverage alone cannot fully capture translation quality, as contextual ellipsis and implicit references can preserve meaning without explicit lexical correspondence. We propose DocGRPO to coordinate coverage and fluency preferences through stage- and length-aware reward weighting, combining group-relative and reference-relative advantages to guide policy updates. Experiments across five translation directions and two model scales show improvements over strong document-level baselines. Human evaluation confirms higher overall quality and fewer over- and under-translation errors. Further analyses support the complementarity of the rewards, the benefits of adaptive weighting, and the fused reward's alignment with aggregate human judgments. Together, these findings underscore the importance of balancing faithful information transfer with coherent discourse realization, rather than maximizing either coverage or fluency alone.

AI USE STATEMENT

Generative AI tools were used to assist with code implementation and manuscript polishing. The authors independently verified all code, technical claims, derivations, citations, and empirical results and take full responsibility for the final content of this work.

REFERENCES

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. Tower: An open multilingual large language model for translation-related tasks, 2024.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 675–718, 2023.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. *ArXiv*, abs/2309.11495, 2023. URL https://api.semanticscholar.org/CorpusID:262062565.

Yvette Graham, Qingsong Ma, Timothy Baldwin, Qun Liu, Carla Parra, and Carolina Scarton. Improving evaluation of document-level machine translation quality estimation. In *Proceedings of EACL*, pp. 356–361, 2017.

Jiaxin Guo, Daimeng Wei, Yuanchang Luo, Xiaoyu Chen, Zhanglin Wu, Huan Yang, Hengchao Shang, Zongyao Li, Zhiqiang Rao, Jinlong Yang, and Hao Yang. Align-then-slide: A complete evaluation framework for ultra-long document-level machine translation, 2025. URL https://arxiv.org/abs/2509.03809.

Hanghai Hong, Yibo Xie, Jiawei Zheng, and Xiaoli Wang. SubDocTrans: Enhancing document-level machine translation with plug-and-play multi-granularity knowledge augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 14490–14506. Association for Computational Linguistics, 2025.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proceedings of ICLR*, 2022.

Hanxu Hu, Jannis Vamvas, and Rico Sennrich. Source-primed multi-turn conversation helps large language models translate documents. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23702–23712. Association for Computational Linguistics, 2025.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. Does neural machine translation benefit from larger context? *Computing Research Repository*, arXiv:1704.05135, 2017.

Baijun Ji, Zixuan Zhou, Xiangyu Duan, Yu Liu, Longbo Sun, Rupu Wei, and Bohong Zhao. G^2C-MT: Graph-guided context selection for document-level machine translation. *arXiv preprint arXiv:2606.03078*, 2026.

Yuchen Eleanor Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. BlonDe: An automatic evaluation metric for document-level machine translation. In *Proceedings of NAACL*, pp. 1550–1565, 2022.

Ireh Kim, Tesia Sker, and Chanwoo Kim. Enhancing document-level machine translation via filtered synthetic corpora and two-stage LLM adaptation. *arXiv preprint arXiv:2603.22186*, 2026.

Xinglin Lyu, Junhui Li, Yanqing Zhao, Min Zhang, Daimeng Wei, Shimin Tao, and Hao Yang. Dempt: Decoding-enhanced multi-phase prompt tuning for making llms be better context-aware translators. In *Proceedings of EMNLP*, pp. 20280–20295, 2024.

Valentin Mace and Christophe Servan. Using whole document context in neural machine translation. In *Proceedings of IWSLT*, 2019.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of WMT*, pp. 186–191, 2018.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In *Proceedings of EMNLP*, pp. 2685–2702, 2020.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Yirong Sun, Dawei Zhu, Yanjun Chen, Erjia Xiao, Xinghao Chen, and Xiaoyu Shen. Fine-grained and multi-dimensional metrics for document-level machine translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pp. 1–17. Association for Computational Linguistics, 2025.

Zilu Tang, Rajen Chatterjee, and Sarthak Garg. Mitigating hallucinated translations in large language models with hallucination-focused preference optimization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3410–3433. Association for Computational Linguistics, 2025.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. Document-level machine translation with large language models. In *Proceedings of EMNLP*, pp. 16646–16661, 2023.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. Adapting large language models for document-level machine translation. *Computing Research Repository*, arXiv:2401.06468, 2024a.

Qiyu Wu, Masaaki Nagata, Zhongtao Miao, and Yoshimasa Tsuruoka. Word alignment as preference for machine translation. In *Proceedings of EMNLP*, pp. 3223–3239, 2024b.

Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. Modeling coherence for discourse neural machine translation. In *Proceedings of AAAI*, pp. 7338–7345, 2019.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*, 2024.

Hongfei Xu, Deyi Xiong, Josef Van Genabith, and Qiuhui Liu. Efficient context-aware neural machine translation with layer-wise weighting and input-aware gating. In *Proceedings of IJCAI*, pp. 3933–3940, 2021.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren's song in the ai ocean: A survey on hallucination in large language models. *ArXiv*, abs/2309.01219, 2023. URL https://api.semanticscholar.org/CorpusID:261530162.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*, 2023.

A Dataset Details

Following DeMPT Lyu et al. (2024), we use its processed document-aligned News Commentary v18 corpus for Zh/Fr/De/Es/Ru–En translation. We retain the released splits: 80% of document pairs for training and 150 document pairs each for validation and testing. Table 2 reports the document and sentence-pair counts. We preserve document boundaries and sentence order, tokenize the released data with the Qwen3 tokenizer, and apply no additional corpus-level preprocessing.

Table 2: Statistics of the document-level training, validation, and test sets adopted from DeMPT. Doc. and Sent. denote the numbers of document pairs and sentence pairs, respectively.

	Zh ⇒ En		**Fr ⇒ En**		**De ⇒ En**		**Es ⇒ En**		**Ru ⇒ En**	
Split	**Doc.**	**Sent.**	**Doc.**	**Sent.**	**Doc.**	**Sent.**	**Doc.**	**Sent.**	**Doc.**	**Sent.**
Training	8,622	342,495	7,915	310,489	8,417	333,201	9,677	378,281	7,255	272,100
Validation	150	6,061	150	5,890	150	5,866	150	5,782	150	5,691
Test	150	5,747	150	5,795	150	5,967	150	5,819	150	5,619

B Training Details

Model Initialization and LoRA Adaptation. We use Qwen3-4B and Qwen3-8B as the foundation models. All Qwen-based systems, including the supervised baselines, reproduced preference-learning baselines, and DocGRPO variants, are fine-tuned with Low-Rank Adaptation (LoRA) Hu et al. (2022). The backbone parameters remain frozen and only the LoRA parameters are updated. We use the same LoRA configuration for all systems at a given model scale so that performance differences are attributable to the training objective rather than the number of trainable parameters. All systems also share the same tokenizer, data splits, maximum sequence lengths, and decoding configuration.

Cold Start and Reinforcement Learning. Before reinforcement-learning optimization, we perform a cold-start stage on the document-level training set to obtain a stable translation policy. DocGRPO is initialized from this document-level supervised checkpoint. We implement DocGRPO with EasyR1, an efficient reinforcement-learning framework based on veRL, [10] and extend its GRPO pipeline with our coverage and fluency rewards, stage- and length-aware reward combination, and reference-relative advantage. For each source document, the policy samples $K = 8$ candidate translations using temperature 0.7 and nucleus sampling with $p = 0.95$. The human translation is scored by the same reward functions but is excluded from the on-policy group statistics and policy-ratio computation.

Optimization and Reward Configuration. We optimize the LoRA parameters with AdamW using a learning rate of 1×10^{-6} and an effective batch size of 32 documents. Source and target sequences are truncated to at most 2,048 tokens. The GRPO clipping threshold is $\varepsilon = 0.2$, the KL coefficient is $\beta = 0.04$, and the numerical-stability constant is 10^{-6}. For fluency estimation, the sentence-level and token-level components are combined with $\alpha = 0.7$. The adaptive reward uses $\lambda_0 = 0.2$, $\lambda_1 = 0.7$,

[10] https://github.com/hiyouga/EasyR1

Table 3: Training and optimization configurations used in our experiments.

Configuration	Value
Foundation models	Qwen3-4B, Qwen3-8B
Parameter adaptation	LoRA
Candidate translations K	8
Sampling temperature	0.7
Nucleus probability p	0.95
Learning rate	1×10^{-6}
Effective batch size	32 documents
Maximum source/target length	2,048 tokens
Clipping threshold ε	0.2
KL coefficient β	0.04
Fluency coefficient α	0.7
Stage boundaries (λ_0, λ_1)	$(0.2, 0.7)$
Length parameters (γ, L)	$(0.2, 256)$
Precision	bfloat16
Hardware	8× NVIDIA A100

$\gamma = 0.2$, and reference length $L = 256$. Training uses bfloat16 precision and gradient checkpointing on eight NVIDIA A100 GPUs.

Controlled Comparison. All Qwen-based baselines use identical training and evaluation splits. For WPO and DeMPT, we adapt the released implementations to the same Qwen3 checkpoints and LoRA-based training setup. GRPO-Cov and GRPO-Flu use the same initialization, candidate sampling, optimizer, and policy objective as DocGRPO, differing only in the reward used for optimization. The adaptive-weighting ablations additionally share all settings except for the specified stage or length component.

C Human Evaluation Details

Evaluation Scope and Recruitment. We conduct human evaluation on Zh–En and De–En translations generated with Qwen3-8B. The evaluated systems include DocMT-SFT, DeMPT, DocGRPO-Cov, DocGRPO-Flu, and DocGRPO. We recruit bilingual evaluators with professional translation experience in the corresponding language pair. For each document, evaluators are shown the complete source document, its human reference, and one anonymized system translation. System identities are hidden, and translations are presented in randomized order.

Direct Assessment Criterion. Following the direct assessment protocol Graham et al. (2017), evaluators assign each translated document a score from 0 to 100. They jointly consider whether the translation preserves the meaning of the complete source document and whether it forms a fluent and coherent target-language document. Scores of 90–100 indicate accurate, complete, and natural translations; 70–89 indicate generally correct translations with minor errors; 50–69 indicate major errors while retaining part of the source meaning; and 0–49 indicate severe mistranslation, omission, unsupported generation, or loss of document-level coherence.

Over- and Under-Translation Annotation. Evaluators also independently identify and count adequacy errors in each document. **Under-translation** refers to semantically necessary source content that is absent from the translation, including omitted events, entities, attributes, or discourse relations. **Over-translation** refers to unsupported or unnecessarily repeated target content, including hallucinated information, duplicated meaning, and additions that cannot be inferred from the source context. Legitimate ellipsis, target-language restructuring, and contextually recoverable implicit information are not counted as under-translation. Explicit target expressions required for grammaticality or discourse coherence are not treated as over-translation.

Quality Control and Aggregation. Evaluators annotate each document independently and mark the source or target span associated with every adequacy error. Disagreements concerning legitimate implicit realization and genuine omission or addition are resolved through adjudication. We average DA scores across evaluators and documents and report the mean numbers of under- and over-translation errors per document.

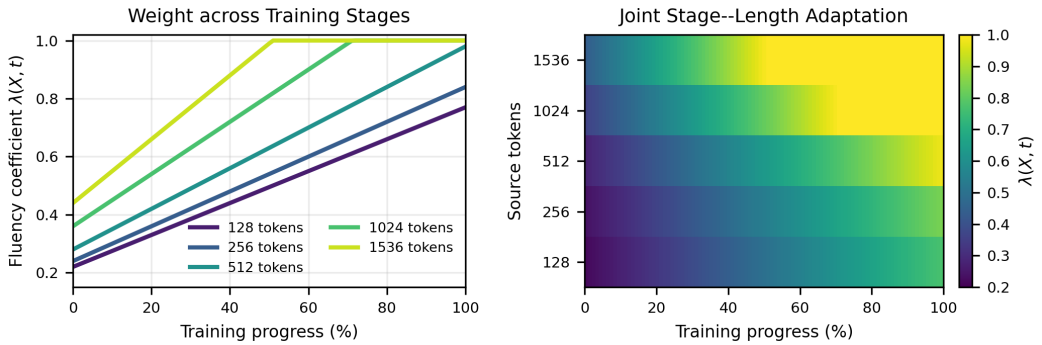


Figure 8: Stage- and length-aware fluency coefficient with $\lambda_0 = 0.2$, $\lambda_1 = 0.7$, $\gamma = 0.2$, and $L = 256$. The left panel shows the coefficient across training stages for different source lengths, and the right panel presents their joint effect.

D Visualization of Adaptive Reward Weighting

Figure 8 visualizes the adaptive fluency coefficient used in our experiments. The coefficient increases as training proceeds and assigns greater fluency emphasis to longer documents, while clipping the maximum value at 1.

E Discourse Phenomena and Case Studies

We analyze pronoun resolution, lexical cohesion, ellipsis, and entity consistency. Table 4 presents two representative examples that illustrate how the complementary rewards affect document-level translation behavior. The examples contrast genuine omissions and unnecessary repetition with contextually appropriate implicit realization.

Table 4: Representative case studies of discourse-sensitive translation. Bold text marks omissions, unsupported or repetitive realizations, lexical inconsistencies, and contextually appropriate discourse choices. These examples are not sampled model outputs.

System	Zh ⇒ En Example	De ⇒ En Example
Source	*English gloss:* The U.S. representative said the talks would continue on Friday. Meanwhile, the EU delegation will submit a revised proposal. The U.S. delegation will respond next week.	Die Kommission legte den Bericht am Montag vor. Sie erklaerte, die Massnahmen wuerden erst nach einer weiteren Pruefung umgesetzt. Fuer die Mitgliedstaaten sei keine sofortige Reaktion erforderlich.
Reference	The U.S. representative said the talks would continue on Friday. For its part, the EU delegation will submit a revised proposal. The U.S. delegation will respond next week.	The Commission presented the report on Monday. It said the measures would be implemented only after further review. No immediate response was required from the member states.
DocMT-SFT	The U.S. representative said the talks would continue. Meanwhile, the EU delegation will submit a proposal. The U.S. delegation will respond next week. **(omits "on Friday" and "revised")**	The Commission presented the report on Monday. **She** said the measures would be implemented after review. The member states did not need to respond immediately. **(incorrect pronoun; omits "only" and "further")**
DeMPT	The **U.S.** representative said the talks would continue on Friday. At the same time, the EU delegation will submit a revised proposal. The **American** delegation will respond next week. **(inconsistent entity realization)**	The Commission presented the report on Monday. It said the measures would be implemented after further review. The member states did not need to respond immediately. **(omits "only")**
DocGRPO-Cov	The U.S. representative said the talks would continue on Friday. At the same time, **the EU delegation** will submit a revised proposal, and **the U.S. delegation** will respond next week. **(adequate but overly explicit)**	The Commission presented the report on Monday. **The Commission** said the measures would be implemented only after further review. **The member states** required no immediate response. **(repetitive realization)**
DocGRPO-Flu	The U.S. representative said the talks would continue on Friday. For its part, the EU delegation will submit a proposal, with Washington responding later. **(omits "revised" and "next week")**	The Commission presented the report on Monday. It said the measures would be implemented after further review, with no immediate response required. **(omits "only" and the member-state argument)**
DocGRPO	The U.S. representative said the talks would continue on Friday. **For its part**, the EU delegation will submit a revised proposal. **The U.S. delegation** will respond next week.	The Commission presented the report on Monday. **It** said the measures would be implemented only after further review. No immediate response was required from the member states.