

Style-Ruler: Inference-Only Stylized Translation via Unsupervised Extraction and Agentic Collaboration

Hong Ren, Jiahe Li, Shaolin Zhu, Deyi Xiong, Nannan Wu*

School of Computer Science and Technology, Tianjin University, China
{rhong,ljh_acking1,zhushaolin,dyxiong,nannan.wu}@tju.edu.cn

Abstract

While Large Language Models have achieved state-of-the-art performance in general machine translation, they frequently struggle to adhere to complex stylistic nuances and lexical constraints strictly and often revert to a neutral style. The scarcity of style-specific parallel data hinders existing solutions, whereas standard prompting lacks rigorous verification mechanisms. To bridge this gap, we propose **Style-Ruler**, an inference-only agentic framework that operationalizes style via unsupervised extraction and neuro-symbolic verification. Style-Ruler orchestrates three specialized agents: (1) a Style-Reverse Analyst (SRA) that mines quantifiable linguistic constraints from monolingual reference texts via unsupervised extraction; (2) a Constraint-Adaptive Translator (CAT) that employs a Style-Chain-of-Thought mechanism to plan stylistic adherence before generation explicitly; and (3) a Sandbox-Augmented Auditor (SAA) that provides heterogeneous feedback through deterministic code execution. These components form a closed-loop refinement mechanism where symbolic logic guides neural generation. Extensive experiments on the GuoFeng Webnovel dataset and WMT24++ benchmark demonstrate the superiority of our approach. Style-Ruler achieves a CSR of 82.4% on stylized translation, surpassing standard Chain-of-Thought prompting by over 50%, while improving BLEU scores by +1.0 on En→Zh.

1 Introduction

Large Language Models (LLMs) have achieved remarkable advancements in machine translation (MT), demonstrating strong performance even in low-resource settings (Alves et al., 2023; Manakhimova et al., 2024; Zhu et al., 2024c). Unlike traditional Neural Machine Translation (NMT) systems that rely on massive parallel corpora, LLMs

leverage their emergent abilities to perform translation via in-context learning (Agrawal et al., 2023; Cui et al., 2024; Zhu et al., 2024a). As the scale of LLMs increases, they have shown the potential to bridge linguistic gaps with high fluency and coherence (Dong et al., 2025; He et al., 2024).

However, deploying LLMs for controllable and stylized translation¹ remains a significant challenge (Tao et al.; Havaladar et al., 2025). Recent evaluations reveal that while LLMs excel at capturing the gist of a source text, they often fail to preserve specific stylistic nuances or adhere to strict lexical constraints (Kim et al., 2024). For instance, Havaladar et al. (2025) observed a tendency towards neutrality in LLM-generated translations where the distinct voice of the original author is often washed out. Similarly, Manakhimova et al. (2024) found that general-purpose LLMs still lag behind specialized NMT systems in handling fine-grained linguistic phenomena, such as terminology consistency, negation, and specific syntactic structures (e.g., stripping). This limitation is particularly critical in professional scenarios (e.g., legal or literary translation) where adherence to a specific style guide or glossary is not merely a preference but a requirement (Kim et al., 2024).

Existing approaches to address these issues generally fall into two categories: supervised fine-tuning (SFT) and prompt engineering (Zhang et al., 2023; Zheng et al., 2024). SFT-based methods, such as constructing parallel datasets for style transfer (Tao et al.), can effectively adapt models to a target domain. However, they suffer from two significant drawbacks: (1) Data Scarcity: high-quality parallel corpora aligned with specific styles are expensive or impossible to obtain (Coleman et al., 2024); and (2) Catastrophic Forgetting: fine-tuning of a particular style often degrades the

*Corresponding authors.

¹A detailed delineation of our scope of stylized translation is provided in Appendix A.4

model’s general capabilities (Dong et al., 2025; Zhu et al., 2024b). Prompt engineering and retrieval-augmented generation offer a lightweight alternative (Pourkamali and Sharifi, 2024). Yet, LLMs often hallucinate or ignore complex instructions in the prompt, especially when multiple constraints (e.g., length, terminology, and tone) are imposed simultaneously (Benkirane et al., 2024; Coleman et al., 2024).

To bridge this gap, we propose **Style-Ruler**, an inference-only structured agentic workflow that disentangles style definition from content generation. Style-Ruler explicitly extracts stylistic features and enforces them via a multi-agent workflow. Specifically, our framework comprises three innovative modules: (1) Style-Reverse Analyst (SRA): It functions as an unsupervised feature miner to move beyond simple prompt instructions. By leveraging linguistic probing tools, SRA extracts fine-grained metrics to translate abstract style concepts into explicit, machine-readable constraints. (2) Constraint-Adaptive Translator (CAT): It employs a Style-Chain-of-Thought (Style-CoT) mechanism, which explicitly plans its translation strategy based on the SRA’s constraints. (3) Sandbox-Augmented Auditor (SAA): It recognizes that LLMs struggle with precise counting and logic verification and operates within a Python code sandbox to audit the translation against the style brief.

Our main contributions are summarized as follows: (i) We propose a training-free methodology that explicitly mines quantifiable linguistic constraints from monolingual reference texts. (ii) We design a new mechanism to bridge the gap between neural generation and symbolic logic and achieve rigorous adherence to hard constraints where purely probabilistic models typically fail. (iii) Extensive experiments on multiple benchmarks demonstrate that Style-Ruler surpasses strong baselines in constraint satisfaction rate and style alignment.

2 Related Work

Stylized and Constrained Translation Controlling the style (e.g., formality, literary tone) and lexical constraints (e.g., terminology) of translations is a longstanding challenge (Wang et al., 2023; Coleman et al., 2024). Traditional approaches rely on constructing large-scale parallel datasets aligned with specific styles (Banerjee and Banik,

2025). Tao et al. (2025) proposed CAT-LLM, which utilizes a text style definition module but necessitates training on ten parallel datasets to learn style representations. Similarly, Havaladar et al. (2025) introduced RASTA, a retrieval-augmented method that aligns style by retrieving native exemplars. While effective, these methods depend heavily on the availability of reference corpora or retrieval databases. Regarding lexical constraints, Kim et al. (2024) demonstrated that fine-tuning on terminology pairs can improve consistency in specialized domains like patents. However, Manakhimova et al. (2024) indicates that even advanced LLMs struggle with fine-grained linguistic constraints (e.g., stripping, negation) and often default to a neutral and hallucinated style. Unlike these works, Style-Ruler operationalizes style not through implicit embeddings or parallel data, but through unsupervised extraction of explicit linguistic rules from monolingual text.

Agentic Collaboration and Verification The deployment of LLMs as specialized LLM modules has shown promise in complex reasoning tasks (Moslem et al., 2023; Qian et al., 2024; Briva-Iglesias, 2025; Feng et al., 2025). Structured agentic workflows, such as TransAgents (Wu et al., 2024), simulate the workflow of a translation company (translator, reviewer, proofreader) to improve quality through iterative dialogue. Similarly, Ren et al. (2026) proposed AdaDPI, an adaptive agentic framework for document-level translation that dynamically internalizes discourse constraints directly into model parameters, offering a complementary perspective on agentic translation control. However, most existing agentic frameworks rely on homogeneous feedback, in which one LLM critiques another in natural language. This leads to error propagation, as LLMs are prone to hallucinations when verifying logical or arithmetic constraints (Guerreiro et al., 2023; Benkirane et al., 2024; Gogoulou et al., 2025). To address this, recent studies have explored tool-augmented agents. Coleman et al. (2024) utilized rule-based sentence builders to assist LLMs in low-resource translation. Building on this line of work, our framework introduces a multi-agent collaboration mechanism. Instead of relying solely on probabilistic critique, our method executes deterministic code within a sandbox to verify hard constraints. This heterogeneous feedback loop ensures rigorous adherence to the style brief, distinguishing our approach from purely neural revision methods.

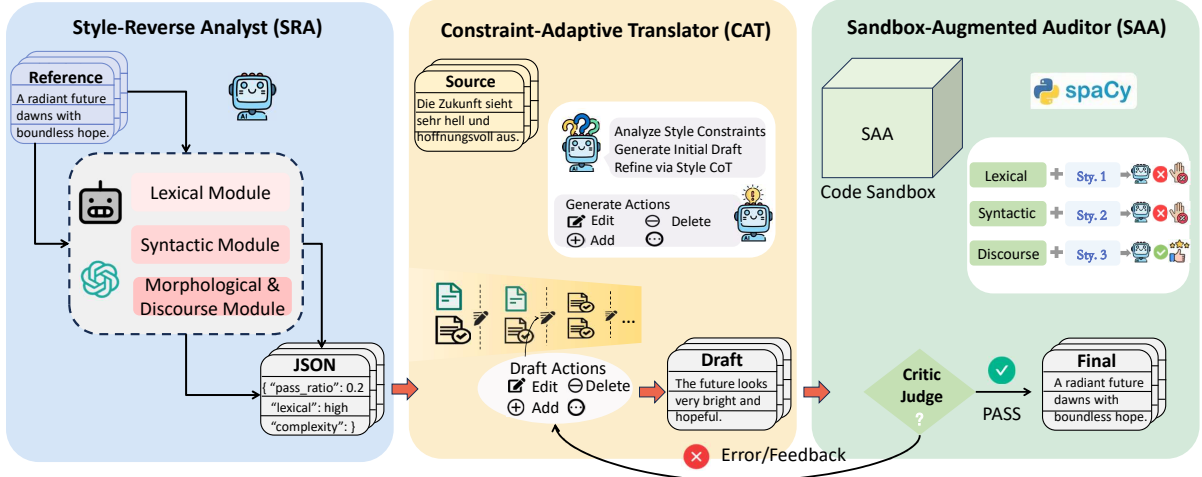


Figure 1: Illustration of the proposed Style-Ruler.

3 Methodology

The proposed Style-Ruler framework is illustrated in Figure 1. We address the task of stylized machine translation where, in addition to the source sentence x , a set of monolingual reference texts $\mathcal{R}$ representing a specific target style is provided. The objective is to generate a translation y that preserves the semantics of x while adhering to the linguistic and stylistic patterns inherent in $\mathcal{R}$. Unlike standard translation models that model $P(y|x)$ or rely on implicit style embeddings, we decompose the generation process into an explicit, interpretability-driven pipeline mediated by a discrete style brief $\mathcal{B}$. Our framework coordinates three specialized agents: (1) the Style-Reverse Analyst (SRA), which extracts the explicit style definition $\mathcal{B}$ from $\mathcal{R}$ via unsupervised mining; (2) the Constraint-Adaptive Translator (CAT), which generates the translation conditioned on both x and $\mathcal{B}$; and (3) the Sandbox-Augmented Auditor (SAA), which verifies the output using deterministic code execution. These components form a collaborative refinement loop in which heterogeneous feedback from the SAA guides the CAT to optimize the translation until rigorous constraint satisfaction is achieved iteratively.

3.1 Style-Reverse Analyst (SRA)

The primary objective of the SRA is to extract an explicit, machine-readable style definition from a set of monolingual reference texts $\mathcal{R} = \{r_1, r_2, \dots, r_m\}$, without relying on parallel translation pairs. Unlike latent style representations used in previous work (Tao et al.), SRA operationalizes style as a set of quantifiable linguistic

constraints. We define a feature extraction function $\Phi(\cdot)$ that maps a text sequence to a multi-dimensional feature vector. To capture the multifaceted nature of style, Φ is composed of three distinct probing modules:

- **Lexical Module (Φ_{lex}):** This module analyzes vocabulary usage. It computes metrics such as the Type-Token Ratio to measure vocabulary diversity and calculates the density of specific word classes to determine the register.
- **Syntactic Module (Φ_{syn}):** This module focuses on sentence structure. We parse each sentence into a dependency tree to extract structural features. Key metrics include the average dependency distance, which correlates with syntactic complexity, and the passive voice ratio.
- **Morphological & Discourse Module (Φ_{mor}):** This module captures the texture and rhythm of the text. It tracks the distribution of verb tenses, grammatical person perspectives, and sentence-length variance, which reflects the rhythmic pacing of the writing.

For the entire reference set $\mathcal{R}$, the SRA first computes the feature vector for each sentence r_i and then aggregates them to form a global style profile $\mathbf{v}_{\mathcal{R}}$. The aggregation is performed via mean pooling to minimize the noise from outliers:

$$\mathbf{v}_R = \frac{1}{m} \sum_{i=1}^m \text{Concat}(\Phi_{\text{lex}}(r_i), \Phi_{\text{syn}}(r_i), \Phi_{\text{mor}}(r_i)) \quad (1)$$

While $\mathbf{v}_R$ provides a numerical representation, these raw numbers are not directly actionable for a language model. Therefore, the SRA performs a final summarization step. It takes both the raw text $\mathcal{R}$ and the calculated profile $\mathbf{v}_R$ as input to generate a style brief ($\mathcal{B}$). The ($\mathcal{B}$) is a natural language description formatted as a JSON object, containing specific constraints (detailed in Appendix A.1).

3.2 Constraint-Adaptive Translator (CAT)

The CAT is responsible for generating the target text. A common issue in standard LLM prompting is instruction skipping, where the model ignores specific constraints when the context becomes complex. To address this, we introduce an intermediate planning step z , termed Style Chain-of-Thought (Style-CoT). We decompose the generation probability $P(y|x, \mathcal{B})$ into two sequential stages: planning and execution. Formally, given x and $\mathcal{B}$ derived by the SRA, the joint probability is factorized as:

$$P(y, z|x, \mathcal{B}) = \underbrace{P(z|x, \mathcal{B})}_{\text{Planning}} \cdot \underbrace{P(y|x, \mathcal{B}, z)}_{\text{Execution}} \quad (2)$$

Phase 1: Planning (z). The CAT agent analyzes $\mathcal{B}$ against x . It generates a natural language plan z that explicitly identifies which words in x require terminological replacement and which syntactic structures must be transformed. This step forces the model to attend to the constraints before committing to a translation.

Phase 2: Execution (y). Conditioned on the plan z , the model generates the final translation y . By explicitly conditioning on z , the generation process is grounded in the style constraints, reducing the likelihood of reverting to the model’s default training distribution.

We implement this two-stage process within a single inference call using a structured prompt. The prompt template and examples of the Style-CoT reasoning process are detailed in Appendix A.2.

3.3 Sandbox-Augmented Auditor (SAA)

Existing agent workflows often employ one LLM to critique another’s output. Since both the translator and the critic are neural probabilistic models, they share the same underlying weaknesses. If an LLM struggles to count tokens or verify strict set inclusion during generation accurately, a critic LLM will likely fail to detect these errors during evaluation. To address this, SAA introduces a heterogeneous feedback mechanism. We combine the semantic understanding of neural models with the rigor of symbolic logic. In this setup, the auditor does not directly evaluate the text. Instead, it translates the constraints into a deterministic Python script. The feedback comes from the execution of this script rather than the model’s internal probabilities. Formally, given the generated translation y and $\mathcal{B}$, the SAA generates a verification code snippet c :

$$c = \text{CodeGen}(y, \mathcal{B}) \quad (3)$$

This script c contains assertion statements mapping to the constraints in $\mathcal{B}$. We then execute c in an isolated Python interpreter (Sandbox). We define the execution function $E(\cdot)$ which returns a feedback signal f :

$$f = E(c) = \begin{cases} \text{PASS}, & \text{no error} \\ \text{FAIL} + \text{Message}(e), & \text{error} \end{cases} \quad (4)$$

This approach ensures that hard constraints are verified by a system that effectively prevents logical errors, thereby complementing the probabilistic nature of the generation model (Details in Appendix A.3).

3.4 Collaborative Refinement Mechanism

The individual agents described above are integrated into a closed-loop system. We treat the translation process not as a one-shot generation, but as a trajectory optimization problem over a multi-turn dialogue. The collaboration is driven by the interaction between CAT and SAA, initialized by the style definition from the SRA. The process begins with SRA extracting $\mathcal{B}$ from the reference set $\mathcal{R}$. At the initial step ($t = 0$), CAT generates a draft translation y_0 conditioned on the source text x and $\mathcal{B}$. For each subsequent iteration t , SAA takes the current draft y_t and generates a verification script c_t . Executing this script yields

a deterministic feedback signal f_t . The workflow branches based on the content of f_t :

1. **Termination (Success):** If f_t returns PASS, the system terminates, and y_t is output as the final translation y^* .
2. **Iteration (Refinement):** If f_t contains error messages, these errors are treated as negative constraints. We update the dialogue history H_t to include the rejected draft and the specific feedback.

Formally, we define refinement as a conditional-probability update. Let H_t denote the context available to the CAT agent at step t . At $t = 0$, $H_0 = \{x, \mathcal{B}\}$. If the verification fails, the context for the next step H_{t+1} is augmented with the error information:

$$H_{t+1} = H_t \cup \{y_t, f_t, I_{\text{fix}}\} \quad (5)$$

where I_{fix} is a system instruction prompting the model to correct the specific errors identified in f_t . CAT agent then generates a revised translation y_{t+1} by maximizing the probability conditioned on this updated history:

$$y_{t+1} = \underset{y}{\operatorname{argmax}} P(y|H_{t+1}) \quad (6)$$

This iterative loop continues until the feedback is PASS or a maximum retry limit T_{max} is reached. If the loop terminates at T_{max} without verification, we select the candidate y_t that incurred the fewest assertion errors as the fallback output. This mechanism effectively converts the discrete symbolic feedback from the code sandbox into semantic guidance for the neural model. It forces the translation to converge towards the constraints defined in $\mathcal{B}$.

4 Experiments

We conducted extensive experiments and compared them with a series of strong baselines.

4.1 Datasets

To comprehensively evaluate Style-Ruler across varying levels of stylistic complexity, we conducted experiments on two distinct datasets, one representing stylized and the other general domains. First, we employed the Chinese $\leftrightarrow$ English GuoFeng Webnovel dataset (Wang et al., 2024a)

to evaluate the framework’s ability to capture high-resource stylistic nuances. Aligned with the WMT 2024 Discourse-Level Literary Translation task standards, this dataset consists of web novel chapters in the “GuoFeng” genre, which is rich in semi-classical grammar, archaic terminology, and four-character idioms. This stylized corpus provided the necessary monolingual references for our Style-Reverse Analyst to mine explicit linguistic constraints and verify the agent’s adherence to a specific literary voice. In contrast, we utilized WMT24++ (Deutsch et al., 2025), a large-scale evaluation set that expands the language coverage of WMT24 for the general domain. Specifically, we evaluated the English $\leftrightarrow$ {Chinese, German, Russian, Finnish} directions. This served as our baseline for assessing standard translation quality and ensuring that style constraints did not degrade general linguistic proficiency.

4.2 Settings and Baselines

Settings. We employed DeepSeek-V3.2 (Liu et al., 2025) as the backbone LLM for all agents in our framework, owing to its strong multilingual capabilities and instruction-following performance. For the SRA and the SAA, we set the sampling temperature to 0.1 to ensure deterministic extraction and code generation. Conversely, for the CAT, we set the temperature to 0.7 to balance stylistic diversity with coherence. The maximum retry limit for the collaborative refinement loop (T_{max}) was set to 3. For baselines or modules utilizing In-Context Learning (ICL), the number of few-shot exemplars (k) was set to 3.

Baselines. We compared Style-Ruler against a diverse set of baselines covering prompting, self-reflection, and hybrid paradigms. These included Zero-shot and Few-shot (ICL) (Dong et al., 2024) prompting, which served as fundamental baselines to gauge the model’s innate capabilities with and without in-context exemplars, and Chain-of-Thought (CoT) (Wei et al., 2022), which encouraged step-by-step reasoning to plan the translation before execution. Furthermore, we evaluated against specialized frameworks such as TasTe (Wang et al., 2024b), a self-reflection method that improved translation quality through a two-stage inference process of generation and refinement; LLM-RBMT (Coleman et al., 2024), which integrated rule-based constraints to assist LLMs, particularly effective for enforcing strict linguistic rules; and RASTA (Havaldar et al., 2025), a

Methods	En $\Rightarrow$ Zh					Zh $\Rightarrow$ En				
	BLEU	COMET	CSR _{PAS}	CSR _{ADD}	CSR _{ALL}	BLEU	COMET	CSR _{PAS}	CSR _{ADD}	CSR _{ALL}
Zero-shot	27.62	88.59	95.6	66.8	20.2	12.95	80.94	80.8	22.1	5.1
ICL (Dong et al., 2024)	29.23	89.20	95.8	68.5	24.5	13.42	81.52	83.2	28.4	8.6
CoT (Wei et al., 2022)	26.24	89.23	96.0	70.2	28.1	13.83	81.91	85.5	35.6	12.4
TasTe (Wang et al., 2024b)	28.13	89.10	96.2	72.5	35.4	14.51	82.42	88.0	45.2	20.5
LLM-RBMT (Coleman et al., 2024)	29.32	89.32	96.5	75.2	55.8	14.17	82.15	92.5	85.4	45.3
RASTA (Havaladar et al., 2025)	29.78	89.40	96.1	73.8	40.2	15.25	83.14	90.1	60.5	32.8
Style-Ruler	30.78	89.48	96.8	77.8	82.3	15.72	83.74	96.1	98.7	82.4

Table 1: Results on the GuoFeng Webnovel dataset for both En $\Rightarrow$ Zh and Zh $\Rightarrow$ En directions. We report BLEU and COMET for translation quality, and Constraint Satisfaction Rate (CSR) for style adherence. Specifically, CSR_{PAS}, CSR_{ADD}, and CSR_{ALL} denote the satisfaction rates for passive voice ratio, average dependency distance (syntactic complexity) and the overall constraint set, respectively.

Methods	En $\Rightarrow$ Zh			En $\Rightarrow$ Ru			En $\Rightarrow$ De			En $\Rightarrow$ Fi		
	BLEU	COMET	CSR	BLEU	COMET	CSR	BLEU	COMET	CSR	BLEU	COMET	CSR
Zero-shot	43.95	84.99	21.20	24.61	89.22	0.00	30.54	85.88	0.00	23.41	91.42	0.00
ICL	44.53	85.12	22.48	24.93	90.38	1.45	31.22	86.45	2.53	24.13	91.55	1.18
CoT	44.21	85.05	24.12	24.81	89.25	2.13	30.87	86.10	3.84	23.86	91.48	2.47
TasTe	44.87	85.18	25.49	25.12	89.31	3.52	31.43	86.62	5.21	24.52	91.60	3.83
LLM-RBMT	43.82	84.92	28.53	24.51	90.15	8.18	29.54	85.51	15.52	23.21	90.52	10.49
RASTA	45.12	85.20	26.04	25.26	90.32	4.53	31.85	86.80	6.53	25.12	91.75	4.22
Style-Ruler	45.52	85.24	32.45	25.52	89.35	10.50	22.72	84.10	21.80	26.21	90.85	13.56

Table 2: Performance on English-to-X translation directions from WMT24++.

retrieval-augmented approach that aligned translation style by retrieving and referencing stylistic exemplars from a datastore. The prompt formulations of baselines are as in Appendix A.5.

Evaluation Metrics. We evaluated Style-Ruler using both semantic and constraint-centric metrics. For translation quality, we reported BLEU (Post, 2018) and COMET (Rei et al., 2020) to assess n-gram overlap and semantic consistency, respectively. To quantify style adherence, we proposed the CSR, a metric grounded in our Sandbox-Augmented Auditor’s verification logic. CSR deterministically calculated the percentage of samples satisfying specific constraints by executing python scripts: we utilized the spaCy² library for dependency parsing to verify syntactic requirements (e.g., detecting auxpass labels for Passive Voice CSR, CSR_{PAS}). We calculated Average Dependency Distance for syntactic complexity (CSR_{ADD}), while using string operations for lexical constraints. We reported these fine-grained metrics alongside the strict Overall CSR (CSR_{ALL}), which requires simultaneous satisfaction of all constraints, on the stylized GuoFeng dataset. At the same time, we focused on CSR_{ALL} for the general WMT24++ benchmark.

²<https://github.com/explosion/spaCy>

4.3 Main Results

Table 1 reports the results on the GuoFeng Webnovel dataset. This task presents a dual challenge: maintaining the semantic integrity of the source narrative while strictly adhering to the specific literary style defined in the references. Style-Ruler achieves the best performance across all metrics for both En $\rightarrow$ Zh and Zh $\rightarrow$ En directions. In terms of translation quality, it surpasses the strongest baseline (RASTA) by +1.0 BLEU on En $\rightarrow$ Zh and +0.47 BLEU on Zh $\rightarrow$ En. The improvement in COMET scores confirms that our method preserves semantic meaning more effectively than standard in-context learning approaches. A critical observation lies in the CSR. Standard prompting methods (Zero-shot, CoT) and self-reflection methods (TasTe) struggle to satisfy the complex stylistic constraints, with CSR scores generally remaining low (e.g., CoT achieves only 28.1% CSR_{ALL} on Zh $\rightarrow$ En). This indicates that, although LLMs can understand instructions, they often fail to simultaneously enforce fine-grained lexical and syntactic constraints during a single inference pass. In contrast, Style-Ruler improves the CSR_{ALL} to 82.4%. This substantial margin validates the effectiveness of the SAA. A rigorous human evaluation (Appendix A.6) further corroborates this re-

Methods	Zh→En			Ru→En			De→En			Fi→En		
	BLEU	COMET	CSR	BLEU	COMET	CSR	BLEU	COMET	CSR	BLEU	COMET	CSR
Zero-shot	25.83	84.02	10.00	23.61	84.35	12.20	34.12	83.12	7.58	36.10	88.20	10.00
ICL	26.05	84.52	11.48	24.53	84.85	15.52	34.82	84.51	10.23	36.52	88.45	12.48
CoT	25.96	84.45	12.83	24.22	84.71	20.14	34.53	84.22	15.62	36.26	88.35	15.83
TasTe	26.21	84.81	13.52	25.13	85.12	25.43	35.24	85.81	22.54	36.81	88.60	20.22
LLM-RBMT	25.72	84.21	14.83	24.02	84.52	55.18	33.52	84.02	65.42	35.52	87.51	52.43
RASTA	26.36	84.95	13.22	25.82	85.31	30.53	35.50	86.20	28.53	37.10	88.72	25.62
Style-Ruler	26.40	85.02	15.20	26.68	85.52	73.10	34.42	87.11	82.80	33.88	88.82	65.24

Table 3: Performance on X-to-English translation directions from WMT24++.

sult: Style-Ruler achieves a human Style Appropriateness score of 4.45 and a strong Spearman correlation ($\rho = 0.72$, $p < 0.01$) between the automated CSR and human stylistic judgments.

Tables 2 and 3 present the performance on the WMT24++ benchmark. Unlike the stylized dataset, this setting evaluates the model’s ability to handle general domains under extracted stylistic guidance from reference sets. In the English-to-X task, Style-Ruler consistently outperforms baselines, including retrieval-augmented RASTA and fine-tuning-based LLM-RBMT. For instance, Style-Ruler achieves a BLEU score of 26.21, significantly higher than the Zero-shot baseline (23.41) and RASTA (25.12) in En→Fi. This suggests that the constraints extracted by the SRA do not introduce noise or degrade the general translation capability. Instead, they serve as effective guide rails that steer the model toward more native-like expressions in the reference set. Results in the X-to-English directions show a similar trend. Our method maintains high performance even when translating between linguistically distant languages such as Chinese and Russian. While some baselines, such as TasTe, achieve competitive BLEU scores on specific pairs (e.g., Ru→En), they lag significantly in CSR. It implies that their translations may not accurately reflect the target domain’s specific linguistic features. Style-Ruler effectively balances this trade-off, delivering translations that are both semantically accurate and stylistically compliant.

We also observe a pronounced target-language asymmetry in CSR across Tables 2 and 3: constraint satisfaction rates are significantly higher when translating *into* high-resource languages (e.g., X→En, En↔Zh) than into morphologically complex targets such as Finnish or Russian (En→Fi/Ru). This discrepancy stems from two compounding factors: (1) the backbone LLM’s

Methods	BLEU	COMET	CSR _{ALL}
Style-Ruler	15.72	83.74	82.4
w/o SRA	14.05	81.20	58.5
w/o Style-CoT	14.88	82.55	55.6
w/ LLM-Critic	15.10	82.90	63.8

Table 4: Ablation results on the GuoFeng dataset.

diminished instruction-following capacity when generating in lower-resource target languages, causing it to prioritize basic semantic fidelity over structural stylistic constraints; and (2) a typological bias in our verification tooling, as the underlying spaCy parsers are less robust at verifying complex syntactic dependencies in agglutinative or morphologically rich languages, resulting in higher rates of sandbox verification failures.

Furthermore, to show the generalizability of our framework beyond the archaic literary genre, we conducted extensive evaluations in Appendix A.9. An analysis of inference efficiency and computational cost is provided in Appendix A.10, and robustness to reference sparsity is evaluated in Appendix A.8.

4.4 Ablation Study

To validate the effectiveness of Style-Ruler, we conducted an ablation study on the GuoFeng (Zh→En). We compared the complete Style-Ruler framework with three distinct variants to isolate the contributions of SRA, CAT, and SAA as follows: (i) **w/o SRA (Generic Prompt)**: We replaced the fine-grained Style Brief extracted by SRA with a generic stylistic instruction. (ii) **w/o Style-CoT (Direct Gen)**: We removed the intermediate planning step from the CAT agent. (iii) **w/ LLM-Critic (Homogeneous)**: We replaced the SAA with a standard LLM-based evaluator.

Table 4 presents the results. We observe that removing the SRA (w/o SRA) results in the most sig-

nificant decline in both translation quality (-1.67 BLEU) and constraint satisfaction (CSR_{ALL} drops to 58.5). Without the explicit Style Brief mined from reference texts, the model fails to capture specific linguistic nuances, such as passive-voice ratios and syntactic complexity. This confirms that unsupervised rule mining is foundational for defining the target style. Then, we can find that “w/o Style-CoT” leads to a substantial decrease in CSR_{ALL} (55.6 vs. 82.4). While the BLEU score remains relatively competitive, the model struggles to satisfy multiple constraints simultaneously. The planning agent effectively decomposes the complex generation task and makes the model allocate attention to specific constraints before generation. Finally, “w/ LLM-Critic” yields a CSR_{ALL} of 63.8, which is notably lower than the complete model’s 82.4. While the LLM-Critic improves upon the direct generation baseline, it suffers from homogeneous failure modes. In contrast, the SAA provides rigorous, irrefutable error signals that compel the translator to converge on the correct style. This demonstrates that heterogeneous feedback is critical for high-precision constraint satisfaction.

4.5 Fine-grained Style Controllability Analysis

A distinguishing feature of Style-Ruler is its ability to operationalize abstract style into quantifiable constraints. To assess the precision of this control mechanism, we examined the correlation between the target linguistic metrics specified in the Style Brief and the actual metrics observed in the generated translations. We designed a controlled experiment in which the passive voice ratio was the target variable, ranging from 0.0 (fully active) to 1.0 (entirely passive) in increments of 0.1. We compared Style-Ruler against a strong prompting baseline (DeepSeek-V3.2 with explicit instructions, e.g., “Use 30% passive voice”). We computed the Pearson Correlation Coefficient (ρ) to quantify the linearity of the control.

Figure 2 illustrates the results. The baseline model exhibits a weak correlation ($\rho = 0.18$), with generated values clustering around $0.15 \sim 0.25$. This suggests that LLMs exhibit a neutrality collapse, defaulting to the statistical mean of their pretraining data regardless of numerical instructions. The model fails to map abstract percentage constraints to token-level generation probabilities. In contrast, Style-Ruler demonstrates a high de-

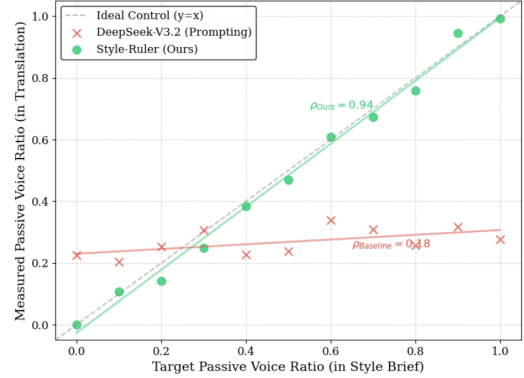


Figure 2: Controllability analysis of the Passive Voice Ratio.

Methods	En→Fi (Legal)		En→De (OOD)	
	BLEU	CSR	COMET	CSR
Zero-shot	34.2	15.6%	85.1	8.5%
CoT	34.8	31.2%	84.9	33.1%
Style-Ruler	35.5	94.2%	83.6	95.4%

Table 5: Generalization performance on WMT24 across different languages and OOD constraints DeepSeek-V3.2.

gree of controllability ($\rho = 0.94$). As the target ratio increases, the measured ratio in Style-Rulers output follows a near-linear trajectory along the $y = x$ diagonal ($\rho = 0.94$). This high degree of controllability emerges from a synergistic dynamic between the CAT’s planning and the SAA’s verification. While the intermediate Style-CoT planning effectively guides the model to satisfy constraints on the first attempt for a majority of straightforward cases, LLMs still inevitably suffer from “neutrality collapse” on linguistically complex sentences, resisting the target distribution.

4.6 Cross-Lingual & Out-of-Distribution Generalization

A unique challenge in agentic translation workflows is whether stylistic rules extracted in one language can be universally applied to translating into another. In this section, we reframe the traditional concept of cross-lingual transfer into an inference-time cross-Lingual zero-Shot generalization setting. Our motivation is highly practical: high-quality stylistic reference texts (e.g., formal legal templates) are often abundant in high-resource languages like English but scarce in others. We test whether the SRA can extract abstract stylistic constraints from an *English* reference set, and whether the CAT and SAA can subsequently

enforce these exact rules when translating into a morphologically different target language (e.g., *Finnish*). Because our framework relies on Python logic and multilingual dependency parsers (spaCy) rather than language-specific neural weights, the extracted constraints act as an interlingua. Table 5 demonstrates that Style-Ruler effectively enforces English-derived legal constraints (high passive ratio, formality) onto Finnish outputs, significantly outperforming CoT baselines. This confirms that our neuro-symbolic approach harnesses the inherent multilingual alignment of the backbone LLM, successfully generalizing explicit stylistic logic across language barriers without requiring any cross-lingual fine-tuning.

5 Conclusion

We introduced Style-Ruler, an inference-only framework that addresses LLMs’ limitations in adhering to complex stylistic and lexical constraints. By decomposing the translation process into unsupervised rule extraction, explicit planning, and code-based verification, our approach effectively decouples style definition from content generation. The integration of a heterogeneous feedback mechanism ensures rigorous compliance with hard constraints, where purely probabilistic models often fail. Empirical results on the GuoFeng Webnovel and WMT24++ benchmarks confirm that Style-Ruler significantly outperforms strong baselines in both constraint satisfaction and style alignment. Style-Ruler bridges the gap between probabilistic neural generation and deterministic constraints.

Limitations

The iterative nature of the agentic workflow inevitably increases inference latency and computational cost compared to single-pass prompting. While Style-Ruler effectively ensures quality for high-stakes offline tasks (e.g., literary or legal translation), it may not be suitable for real-time applications that require millisecond-level response times. Our current framework focuses on quantifiable stylistic features (e.g., lexical density, syntactic structure). Highly abstract or subjective stylistic elements, such as sarcasm, humor, or specific cultural metaphors, remain challenging to verify deterministically through code execution and may still benefit from human-in-the-loop evaluation. Furthermore, while our quantifiable metrics effectively steer stylistic generation, the cur-

rent SRA module exhibits an Indo-European typological bias inherited from the underlying dependency parsers (e.g., spaCy). For instance, computing tense ratios is not meaningfully applicable to isolating languages like Chinese, which lack inflectional tense morphology. Similarly, passive voice ratios do not seamlessly translate to languages with ergative-absolutive alignments. Future iterations of Style-Ruler must incorporate typologically-aware or language-specific feature extractors to ensure true cross-lingual universality.

Acknowledgments

The present research was supported by the National Key Research and Development Program (Grant No. 2023YFE0116400), National Natural Science Foundation of China Youth Fund (Grant No. 62306210). We would like to thank the anonymous reviewers for their insightful comments.

References

- Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873.
- Duarte Alves, Nuno Guerreiro, João Alves, José Pomal, Ricardo Rei, José de Souza, Pierre Colombo, and André FT Martins. 2023. Steering large language models for machine translation with finetuning and in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11127–11148.
- Anasua Banerjee and Dr Debajyoty Banik. 2025. A comprehensive survey on transformer-based machine translation: Identifying research gaps and solutions for large language models. *ACM Computing Surveys*.
- Kenza Benkirane, Laura Gongas, Shahar Pelles, Naomi Fuchs, Joshua Darmon, Pontus Stenetorp, David Adelani, and Eduardo Sánchez. 2024. Machine translation hallucination detection for low and high resource languages using large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9647–9665.
- Douglas Biber. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press.
- Vicent Briva-Iglesias. 2025. Are ai agents the new machine translation frontier? challenges and opportunities of single-and multi-agent systems for multilingual digital communication. *arXiv preprint arXiv:2504.12891*.

- Jared Coleman, Bhaskar Krishnamachari, Ruben Rosales, and Khalil Iskarous. 2024. Llm-assisted rule based machine translation for low/no-resource languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 67–87.
- Menglong Cui, Jiangcun Du, Shaolin Zhu, and Deyi Xiong. 2024. Efficiently exploring large language models for document-level machine translation with in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10885–10897.
- 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.
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, and 1 others. 2024. A survey on in-context learning. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 1107–1128.
- Tianyu Dong, Bo Li, Jinsong Liu, Shaolin Zhu, and Deyi Xiong. 2025. Mlas-lora: Language-aware parameters detection and lora-based knowledge transfer for multilingual machine translation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15645–15660.
- Zhaopeng Feng, Jiayuan Su, Jiamei Zheng, Jiahao Ren, Yan Zhang, Jian Wu, Hongwei Wang, and Zuozhu Liu. 2025. M-mad: Multidimensional multi-agent debate for advanced machine translation evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7084–7107.
- Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.
- Evangelia Gogoulou, Shorouq Zahra, Liane Guillou, Luise Dürlich, and Joakim Nivre. 2025. Can llms detect intrinsic hallucinations in paraphrasing and machine translation? *arXiv preprint arXiv:2504.20699*.
- Nuno M Guerreiro, Duarte M Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André FT Martins. 2023. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.
- Shreya Havaldar, Adam Stein, Eric Wong, and Lyle Ungar. 2025. Towards style alignment in cross-cultural translation. *arXiv preprint arXiv:2507.00216*.
- Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujia Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.
- Justine Kao and Dan Jurafsky. 2012. A computational analysis of style, affect, and imagery in contemporary poetry. In *Proceedings of the NAACL-HLT 2012 workshop on computational linguistics for literature*, pages 8–17.
- Sejoon Kim, Mingi Sung, Jeonghwan Lee, Hyunkuk Lim, and Jorge Gimenez Perez. 2024. Efficient terminology integration for llm-based translation in specialized domains. In *Proceedings of the Ninth Conference on Machine Translation*, pages 636–642.
- Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 923–929.
- Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.
- Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191.
- 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.
- Yasmin Moslem, Rejwanul Haque, John Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237.
- Ingo Plag, Christiane Dalton-Puffer, and Harald Baayen. 1999. Morphological productivity across speech and writing. *English Language & Linguistics*, 3(2):209–228.
- Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

- Nooshin Pourkamali and Shler Ebrahim Sharifi. 2024. Machine translation with large language models: Prompt engineering for persian, english, and russian directions. *arXiv preprint arXiv:2401.08429*.
- Shenbin Qian, Archchana Sindhuja, Minnie Kabra, Diptesh Kanojia, Constantin Orasan, Tharindu Ranasinghe, and Fred Blain. 2024. What do large language models need for machine translation evaluation? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3660–3674.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.
- Hong Ren, Liting Deng, Shaolin Zhu, and Deyi Xiong. 2026. Adadpi: Document-level translation adaptive agent via dynamic parametric internalization. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 37912–37929.
- Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Dániel Varga. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. *arXiv preprint cs/0609058*.
- Zhen Tao, Dinghao Xi, Zhiyu Li, Liumin Tang, and Wei Xu. Cat-llm: Style-enhanced large language models with text style definition for chinese article-style transfer. *ACM Transactions on Knowledge Discovery from Data*.
- Zhen Tao, Dinghao Xi, Zhiyu Li, Liumin Tang, and Wei Xu. 2025. Cat-llm: Style-enhanced large language models with text style definition for chinese article-style transfer. *ACM Transactions on Knowledge Discovery from Data*, 19(7):1–33.
- Longyue Wang, Siyou Liu, Chenyang Lyu, Wenxiang Jiao, Xing Wang, Jiahao Xu, Zhaopeng Tu, Yan Gu, Weiyu Chen, Minghao Wu, and 1 others. 2024a. Findings of the wmt 2024 shared task on discourse-level literary translation. *arXiv preprint arXiv:2412.11732*.
- Yifan Wang, Zewei Sun, Shanbo Cheng, Weiguo Zheng, and Mingxuan Wang. 2023. Controlling styles in neural machine translation with activation prompt. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2606–2620.
- Yutong Wang, Jiali Zeng, Xuebo Liu, Fandong Meng, Jie Zhou, and Min Zhang. 2024b. Taste: Teaching large language models to translate through self-reflection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6144–6158.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.
- Minghao Wu, Jiahao Xu, and Longyue Wang. 2024. Transagents: Build your translation company with language agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 131–141.
- Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with qlora. In *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481.
- Jaawei Zheng, Hanghai Hong, Feiyan Liu, Xiaoli Wang, Jingsong Su, Yonggui Liang, and Shikai Wu. 2024. Fine-tuning large language models for domain-specific machine translation. *arXiv preprint arXiv:2402.15061*.
- Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.
- Shaolin Zhu, Menglong Cui, and Deyi Xiong. 2024a. Towards robust in-context learning for machine translation with large language models. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)*, pages 16619–16629.
- Shaolin Zhu, Leiyu Pan, Bo Li, and Deyi Xiong. 2024b. Landermt: Detecting and routing language-aware neurons for selectively finetuning llms to machine translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12135–12148.
- Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024c. 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 Appendix

A.1 Implementation Details of Style-Reverse Analyst

Rather than relying on latent embeddings, the SRA operationalizes the stylistic skeleton into quantifiable linguistic constraints rooted in computational stylistics. Specifically, we extract three categories of macroscopic features: 1) **Lexical Metrics**: Lexical density and domain-specific terminology ratios, which dictate the formal register and information packaging. 2) **Syntactic Metrics**: Average Dependency Distance (ADD) and passive voice ratios. These govern the structural complexity and rhetorical stance (e.g., objective distancing in legal texts). 3) **Discourse Metrics**: Sentence length variance and tense distributions, which control the narrative rhythm.

We rely on this explicit rule mining because prior literature and our preliminary observations indicate an inherent limitation in general-purpose LLMs: they suffer from “instruction neglect” and “neutrality collapse” when prompted with complex or abstract constraints during a single inference pass, frequently reverting to the statistical mean of their pretraining corpora (He et al., 2024).

A.1.1 Linguistic Metrics Definition

To implement the feature extraction function Φ , we use the spaCy library for dependency parsing and Part-of-Speech tagging. The specific metrics for the three modules are defined as follows:

1. Lexical Metrics (Φ_{lex})

- **Lexical Density**: Calculated as $N_{\text{content}}/N_{\text{total}}$, where N_{content} is the count of nouns, verbs, adjectives, and adverbs, and N_{total} is the total token count. High density typically indicates an information-heavy or academic style.
- **Connector Frequency**: The frequency of explicit discourse markers normalized by sentence length. This is computed via a hybrid method: tokens are matched against a language-specific, manually-curated list of transitional markers (e.g., “furthermore”, “however”) and their structural role is verified using spaCy dependency tags (e.g., ensuring they are parsed as advmod or mark modifying a clausal root).

2. Syntactic Metrics (Φ_{syn})

- **Average Dependency Distance (ADD)**: This metric measures syntactic complexity. For a sentence of length n , it is defined as:

$$\text{ADD} = \frac{1}{n-1} \sum_{i=1}^n |i - h(i)| \quad (7)$$

where i is the position of a word and $h(i)$ is the position of its syntactic head. Dividing by $n-1$ normalizes the sum by sentence length, preventing longer sentences from spuriously inflating the score.

- **Passive Ratio**: The number of sentences containing nsubjpass or auxpass dependency labels divided by the total number of sentences.

3. Morphological & Discourse Metrics (Φ_{mor})

- **Tense Distribution**: We calculate the ratio of past tense verbs (POS tags VBD, VBN) to present tense verbs (VBP, VBZ).
- **Sentence Length Variance**: We compute the normalized variance of sentence lengths across the reference set $\mathcal{R}$, defined as the raw variance divided by the squared mean sentence length $\bar{L}^2$:

$$\text{NormVar}(L) = \frac{\text{Var}(L)}{\bar{L}^2} \quad (8)$$

This normalization prevents styles with intrinsically longer sentences from exhibiting spuriously higher variance. A high value indicates a style that mixes short and long sentences (e.g., literary), while a low value indicates uniformity (e.g., technical manuals).

A.1.2 SRA Prompt Template

Table 6 shows the exact prompt used to convert the calculated statistics into the JSON Style Brief.

A.2 Implementation Details of Constraint-Adaptive Translator

A.2.1 Style-CoT Prompt Design

The CAT agent uses a structured prompt to enforce the generation of the planning step (z) before the translation (y). Table 7 presents the prompt template.

System Instruction for SRA

You are a linguistic analysis tool. Your goal is to convert statistical features into a structured style definition.

Input Data:

- **Reference Text Snippets:** [Insert 3-5 examples from $\mathcal{R}$]
- **Computed Stats:**
 - Passive Voice Ratio: {val} (High if > 0.3)
 - Lexical Density: {val}
 - Sentence Length Std Dev: {val}

Task:

Generate a JSON object describing this style. The tone field should summarize the register. The constraints list should specify hard rules derived from the stats.

Output Schema:

```
{
  "tone": "...",
  "mandatory_constraints": [...],
  "negative_constraints": [...]"}
```

Table 6: The prompt template used by the Style-Reverse Analyst.

A.2.2 Inference Configuration

For the CAT module, we use a temperature setting of 0.7 to allow for stylistic flexibility while maintaining coherence. The maximum token limit for the plan z is set to 100 tokens to encourage concise reasoning. We parse the final output using regular expressions to separate the plan from the final translation.

A.3 Implementation Details of Sandbox-Augmented Auditor

A.3.1 Verification Logic Types

The SAA bridges natural-language constraints and symbolic verification via specific logical patterns. We classify the verification scripts into three categories based on the tool required:

1. **String Operations (Lexical Check):** For terminology constraints, the SAA generates code using basic string methods or regular expressions.
 - *Example Constraint:* “Must include semiconductor”.
 - *Code:* `assert “semiconductor” in translation, “Missing term: semiconductor”`
2. **Arithmetic Logic (Length Check):** Neural models often fail at precise counting. The SAA delegates this to the Python interpreter.

System Instruction for CAT

You are a professional translator. You must translate the source text into the target language while strictly adhering to the provided “Style Brief”.

Input Data:

- **Source Text (x):** [Insert Source Sentence]
- **Style Brief ($\mathcal{B}$):** [Insert JSON from SRA]

Instruction:

Perform the task in two steps:

1. **Plan (z):** Analyze the source text. List specific changes needed to satisfy the Style Brief (e.g., “I must use the term X for Y”, “I must restructure the sentence to be passive”).
2. **Translate (y):** Generate the final translation based on the plan.

Output Format:

[Plan]: ...

[Translation]: ...

Table 7: The prompt template for the Constraint-Adaptive Translator using Style-CoT.

- *Example Constraint:* “Sentence length under 20 words.”
- *Code:* `assert len(translation.split()) < 20, “Text too long”`

3. **Linguistic Tooling (Syntactic Check):** For abstract constraints like “passive voice”, the SAA generates code that calls the spaCy library within the sandbox to parse the dependency tree of the generated text, ensuring the linguistic structure matches the requirement.

A.3.2 SAA System Prompt

Table 8 shows the instructions given to the SAA agent. Note that the agent is explicitly instructed to write assert statements that trigger the heterogeneous feedback mechanism when executed.

A critical vulnerability in code-augmented agents is the generation of syntactically invalid or logically flawed scripts. To ensure systemic robustness, the execution function $E(c)$ encapsulates the code within a strict try-except block. If the script crashes due to a runtime error (e.g., invalid syntax or API failures) rather than an assertion failure, the framework captures the stack trace and prompts the SAA to self-correct its script. If the SAA fails to produce executable code after two attempts, Style-Ruler gracefully degrades: it bypasses the deterministic verification for that specific turn and falls back to the latest generated draft from the CAT. This fail-safe mechanism ensures that translation pipelines are never halted by anomalous script generation.

System Instruction for SAA

You are a QA Engineer. Your task is to write a Python script to verify if a translation strictly meets the requirements in the Style Brief.

Input Data:

1. **Translation** (y): [Generated Text]
2. **Style Brief** (β): [JSON Object]

Task:

Write a Python function `verify(text)` that executes logical checks.

- Use `assert` statements for each constraint.
- If a constraint fails, the assertion message must clearly state what was expected vs. what was found.
- For linguistic checks (e.g., passive voice), import and use `spacy`.

Output Format:

```
def verify(text):  
    # Check 1: Terminology  
    assert "term" in text, "Missing term..."  
    # Check 2: Length  
    ...
```

Table 8: The prompt template for the SAA agent. The generated code produces deterministic feedback when executed.

A.4 Scope of Stylized Translation and Theoretical Justification

In response to the multifaceted nature of “style” in natural language, we explicitly delineate the boundaries of stylized translation addressed in Style-Ruler. The concept of style encompasses a broad spectrum, ranging from abstract literary devices (such as authorial voice, sarcasm, and cultural metaphors) to structural linguistic patterns (such as formality, register, and syntactic complexity).

The SRA and SAA modules in our framework are designed to operate within the paradigm of *computational stylistics* and *register variation* (Plag et al., 1999). We do not claim to deterministically map or verify highly abstract literary semantics. Instead, we hypothesize that the structural “skeleton” of a style can be rigorously modeled and constrained. This scoping aligns with established sociolinguistic frameworks, most notably Biber’s Multidimensional Analysis (MDA), which posits that text genres and styles can be robustly differentiated using co-occurring, quantifiable linguistic features (Biber, 1995). To address whether these statistical indicators align with human judgments of stylized translation, we anchor our metric selection in linguistic literature. The specific constraints extracted by SRA serve as established proxies for human stylistic perception:

Lexical Density and Sentence Length Variance.

Lexical density is a primary indicator of text formality and information packaging. High lexical density is cognitively associated with academic, legal, or formal written styles, whereas lower density characterizes conversational or colloquial text (Plag et al., 1999). Similarly, sentence length variance dictates the rhythmic pacing of a narrative. Enforcing these metrics does not merely alter token counts; it fundamentally shifts the register of the generated text, which is the most immediate stylistic cue perceived by human readers (Kao and Jurafsky, 2012).

Syntactic Complexity (Dependency Distance).

Average Dependency Distance (ADD) is deeply tied to cognitive load and language processing (Futrell et al., 2015; Liu, 2008). Specialized domains (e.g., archaic literature in the GuoFeng dataset or bureaucratic documents) frequently employ complex, long-distance dependencies that are atypical in general-domain pretraining corpora. By enforcing ADD constraints, SAA compels the neural model to adopt the syntactic scaffolding inherent to the target domain, effectively preventing the LLM from reverting to its default, flattened syntactic distribution.

Passive Voice and Discourse Markers. The distribution of voice and explicit connectors dictates the rhetorical stance of a text. For instance, scientific and legal translations strictly mandate high passive ratios to maintain objective distancing (Plag et al., 1999). Correcting these ratios via SAA’s deterministic feedback ensures that the translation respects the stylistic conventions of the target discipline.

In summary, while our quantifiable metrics do not capture the entirety of literary style, they govern the structural and syntactic dimensions that form the prerequisite for any stylized translation. Successfully enforcing these constraints steers the LLM into the correct stylistic subspace, thereby significantly elevating the human perception of domain adherence.

A.5 Detail Baselines and Prompt Fairness

To ensure a strictly fair comparison and avoid evaluating baselines on unseen criteria, we standardize the prompt formulations across all zero-shot and CoT baselines. Specifically, we do not provide them with vague stylistic descriptions (e.g., “translate in a literary tone”); instead, we inject the ex-

act explicit constraints derived from our SRA module (e.g., specific passive voice ratios and strict vocabulary inclusions) directly into their system prompts. The failure of baseline models to satisfy these constraints stems from the inherent “instruction neglect” phenomenon during single-pass neural generation (Zhou et al., 2023), rather than an asymmetry in the provided information.

A.6 Human Evaluation and Correlation Analysis

A critical concern regarding our unsupervised feature extraction is whether quantifiable metrics (e.g., syntactic complexity, passive ratios) genuinely reflect human perception of a stylized text. To empirically validate the alignment between our deterministic Constraint CSR and human subjective judgments, we conducted a rigorous human evaluation on a stratified sample of 200 sentences from the GuoFeng (Zh→En) translation task.

Setup and Metrics. Three professional bilingual linguists, blinded to the model identities, evaluated the outputs. Following standard WMT protocols, they rated each translation on three dimensions using a 1-5 Likert scale: **Adequacy** (preservation of source meaning), **Fluency** (grammatical correctness), and **Style Appropriateness** (how well the translation captures the archaic, literary flavor of the reference). To measure inter-annotator agreement (IAA), we calculated Fleiss’ Kappa (κ). Finally, we computed the Spearman’s rank correlation (ρ) between the automated CSR_{ALL} and the human Style Appropriateness scores.

Table 9 details the evaluation outcomes. While baselines like CoT and RASTA perform adequately in meaning preservation, they struggle significantly with stylistic alignment (scoring 2.58 and 3.42, respectively). Style-Ruler achieves a superior Style score of 4.45 with the lowest variance (± 0.31), demonstrating consistent stylistic enforcement. Most importantly, the correlation analysis directly addresses the validity of our quantifiable metrics. We observed a strong, statistically significant positive correlation ($\rho = 0.72, p < 0.01$) between the automated CSR_{ALL} and human stylistic judgments for Style-Ruler. This provides concrete evidence that while lexical density and dependency distances act as structural constraints, successfully enforcing them effectively steers the LLM into the target stylistic subspace, resulting in

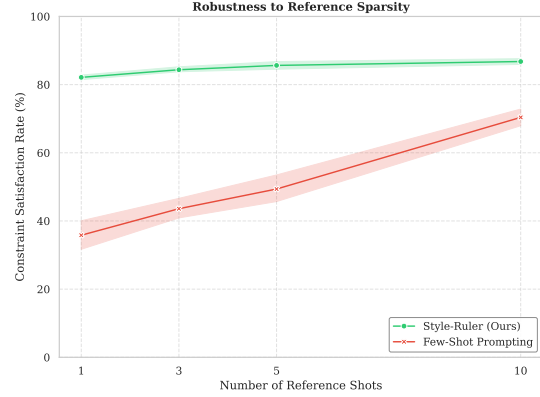


Figure 3: CSR under varying reference set sizes ($|\mathcal{R}|$). Style-Ruler maintains superior performance and stability even in the extreme 1-shot setting, whereas the prompting baseline exhibits high variance and poor adherence when data are scarce.

translations that human readers perceive as significantly more authentic.

A.7 Independent Style Consistency Evaluation

A potential critique of the CSR is that it intrinsically favors the Style-Ruler framework, which explicitly optimizes against these exact Python assertions. To corroborate the CSR findings and ensure evaluation fairness, we introduced two independent, reference-free style metrics that operate entirely outside our SRA/SAA loop. First, we deployed an independent GPT-5.4 evaluator (**GPT-5.4**) prompted solely with the source text and the monolingual reference domain (without exposing it to the discrete SRA constraints). It was tasked to blindly score the stylistic resemblance of the outputs on a 1-100 scale. Second, we utilized a pre-trained domain classifier (**BERT-Style-Cls**), fine-tuned on general vs. domain-specific corpora, to output the probability (P_{style}) that a translated sentence belongs to the target stylized distribution. As detailed in Table 10, Style-Ruler significantly outperforms standard CoT and RASTA across both independent metrics. The BERT-Style-Cls assigns a 0.88 probability to our outputs, compared to 0.45 for zero-shot prompts. This confirms that the stylistic advantages captured by the CSR metric translate robustly to broader, constraint-agnostic stylistic evaluations.

A.8 Robustness to Reference Sparsity

In real-world scenarios, obtaining a large corpus of reference texts for a specific niche style (e.g.,

Methods	Human Evaluation (1-5)			Correlation
	Adequacy	Fluency	Style	CSR vs. Style (ρ)
Zero-shot	3.82 ± 0.41	4.10 ± 0.35	2.15 ± 0.52	0.31
CoT	3.95 ± 0.38	4.25 ± 0.30	2.58 ± 0.61	0.45
RASTA	4.12 ± 0.33	4.30 ± 0.28	3.42 ± 0.45	0.58
Style-Ruler	4.28 ± 0.25	4.35 ± 0.22	4.45 ± 0.31	0.72
<i>Fleiss' κ</i>	<i>0.68</i>	<i>0.71</i>	<i>0.64</i>	-

Table 9: Human evaluation results on the GuoFeng dataset (200 samples). Style-Ruler not only achieves the highest human ratings across all dimensions but also exhibits a strong positive correlation ($\rho = 0.72$) between its deterministic CSR and human stylistic judgments.

Methods	GPT-5.4 $\uparrow$	BERT-Style-Cls (P_{style}) $\uparrow$
Zero-shot	45.2	0.45
CoT	52.8	0.51
RASTA	68.4	0.63
Style-Ruler	84.5	0.88

Table 10: Independent style consistency evaluation. Both metrics are completely agnostic to the explicit constraints used in our CSR metric, confirming the generalizable stylistic superiority of Style-Ruler.

a particular author’s idiolect) is often impractical. Therefore, data efficiency is a critical metric. We evaluate the robustness of Style-Ruler by varying the size of the reference set $|\mathcal{R}|$ (number of shots) from 1 to 10 on the GuoFeng dataset. We compare our method against standard Few-Shot Prompting, in which the reference texts are provided directly within the context window as demonstrations. Figure 3 illustrates the performance trends in terms of CSR. We observe two distinct behaviors:

- **High Efficiency of Style-Ruler:** Our method achieves a high CSR ($> 80\%$) even in the 1-shot setting. This stability stems from the SRA. Even with a single reference document, the SRA can successfully extract abstract linguistic statistics (e.g., "sentence length variance = high") and formulate them into explicit constraints. The subsequent neuro-symbolic verification ensures that these constraints are enforced, thereby reducing the system’s dependence on the volume of reference data.
- **Instability of Few-Shot Prompting:** The baseline exhibits high variance and poor performance in low-resource settings ($|\mathcal{R}| <$

3). Standard LLMs struggle to infer complex stylistic patterns solely from one or two examples, often defaulting to their pre-trained generic style. While the baseline’s performance improves as $|\mathcal{R}|$ increases, it eventually plateaus below Style-Ruler, limited by the context window and by the model’s inherent difficulty in attending to long-tail stylistic features.

These results confirm that Style-Ruler effectively decouples the quality of style transfer from the quantity of reference data, offering a robust solution for low-resource style adaptation.

A.9 Generalization to Diverse Stylistic Typologies

A fundamental prerequisite for a controllable generation framework is its domain agnosticism. To empirically validate that the performance gains of Style-Ruler are not confined to the specific linguistic features of the GuoFeng Webnovel dataset (i.e., archaic and literary tones), we expanded our evaluation paradigm to include two diametrically opposed stylistic extremes. We selected two new translation settings that impose orthogonal linguistic constraints:

- **Legal/Bureaucratic (En $\rightarrow$ De):** Sourced from the JRC-Acquis corpus (Steinberger et al., 2006). This domain exemplifies maximum structural rigidity. Unsupervised extraction via our SRA module identified requirements for high passive voice ratios ($> 40\%$), extensive dependency distances (due to complex subordinate clauses), and strict adherence to formal terminologies.

- **Colloquial Subtitles (En→Fr):** Sourced from the OpenSubtitles dataset (Lison and Tiedemann, 2016). This domain represents informal spoken register. The SRA dynamically extracted rules enforcing brevity (strict token limits), high active voice ratios, and the preservation of specific discourse markers (e.g., colloquial contractions and morphological reductions).

Table 11 summarizes the translation quality (BLEU, COMET) alongside the specific Constraint Satisfaction Rates (CSR) across these new domains. The empirical results reveal a consistent trajectory. Baseline paradigms, including Chain-of-Thought and retrieval-augmented systems (RASTA), exhibit severe performance degradation when asked to adhere to strict stylistic blueprints. They frequently default to the statistical mean of their pretraining distribution, outputting a “neutral” translation regardless of the prompt. In contrast, Style-Ruler doubles the CSR across both extremes (reaching 88.2% for Legal and 91.5% for Subtitles). This robust generalizability is directly attributable to the architectural decoupling within our framework. Because the SRA functions purely as an analytical probe and the SAA operates as a deterministic execution engine, the agents themselves hold no domain bias. They systematically operationalize whatever structural logic is present in the reference texts, confirming that Style-Ruler is a universal mechanism for stylistic alignment rather than a domain-specific heuristic.

A.10 Inference Efficiency and Cost

Agentic frameworks inherently introduce computational overhead compared to single-pass prompting. To verify whether Style-Ruler remains practical for real-world deployment, we analyzed the trade-off between translation quality (CSR) and inference cost (Latency and Token Usage). We measured the average wall-clock time per sentence on the GuoFeng dataset using the same hardware environment. Figure 4 visualizes the Quality-Latency landscape. As expected, the Zero-shot baseline is the fastest (1.2s/sentence) but yields the lowest compliance (21.2% CSR). The CoT baseline increases latency to 2.5s due to the generation of reasoning steps, yet only marginally improves compliance.

Style-Ruler operates with an average latency of

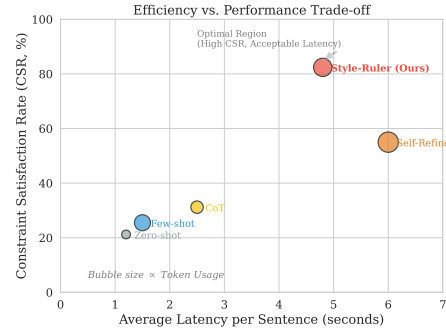


Figure 4: Efficiency-Performance Trade-off Analysis. The x-axis denotes inference latency, and the y-axis represents the CSR. Bubble sizes are proportional to the number of tokens consumed. Style-Ruler achieves a substantial leap in performance with a manageable increase in latency compared to iterative baselines like Self-Refine.

4.8s per sentence. While this is approximately $2\times$ slower than CoT, it achieves a decisive CSR of 82.4%. Further breakdown of the time budget reveals that Python code execution within the SAA accounts for less than 5% of the total latency, indicating that the overhead is primarily due to the iterative generation of the Translator agent. Crucially, our analysis of the **iteration distribution** shows that the system is highly convergent. On average, 68% of samples pass verification in the first round (requiring no refinement), and 92% converge within two rounds. The average number of inference calls is 1.4 per sentence. This demonstrates that Style-Ruler does not rely on brute-force resampling. Instead, the neuro-symbolic feedback efficiently guides the model to the target distribution, making the computational cost predictable and manageable for high-stakes offline applications where precision is paramount.

A.11 Impact of Backbone Models

To verify the universality of our framework beyond open-source models, we applied Style-Ruler to GPT-4o, one of the strongest proprietary LLMs available. We conducted experiments on the GuoFeng Webnovel dataset (En↔Zh) and compared our method against Zero-shot, Few-shot ICL (5-shot), and Supervised Fine-Tuning (LoRA). Figure 5 presents the performance profile across three dimensions: COMET, d-BLEU, and LTCR. The results indicate that Style-Ruler consistently encompasses the baselines, achieving the largest area on the radar chart. Specifically, while GPT-4o already possesses strong zero-shot capabili-

Methods	Legal (En→De)			Subtitles (En→Fr)		
	BLEU	COMET	CSR _{ALL}	BLEU	COMET	CSR _{ALL}
Zero-shot	28.40	81.20	14.5	30.12	82.50	18.2
CoT	29.20	82.05	28.3	30.80	82.95	31.4
RASTA	31.80	83.45	48.6	32.60	84.15	51.2
Style-Ruler	33.40	85.12	88.2	34.15	85.80	91.5

Table 11: Performance across structurally distinct stylistic domains. Style-Ruler consistently maintains high constraint adherence regardless of whether the target style requires extreme syntactic complexity (Legal) or informal brevity (Subtitles).

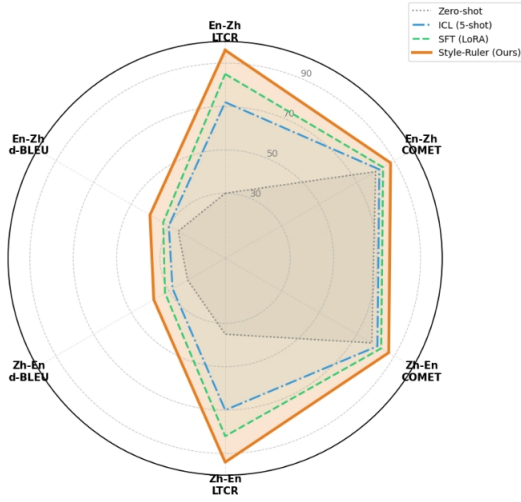


Figure 5: Performance comparison on the GuoFeng dataset using GPT-4o as the backbone. The radar chart illustrates metrics for translation quality (COMET, d-BLEU) and constraint adherence (LTCR).

ties, it still benefits significantly from our neuro-symbolic mechanism. On the En-Zh task, Style-Ruler improves the LTCR from approximately 75% (ICL) to over 95%. This confirms that even for state-of-the-art models, internal probabilistic reasoning is insufficient for strict constraint satisfaction. The externalized verification loop provided by the Sandbox-Augmented Auditor effectively complements the reasoning power of GPT-4o, unlocking superior performance in specialized domains without requiring parameter updates.