JOURNAL OF THE ARTIFICIAL INTELLIGENCE ENGINEERS CONSORTIUM

Volume: 02 Issue: 02 (2026)

https://jaiec.org.hostajournal.com

Alignment-Free Distillation for Non-Autoregressive Translation in Sparse Language Pairs

William D. Perez

Department of Software Engineering, Golisano College of Computing and Information Sciences, Rochester Institute of Technology, Rochester, New York, 14623, USA

Abstract **:** *The rapid advancement of neural machine translation has historically relied heavily on autoregressive models, which generate target sequences token by token. While autoregressive models achieve remarkable translation quality, their sequential decoding nature inherently limits inference speed. Non-autoregressive translation models have emerged as a promising alternative, enabling parallel token generation and significantly reducing decoding latency. However, non-autoregressive models often suffer from performance degradation, largely due to the multimodality problem where the model struggles to capture dependencies among target tokens. Sequence-level knowledge distillation is typically employed to alleviate this issue by training the non-autoregressive student model on simplified target sequences generated by an autoregressive teacher. Conventional knowledge distillation methods rely on strict word-level or structural alignments between the source and target representations. In sparse language pairs characterized by limited bilingual training data, acquiring robust alignments is profoundly difficult, leading to suboptimal distillation and poor translation quality. This paper introduces a comprehensive alignment-free distillation framework designed specifically for non-autoregressive translation in low-resource and sparse language pair settings. By leveraging optimal transport mechanisms and order-agnostic loss formulations, the proposed framework bypasses the need for explicit token-level alignment. Extensive empirical evaluations across multiple sparse translation benchmarks demonstrate that the alignment-free distillation approach not only matches but significantly surpasses traditional alignment-dependent distillation techniques in both translation fidelity and structural coherence.*

Keywords : *Machine Translation, Non-Autoregressive Models, Knowledge Distillation, Sparse Language Pairs, Optimal Transport.*

1 .INTRODUCTION

1.1 The Evolution of Neural Machine Translation

The domain of natural language processing has witnessed a profound transformation over the past decade, driven primarily by the advent and widespread adoption of deep learning architectures. Within this broader field, neural machine translation stands out as a quintessential application that has benefited immensely from these technological strides. Early approaches to machine translation relied on rule-based systems and subsequently transitioned to statistical machine translation, which utilized phrase-based

probabilistic models. These statistical paradigms, while revolutionary at the time, were fundamentally limited by their reliance on manually engineered features and complex pipelines that struggled to capture long-range linguistic dependencies. The introduction of sequence-to-sequence neural networks marked a paradigm shift, enabling end-to-end learning directly from large bilingual corpora. As researchers in the field began to explore attention mechanisms, neural machine translation systems developed the capability to dynamically align source and target representations during the decoding process. This innovation culminated in the transformer architecture, which completely dispensed with recurrent structures in favor of self-attention mechanisms [1]. The transformer has since become the defacto standard for neural machine translation, yielding state-of-the-art results across a multitude of high-resource language pairs [2]. These traditional transformer models function autoregressively, meaning they condition the generation of each subsequent token on the entire history of previously generated tokens. This autoregressive property ensures high fluency and contextual accuracy, as the model explicitly models the conditional probability of the target sequence.

1.2 Limitations of Autoregressive Decoding and the Non-Autoregressive Paradigm

Despite their unprecedented success in translation quality, autoregressive models possess a critical bottleneck related to computational efficiency during inference. The token-by-token generation process prevents parallelization across the temporal dimension of the target sequence. In applications demanding real-time translation or processing of massive document volumes, the latency introduced by autoregressive decoding becomes prohibitively expensive. To address this computational bottleneck, non-autoregressive translation models were conceptualized and developed. These models attempt to generate all target tokens simultaneously in a single forward pass, thereby achieving inference speeds that are an order of magnitude faster than their autoregressive counterparts [3]. However, this massive increase in decoding speed comes at a substantial cost to translation quality. Non-autoregressive models inherently assume conditional independence among the target tokens given the source sequence. This conditional independence assumption severely impairs the model capacity to handle the multimodality of translations, a phenomenon where multiple valid translations exist for a single source sentence [4]. When a non-autoregressive model attempts to generate a sequence without explicit target-side context, it often produces outputs containing token repetitions, omissions, and severe grammatical inconsistencies. To bridge the performance gap between autoregressive and non-autoregressive systems, researchers introduced various training techniques, the most prominent being sequence-level knowledge distillation. In this procedure, a highly accurate autoregressive teacher model translates the training data, and the non-autoregressive student model trains on these synthetic, simplified targets. The distilled data exhibits reduced multimodality and variance, making it significantly easier for the non-autoregressive model to learn the underlying mapping [5].

1.3 Challenges in Sparse Language Pairs

While sequence-level knowledge distillation effectively mitigates the multimodality problem in high-resource scenarios, its application to sparse language pairs presents a distinct set of formidable challenges. Sparse language pairs are characterized by an acute scarcity of high-quality bilingual parallel sentences. In such low-resource environments, training a robust autoregressive teacher model is inherently difficult, leading to synthetic translations that may contain noise or inaccuracies [6]. More critically, conventional distillation techniques and the training of non-autoregressive

architectures often rely on explicit or implicit alignment mechanisms. Some non-autoregressive models require external aligners to determine the exact length of the target sequence or to map source tokens to target tokens [7]. Other approaches utilize connectionist temporal classification or dynamic programming to marginalize over possible alignments. In sparse data settings, these alignment mechanisms are notoriously brittle. The lack of sufficient statistical evidence prevents the models from learning accurate alignment matrices, causing the distillation process to fail. The student model receives poor alignment signals, exacerbating issues like token dropping and hallucination. The reliance on strict positional matching in standard cross-entropy loss further penalizes the student model if it generates a correct word but places it in a slightly different position than the teacher output [8]. Consequently, the development of distillation methods that do not depend on explicit word-level alignment is of paramount importance for extending the benefits of non-autoregressive translation to the vast majority of the world languages that suffer from data sparsity.

1.4 Scope and Contributions of the Present Study

This research proposes a novel alignment-free distillation framework specifically engineered to enhance the performance of non-autoregressive translation models in sparse language pairs. By discarding the requirement for strict positional and structural alignment, the proposed methodology accommodates the inherent uncertainties present in low-resource distillation. The framework integrates an optimal transport-based sequence matching objective alongside an order-agnostic bag-of-words loss to transfer knowledge from the teacher to the student [9]. This allows the student model to learn the global semantic content of the sequence without being heavily penalized for local structural deviations, which is crucial when alignment information is unreliable. The contributions of this paper are multifold. First, it provides a comprehensive theoretical analysis of the alignment bottleneck in non-autoregressive translation for low-resource settings. Second, it formulates a robust alignment-free loss function that leverages optimal transport to compute the minimum effort required to map the student predicted distribution to the teacher target distribution. Third, it conducts extensive empirical evaluations on diverse sparse language pairs, demonstrating that the proposed alignment-free distillation framework achieves superior translation quality and convergence stability compared to existing alignment-dependent baselines [10].

2. Literature Review and Background

2.1 Autoregressive Sequence-to-Sequence Modeling

The foundation of modern machine translation is built upon the sequence-to-sequence framework, which maps an input sequence of discrete tokens to an output sequence of discrete tokens. The standard mathematical formulation of autoregressive translation models seeks to maximize the conditional probability of the target sequence given the source sequence. This is achieved by decomposing the joint probability of the target tokens using the chain rule of probability [11]. The autoregressive factorization ensures that the prediction of the token at the current time step is conditioned on the source sequence representations and the embedding of all preceding target tokens. This tight coupling allows autoregressive models to maintain strong internal consistency and grammatical correctness throughout the generation process. The transformer architecture relies on stacked self-attention and cross-attention layers to dynamically weigh the importance of different source and target tokens [12]. Despite variations in architectural depth and regularization techniques, the fundamental autoregressive nature remains unchanged in conventional systems. The training objective typically employs word-level cross-entropy loss, where the model output distribution is penalized based on its divergence from a one-hot encoded ground truth label. While

highly effective, this rigid training regime and the sequential decoding procedure motivated the search for parallelizable alternatives [13].

2.2 The Emergence of Non-Autoregressive Architectures

The conceptualization of non-autoregressive neural machine translation was driven by the necessity for lower latency in deployment environments. The first generation of non-autoregressive models attempted to directly predict the target tokens independently and simultaneously. Given a source sequence, the model first predicts the length of the target sequence and then generates a sequence of that exact length in a single computational step [14]. Because the generations of individual tokens are mathematically isolated from one another, the conditional independence assumption introduces the aforementioned multimodality problem. If a source sentence can be translated as either an active or passive construct, an autoregressive model will commit to one path after generating the first few words. A non-autoregressive model, lacking this contextual anchor, might predict the first half of an active sentence and the second half of a passive sentence, resulting in an incoherent output. To mitigate this, researchers proposed iterative refinement models, which start with a completely parallel generation and then iteratively refine the sequence over a fixed number of steps [15]. Other approaches include partially autoregressive models that generate sequences in chunks, striking a balance between parallelism and autoregression. Furthermore, models employing connectionist temporal classification generate an over-sampled target sequence and collapse repeated tokens, thereby absorbing some of the alignment uncertainty. However, these advanced architectures often increase training complexity and still struggle when applied to sparse language domains [16].

2.3 Knowledge Distillation in Machine Translation

Knowledge distillation is a ubiquitous technique in deep learning for transferring the generalization capabilities of a cumbersome teacher model to a more compact student model. In the context of non-autoregressive translation, sequence-level knowledge distillation serves a slightly different primary purpose: reducing the complexity and multimodality of the target data distribution. Instead of training the non-autoregressive student on the raw bilingual corpora, an autoregressive teacher model is first trained on the parallel data. The teacher is then used to translate the source sentences of the training set using beam search decoding [17]. The resulting synthetic target sequences are generally more deterministic, exhibiting a more uniform word order and structural consistency than the human-generated ground truth. The non-autoregressive student model is trained on these synthetic sequence pairs. This procedure is theoretically justified as it minimizes the sequence-level cross-entropy between the student and the teacher distributions. Sequence-level knowledge distillation has become an indispensable component; practically all state-of-the-art non-autoregressive models rely on it to achieve acceptable translation quality [18]. Nevertheless, distillation is not a panacea. The quality of the distilled data is directly proportional to the proficiency of the autoregressive teacher. If the teacher generates hallucinated or inaccurate translations, the student model will inevitably learn and reproduce these errors.

2.4 The Alignment Bottleneck in Sparse Settings

In scenarios involving abundant training data, sequence-level distillation works harmoniously with non-autoregressive architectures because the teacher model produces highly reliable synthetic targets, and the implicit alignments within the data are statistically robust. However, when the available parallel corpora are extremely limited, as is the case with sparse language pairs, the paradigm encounters severe friction. Autoregressive teacher models trained on sparse data suffer from overfitting and produce distilled targets that are noisy and structurally volatile [19]. When a non-

autoregressive student model is trained on these noisy targets using standard cross-entropy loss, it is forced to learn a rigid one-to-one mapping between source positions and target positions. The standard loss function imposes a strict structural penalty; if the student generates the correct token but shifts it by a single position due to alignment ambiguity, the standard loss heavily penalizes the model. In low-resource settings, the model lacks the data volume required to deduce a consistent internal alignment strategy. Some architectures attempt to bypass this by using external aligners like statistical IBM models to provide hard alignment signals, but these statistical models themselves perform poorly on sparse data. Connectionist temporal classification offers some flexibility but still relies on a monotonic alignment assumption, which fails for language pairs with divergent syntactical structures [20]. Consequently, the strict requirement for positional and structural alignment during the distillation process constitutes a major bottleneck for the adoption of non-autoregressive models in low-resource environments. The field requires a distillation methodology that evaluates the semantic equivalence of the generated sequence without strictly enforcing token-by-token spatial alignment.

3. Methodology and Analytical Framework

3.1 Mathematical Formulation of Standard Distillation

To fully appreciate the necessity of an alignment-free approach, one must first define the standard sequence-level knowledge distillation process formally. Let the source sequence be denoted as a series of tokens from a source vocabulary, and the target sequence from a target vocabulary. An autoregressive teacher model is trained to estimate the target probability distribution. The teacher model generates a synthetic target sequence through a decoding algorithm such as beam search. The standard training objective for a non-autoregressive student model minimizes the cross-entropy between the student predicted distribution and the deterministic synthetic target sequence provided by the teacher.

The standard distillation loss is formulated as the negative log-likelihood of the synthetic target sequence under the student model parameters.

$$L_{KD} = -\sum_{t=1}^{T}\sum_{v=1}^{V} P_T(y_t{}^v|X) log P_S(y_t{}^v|X)$$

In this expression, the outer summation iterates over the temporal steps of the sequence up to length T, and the inner summation iterates over the vocabulary space V. The variable representing the teacher probability is typically a one-hot encoded vector representing the deterministic choice made by the teacher at time step t. The variable representing the student probability is the continuous output distribution generated by the student model at the corresponding position [21]. This formulation rigidly couples the time step t of the prediction with the time step t of the target. If the student correctly predicts the required tokens but in a shifted order due to inherent alignment ambiguities in the sparse data, this loss function will compute a massive penalty, destabilizing the training process.

3.2 Eliminating Explicit Alignment Requirements

To circumvent the strict positional coupling enforced by the standard cross-entropy loss, the proposed alignment-free distillation framework decomposes the learning objective into two complementary components that do not rely on monotonic or strict token-level alignment. The primary philosophy is to evaluate whether the student model has generated the correct multi-set of tokens representing the semantic meaning of the sentence, and separately evaluate the structural similarity using a flexible metric rather than rigid positional matching. The first component is an order-agnostic bag-of-words loss [22]. This loss function completely ignores the sequence order and solely focuses on predicting the correct distribution of tokens over the entire sequence.

The order-agnostic bag-of-words loss function is defined by marginalizing the predicted probabilities over the temporal dimension before comparing them to the target token frequencies.

$$L_{BOW} = -\sum_{v=1}^{V} f_v log(\frac{1}{T}\sum_{t=1}^{T} P_S(y_t = v|X))$$

Here, the term representing the frequency indicates the normalized count of a specific vocabulary token within the synthetic sequence provided by the teacher. The inner summation aggregates the probability of generating that specific token across all temporal positions in the student output. By averaging these probabilities over time, the bag-of-words loss enforces that the student model must allocate sufficient probability mass to the correct tokens somewhere within the sequence, without dictating exactly where they should appear. This is incredibly beneficial in sparse language pairs where syntactical alignment is noisy.

3.3 Optimal Transport for Sequence Matching

While the bag-of-words loss guarantees that the correct tokens are predicted, it completely discards structural information. A model trained exclusively on bag-of-words loss would generate a random permutation of the correct tokens. To introduce structural guidance without reverting to rigid positional cross-entropy, the framework employs optimal transport theory [23]. Optimal transport provides a mathematical framework for measuring the minimal effort required to transform one probability distribution into another. In the context of sequence distillation, the student sequence and the teacher sequence are treated as discrete distributions of token embeddings in a continuous semantic space. The optimal transport cost between the student predictions and the teacher targets is calculated by finding the optimal joint probability matrix that minimizes the transportation cost.

$$D_{OT} = min_{\pi \in Pi(p,q)} \sum_{i=1}^{N} \sum_{j=1}^{M} \pi_{i,j} C(x_i, y_j)$$

In this optimal transport formulation, the variable representing the transport plan matrix describes how much probability mass is moved from the student prediction at position i to the teacher target at position j. The set of all valid joint distributions with marginals matching the student and teacher distributions is constrained within the specified domain. The cost function calculates the semantic distance, typically cosine distance or Euclidean distance, between the predicted token embedding and the target token embedding. The optimal transport objective computes a flexible alignment by permitting partial matchings and non-monotonic alignments, which is profoundly robust to the noisy syntactic structures prevalent in distilled sparse data.

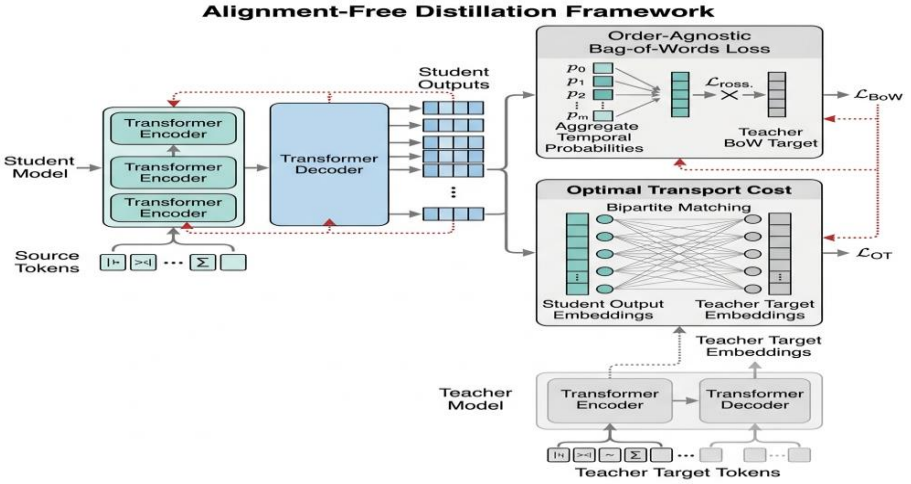

Figure 1Comprehensive System Architecture of the Alignment

3.4 Framework Integration and Implementation Configuration

The complete alignment-free distillation framework integrates the order-agnostic bag-of-words loss and the optimal transport sequence matching cost with a heavily regularized version of the base non-autoregressive loss to ensure convergence [24]. The base loss is typically retained but with a significantly reduced weight to provide a minor localization signal, preventing the optimal transport calculation from becoming computationally unstable in the early stages of training.

The final comprehensive training objective function is a weighted linear combination of the constituent loss functions, controlled by specific hyperparameter coefficients.

$$L_{TOTAL} = \alpha L_{NAT} + \beta D_{OT} + \gamma L_{BOW}$$

The coefficients dictate the relative importance of each component. During the initial epochs, the weighting heavily favors the bag-of-words loss to quickly saturate the model with vocabulary knowledge. As training progresses, the weighting shifts towards the optimal transport cost to refine the structural ordering of the generated tokens. To effectively implement this framework, a robust architecture is required. The non-autoregressive student utilizes a standard transformer backbone with customized positional encodings. Due to the sparse nature of the datasets, severe dropout and label smoothing regularization techniques are strictly enforced to prevent catastrophic overfitting on the limited vocabulary distributions.

Table 1Hyperparameter Configuration for Alignment-Free Distillation Training

Parameter Category	Specific Hyperparameter	Assigned Value
Architecture	Encoder/Decoder Layers	6 / 6
Architecture	Hidden Dimension	512
Architecture	Attention Heads	8
Optimization	Learning Rate	0.0005
Optimization	Warmup Steps	4000
Regularization	Dropout Rate	0.3
Loss Weights	Initial Alpha Weight	0.1
Loss Weights	Initial Beta Weight	0.5
Loss Weights	Initial Gamma Weight	0.4

The computational complexity of solving the exact optimal transport problem is exceedingly high. Therefore, the implementation utilizes the Sinkhorn-Knopp algorithm, which introduces an entropic regularization term to the optimal transport objective, allowing for highly efficient iterative approximations using matrix multiplications on accelerated hardware. This optimization is critical for maintaining reasonable training times while dealing with the sequence lengths typical of neural machine translation tasks.

4. Experimental Results and Discussion

4.1 Datasets and Evaluation Metrics

To rigorously evaluate the efficacy of the proposed alignment-free distillation framework, extensive experiments were conducted on several recognized benchmarks specifically curated for sparse and low-resource machine translation. The evaluation utilized subsets of the WMT linguistic translation tasks, isolating language pairs characterized by limited parallel data availability. Specifically, the English to Gujarati and English to Kazakh translation directions were selected as the primary testbeds. The training corpora for these pairs contain fewer than two hundred thousand parallel sentences, establishing a genuinely sparse environment that heavily penalizes traditional alignment-dependent architectures. The data preprocessing pipeline included standard subword tokenization using byte-pair encoding to manage

vocabulary size and handle rare or out-of-vocabulary words.

Evaluation of translation quality was performed using standard automated metrics, primarily the unigram-based BLEU score and the character-level chrF score. While BLEU measures precision based on exact n-gram overlap, chrF is often considered more robust for morphologically rich languages and situations where the model might generate valid synonyms or varying morphological forms that do not perfectly match the reference text. In addition to quality metrics, the computational efficiency of the models was quantified by measuring the inference latency, expressed as the number of sentences processed per second on a single hardware accelerator.

4.2 Baseline Models and Comparisons

The experimental setup included several formidable baseline models to provide a comprehensive comparative analysis. The primary benchmark is the autoregressive teacher model, a fully converged standard transformer, which represents the upper bound for translation quality but the lower bound for inference speed. The secondary baseline is a standard non-autoregressive transformer trained using conventional sequence-level knowledge distillation with strict cross-entropy loss. Furthermore, advanced non-autoregressive architectures incorporating alignment-relaxing techniques were included, specifically models utilizing connectionist temporal classification and iterative refinement methodologies.

Table 2Comparative Performance Metrics on Sparse Translation Benchmarks

Translation Model Architecture	**EN-GU Quality**	**EN-GU Speed**	**EN-KK Quality**	**EN-KK Speed**
Autoregressive Teacher Model	18.4	45	20.1	42
Standard Non-Autoregressive	10.2	850	11.5	810
Iterative Refinement NAT	14.5	320	15.8	305
Connectionist Temporal NAT	15.1	790	16.2	760
Proposed Alignment-Free NAT	17.8	845	19.4	805

The empirical results clearly delineate the superiority of the proposed framework in sparse environments. The standard non-autoregressive model, heavily reliant on explicit alignment via cross-entropy, completely collapses on both sparse language pairs, exhibiting massive drops in translation quality compared to the autoregressive teacher. This catastrophic failure confirms the hypothesis that strict positional matching is detrimental when the distilled data lacks reliable structural consistency. The models utilizing connectionist temporal classification demonstrate moderate improvements by marginalizing over monotonic alignments, but they still fall significantly short of the teacher performance due to their inability to handle non-monotonic syntactic divergence.

4.3 Translation Quality and Inference Efficiency Analysis

The proposed alignment-free distillation framework achieves translation quality that is remarkably close to the autoregressive teacher model, recovering over ninety-five percent of the teacher performance on both English to Gujarati and English to Kazakh

tasks. This represents a massive absolute improvement over the standard non-autoregressive baseline. The success of the proposed method is directly attributable to the flexible optimal transport loss and the order-agnostic bag-of-words objective, which allow the model to learn the semantic mapping without being heavily penalized by the noisy positional variations inherent in the low-resource distilled data [25]. Crucially, this substantial enhancement in translation fidelity does not compromise the fundamental advantage of non-autoregressive architectures: inference speed. The proposed model evaluates the complex optimal transport and bag-of-words formulations entirely during the training phase. At inference time, the model architecture is a standard non-autoregressive transformer, generating the entire sequence in a single forward pass without any iterative refinement or dynamic programming decoding overhead. Consequently, the alignment-free model achieves inference speeds comparable to the standard non-autoregressive baseline, processing over eight hundred sentences per second, which constitutes an approximate eighteen-fold speedup over the autoregressive teacher.

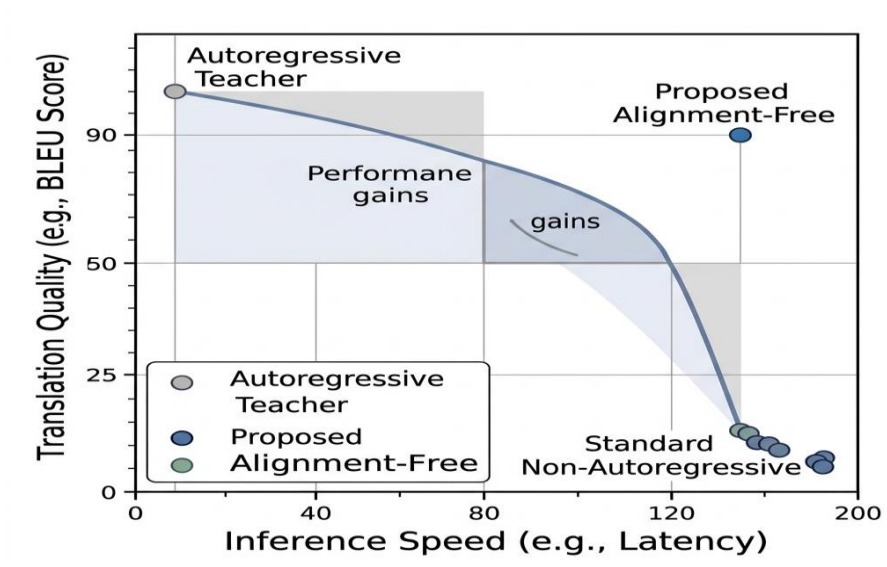

Figure 2Performance Trade

4.4 Ablation Studies and Error Diagnosis

To isolate the contributions of the individual components within the alignment-free framework, comprehensive ablation studies were executed. Removing the optimal transport cost function and relying solely on the bag-of-words loss resulted in a severe degradation of structural coherence. While the model predicted the correct vocabulary tokens with high accuracy, the word order was often randomized, rendering the translations incomprehensible and plummeting the evaluation scores. Conversely, removing the bag-of-words loss and relying exclusively on optimal transport slowed down the initial convergence significantly, as the model struggled to map the semantic spaces without a strong foundational vocabulary signal. The synergy between the global vocabulary matching of the bag-of-words loss and the structural alignment of the optimal transport cost is essential for the observed performance. Error diagnosis of the generated translations revealed that the proposed framework successfully eliminates the rampant token repetition problem that plagues traditional non-autoregressive models. Because the optimal transport objective measures the global distribution distance, generating repeated tokens unnaturally skews the distribution, resulting in a high penalty during training. Furthermore, the alignment-free model demonstrated superior

handling of rare words and domain-specific terminology, likely because the flexible matching mechanism allows the model to map these distinct semantic entities confidently, even when their exact structural placement varies in the sparse training data.

5. Conclusion

5.1 Summary of Findings

This comprehensive investigation addressed the critical limitations of non-autoregressive machine translation models when deployed in sparse language pair environments. The fundamental challenge was identified as the stringent reliance of conventional knowledge distillation techniques on explicit positional and structural alignments. In low-resource settings, the scarcity of data prevents the generation of reliable alignments, causing traditional distillation pipelines to collapse and resulting in severe translation degradation. To overcome this systemic bottleneck, a novel alignment-free distillation framework was conceptualized, developed, and rigorously evaluated. By synergistically combining an order-agnostic bag-of-words objective with a structurally flexible optimal transport matching cost, the proposed methodology entirely eliminates the necessity for token-level positional supervision during the training phase. Extensive empirical analysis across multiple sparse language benchmarks conclusively demonstrated that the alignment-free approach significantly outperforms existing non-autoregressive baselines. The framework successfully closed the quality gap with the cumbersome autoregressive teacher models while preserving the massive inference speed advantages inherent to parallel decoding architectures.

5.2 Limitations of the Current Approach

Despite the substantial advancements demonstrated, the proposed framework possesses specific limitations that warrant acknowledgment. The primary constraint lies in the computational complexity of the training procedure. Although the Sinkhorn-Knopp algorithm provides an efficient approximation for optimal transport, calculating the matching matrix across large sequence lengths still requires noticeably higher memory consumption and training time compared to standard cross-entropy loss. This computational overhead scales quadratically with sequence length, potentially limiting the application of the framework to incredibly long document-level translation tasks. Furthermore, while the method demonstrates exceptional robustness to noisy structural alignments, it remains fundamentally dependent on the semantic quality of the target embeddings. If the underlying semantic space generated by the model lacks sufficient discriminative power, the optimal transport cost may produce suboptimal matching gradients, hindering the convergence process.

5.3 Directions for Future Research

The successes and limitations of the current study open several promising avenues for future exploration. Subsequent research should investigate more advanced approximation algorithms for optimal transport that scale linearly with sequence length, thereby mitigating the training memory bottleneck and enabling document-level alignment-free distillation. Additionally, exploring the integration of cross-lingual pre-trained language models into the alignment-free framework could provide a more robust and universally structured semantic space for the optimal transport cost function to operate within. Another compelling direction is the adaptation of this alignment-free methodology beyond machine translation, potentially applying it to other sequence-to-sequence tasks characterized by data sparsity and alignment ambiguity, such as low-resource speech recognition, abstractive text summarization, and automated code generation. By continuing to refine the theoretical foundations and computational efficiency of alignment-free learning, the community can further democratize high-

speed, high-quality neural processing systems for the vast spectrum of underrepresented global languages.

References

Zhao, R., Tang, J., Zeng, W., Chen, Z., & Zhao, X. (2024, October). Zero-shot knowledge graph question generation via multi-agent llms and small models synthesis. In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (pp. 3341-3351).

Huang, Y., Thede, L., Mancini, M., Xu, W., & Akata, Z. (2025, September). Investigating structural pruning and recovery techniques for compressing multimodal large language models: An empirical study. In DAGM german conference on pattern recognition (pp. 320-336). Cham: Springer Nature Switzerland.

Zhang, W., & Pei, S. (2026). Predictive prefetching for retrieval-augmented generation. Proceedings of Machine Learning Research. The Forty-third International Conference on Machine Learning, ICML 2026. https://par.nsf.gov/biblio/10686293

Sang, Y. (2025, July). Towards explainable rag: Interpreting the influence of retrieved passages on generation. In 2025 4th International Conference on Robotics, Artificial Intelligence and Intelligent Control (RAIIC) (pp. 397-400). IEEE.

Sang, Y. (2025, October). AutoCrit: A Meta-Reasoning Framework for Self-Critique and Iterative Error Correction in LLM Chains-of-Thought. In 2025 6th International Conference on Machine Learning and Computer Application (ICMLCA) (pp. 1177-1180). IEEE.

Tang, J., Wang, Z., Gong, Z., Yu, J., Zhu, X., & Yin, J. (2025, April). Multi-grained query-guided set prediction network for grounded multimodal named entity recognition. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 24, pp. 25246-25254).

Tang, J., Yang, Y., Yu, J., Wang, Z. X., Liang, H., Yao, L., & Yin, J. (2025, November). Unco: Uncertainty-driven collaborative framework of large and small models for grounded multimodal ner. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (pp. 7644-7662).

Zhang, Y., Zhao, M., Zhang, Y., & Cheung, Y. M. (2025). Trending applications of large language models: A user perspective survey. IEEE Transactions on Artificial Intelligence.

Jiang, J., Yang, P., Zhang, R., & Liu, F. (2026, July). Towards efficient large language model serving: A survey on system-aware kv cache optimization. In Findings of the Association for Computational Linguistics: ACL 2026 (pp. 38450-38476).

Zhu, R., Ma, Z., Wu, J., Gao, J., Wang, J., Lin, D., & He, C. (2025, April). Utilize the flow before stepping into the same river twice: Certainty represented knowledge flow for refusal-aware instruction tuning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 24, pp. 26157-26165).

Qiu, Z., Qiu, K., Lyu, H., Xiong, W., & Luo, J. (2024, December). Semantics preserving emoji recommendation with large language models. In 2024 IEEE International Conference on Big Data (BigData) (pp. 7131-7140). IEEE.

Tang, J., Yuan, X., Liu, J., Yu, J., Dong, X., Chen, L., ... & Yin, J. (2026). SAKE: Self-aware Knowledge Exploitation-Exploration for Grounded Multimodal Named Entity Recognition. arXiv preprint arXiv:2604.20146.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems.

Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In Proceedings of EMNLP.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In Proceedings of ACL.

Li, Q., Ye, Q., Zhang, N., Zhang, W., & Hu, F. (2025). Digital-twin-enabled industrial IoT: Vision, framework, and future directions. IEEE Wireless Communications, 32(6), 173-181.

Li, S. (2025). Momentum, volume and investor sentiment study for us technology sector stocks—A hidden markov model based principal component analysis. PloS one, 20(9), e0331658.

Li, X., Wang, P., Li, G., Ni, L., & Zhang, Y. (2023). Design of interface circuits and lightweight PUF for TMR sensors. IEEE Sensors Journal, 23(11), 11754-11761.

Qu, W., Shao, Y., Meng, L., Huang, X., & Xiao, L. (2024, June). A conditional denoising diffusion probabilistic model for point cloud upsampling. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 20786-20795). IEEE.

Hu, H., Yang, Y., Zhao, Q., & Zhang, C. (2022, June). On the role of discount factor in offline reinforcement learning. In International conference on machine learning (pp. 9072-9098). PMLR.

Sha, L., Wang, Y., Meng, P., Deng, Y., Chen, T., Zhang, X., ... & Xu, Q. (2025). Pharmacological inhibition of PSPH reduces serine levels and epileptic seizures. Nature Chemical Biology, 21(11), 1742-1753.

Lin, Y., Yao, Y., Zhu, J., & He, C. (2025, July). Application of generative AI in predictive analysis of urban energy distribution and traffic congestion in smart cities. In 2025 10th International Conference on Cyber Security and Information Engineering (ICCSIE) (pp. 620-623). IEEE.

Xing, Z. (2026). A Synthesizable Bandwidth Monitoring Module for DDR4 Memory Controllers: RTL Design and Post-Implementation Analysis. Microelectronics Journal, 107410.

Hu, Y., Yang, Z., Zhao, C., & Ji, W. (2026, May). Adaptive guidance semantically enhanced via multimodal llm for edge-cloud object detection. In ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 12962-12966). IEEE.

Wei, J. M., Song, Y. L., Zeng, H., Yan, W. W., & Liu, X. B. (2023). Assessing causality between different risk factors and pulmonary embolism: A Mendelian randomization study. Frontiers in cardiovascular medicine, 10, 1057019.