Context-Aware Iterative Token Detection and Masked Transmission for Wireless Token Communication

Junyong Shin[1], Joohyuk Park[1], Jihong Park[2], Jinho Choi[3], Yo-Seb Jeon[1]

[1]Pohang University of Science and Technology
[2]Singapore University of Technology and Design
[3]The University of Adelaide
{sjyong, joohyuk.park, yoseb.jeon}@postech.ac.kr, jihong_park@sutd.edu.sg, jinho.choi@adelaide.edu.au

Abstract

The success of large-scale language models has established tokens as compact and meaningful units for natural-language representation, which motivates token communication over wireless channels, where tokens are considered fundamental units for wireless transmission. We propose a context-aware token communication framework that uses a pretrained masked language model (MLM) as a shared contextual probability model between the transmitter (Tx) and receiver (Rx). At Rx, we develop an iterative token detection method that jointly exploits MLM-guided contextual priors and channel observations based on a Bayesian perspective. At Tx, we additionally introduce a context-aware masking strategy which skips highly predictable token transmission to reduce transmission rate. Simulation results demonstrate that the proposed framework substantially improves reconstructed sentence quality and supports effective rate adaptation under various channel conditions.

Introduction

Recent advances in natural-language processing have demonstrated the effectiveness of processing information through discrete tokens, which represent context-aware linguistic units. As wireless communications increasingly connects devices executing token-based applications, directly transmitting tokens over the air has emerged as a relevant and promising approach, which we refer to as *token communication* (3GPP TSG RAN1 WG1 2025; Qiao et al. 2025; Lee et al. 2025; Oh et al. 2025).

Conventional communication systems treat symbols as independent entities and overlook the contextual relationships within token sequences. As a result, all tokens are transmitted with equal priority—even when some are easily predictable—and channel errors become difficult to resolve without contextual redundancy. Although recent studies have explored machine learning–based symbol detection (Fan, Yang, and Shen 2025), these methods remain confined to minimizing local detection errors without leveraging inter-token dependencies. In contrast, token communication exploits contextual relationships among tokens, allowing receivers to infer missing or corrupted tokens from surrounding context. As token-based applications continue to expand,

this limitation highlights the need for a shift from context-agnostic to context-aware wireless communications.

Context modeling provides a means to bridge this gap. By estimating how tokens relate to and depend on each other, context models can guide both which tokens must be transmitted and how lost information can be recovered. At the transmitter (Tx) side, context can identify predictable tokens that do not need explicit transmission, directly translating contextual redundancy into rate savings. At the receiver (Rx), context can assist in reconstructing missing or unreliable tokens from surrounding evidence, improving robustness without solely relying on stronger error protection.

Recent advances in artificial intelligence enable highly capable context models trained from large corpora of token sequences. Among them, masked language models (MLMs) such as bidirectional encoder representations from Transformers (BERT) (Devlin et al. 2019) can capture both local and long-range dependencies while inferring masked tokens based on the context. This predictive capability makes such models naturally suited for token communication: the Rx can leverage contextual knowledge as a prior, which can be combined with channel observations in a Bayesian manner. Likewise, the Tx can use the same prior to estimate which tokens are predictable and thus less critical for explicit transmission. This leads to a unified view where context models explicitly guide both transmission decisions and token inference.

Overall, we aim to address the following key question: *How can we jointly exploit contextual priors and channel observations to improve token transmission and reception?* To address this, we propose a Bayesian-inspired token communication framework with rate adaptivity built on MLM-based contextual priors, consisting of:

- **Iterative Token Detection at Rx:** An iterative maximum a posteriori (MAP)-based token detection algorithm that refines token decisions using contextual probability and channel likelihood.

- **Context-Aware Token Masking at Tx:** A context-aware masking strategy that suppresses predictable tokens with low entropy, reducing transmission overhead while maintaining sentence quality.

Our simulation results show clear gains in sentence quality under reduced-rate transmission. Over conventional ML

detection, the proposed iterative token detection improves cosine similarity by up to 0.1769 (Europarl) and 0.1558 (WikiText-103). Additionally, under equal masking ratios, the proposed context-aware masking yields up to 0.0769 and 0.0461 improvement over random masking on the two datasets, respectively. These results highlight the effectiveness of leveraging contextual priors for both robust detection and rate adaptation.

System Model

We consider a wireless token communication system where a sequence of discrete language tokens is transmitted over a wireless fading channel. Let $\mathbf{w}$ denote the original token sequence of length T, given as:

$$\mathbf{w} = [w_1, w_2, ..., w_T]. \tag{1}$$

To enable rate-adaptive transmission, the Tx employs a masking-based strategy that selectively masks less informative tokens according to their contextual predictability or uncertainty, producing a masked sequence $\mathbf{w}_\mathrm{m}$. Each masked token is treated as omitted from physical transmission, and only the remaining unmasked tokens are transmitted over the channel. Hence, the masking ratio directly determines the effective transmission rate, as higher masking ratios correspond to lower physical-layer symbol usage while maintaining recoverability through contextual inference at the Rx.

Each token w_i belongs to a vocabulary of size V and can be represented by a binary vector of length $\lceil \log_2(V) \rceil$ through a tokenizer. Accordingly, each token element w_i is converted into a bit sequence $\mathbf{b}_i$, defined as

$$\mathbf{b}_i = [b_{i,1}, b_{i,2}, ..., b_{i,\lceil \log_2(V) \rceil}], \quad b_{i,n} \in \{0, 1\}, \ \forall n. \tag{2}$$

The resulting bit sequence $\mathbf{b}(w_i)$ is then grouped and modulated into complex-valued symbols as

$$\mathbf{s}_i = [s_{i,1}, s_{i,2}, ..., s_{i,\lceil \lceil \log_2(V) \rceil / m \rceil}], \quad s_{i,k} \in \mathcal{S}, \ \forall k, \tag{3}$$

where each symbol $s_{i,k}$ represents a group of m consecutive bits that are jointly mapped onto a complex symbol constellation $\mathcal{S}$. For example, a 2^m-ary Quadrature Amplitude Modulation (QAM) scheme can be employed for symbol mapping, which is widely used in digital communication systems. The modulated symbols are transmitted over a Rayleigh block-fading channel (Goldsmith 2005; Shin et al. 2025; Kang et al. 2025) modeled as

$$\mathbf{y}_i = h\sqrt{p}\mathbf{s}_i + \mathbf{n}_i, \tag{4}$$

where $h \in \mathbb{C}$ denotes the complex fading coefficient that remains constant over one transmission block, p is the transmit power, and $\mathbf{n}_i \sim \mathcal{CN}(0, \sigma^2 \mathbf{I})$ is additive white Gaussian noise (AWGN). The signal-to-noise ratio (SNR) in dB is defined as

$$\mathrm{SNR\ (dB)} = 10 \log_{10} \left(\frac{p \mathbb{E}[|h|^2]}{\sigma^2} \right). \tag{5}$$

Given the entire observation $\mathbf{y} = [\mathbf{y}_1, ..., \mathbf{y}_T]$, the Rx computes the symbol-wise likelihood as

$$P(y_{i,k}|\hat{s}_{i,k}) = \frac{1}{\pi\sigma^2}\exp\left(-\frac{|y_{i,k} - h\sqrt{p}\hat{s}_{i,k}|^2}{\sigma^2}\right),$$
$$\hat{s}_{i,k} \in \mathcal{S}, \ \forall i, k. \tag{6}$$

From these, the token-level likelihoods $P(\mathbf{y}_i|w_i)$ are derived as

$$P(\mathbf{y}_i|w_i) = \prod_{k=1}^{\lceil \lceil \log_2(V) \rceil / m \rceil} P(y_{i,k}|\hat{s}_{i,k}). \tag{7}$$

This probabilistic representation $P(\mathbf{y}_i|w_i)$ expresses the likelihood of receiving token w_i under channel distortion. Because no channel information is conveyed for masked positions, their likelihoods are set as a uniform distribution over the vocabulary, i.e., $P(\mathbf{y}_{i^*}|w_{i^*}) = 1/V$ for any masked token position i^*.

The detected token sequence is expressed as

$$\hat{\mathbf{w}} = [\hat{w}_1, \hat{w}_2, ..., \hat{w}_T]. \tag{8}$$

If a conventional wireless communication system were applied, token detection would rely solely on maximizing this channel likelihood, following the maximum likelihood (ML) detection principle as

$$\hat{w}_i = \operatorname*{argmax}_{w_i} P(\mathbf{y}_i|w_i). \tag{9}$$

However, this approach ignores the contextual dependencies among tokens—each token is treated independently, even when certain tokens are highly predictable given their neighbors. As a result, ML detection becomes suboptimal in wireless token communication, where contextual information can play a critical role in recovering omitted or corrupted tokens.

Iterative Token Detection for Context-Aware Wireless Token Communication

To overcome the limitations of conventional ML detection, which relies solely on channel observations and ignores contextual dependencies among tokens, we reformulate token detection as a MAP problem. This formulation enables the Rx to jointly consider both the channel likelihood and a **contextual prior** that captures inter-token dependencies. The contextual prior represents the conditional probability of a token given its surrounding tokens, serving as a probabilistic reference that guides detection when channel information is unreliable or missing.

To compute this contextual prior, we employ a MLM, which estimates token-level conditional probabilities in masked position by leveraging contextual information from the surrounding tokens. Leveraging this shared prior knowledge, the proposed framework establishes a unified design for both Tx and Rx: the Rx iteratively refines uncertain token estimates using MLM-derived contextual priors, while the Tx exploits the same priors in reverse to mask predictable tokens and thereby achieve effective rate adaptation. The overall proposed framework is illustrated in Fig. 1.

MLM as Contextual Token Prior Model

In our token communication framework, MLM is adopted as a contextual token prior model shared by both the Tx and Rx. The MLM produces a contextual probability distribution over the vocabulary for any masked token position, leveraging surrounding tokens to infer the most plausible candidate. Formally, given an input sequence of tokens
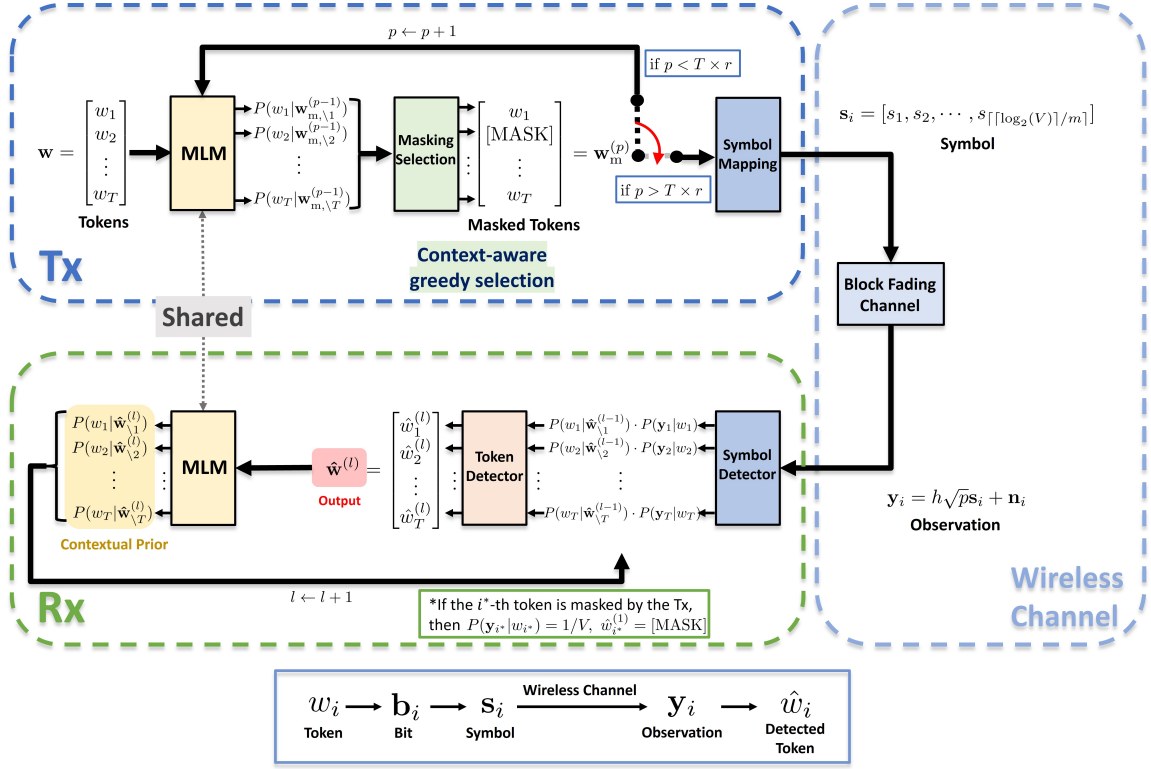
Figure 1: An illustration of the proposed token communication framework.

$\mathbf{x} = [x_1, ..., x_T]$, MLM estimates a conditional categorical distribution over the vocabulary at each position:

$$P(x_i|\mathbf{x}_{\setminus i}) = [\text{MLM}(\mathbf{x}_{\setminus i})]_i, \qquad (10)$$

where $\text{MLM}(\cdot)$ represents the MLM function, and $\mathbf{x}_{\setminus i} \triangleq [x_1, ..., x_{i-1}, x_{i+1}, ..., x_T]$ denotes the sequence with i-th element ommited. This operation intrinsically models token dependency and contextual semantics.

Let [MASK] denote the special mask token used by MLM to hide selected positions. For any masking set $\mathcal{M} \subset \{1, ..., T\}$, we define a masking operator:

$$[\text{Mask}(\mathbf{w}; \mathcal{M})]_i = \begin{cases} [\text{MASK}], & i \in \mathcal{M}, \\ w_i, & i \notin \mathcal{M}, \end{cases} \qquad (11)$$

and express the masked token sequence used for inference as $\mathbf{w}_m = \text{Mask}(\mathbf{w}; \mathcal{M})$. If $\mathcal{M} = \varnothing$, this operation reduces to the original sequence, i.e., $\mathbf{w} = \mathbf{w}_m = \text{Mask}(\mathbf{w}; \mathcal{M})$. For a specific index i, the conditional distribution of the true token w_i is computed by additionally masking the i-th position of the masked token sequence $\mathbf{w}_m$, yielding $\mathbf{w}_{m, \setminus i}$, and feeding it into the MLM as input: replacing only the i-th position with [MASK]:

$$P(w_i|\mathbf{w}_{m,\setminus i}) = [\text{MLM}(\mathbf{w}_{m,\setminus i})]_i, \qquad (12)$$

which provides a prior belief about w_i given the observed or unmasked context tokens. We refer to this prior belief as the **contextual prior**, as it captures the token-level probability distribution conditioned on the surrounding context.

In the proposed wireless token communication system, this contextual prior plays two complementary roles:

- **Tx-side:** The contextual prior $P(w_i|\mathbf{w}_{m,\setminus i})$ provides an entropy metric to determine which tokens are more redundant and can be masked for rate adaptation.

- **Rx-side:** The same contextual prior $P(w_i|\hat{\mathbf{w}}_{\setminus i}^{(l-1)})$ serves as the probabilistic reference during iterative detection, refining token estimates by combining contextual knowledge with channel observations.

Thus, MLM establishes a shared contextual probabilistic model between Tx and Rx, enabling semantic-aware encoding and robust inference over wireless channels.

Rx Strategy: Context-Aware Iterative Token Detection

This part builds on the contextual prior established in the last section to implement Bayesian token detection. By iteratively refining token estimates using contextual priors derived from MLM, the Rx progressively approaches the MAP token detection.

At the Rx, the channel observation is first converted into a set of token-wise likelihoods representing the equivalent received token sequence $\mathbf{y}$. The objective of iterative token detection is to solve a MAP token detection problem:

$$\hat{w}_i = \underset{w_i}{\arg\max} \, P(w_i|\mathbf{y}). \qquad (13)$$

Applying Bayes' rule and using the total probability theorem over $\mathbf{w}_{\setminus i}$ yields

$$\hat{w}_i = \underset{w_i}{\arg\max} \; P(\mathbf{y}|w_i)P(w_i) \tag{14}$$

$$= \underset{w_i}{\arg\max} \sum_{\mathbf{w}_{\setminus i}} P(\mathbf{y}|\mathbf{w})P(\mathbf{w}). \tag{15}$$

Moreover, using the token-conditional independence of observations, the MAP rules becomes

$$\hat{w}_i = \underset{w_i}{\arg\max} \sum_{\mathbf{w}_{\setminus i}} \prod_{j=1}^{T} P(\mathbf{y}_j|w_j)P(\mathbf{w}) \tag{16}$$

$$= \underset{w_i}{\arg\max} \; P(\mathbf{y}_i|w_i) \sum_{\mathbf{w}_{\setminus i}} \prod_{j\neq i} P(\mathbf{y}_j|w_j)P(\mathbf{w}) \tag{17}$$

$$= \underset{w_i}{\arg\max} \; P(\mathbf{y}_i|w_i) \sum_{\mathbf{w}_{\setminus i}} \prod_{j\neq i} P(\mathbf{y}_j|w_j)P(w_i|\mathbf{w}_{\setminus i})P(\mathbf{w}_{\setminus i}). \tag{18}$$

To avoid the entanglement and intractability arising from marginalizing over all combinations of $\mathbf{w}_{\setminus i}$ — since each token's probability is inherently coupled with others in the sequence, forming probabilistic loops — we adopt a widely used single-sequence approximation in iterative detection and decoding frameworks (Tüchler, Koetter, and Singer 2002). The key insight is that the posterior probability mass $P(\mathbf{w}_{\setminus i})$ tends to be highly concentrated around the most likely token sequence estimate from the previous iteration. Therefore, we approximate

$$P(\mathbf{w}_{\setminus i}) \approx \begin{cases} 1, & \text{if } \mathbf{w}_{\setminus i} = \hat{\mathbf{w}}_{\setminus i}^{(l-1)}, \\ 0, & \text{otherwise,} \end{cases} \tag{19}$$

which allows the summation over $\mathbf{w}_{\setminus i}$ to be replaced by the single detected sequence from iteration $l-1$. This leads to the following tractable MAP detection rule:

$$\hat{w}_i^{(l)} = \underset{w_i}{\arg\max} \; P(\mathbf{y}_i|w_i)P(w_i|\hat{\mathbf{w}}_{\setminus i}^{(l-1)}). \tag{20}$$

Here, the first term $P(\mathbf{y}_i|w_i)$ corresponds to the channel likelihood introduced in the previous section, while the second term $P(w_i|\hat{\mathbf{w}}_{\setminus i}^{(l-1)})$ represents the contextual prior derived from MLM. Unlike conventional ML detection, which relies solely on $P(\mathbf{y}_i|w_i)$, the proposed MAP formulation jointly considers both the channel likelihood and the context-based prior that captures inter-token dependencies. This integration enables more reliable token inference, particularly in cases where the received symbols are corrupted or when certain tokens are omitted by the Tx, since contextual knowledge can compensate for incomplete or noisy channel observations.

For the first iteration $l = 1$, no prior information is available at the Rx. Therefore, the prior is uniformly defined:

$$P(w_i|\hat{\mathbf{w}}_{\setminus i}^{(0)}) = \frac{1}{V}, \; \forall i. \tag{21}$$

If the i^*-th token is masked by the Tx, the corresponding received token at the first iteration is set to the mask token,

i.e., $\hat{w}_{i^*}^{(1)} = [\text{MASK}]$. For $l \geq 2$, priors are refined via the shared MLM. Specifically, the Rx masks position i in $\hat{\mathbf{w}}^{(l)}$, thereby constructing $\hat{\mathbf{w}}_{\setminus i}^{(l)}$, and feeds it as MLM input:

$$P(w_i|\hat{\mathbf{w}}_{\setminus i}^{(l)}) = [\text{MLM}(\hat{\mathbf{w}}_{\setminus i}^{(l)})]_i, \tag{22}$$

which is then used as the prior for the next iteration. This iterative process continues until a maximum iteration count, gradually improving its token prior and detection output.

Tx Strategy: Context-Aware Masking for Rate-Adaptive Token Transmission

The Tx side complements the iterative detection by exploiting the same contextual prior in reverse: tokens that exhibit low uncertainty under the MLM are masked and not physically transmitted. This mechanism enables effective rate adaptation, as predictable tokens rely on contextual inference at the Rx, while only uncertain tokens are explicitly transmitted.

To formalize this process, the Tx quantifies the transmission rate by the masking ratio $r \in [0, 1]$, which determines how many tokens are omitted from physical transmission. The Tx exploits the shared MLM to decide which tokens can be reliably inferred by the Rx—thus enabling rate adaptation under a given transmission budget. Let r denote the masking ratio, resulting in a masking budget of $p = \lfloor Tr \rfloor$.

We identify the masking set $\mathcal{M}^{(p)} \subset \{1, ..., T\}$ through a greedy selection process performed over p iterations. At iteration $p \in \{1, ..., \lfloor Tr \rfloor\}$, the current masked sequence is denoted as $\mathbf{w}_{\mathrm{m}}^{(p-1)}$. Given $\mathbf{w}_{\mathrm{m}}^{(p-1)}$, the Tx evaluates, for every unmasked position $i \notin \mathcal{M}^{(p-1)}$, the MLM predictive distribution obtained as

$$P(w_i|\mathbf{w}_{\mathrm{m},\setminus i}^{(p-1)}) = [\text{MLM}(\mathbf{w}_{\mathrm{m},\setminus i}^{(p-1)})]_i. \tag{23}$$

The predictability of position i is quantified by its token-wise entropy

$$H_i^{(p)} = -\sum_v P(w_i = v|\mathbf{w}_{\mathrm{m},\setminus i}^{(p-1)})\log_2 P(w_i = v|\mathbf{w}_{\mathrm{m},\setminus i}^{(p-1)}), \tag{24}$$

where a lower entropy implies that the token is easier for the Rx to estimate without explicit transmission. Utilizing this formulation, the Tx selects the most predictable token index to be masked next:

$$i_p^* = \underset{i \notin \mathcal{M}^{(p-1)}}{\arg\max} \; H_i^{(p)}. \tag{25}$$

Accordingly, the masking set and masked sequence are updated as $\mathcal{M}^{(p)} = \mathcal{M}^{(p-1)} \cup \{i_p^*\}$, $\mathbf{w}_{\mathrm{m}}^{(p)} = \text{Mask}(\mathbf{w}; \mathcal{M}^{(p)})$, and this procedure iterates until $p > Tr$. By construction, $\mathbf{w}_{\mathrm{m}}^{(0)} = \mathbf{w}$ and $|\mathcal{M}^{(p)}| = p$.

Since each transmitted (i.e., unmasked) token requires $\lceil \log_2(V) \rceil$ bits, masking reduces the physical-layer transmission rate while relying on the shared language model to restore the missing semantics at the Rx. This tightly couples the communication rate with the linguistic predictability of the source content.
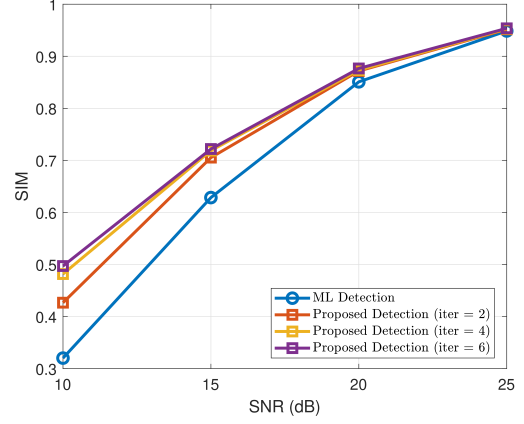
Simulation Results

We evaluate the proposed context-aware token communication framework using a text transmission task. The Europarl corpus (Koehn 2005) and the WikiText-103 dataset (Merity et al. 2016) are used as the source text dataset, and tokens are generated by WordPiece (Wu et al. 2016)-based tokenization. The BERT model in (Devlin et al. 2019) is employed as the shared MLM at both the Tx and Rx. We consider packets consisting of $T = 128$ tokens, where each token is represented by 15 bits. For physical-layer transmission, 16-QAM modulation is applied over a Rayleigh block fading channel. To assess the semantic correctness of the received text, we use a cosine similarity metric, denoted as *SIM*, between sentence embeddings obtained from the "all-MiniLM-L6-v2" SentenceTransformer model (Reimers and Gurevych 2019). This embedding-based similarity effectively measures meaning preservation, which aligns with the goal of token communication.

Fig. 2 illustrates the performance of the proposed iterative token detection strategy without applying the Tx masking scheme (i.e., full token transmission). The baseline *ML Detection* performs token estimation solely based on the channel likelihood $P(\mathbf{y}_i|w_i)$, without incorporating the contextual prior $P(w_i|\hat{\mathbf{w}}_{\backslash i}^{(l-1)})$ derived from MLM. In contrast, the *Proposed Detection* incorporates contextual priors from MLM and iteratively refines the token estimates using eq. (20). Fig. 2 shows that even a single iteration significantly improves performance over the baseline, particularly in low-SNR regimes. As the number of iterations increases from iter $= 2$ to iter $= 6$, the SIM value improves steadily and nearly converges at five iterations. Overall, the proposed Rx strategy clearly outperforms the baseline, indicating that semantic-aware iterative refinement can effectively compensate for channel-induced distortion.
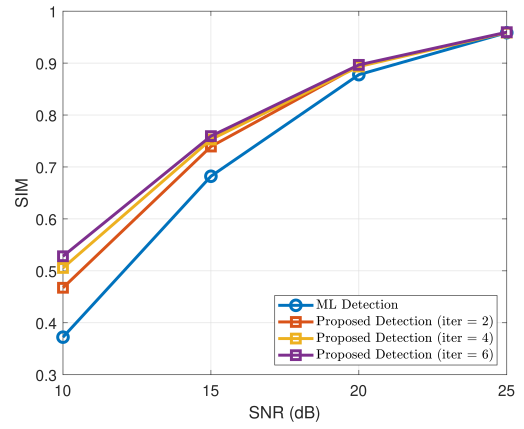
Fig. 3 presents the performance when the Tx adopts the proposed context-aware masking strategy to reduce the transmission rate. Except for the *ML Detection*, the proposed iterative token detection method is adopted, and the number of iterations iter $= 6$ and masking ratios of $r \in \{0.1, 0.3\}$ are adopted. Aditionally, as Tx strategies, two masking policies are compared:

- **Random Masking**: the Tx removes tokens at arbitrary positions without considering the predictability of Rx, serving as a lower-bound inference.

- **Context-Aware Masking**: the proposed greedy masking strategy that selectively masks highly predictable tokens based on MLM inference.

As shown in the figure, random masking causes noticeable performance degradation due to loss of important semantic information. In contrast, the proposed context-aware masking achieves significantly higher SIM levels than that of random masking, for each masking ratios. Particularly, when $r = 0.1$, the performance of proposed context-aware masking remains very close to the *Rx Strategy only* (without masking) case, demonstrating strong rate-adaptation capability with minimal degradation in context quality. These results confirm that the proposed joint Tx–Rx strategy suc-

(a) The Europarl corpus dataset

(b) The WikiText-103 dataset

Figure 2: Performance of the proposed iterative token detection strategy (Rx-side only) under different iteration counts.

cessfully harnesses context redundancy to reduce physical-layer transmission cost while maintaining high-quality text reconstruction at the Rx.

Conclusion

We presented a Bayesian-inspired token communication framework that integrates MLM as a contextual prior model. In this framework, the Tx selectively masks predictable tokens to reduce rate, while the Rx performs iterative MAP-based detection using contextual priors and channel likelihoods. Experiments demonstrate significant improvements in reconstructed sentence quality and strong rate-adaptive behavior under noisy channels. While this study focuses on text token transmission, extending the framework to multimodal token scenarios could be an interesting topic for future research.
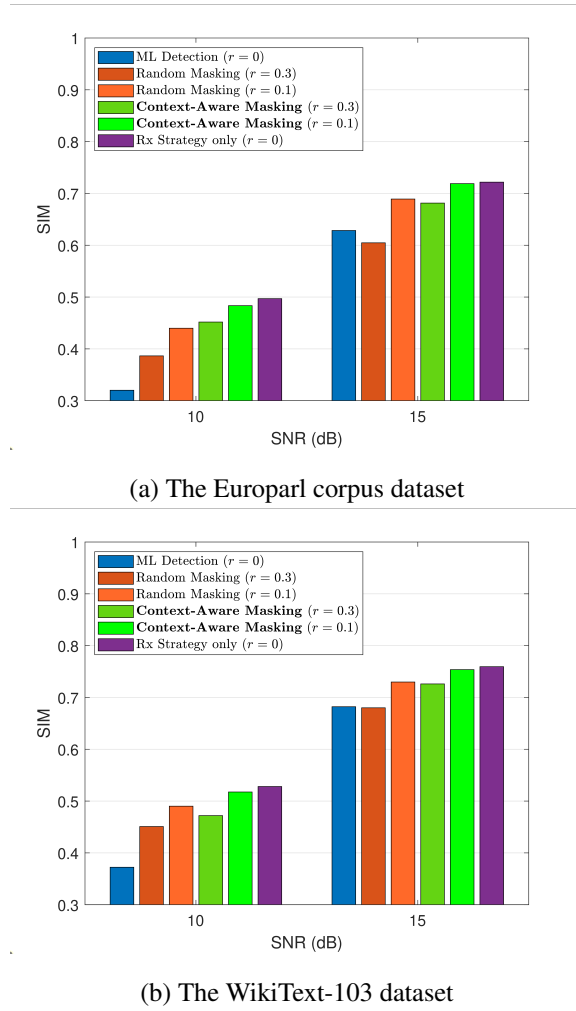
(a) The Europarl corpus dataset

(b) The WikiText-103 dataset

Figure 3: Performance comparison of the joint Tx–Rx strategy with different masking ratios and masking policies.

Acknowledgments

References

3GPP TSG RAN1 WG1. 2025. FLS on Evaluation Assumptions for 6GR Air Interface. Meeting report (#122, bengaluru, india). Document R1-2506548, "FLS#2 on evaluation assumptions for 6GR air interface".

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 4171–4186.

Fan, L.; Yang, J.; and Shen, C. 2025. Transformer-Based Wireless Symbol Detection Over Fading Channels. *arXiv preprint arXiv:2503.16594*. Version 1.

Goldsmith, A. 2005. *Wireless Communications*. Cambridge University Press.

Kang, Y.; Jeon, S.; Shin, J.; Jeon, Y.-S.; and Poor, H. V. 2025. MIMO Detection under Hardware Impairments: Data Augmentation with Boosting. *IEEE Transactions on Communications*. Early Access.

Koehn, P. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the MT Summit*.

Lee, S.; Park, J.; Choi, J.; and Park, H. 2025. Low-Complexity Semantic Packet Aggregation for Token Communication via Lookahead Search. *arXiv preprint arXiv:2506.19451*. Version 1.

Merity, S.; Xiong, C.; Bradbury, J.; and Socher, R. 2016. Pointer Sentinel Mixture Models. *arXiv preprint arXiv:1609.07843*.

Oh, S.; Kim, J.; Park, J.; Ko, S.-W.; Choi, J.; Quek, T. Q. S.; and Kim, S.-L. 2025. Communication-Efficient Hybrid Language Model via Uncertainty-Aware Opportunistic and Compressed Transmission. *arXiv preprint arXiv:2505.11788*. Version 1.

Qiao, L.; Mashhadi, M. B.; Gao, Z.; and Gündüz, D. 2025. Token-Domain Multiple Access: Exploiting Semantic Orthogonality for Collision Mitigation. *arXiv preprint arXiv:2502.06118*. Version 1.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings Using Siamese BERT Networks. In *2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992.

Shin, J.; Oh, Y.; Park, J.; Park, J.; and Jeon, Y.-S. 2025. ESC-MVQ: End-to-End Semantic Communication with Multi-Codebook Vector Quantization. *IEEE Transactions on Wireless Communications*. Early Access.

Tüchler, M.; Koetter, R.; and Singer, A. C. 2002. Turbo Equalization: Principles and New Results. *IEEE Transactions on Communications*, 50(5): 754–767.

Wu, Y.; Schuster, M.; Chen, Z.; et al. 2016. Google's Neural Machine Translation System: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.