Robustifying Token Communication Systems through Conformal Risk Control

Chenhao Wang[1], Zihan Chen[2], Tony Q.S. Quek[2], Howard H. Yang[1] *

[1]ZJU-UIUC Institute, Zhejiang University, Haining, China
[2]Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore
chenhao.22@intl.zju.edu.cn, zihan_chen@sutd.edu.sg, tonyquek@sutd.edu.sg, haoyang@intl.zju.edu.cn

Abstract

Token communication (TokCom) systems leverage the powerful semantic understanding of large language models (LLMs) to recover lost tokens, significantly enhancing communication efficiency and robustness against bursty losses. However, existing TokCom frameworks rely solely on LLM completion and completely fail to validate the correctness of the recovered tokens, leading to unacceptable error risks in critical tasks. To address this gap, we propose the *first* retransmission mechanism for TokCom systems. Unlike conventional retransmission mechanisms that unconditionally retransmit error tokens, our mechanism is based on Conformal Risk Control to learn a decision threshold with theoretical guarantees, enabling selective retransmission without any assumptions about the internal structure of the LLMs. Extensive simulation results on text and image tasks show that our method significantly improves communication robustness compared with conventional TokCom systems that rely solely on LLM completion, while requiring significantly fewer retransmission requests than traditional protocols.

Introduction

Motivation

With the development of 6G technology, the communication paradigm is shifting from traditional bit-level transmission to a new paradigm of semantic communication (Xie et al. 2021; Guo et al. 2025) focused on conveying meaning. This shift is largely driven by the rise of large language models (LLMs), which have provided unprecedented capabilities for semantic understanding, reasoning, and generation (Xia et al. 2025). These foundational models are not only used to comprehend semantics but also to enhance communication effectiveness at the semantic level, enabling generative semantic communication (Liang et al. 2025; Qiao et al. 2024). However, the generative semantic communication paradigm still relies on representing information as continuous, high-dimensional semantic feature vectors. While these vectors are effective at capturing high-level meanings, they suffer from significant drawbacks. They not only incur high communication overhead but are also vulnerable to factors like channel noise and model biases during transmission and reconstruction, which can ultimately lead to a loss of semantic fidelity.

A more promising and structured approach is to transmit information as discrete tokens. As the digital world increasingly relies on LLMs, this token-based data flow is poised to become a significant component of traffic in future communication systems. These tokens are the native data-processing units for Transformer-based LLMs (Vaswani et al. 2017), effectively bridging the gap between raw, unstructured data and high-level semantic abstractions. This fundamental insight gives rise to Token Communication (TokCom) (Qiao et al. 2025b; Wei et al. 2025; Qiao et al. 2025a). Within the TokCom framework, tokens serve a dual role as universal units for both information processing and wireless transmission. By sharing a common codebook, the transmitter only needs to send a concise sequence of discrete token IDs. Upon receipt, a powerful LLM at the receiver uses this sequence to reconstruct the original high-fidelity content. This architecture is also inherently robust to data loss. If a packet is corrupted or lost, resulting in missing tokens, the receiving LLM can infer the most likely values based on the surrounding context, effectively performing semantic completion.

However, relying on LLMs for semantic completion is not without its pitfalls. These preliminary approaches suffer from fundamental limitations that undermine system robustness, which define the primary challenges that TokCom system should overcome. First, a critical limitation of existing methods (Qiao et al. 2025b; Wei et al. 2025; Qiao et al. 2025a) is their fragility in noisy wireless environments where significant tokens are lost, erasing semantically vital information during the transmission. This limitation makes it challenging to ensure reliability when the LLM is forced to generate completions from highly ambiguous and fragmented input. When the model is forced to guess based on insufficient clues, it may fabricate contents, which results in outputs that appear plausible and grammatically sound, but are semantically detached from the original truth. This phenomenon is arguably more detrimental than traditional bit errors, as it creates a coherent but fundamentally incorrect understanding at the receiver. Second, a fundamental limitation of existing methods (Qiao et al. 2025b; Wei et al. 2025; Qiao et al. 2025a) is the inherent uncertainty regarding an

*Corresponding Author: Howard H. Yang
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

LLM's generalization capabilities. The semantic content of any real-time transmission is unknown in advance and may involve specialized, timely, or novel information that was not well-represented in the model's original training corpus. This limitation undermines the robustness of TokCom: we simply cannot guarantee that the LLM's pre-trained completion abilities will be effective or relevant for the specific data being transmitted. This unpredictable performance can easily lead to the model generating completions that are logically flawed or fundamentally inconsistent with the original semantics when encountering new contents.

Consequently, to guarantee end-to-end communication reliability, a mechanism for retransmitting specific tokens remains essential. The conflict lies in that conventional retransmission protocols designed for bit-level integrity are ill-suited for TokCom. These traditional protocols demand bit-perfect replication, forcing a full retransmission of any token that contains errors, without making any judgment on the uncertainty of the completed tokens. Such approaches negate the TokCom design, wasting the LLM's core ability for semantic completion and leaving the semantic information in the tokens unused. Furthermore, the black-box nature of these LLMs prevents us from directly determining which tokens have the most semantic impact and should be prioritized for retransmission. This raises the critical research question: *how can we design a retransmission scheme to achieve robust TokCom while minimizing retransmission overhead?*

Contribution

To answer the question, we proposes a robust retransmission mechanism for TokCom, innovatively applying the Conformal Risk Control (CRC) framework (Angelopoulos et al. 2024; Feldman et al. 2023) to token-level error recovery. This mechanism achieves highly reliable semantic communication with minimal retransmission cost under rigorous theoretical guarantees. To the best of our knowledge, this paper is the *first* to systematically study the retransmission problem in TokCom, aiming to design a token-level retransmission scheme that balances both reliability and efficiency.

The core challenge of our work lies in the fact that traditional CRC methods are primarily designed for classification or regression tasks and cannot be directly applied to TokCom's unique token-level retransmission problem. To address this, we perform a critical redesign of the CRC framework, enabling it to effectively assess and control the downstream task risk posed by token errors. Our specific contributions are as follows:

- We design a theoretically guaranteed token retransmission mechanism, achieving a robust TokCom system with only a partial retransmission of the erroneous tokens. This retransmission mechanism minimizes retransmissions as much as possible by fully leveraging the completion capability of TokCom while ensuring theoretical coverage.
- We leverage the CRC framework to formally derive a theoretical performance guarantee, ensuring that the task accuracy of our proposed algorithm meets a predefined confidence level.
- Through extensive simulations across multiple data modalities like text and images, we demonstrate that our proposed mechanism achieves reliable TokCom under different communication quality and requires significantly fewer retransmission requests than traditional retransmission protocols.

System Model

This section presents the system model for a typical task-oriented TokCom system, in which we shall introduce the transmitter, the wireless channel, and the receiver, respectively. Although TokCom systems can tolerate channel errors due to their powerful semantic completion capabilities, this completion is unreliable under noisy conditions or when critical information is lost. Therefore, our system introduces an additional retransmission decision unit at the receiver, which performs an extra determination on whether a token needs to be retransmitted. The overall objective is to obtain high-reliability task outcomes at the receiver side.

Transmitter

The transmitter is composed of a tokenizer and channel encoding and modulation components. We consider transmitting a dataset $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{N_d}\}$ of size N_d over a wireless communication system for the downstream task. For any source datum $\boldsymbol{x}_i$ in the dataset (where $i \in \{1, 2, \ldots, N\}$), the system first uses a pre-trained tokenizer $f(\cdot)$ to convert it into a discrete token sequence $\boldsymbol{T}_i = [T_{i,1}, T_{i,2}, \ldots, T_{i,L_i}]$, whose length is L_i corresponding to the given $\boldsymbol{x}_i$. Specifically, the tokenization process projects $\boldsymbol{x}_i$ from the original domain into a discrete space, then maps these representations to their corresponding indices in a learned codebook, yielding a sequence of discrete token representations. Through encoding and modulation, the obtained tokens are converted into the transmission signal $\boldsymbol{s}_i$, which is given by:

$$\boldsymbol{s}_i = g(f(\boldsymbol{x}_i)), \tag{1}$$

where $g(\cdot)$ represents the encoding and modulation process.

Wireless channel

The signal $\boldsymbol{s}_i$ is then transmitted through a wireless channel:

$$\boldsymbol{u}_i = h_c \boldsymbol{s}_i + \boldsymbol{n}, \tag{2}$$

where $\boldsymbol{n} \sim \mathcal{CN}\left(0, \sigma^2 \mathbf{I}\right)$ represents the additive white Gaussian noise with the noise variance σ^2 and h_c denotes the Rayleigh channel fading.

Receiver

The receiver consists of a demodulation and decoding module, a token completion LLM, a token conversion and task execution module, and a retransmission unit. The receiver first processes the incoming signal $\boldsymbol{u}_i$ with a demodulation and decoding function $\mathbf{d}(\cdot)$, resulting in a sequence of tokens $\hat{\boldsymbol{T}}_i$. Transmission errors may cause some tokens to be undecodable, which are marked as `[LOST]`.

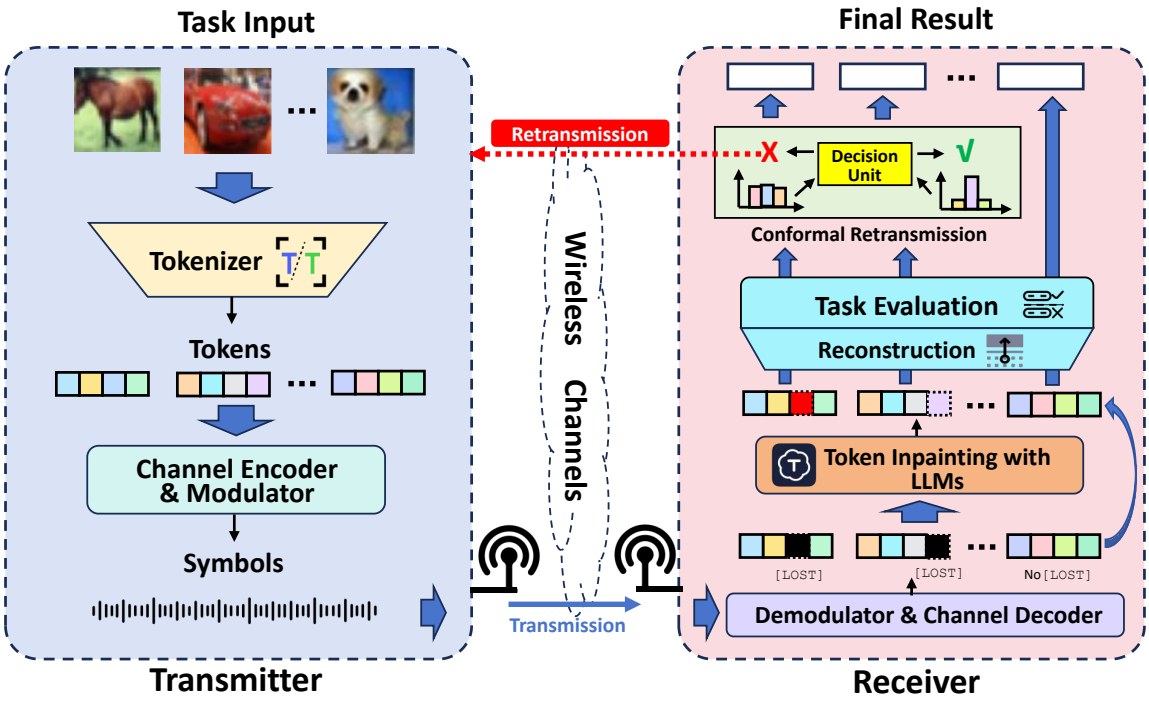

Figure 1: An overview of the TokCom system with conformal retransmission mechanism.

Rather than demanding retransmission, our token-based communication system uses an LLM to intelligently infer and restore these `[LOST]` tokens, saving bandwidth. The model, denoted as $M(\cdot)$, takes the incomplete sequence $\hat{\boldsymbol{T}}_i$ and outputs a complete version $\tilde{\boldsymbol{T}}_i$:

$$\tilde{\boldsymbol{T}}_i = M(\hat{\boldsymbol{T}}_i). \tag{3}$$

These restored tokens are then passed through a detokenizer to revert into the original data modality, which can be used to finish the intended tasks. We denote the above process as $h(\cdot)$. The entire recovery process for getting the task output can be summarized as:

$$(\hat{\boldsymbol{y}}_i, \boldsymbol{p}_i) = h\left(M\left(\mathbf{d}\left(\boldsymbol{u}_i\right)\right)\right), \tag{4}$$

where $\hat{\boldsymbol{y}}_i = \{\hat{y}_{i,1}, \hat{y}_{i,2}, \ldots, \hat{y}_{i,t}\}$ denotes the sequence of predicted labels for the task contents, with $t \in \{1, \ldots, L_i\}$ denotes the number of task targets in sample i, and $\boldsymbol{p}_i$ contains the corresponding confidence scores for those predictions. For example, if the task is token correctness prediction, $h(\cdot)$ outputs each completed token's predicted label together with its probability; if the task is image recognition, $h(\cdot)$ first decodes the completed token sequence into an image $\hat{\boldsymbol{x}}_i$ and then applies an image classifier to produce a predicted class label $\hat{\boldsymbol{y}}_i$ with $t = 1$ and its probability $\boldsymbol{p}_i$.

Retransmission decision unit

Although LLMs possess the capability to complete missing tokens at the receiver, these completions are inherently uncertain. The method to improve reliability is to retransmit every token detected as `[LOST]`, but this strategy incurs significant communication and computational overhead and also wastes contextual semantic information. Therefore, we introduce a robust token completion mechanism with a retransmission decision unit at the receiver. By leverage our CRC algorithm, this unit establishes a retransmission decision boundary based on the test results from the downstream task's calibration set, thereby preserving high-reliability tokens and retransmitting low-reliability ones. The process is then expressed as:

$$\tilde{\boldsymbol{T}}_i = M_{Re}(\hat{\boldsymbol{T}}_i), \tag{5}$$

where $M_{Re}(\cdot)$ denotes our robust token completion process with retransmission. This ensures that task accuracy requirements are met while substantially reducing unnecessary retransmissions.

Conformal Retransmission with Robustness

In this part we present our CRC algorithm used in the retransmission decision unit. CRC extends the principles of conformal prediction (Vovk, Gammerman, and Shafer 2022) to risk minimization problems, aiming to control the expected loss at a specified confidence level. It offers probabilistic guarantees on a given risk metric using a held-out calibration set to define decision thresholds, all without making distributional assumptions and effective even with finite samples. In particular, CRC can offer reliability guarantees without relying on the details of a model. This is especially important for scenarios where the receiver uses LLMs whose behaviors are often uninterpretable. The specific details of the algorithm are outlined below.

Threshold calibration

Algorithm 1: Calibration

Input: $\mathcal{C} = \{(\mathbf{c}_i, y_i^*)\}_{i=1}^N$, tokenizer f, position-error distribution E, LLM M, task operator h, confidence level α, boundary D
Output: λ^*

$\mathcal{P}, \mathcal{W} \leftarrow []$
for $i = 1, \ldots, N$ **do**
  $\mathbf{B}_i \leftarrow f(\mathbf{c}_i)$; $\mathcal{M}_i \sim E$
  $\widehat{\mathbf{B}}_i \leftarrow \text{mask}(\mathbf{B}_i, \mathcal{M}_i)$
  $(\hat{y}_i, \mathbf{p}_i) \leftarrow h\big(M(\widehat{\mathbf{B}}_i)\big)$
  for each prediction index j **do**
    append $p_{i,j}$ to $\mathcal{P}$; append $\mathbb{I}(\hat{y}_{i,j} \neq y_{i,j}^*)$ to $\mathcal{W}$
  end for
end for
$\mathcal{T} \leftarrow$ sorted grid $[0, 1]$
Define $l(\lambda') \leftarrow \sum_{k \in \mathcal{S}(\lambda')} \mathcal{W}[k] - \alpha|\mathcal{S}(\lambda')|$
for $\lambda \in \mathcal{T}$ **do**
  $\widehat{R} \leftarrow \sup\{l(\lambda') \mid \lambda' \in \mathcal{T}, \lambda' > \lambda\}$
  if $\frac{1}{N+1}\widehat{R} + \frac{D}{N+1} \leq 0$ **then**
    return λ
  end if
end for
return $+\infty$

In this section, we describe the procedure for determining a decision threshold via calibration. We begin with a calibration dataset $\mathcal{C} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_N\}$, which is exchangeable with the data intended for transmission. Each sample $\boldsymbol{c}_i$ in the dataset is converted into a token sequence $\boldsymbol{B}_i$ by a tokenizer f. Assume we have access to estimate E, the probability distribution of error locations for tokens that fail to decode during an actual wireless transmission. Using the position-error distribution E, we randomly mask tokens in each token sequence $\boldsymbol{B}_i$ from the dataset $\mathcal{C}$. Concretely, for each sample i we draw a mask position set $\mathcal{M}_i \sim E$ and form the masked sequence

$$\widehat{\boldsymbol{B}}_i = \text{mask}\left(\boldsymbol{B}_i, \mathcal{M}_i\right), \tag{6}$$

which means the tokens at positions in $\mathcal{M}_i$ are replaced with the special token `[LOST]`. Next, we feed the masked sequence $\widehat{\boldsymbol{B}}_i$ into the LLM M. The output is then passed through the task-oriented operator $h(\cdot)$ to obtain the predicted labels $\hat{\boldsymbol{y}}_i$ and their corresponding confidence scores $\boldsymbol{p}_i$:

$$(\hat{\boldsymbol{y}}_i, \boldsymbol{p}_i) = h\left(M\left(\widehat{\boldsymbol{B}}_i\right)\right), \tag{7}$$

All predicted target index pairs (i, j) are gathered into the set $\mathcal{H}$, and their corresponding confidence scores $p_{i,j}$ comprise the set $\mathcal{P}$.

To perform the calibration, we introduce a decision threshold $\lambda \in [0, 1]$. For a given λ, we define a set of retained predictions for each sample i, $\mathcal{S}_i(\lambda)$, which includes the index pairs (i, j) of all predictions whose confidence scores $p_{i,j}$ exceed λ. Formally,

$$\mathcal{S}_i(\lambda) = \{(i, j) \mid j \in \{1, \ldots, t\}, p_{i,j} > \lambda\}. \tag{8}$$

Algorithm 2: Robust Conformal Retransmission

Input: Partially received token sequences $\{\hat{\boldsymbol{T}}_i\}_{i=1}^{N_d}$, LLM M, task operator h, calibration dataset $\mathcal{C}$ (used to get λ^*).
Output: Final task outputs $\{\boldsymbol{y}_i^{\text{final}}\}_{i=1}^{N_d}$.

Obtain decision threshold λ^* by running calibration on $\mathcal{C}$
for $i = 1, \ldots, N_d$ **do**
  $(\hat{\boldsymbol{y}}_i, \boldsymbol{p}_i) \leftarrow h\big(M(\hat{\boldsymbol{T}}_i)\big)$
  $\mathcal{R}_i \leftarrow \{j \mid p_{i,j} \leq \lambda^*\}$
  if $\mathcal{R}_i \neq \varnothing$ **then**
    Request retransmission for `[LOST]` tokens that indices in $\mathcal{R}_i$, receive ground-truth tokens $\{T_{i,j}^*\}_{j \in \mathcal{R}_i}$
  end if
  $\tilde{\boldsymbol{T}}_i^{\text{final}} \leftarrow M_{\text{Re}}\big(\hat{\boldsymbol{T}}_i, \{T_{i,j}^*\}_{j \in \mathcal{R}_i}\big)$
  $\boldsymbol{y}_i^{\text{final}} \leftarrow h\big(\tilde{\boldsymbol{T}}_i^{\text{final}}\big)$
end for
return $\{\boldsymbol{y}_i^{\text{final}}\}_{i=1}^{N_d}$

The overall set of retained predictions, $\mathcal{S}(\lambda)$, is then the union of these individual sets over all samples, i.e., $\mathcal{S}(\lambda) = \bigcup_{i=1}^N \mathcal{S}_i(\lambda)$. Our goal is to find the optimal threshold λ^* such that the proportion of correct predictions within the retained set $S(\lambda^*)$ is no less than a predefined confidence level $1 - \alpha$. Simultaneously, we want λ^* to be the minimum threshold to maximize the number of retained predictions, which gives:

$$\lambda^* = \inf\left\{\lambda | \frac{N}{N+1}\widehat{R}(\lambda) + \frac{D}{N+1} \leq 0\right\}, \tag{9}$$

with

$$\widehat{R}(\lambda) = \frac{1}{N} \sup_{\lambda' > \lambda} \left(\sum_{(i,j) \in \mathcal{S}(\lambda')} \mathbb{I}\left(\hat{y}_{i,j} \neq y_{i,j}^*\right) - \alpha|\mathcal{S}(\lambda')| \right), \tag{10}$$

where $\mathbb{I}(\cdot)$ is the indicator function, $|\mathcal{S}(\lambda)|$ denotes the cardinality of $\mathcal{S}(\lambda)$, and $y_{i,j}^*$ is the true label. λ^* is then used as our final decision threshold. D is the boundary of a single sample. In the algorithmic process, we determine the value of λ^* using a grid search. After sorting all possible values, the first value found that satisfies the requirement is the desired solution. If no valid result is found, the algorithm returns infinity as the threshold to act as a safeguard.

Runtime retransmission

Once the calibration phase is concluded, TokCom is ready for data transmission. For every raw data instance $\boldsymbol{x}_i$, the LLM $M(\cdot)$ is first utilized to autocomplete the partially received token sequence $\hat{\boldsymbol{T}}_i$. The task mapping function $h(\cdot)$ then processes this sequence to generate a predicted label $\hat{\boldsymbol{y}}_i$ and its associated confidence score vector $\boldsymbol{p}_i$:

$$(\hat{\boldsymbol{y}}_i, \boldsymbol{p}_i) = h\left(M\left(\hat{\boldsymbol{T}}_i\right)\right). \tag{11}$$

Among the tokens requiring completion, we accept those with a confidence score $p_{i,j}$ exceeding a predefined threshold λ^*, while requesting retransmission for all others. More formally, the set of indices for retransmission, $\mathcal{R}_i$, is given by:

$$\mathcal{R}_i = \{j \mid p_{i,j} \leq \lambda^*\}, \tag{12}$$

and the set of samples retained after completion is $\mathcal{R}_i^c$, respectively. If $\mathcal{R}_i$ is empty, it means no retransmissions are required – the correctly received tokens and the completed tokens are sufficiently reliable.

Finally, a high-fidelity token sequence $\boldsymbol{T}_i^{\text{final}}$ is assembled. This sequence integrates the correctly transmitted tokens with both the high-confidence predicted tokens and the ground-truth tokens obtained via retransmission. This robust mechanism for sequence reconstruction is denoted as $M_{\text{Re}}(\cdot)$:

$$\tilde{\boldsymbol{T}}_i^{\text{final}} = M_{\text{Re}}\left(\hat{\boldsymbol{T}}_i, \left\{T_{i,j}^*\right\}_{j\in\mathcal{R}_i}\right), \tag{13}$$

where $T_{i,j}^*$ represents the ground-truth tokens obtained via retransmission. This error-corrected and verified sequence is then fed into the task mapping function $h(\cdot)$ to produce the final task output $\boldsymbol{y}_i^{\text{final}}$:

$$\boldsymbol{y}_i^{\text{final}} = h\left(\tilde{\boldsymbol{T}}_i^{\text{final}}\right). \tag{14}$$

Robustness Analysis

This section analyzes the robustness of the proposed method.

Performance guarantee target

A fundamental requirement for a reliable TokCom system is to provide a probabilistic performance guarantee. Specifically, for the set of all tokens that necessitate infilling due to decoding errors, we must ensure that the error rate of the associated downstream task is preserved below a predefined threshold with a confidence level of α, such that:

$$\Pr(\hat{y}_{i,j} \neq y_{i,j}^* \mid (i,j) \in \mathcal{R}) \leq \alpha, \tag{15}$$

where $\mathcal{R}$ is the union of $\mathcal{R}_i$ with $i \in \{1, 2, ..., M\}$.

Preliminaries

To formally prove that our method satisfies the guarantee above, we reframe the problem within the framework of CRC. The core idea is to translate the conditional probability guarantee into a problem of controlling the average of a carefully chosen *loss function*. The loss function in CRC quantifies the penalty associated with a particular outcome. Our strategy is to design a loss function $L(\lambda)$, such that its expected value being non-positive, i.e., $\mathbb{E}[L(\lambda)] \leq 0$, is equivalent to our original goal above.

By doing so, we can leverage the powerful theoretical tools of CRC, which provide a practical recipe for finding a threshold λ^* that guarantees control over this expected loss on new data.

Our analysis relies on the framework of CRC. The key assumptions (Angelopoulos et al. 2024) stated below operate on a generic, user-defined loss function $L_i(\lambda)$ for each sample i, which depends on a decision threshold λ.

Assumption 1 (Exchangeability) Calibration sets $\mathcal{C}$ and test sets $\mathcal{X}$ is exchangeable.

Assumption 2 (Right-continuity). For each sample i, the loss function $L_i(\lambda)$ is right-continuous in λ.

Assumption 3 (Boundedness and feasibility). At the maximum possible threshold value $\lambda_{\max}$, the risk L_i must already be at or below the target level α. The risk function $L_i(\lambda)$ is almost surely bounded by D for all λ.

Thus, we present the loss function corresponding to the retransmission method.

Lemma 1. Given the proposed Algorithm 1 & 2, the loss function of our system is

$$L_{i,j}(\lambda) = \mathbb{I}\left(\hat{y}_{i,j} \neq y_{i,j}^* \text{ and } p_{i,j} > \lambda\right) - \alpha \cdot \mathbb{I}\left(p_{i,j} > \lambda\right), \tag{16}$$

and the empirical total loss function

$$\widehat{R}(\lambda) = \frac{1}{N} \sup_{\lambda' > \lambda} \left(\sum_{(i,j)\in\mathcal{H}} L_{i,j}(\lambda') \right) \tag{17}$$

is non-increasing.

Robustness

In this part, we prove the robustness of the proposed method. We begin with presenting the specific form of the loss function for the proposed method.

Next, we prove the robustness guarantee of our method.

Theorem 1 (Robustness Guarantee of Conformal Retransmission). Let Assumption 1-3 hold. For the loss function $L_{i,j}(\lambda)$ and $\widehat{R}(\lambda)$, when $\hat{\lambda}$ satisfies

$$\hat{\lambda} = \inf\left\{\lambda \,\middle|\, \frac{N}{N+1}\widehat{R}(\lambda) + \frac{D}{N+1} \leq 0\right\}, \tag{18}$$

we have $\mathbb{E}\left[L_{new}(\hat{\lambda})\right] \leq 0$, where $L_{new}(\hat{\lambda})$ denotes the new test point in $\mathcal{R}$.

Proof. For simplicity of notation, we denote the size of the calibration set $|\mathcal{H}|$ as K, and index each element by $i \in [K]$. Let $\widehat{R}_{K+1}(\lambda) = \sup_{\lambda_0 \geq \lambda} \left(L_1(\lambda_0) + \ldots + L_K(\lambda_0) + L_{new}(\lambda_0)\right)/(K+1)$ and $\hat{\lambda}' = \inf\left\{\lambda \in \Lambda : \widehat{R}_{K+1}(\lambda) \leq 0\right\}$. Since $L_{new}(\lambda) \leq D$, we know that

$$\widehat{R}_{K+1}(\lambda) \leq \frac{K}{K+1}\widehat{R}_K(\lambda) + \frac{D}{K+1}. \tag{19}$$

By the definition of $\hat{\lambda}$, we clearly have $\hat{\lambda}' \leq \hat{\lambda}$. Let E be the multiset of loss functions $\{L_1, \ldots, L_{new}\}$, then $\hat{\lambda}$ is a constant conditional on E. Additionally, according to the Assumption 1, $L_{new}(\lambda) \mid E \sim \text{Uniform}\left(\{L_1, \ldots, L_{new}\}\right)$ by exchangeability. Then

$$\mathbb{E}\left[L_{new}\left(\hat{\lambda}\right) \mid E\right] = \frac{1}{n+1}\sum_{i=1}^{K+1} L_i\left(\hat{\lambda}\right) \leq \widehat{R}_{K+1}(\hat{\lambda}) \tag{20}$$

According to Lemma 1, with $\hat{\lambda}' \leq \hat{\lambda}$, we have

$$\mathbb{E}\left[L_{new}\left(\hat{\lambda}\right) \mid E\right] \leq \widehat{R}_{K+1}(\hat{\lambda}) \leq \widehat{R}_{K+1}(\hat{\lambda}') \leq 0 \quad (21)$$

The proof is completed by the law of total expectation.

Remark 1. *Theorem 1 shows that for newly arriving tokens outside the calibration set, by selecting a compliant threshold λ, we can still guarantee that its expectation on our loss function is less than 0. Furthermore, by utilizing the properties of the indicator function, we can convert this expectation into a guarantee on the error probability for these new tokens, which is equivalent to the robustness objective set at the beginning of this section.*

Simulation Result

In this section, we conduct experiments to validate the effectiveness of the proposed conformal retransmission method under various conditions. We begin by detailing our experimental setup. Subsequently, we demonstrate the robustness and efficiency of the proposed method by evaluating its performance on both text and image tasks. Finally, we investigate the impact of different system parameters on the algorithm's performance.

Experimental setup

We define the Token Error Rate (TER) as the fraction of transmitted tokens that are received incorrectly. To simplify simulation, token errors are assumed independent, which means a common approximation achievable in practice via sufficient interleaving. Besides, any retransmitted tokens are treated as perfectly corrected.

Text experiments target correct token recovery. We use a pretrained BERT-Large with its WordPiece tokenizer (Devlin et al. 2019) to impute missing tokens. The evaluation set contains 20,000 sentences sampled from the BROWN corpus (Francis and Kučera 1979), split evenly into a 10,000-sentence calibration set and a 10,000-sentence test set.

Image experiments convert images to discrete tokens with a VQ-VAE(Van den Oord, Vinyals, and Kavukcuoglu 2017) at the transmitter; the receiver reconstructs images from the transmitted token indices using the same codebook. Missing image tokens are filled by a pretrained MaskGIT (Chang et al. 2022) and then decoded back into images via the VQ-VAE de-tokenizer. Downstream evaluation is image classification on Imagenette (Howard 2019): a pretrained ResNet-18 (He et al. 2016) has its classification head fine-tuned on 400 images; separately, 400 images are used for calibration and 400 for testing.

Text transmission

Fig. 2 illustrates the performance comparison between conventional TokCom and our proposed TokCom integrated with the conformal retransmission mechanism, evaluated under various TER. As shown in the figure, given a fixed confidence level $\alpha = 0.1$, our method demonstrates significantly superior accuracy compared to conventional TokCom, as evidenced by its performance on both the complete set of tokens and the specific tokens requiring imputation. Furthermore, as the TER increases, indicative of more adverse channel conditions, our approach maintains a high level of robustness. In contrast, the performance of conventional TokCom degrades markedly under these same conditions.

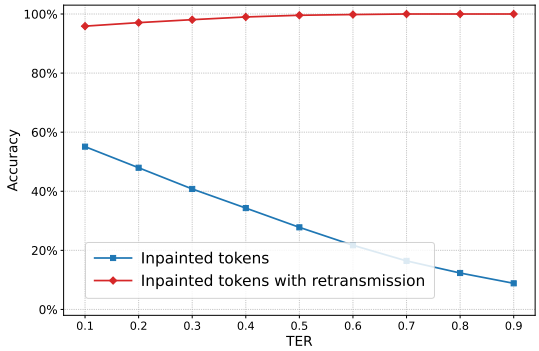

Figure 2: Performance comparison for both accuracy of all tokens and accuracy of inpainted tokens on the text task. Here we choose $\alpha = 0.1$.

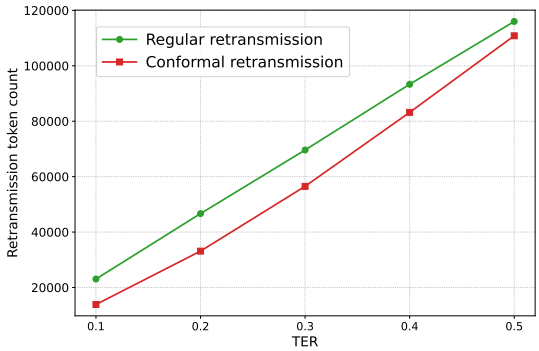

Figure 3: Performance comparison for retransmission token count on the text task. Here we choose $\alpha = 0.1$.

In Fig. 3, we provide a comparison of the resource consumption between traditional retransmission method and the proposed conformal retransmission mechanism. The results demonstrate that under better channel conditions, where the TER is low, the proposed method requires significantly fewer retransmissions than the conventional method. As the TER increases, excessive token errors prevent the transmission of sufficient valid information, rendering the LLM at the receiver incapable of effective imputation. In this situation, the retransmission requests for the proposed method increase, but it still significantly outperforms the traditional retransmission mechanism.

Image recognition

Fig. 4 shows the performance of the conformal retransmission method on the downstream image recognition task. As can be seen, under various TER conditions, the accuracy of our method closely approaches the baseline result (achieved using the original, error-free images) and significantly surpasses that of conventional TokCom. This performance advantage becomes more pronounced as the TER

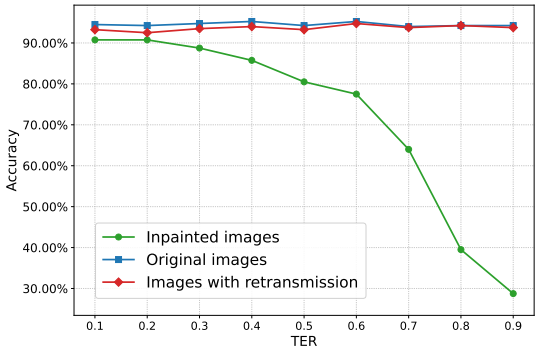

Figure 4: Performance comparison for both accuracy of all tokens and accuracy of inpainted tokens on the image classification task. Here we choose $\alpha = 0.1$.

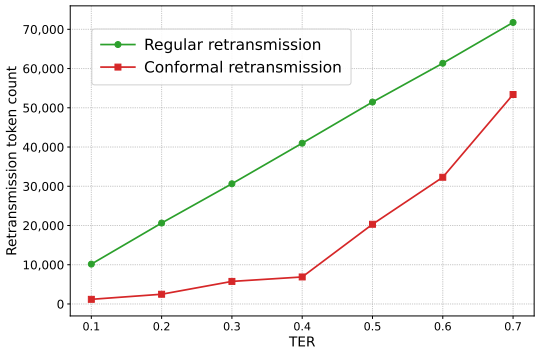

Figure 5: Performance comparison for retransmission token count on the image classification task task. Here we choose $\alpha = 0.1$.

increases, clearly demonstrating the robustness of the proposed method.

Fig. 5 presents results similar to those in Fig. 3. At lower TERs, our method substantially reduces retransmission requests compared to conventional retransmission algorithms. Conversely, at higher TERs, the downstream large model cannot effectively impute the large volume of erroneous tokens based on the few correctly transmitted ones; consequently, the retransmission requests get higher, but still less than conventional method. These findings demonstrate the effectiveness and robustness of the proposed method, highlighting its ability to adapt intelligently to transmission scenarios under diverse channel conditions.

Effects of system parameters

We analyze the impact of the system parameter α on performance. As shown in Fig. 6 for $\alpha = 0.1$, the accuracy of the token subset retained by the conformal retransmission method remains at 90% (a 10% error rate). This result matches the pre-defined α and validates our theoretical robustness analysis.The selection threshold λ increases with the TER, indicating that stricter criteria are necessary to maintain reliability under adverse channel conditions. When the TER is excessively high, λ reaches its upper bound, causing the system to retain no tokens; consequently, no accuracy value is plotted in this regime.

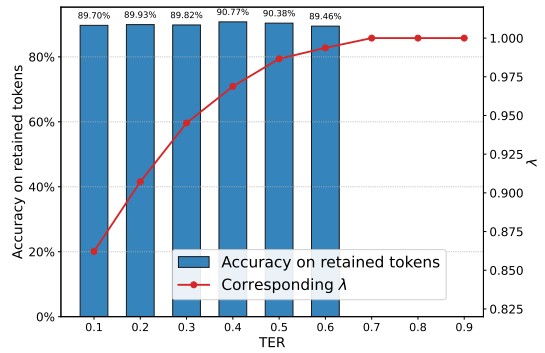

Figure 6: Performance comparison for accuracy on retained tokens and λ with $\alpha = 0.1$.

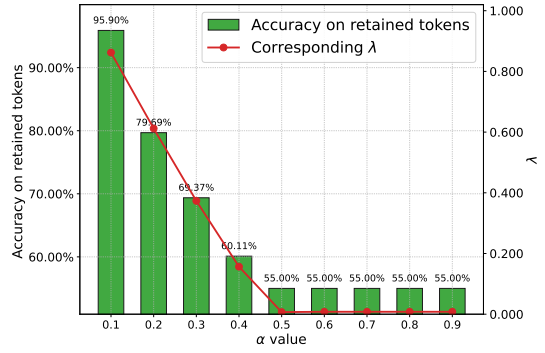

Figure 7: Performance comparison for accuracy on retained tokens and λ on different α.

Fig. 7 presents the accuracy of the retained tokens for various α values under a fixed TER of 0.1. When α is small, the results align with expectations, achieving the specified coverage level $(1 - \alpha)$. However, when α is large, the accuracy obtained before retransmission already exceeds the target $1 - \alpha$. In this scenario, the algorithm does not request any retransmissions, and the threshold λ is set to 0.

Conclusion

We addressed a core robustness problem in TokCom: LLM-based semantic completion for lost tokens could be unreliable. We proposed a novel and robust retransmission mechanism based on CRC. Our method leveraged a calibration set to determine a runtime threshold for filtering low-confidence completions and requesting retransmissions. We proved that this approach provided statistical guarantees for end-to-end task performance. Extensive simulations on text and image tasks confirmed our method maintained high reliability under varying communication quality with far fewer retransmissions than conventional methods. In summary, our work provided an adaptive, efficient solution for making TokCom systems more reliable.

References

Angelopoulos, A. N.; Bates, S.; Fisch, A.; Lei, L.; and Schuster, T. 2024. Conformal Risk Control. In *The Twelfth International Conference on Learning Representations*.

Chang, H.; Zhang, H.; Jiang, L.; Liu, C.; and Freeman, W. T. 2022. MaskGIT: Masked Generative Image Transformer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11305–11315.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NA-ACL)*, 4171–4186. Minneapolis, Minnesota.

Feldman, S.; Ringel, L.; Bates, S.; and Romano, Y. 2023. Achieving Risk Control in Online Learning Settings. *Trans. Mach. Learn. Res.*, 2023.

Francis, W. N.; and Kučera, H. 1979. Brown Corpus of Standard American English (1961). Department of Linguistics, Brown University.

Guo, S.; Wang, Y.; Zhang, N.; Su, Z.; Luan, T. H.; Tian, Z.; and Shen, X. 2025. A Survey on Semantic Communication Networks: Architecture, Security, and Privacy. *IEEE Communications Surveys & Tutorials*, 27(5): 2860–2894.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Howard, J. 2019. Imagenette: A smaller subset of 10 easily classified classes from the ImageNet dataset. fast.ai.

Liang, C.; Du, H.; Sun, Y.; Niyato, D.; Kang, J.; Zhao, D.; and Imran, M. A. 2025. Generative AI-Driven Semantic Communication Networks: Architecture, Technologies, and Applications. *IEEE Transactions on Cognitive Communications and Networking*, 11(1): 27–47.

Qiao, L.; Mashhadi, M. B.; Gao, Z.; Foh, C. H.; Xiao, P.; and Bennis, M. 2024. Latency-Aware Generative Semantic Communications With Pre-Trained Diffusion Models. *IEEE Wireless Communications Letters*, 13(10): 2652–2656.

Qiao, L.; Mashhadi, M. B.; Gao, Z.; Schober, R.; and Gündüz, D. 2025a. ToDMA: Large Model-Driven Token-Domain Multiple Access for Semantic Communications. arXiv:2505.10946.

Qiao, L.; Mashhadi, M. B.; Gao, Z.; Tafazolli, R.; Bennis, M.; and Niyato, D. 2025b. Token Communications: A Large Model-Driven Framework for Cross-Modal Context-Aware Semantic Communications. *IEEE Wireless Communications*, 32(5): 80–88.

Van den Oord, A.; Vinyals, O.; and Kavukcuoglu, K. 2017. Neural discrete representation learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 6309–6318. Red Hook, NY, USA. ISBN 9781510860964.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Vovk, V.; Gammerman, A.; and Shafer, G. 2022. *Algorithmic Learning in a Random World*. Cham, Switzerland: Springer Cham. ISBN 978-3-031-06649-8. EBook ISBN: 978-3-031-06649-8; Hardcover ISBN: 978-3-031-06648-1; Softcover ISBN: 978-3-031-06651-1.

Wei, H.; Ni, W.; Wang, W.; Xu, W.; Niyato, D.; and Zhang, P. 2025. Token Communication in the Era of Large Models: An Information Bottleneck-Based Approach. arXiv:2507.01728.

Xia, L.; Sun, Y.; Liang, C.; Zhang, L.; Imran, M. A.; and Niyato, D. 2025. Generative AI for Semantic Communication: Architecture, Challenges, and Outlook. *IEEE Wireless Communications*, 32(1): 132–140.

Xie, H.; Qin, Z.; Li, G. Y.; and Juang, B.-H. 2021. Deep Learning Enabled Semantic Communication Systems. *IEEE Transactions on Signal Processing*, 69: 2663–2675.