R. Ishizuka, A. E. Kalør, and T. Ohtsuki, "Contextual Soft-Decision Token Detection via LLM-Based Generative Semantic Prior," IEEE Global Communications Conference, 2026.

Contextual Soft-Decision Token Detection via LMM-Based Generative Semantic Prior

Rei Ishizuka
Graduate School of Science and Technology
Keio University, Japan
Email: ishizuka@ohtsuki.ics.keio.ac.jp

Anders E. Kalør
Department of Information and Computer Science
Keio University, Japan
Email: aek@keio.jp

Tomoaki Ohtsuki
Department of Information and Computer Science
Keio University, Japan
Email: ohtsuki@keio.jp

***Abstract*—Token communication has emerged as a promising semantic communication framework that leverages Large Multimodal Models (LMMs) and discrete tokenizers to enable efficient transmission of high-dimensional data, such as text, images, and audio. This paper introduces a token detection technique at the receiver that fuses physical-layer Log-Likelihood Ratios (LLRs) with contextual priors generated by an LMM. A key feature of the proposed method is that the encoder does not require an LMM, making it suitable for edge inference scenarios with resource-constrained transmitters. Simulation results using a visual LMM demonstrate that the proposed method consistently outperforms baseline physical-only decoding, achieving a 2 dB E_b/N_0 gain at a Token Error Rate (TER) below 0.1. This leads to significant enhancements in semantic fidelity, particularly in the medium E_b/N_0 regime (1 dB $\leq E_\mathrm{b}/N_0 \leq 5$ dB), where pixel-level fidelity (PSNR) improves by up to 9.5 dB, perceptual similarity (LPIPS) is reduced by more than 70%, and semantic fidelity (CLIP score) improves by more than 60%. These results indicate that LMM-aided soft-decision token detection is an effective and practical approach for robust semantic communication in future 6G networks, as the transmitter requires only a tokenizer, making it readily deployable in resource-constrained edge devices.**

***Index Terms*—Token Communication, Semantic Communication, Joint Semantic-Channel Decoding, Large Language Models, Large Multimodal Models, 6G Wireless Networks**

I. Introduction

The exponential growth of IoT devices and data-intensive applications, such as autonomous driving and smart-city services, is pushing wireless systems toward the Shannon limit. To address this challenge, Semantic Communication (SemCom) has emerged as a paradigm shift that aims to preserve the *meaning* of transmitted data instead of bit-level fidelity [1].

Early SemCom frameworks, relying on end-to-end trained Deep Joint Source-Channel Coding (DeepJSCC) [2], [3], have demonstrated remarkable reconstruction fidelity and robustness. More recently, Generative AI-driven SemCom approaches utilizing GANs [4] and Diffusion Models [5], [6] have been investigated, achieving a superior perception-rate trade-off by synthesizing high-fidelity content from compact latent representations. However, these frameworks typically rely on analog transmission of continuous-space features, which is not compatible with existing digital infrastructures. Furthermore, they often require extensive task-specific training and are limited to specific data domains.

To bridge this gap, Token-based Communication (TokCom) [7]–[9] has recently been proposed, inspired by the success of tokenization as the universal representation for Generative Foundation Models (GFMs) [10]–[12]. By leveraging pre-trained tokenizers and Large Multimodal Models (LMMs), TokCom enables a training-free and modality-agnostic framework. Its discrete token representation ensures compatibility with digital systems while utilizing the contextual reasoning of LMMs to maintain semantic integrity at the token level.

The advantages of TokCom have been previously demonstrated for semantic error correction. For instance, in [7], additional reliable information, such as text labels or audio, is utilized to assist in the restoration of corrupted image tokens, assuming the additional information is transmitted via an error-free control channel. Additionally, TokCom-UEP [8] employs Expanding Window Fountain codes to provide prioritized protection of semantically critical tokens based on their importance.

Despite the advantages of error correction, existing TokCom architectures usually have a disjointed, sequential design. In these systems, the physical-layer decoder performs hard decision, which effectively transforms the physical channel into an erasure channel from the perspective of the LMM. In this model, any unrecovered tokens are represented by a special [MASK] token. The resulting token sequence is then passed to the LMM, which tries to restore the erased tokens solely on semantic context. However, this disjointed design discards the symbol likelihoods from the physical layer, forcing the LMM to correct errors without any guidance from the underlying signal reliability.

Motivated by these observations, this paper introduces a joint semantic-channel token detection scheme that probabilistically fuses physical-layer likelihoods with LMM-based generative semantic priors through Maximum A Posteriori (MAP) estimation. The scheme effectively resolves noise-induced ambiguities by directly incorporating the predictive power of LMMs into the symbol detection rule without requiring additional parity bits. Extensive simulations reveal that our joint detection scheme yields substantial performance gains in both TER and perceptual metrics over physical-layer likelihood-based approaches.

The rest of this paper is organized as follows. Section II

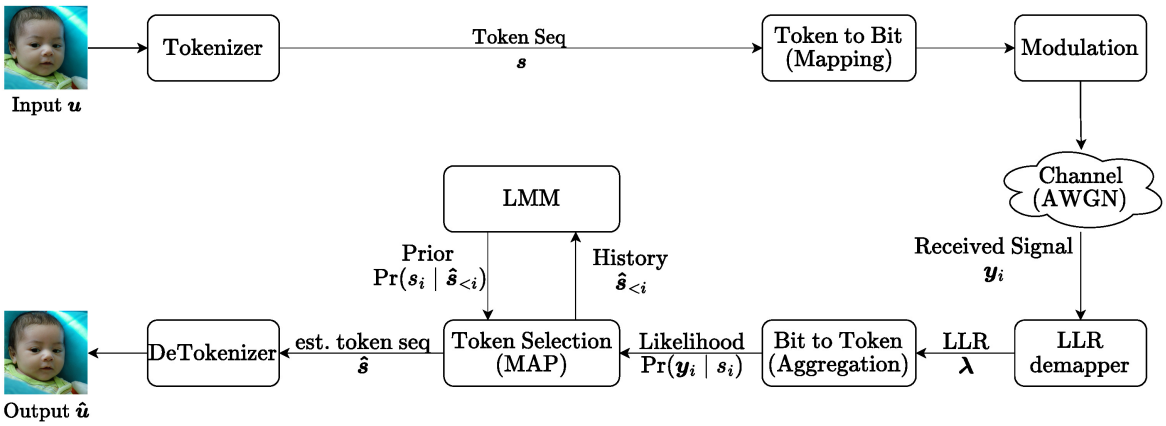


Fig. 1. System model of the proposed joint semantic-channel token detection. The receiver integrates physical-layer LLRs with LMM-based generative semantic priors to enhance token detection performance.

describes the system model. Section III details the proposed joint detection method. Section IV presents the simulation results, and Section V concludes the paper.

II. System Model

We consider the system depicted in Fig. 1, in which a transmitter encodes a generic input source $\mathbf{u}$, such as an image, into a sequence of tokens that are transmitted over an Additive White Gaussian Noise (AWGN) channel. At the receiver, physical-layer soft information in the form of LLRs is extracted from the noisy observations and probabilistically fused with contextual priors generated by an LMM to reconstruct the most likely token sequence $\hat{\mathbf{s}}_{1:N_t}$.

A. Transmitter

Given the generic input $\mathbf{u}$, the transmitter first obtains the discrete token sequence $\mathbf{s}_{1:N_t} = (s_1, \ldots, s_{N_t}) \in \{1, 2, \ldots, 2^L\}^{N_t}$ via a tokenizer, where s_i denotes the index of the i-th token in the codebook, and N_t is the total number of tokens. The tokenization process can be written as:

$$\mathbf{s}_{1:N_t} = \text{tokenizer}(\mathbf{u}), \tag{1}$$

where 2^L denotes the size of the token alphabet.

For digital transmission, each token index s_i is then mapped to a binary sequence $\mathbf{b}_i = (b_{i,1}, \ldots, b_{i,L}) \in \{0, 1\}^L$ of length L. This bit sequence is then encoded by a channel code, such as a low-density parity-check (LDPC) code, with code rate R, producing a coded bit sequence of length L/R. The resulting coded bit sequence is mapped to a complex symbol sequence $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,N_s}) \in \mathbb{C}^{N_s}$ using 2^m-ary modulation, where m denotes the bits per symbol (e.g., $m = 2$ for QPSK). The number of symbols per token is given by $N_s = \lceil L/(R \cdot m) \rceil$. To ensure a fair comparison across different schemes, we normalize the average signal power such that the energy per information bit is $E_\text{b} = 1$. Consequently, the average symbol energy is normalized to $\mathbb{E}[|x_{i,n}|^2] = R \cdot m$.

B. Physical Channel

We assume that the symbols are transmitted over an AWGN channel, which allows us to focus on the LMM's capability to perform soft-decision decoding. The received signal $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,N_s})$ corresponding to the i-th token can be written as:

$$\mathbf{y}_i = \mathbf{x}_i + \mathbf{n}_i, \tag{2}$$

where $\mathbf{n}_i \sim \mathcal{CN}(0, \sigma^2 \mathbf{I})$ and σ^2 denotes the variance of the AWGN noise. To evaluate the system performance, we adopt the ratio of bit energy to noise power spectral density E_b/N_0. Assuming normalized energy per bit $E_\text{b} = 1$, we have $E_\text{b}/N_0 = 1/\sigma^2$.

C. Receiver

The receiver extracts per-information-bit LLRs $\boldsymbol{\lambda}_i = (\lambda_{i,1}, \ldots, \lambda_{i,L})$ from the received signal $\mathbf{y}_i$. For the uncoded case ($R = 1$), these are obtained directly from the symbol demapper. For $R < 1$, the demapper first computes LLRs for the L/R coded bits, which are then passed through a soft-input-soft-output channel decoder (e.g., belief propagation for LDPC codes) to yield the L information bit LLRs $\boldsymbol{\lambda}_i$.

Next, the receiver performs joint semantic-channel detection. To this end, we assume the receiver has access to a pre-trained LMM that takes as input a sequence of tokens and outputs a predictive distribution of the next token in the sequence. The receiver aggregates the LLRs to compute likelihoods for each candidate token s_i and fuses them with the generative semantic priors $\Pr(s_i \mid \hat{\mathbf{s}}_{1:i-1})$ predicted by the LMM. This allows the decoder to map the LLRs to the most likely token sequence $\hat{\mathbf{s}}_{1:N_t}$ by leveraging both signal reliability and semantic context.

Finally, the detokenizer approximately reconstructs the $\hat{\mathbf{u}}$ from the estimated token sequence:

$$\hat{\mathbf{u}} = \text{detokenizer}(\hat{\mathbf{s}}_{1:N_t}). \tag{3}$$

The overall objective of the receiver is to minimize the expected distortion between the original source $\mathbf{u}$ and its reconstruction $\hat{\mathbf{u}}$:

$$\hat{\mathbf{u}} = \arg\min_{\mathbf{u}'} \mathbb{E}[d(\mathbf{u}, \mathbf{u}') \mid \mathbf{y}_{1:N_t}], \tag{4}$$

where $d(\cdot)$ denotes a distortion function. In the context of image transmission, this may capture pixel-level fidelity or higher-level perceptual and semantic quality.

III. LMM-Based Joint Semantic-Channel Token Detection

In this section, we introduce LMM-based joint semantic-channel token detection. We first formulate the problem as a MAP estimation that integrates physical-layer likelihoods with LMM-based generative semantic priors. Then, we calculate the computational and memory complexity of the proposed method.

A. Problem Formulation via MAP Estimation

As established in Section II-C, the receiver aims to minimize the expected source-level distortion $d(\mathbf{u}, \hat{\mathbf{u}})$. To formulate a tractable detection problem, we use the token error as a proxy for this objective. Minimizing the sequence-level error probability is mathematically equivalent to finding the token sequence that maximizes the joint posterior probability. This leads to the MAP estimation for the entire token sequence:

$$\hat{\mathbf{s}} = \arg\max_{\mathbf{s} \in \mathcal{V}^{N_t}} \Pr(\mathbf{s} \mid \mathbf{y}_{1:N_t}), \tag{5}$$

where $\hat{\mathbf{s}} = (\hat{s}_1, \ldots, \hat{s}_{N_t})$ is the estimated token sequence, $\mathbf{y}_{1:N_t} = (\mathbf{y}_1, \ldots, \mathbf{y}_{N_t})$ is the entire sequence of received channel observations, and $\mathcal{V} = \{1, \ldots, V\}$ represents the token vocabulary of size $V = 2^L$.

By applying Bayes' theorem, the joint posterior probability can be decomposed into the channel likelihood and the generative semantic prior of the sequence:

$$\Pr(\mathbf{s} \mid \mathbf{y}_{1:N_t}) \propto \Pr(\mathbf{y}_{1:N_t} \mid \mathbf{s}) \Pr(\mathbf{s}). \tag{6}$$

Assuming a memoryless channel, the sequence likelihood is simply the product of individual token likelihoods, $\Pr(\mathbf{y}_{1:N_t} \mid \mathbf{s}) = \prod_{i=1}^{N_t} \Pr(\mathbf{y}_i \mid s_i)$. Furthermore, using the chain rule of probability, the generative semantic prior of the entire sequence can be expressed as $\Pr(\mathbf{s}) = \prod_{i=1}^{N_t} \Pr(s_i \mid \mathbf{s}_{1:i-1})$. Consequently, the sequence-level MAP objective expands to:

$$\hat{\mathbf{s}} = \arg\max_{\mathbf{s} \in \mathcal{V}^{N_t}} \prod_{i=1}^{N_t} \Pr(\mathbf{y}_i \mid s_i) \Pr(s_i \mid \mathbf{s}_{1:i-1}). \tag{7}$$

Solving (7) exactly requires an exhaustive search over the exponentially large space of V^{N_t} possible sequences, which is computationally intractable. To achieve a practical implementation that aligns with the autoregressive generation mechanism of the LMM, we adopt a greedy decoding approach. By recursively making a hard decision for each token i and conditioning the next step on the previously estimated sequence $\hat{\mathbf{s}}_{1:i-1}$, the sequence-level optimization dramatically simplifies into the following token-by-token joint detection criterion:

$$\hat{s}_i = \arg\max_{k \in \mathcal{V}} \underbrace{\Pr(\mathbf{y}_i \mid s_i = k)}_{\text{symbol likelihood}} \times \underbrace{\Pr(s_i = k \mid \hat{\mathbf{s}}_{1:i-1})}_{\text{generative semantic prior}}. \tag{8}$$

for $i = 1, \ldots, N_t$. This formulation allows us to efficiently integrate the physical-layer soft information with the LMM-based generative semantic prior in a sequential manner.

B. Derivation of Symbol Likelihood

The symbol likelihood $\Pr(\mathbf{y}_i \mid s_i = k)$ represents the probability of observing $\mathbf{y}_i$ given that token k was transmitted. For the channel in (2), assuming that the individual information bits are equally likely (i.e., $\Pr(b_{i,l} = 0) = \Pr(b_{i,l} = 1)$) and mutually independent, the token likelihood is proportional to the product of the individual information bit posteriors:

$$\Pr(\mathbf{y}_i \mid s_i = k) \propto \prod_{l=1}^{L} \Pr(b_{i,l} = b_l^{(k)} \mid \mathbf{y}_i), \tag{9}$$

where $b_l^{(k)} \in \{0, 1\}$ is the l-th information bit of candidate token k. Here, $\Pr(b_{i,l} \mid \mathbf{y}_i)$ denotes the posterior of the information bit, which is obtained directly from the demapper for $R = 1$, or from a soft-output channel decoder (e.g., belief propagation for LDPC codes) for $R < 1$.

For practical implementation in digital receivers, these probabilities are represented as LLRs. The LLR for each information bit l is defined as:

$$\lambda_{i,l} = \ln \frac{\Pr(b_{i,l} = 1 \mid \mathbf{y}_i)}{\Pr(b_{i,l} = 0 \mid \mathbf{y}_i)}. \tag{10}$$

The probability that the l-th information bit equals $b_l^{(k)}$ is given by the inverse logit (sigmoid) of the LLR, denoted as $\sigma((2b_l^{(k)} - 1)\lambda_{i,l})$, where $\sigma(x) = (1 + e^{-x})^{-1}$. Thus, the token likelihood can be calculated using the LLRs as:

$$\Pr(\mathbf{y}_i \mid s_i = k) \propto \prod_{l=1}^{L} \sigma\left((2b_l^{(k)} - 1)\lambda_{i,l}\right). \tag{11}$$

Note that while this independence assumption holds exactly for the uncoded case ($R = 1$), parity constraints introduced by the channel code create statistical dependencies among the decoded bit posteriors for $R < 1$. Therefore, (11) should be interpreted as a practical approximation that is commonly adopted in soft-decision receivers and is sufficiently accurate for MAP token ranking.

C. LMM-Based Generative Semantic Prior Generation

A pre-trained LMM autoregressively predicts the next token from past context, producing the generative semantic prior $\Pr(s_i = k \mid \hat{\mathbf{s}}_{1:i-1})$. To avoid linearly growing computational and memory overhead, we bound the context to a sliding window of size W.

The LMM $\mathcal{F}_\theta$ maps the windowed context to a logit vector $\mathbf{o}_i = \mathcal{F}_\theta(\hat{\mathbf{s}}_{i-W:i-1}) \in \mathbb{R}^V$, from which the prior probability is obtained via:

$$\Pr(s_i = k \mid \hat{\mathbf{s}}_{1:i-1}) = \frac{\exp(o_{i,k}/\tau)}{\sum_{j=1}^{V} \exp(o_{i,j}/\tau)}, \tag{12}$$

where $o_{i,k}$ is the k-th entry of $\mathbf{o}_i$, and $\tau > 0$ controls the sharpness of the prior distribution. Smaller τ concentrates probability on the most likely candidates, while larger τ approaches uniform. Its impact is evaluated in Section IV-D.

D. Joint Token Detection via Log-Likelihood Fusion

To efficiently solve the MAP estimation in (8), we operate in the log domain:

$$\hat{s}_i = \underset{k\in\{1,\dots,V\}}{\arg\max} \left[\log \Pr(\mathbf{y}_i \mid s_i = k) + \log \Pr(s_i = k \mid \hat{\mathbf{s}}_{1:i-1})\right]. \tag{13}$$

By substituting the symbol likelihood (11) and the generative semantic prior (12) into this expression, the softmax normalization term can be discarded from the $\arg\max$ operation since it is independent of the candidate token k. Consequently, the final detection rule simplifies to:

$$\hat{s}_i = \underset{k\in\{1,\dots,V\}}{\arg\max} \left[\frac{o_{i,k}}{\tau} + \sum_{l=1}^{L} \log\sigma\left((2b_l^{(k)} - 1)\lambda_{i,l}\right)\right]. \tag{14}$$

This formulation reveals that the proposed method can be interpreted as a form of log-domain Bayesian fusion, where the LLR term represents channel evidence and the LMM logits act as a structured prior over the token space. From this perspective, the temperature parameter τ controls the relative trust between observation and prior, analogous to a regularization parameter in Bayesian inference.

E. Computational and Memory Complexity

The computational complexity of the proposed joint detection method is primarily driven by the generation of the LMM generative semantic prior and the subsequent MAP score fusion. By employing a sliding window of size W, the self-attention complexity for predicting a single token is bounded by $\mathcal{O}(W \cdot d)$, where d is the dimension of the LMM's hidden layers. Including the linear projection to the vocabulary space of size V, which requires $\mathcal{O}(d \cdot V)$, the total computational complexity for processing a sequence of N_t tokens is given by:

$$\mathcal{O}\Big(N_t \cdot (W \cdot d + d \cdot V)\Big). \tag{15}$$

In practice, the language modeling head projection $\mathcal{O}(d \cdot V)$ dominates the overall complexity because the vocabulary size is typically much larger than the context window ($V \gg W$). The subsequent MAP score fusion requires only $\mathcal{O}(V)$ operations per token. Thus, the physical-layer fusion adds only a small amount of overhead compared to the LMM inference.

A critical factor for the deployment of LMM-based receivers regarding memory complexity is the Key-Value (KV) cache. The KV cache enables efficient autoregressive inference by reusing previously computed key and value tensors. The memory consumption of the KV cache depends on the number of transformer layers N_{layer}, the hidden dimension d, the context window size W, and the attention mechanism.

Many modern foundation models utilize Multi-Query Attention (MQA) or Grouped Query Attention (GQA) [13] to reduce this memory footprint. By employing fewer key-value heads (H_{kv}) relative to the number of query heads (H_q), the KV cache size is compressed by a factor of H_{kv}/H_q. Using *bfloat16* precision (2 bytes per element), the peak KV cache memory M_{kv} required during the detection of the sequence can be generally formulated as:

$$M_{kv} \approx 2 \times N_{\text{layer}} \times d \times W \times \frac{H_{kv}}{H_q} \times 2 \text{ [bytes]}. \tag{16}$$

For instance, with Emu3 [14], which has $N_{\text{layer}} = 32$, $d = 4096$ and a KV compression ratio of $H_{kv}/H_q = 8/32$, the peak KV cache memory is approximately 512 MiB when $W = 4096$.

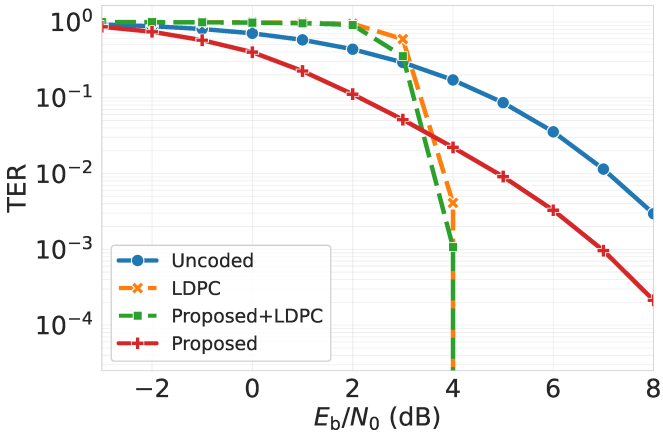


Fig. 2. TER comparison across four conditions: Uncoded, LDPC, Proposed, and Proposed+LDPC.

IV. Simulation Results

In this section, we evaluate the proposed joint semantic-channel token detection method under various AWGN channel conditions.

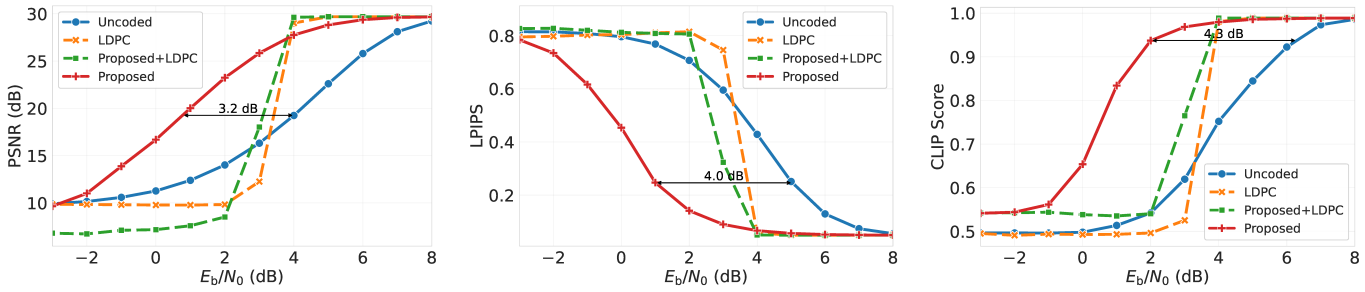
A. Evaluation Setup

1) Model, Dataset, and Evaluation Metrics: We use the pretrained Emu3 [14] as the generative prior, paired with its MoVQ tokenizer [15] with spatial downsampling factor $f = 8$, which maps a 512×512 RGB image to a 64×64 grid of discrete indices.

To align with the autoregressive input format of the LMM, this grid is subsequently flattened into a 1D token sequence of length $N_t = 4096$ in raster-scan order, moving row-wise from the top left corner. Each token is drawn from a codebook of size 32768, corresponding to a token bit-length of $L = 15$.

Unless otherwise specified, we use temperature parameter $\tau = 1$ and window size $W = 4096$. For performance evaluation, we sample 100 images from the FFHQ dataset [16], where all images are resized to 512×512 pixels before tokenization to reduce the evaluation time. We evaluate performance using Peak Signal-to-Noise Ratio (PSNR), LPIPS [17], and CLIP score [18].

2) Channel Setting: We assume an AWGN channel and vary E_b/N_0 from -3 to 8 dB.



(a) PSNR (b) LPIPS (c) CLIP score

Fig. 3. Quantitative evaluation of reconstructed images. The proposed method shows significant gains in the medium E_b/N_0 range.

3) Comparison Methods: As a benchmark, we consider the following four conditions: physical-only uncoded (Uncoded), physical-only with LDPC coding (LDPC), the proposed LMM-based joint detection without coding (Proposed), and with LDPC coding (Proposed+LDPC). We consider $R = 1$ (uncoded) and $R = 1/2$ with a (2480, 1240) 5G NR LDPC code [19] decoded via belief propagation. For the uncoded case, we use the QPSK modulation ($m = 2$) while 16-QAM ($m = 4$) is used for the coded case to maintain the same spectral efficiency.

B. Performance of Joint Semantic-Channel Token Detection

Fig. 2 shows TER across various E_b/N_0 values. When channel coding is not applied, Proposed outperforms Uncoded by approximately 2 dB, demonstrating the benefit of LMM-based semantic prior. When channel coding is applied, both LDPC and Proposed+LDPC exhibit a sharp waterfall effect around $E_b/N_0 = 1$ dB. Below this threshold, LDPC decoding fails catastrophically and yields higher TER than Uncoded, whereas above it, error-free decoding is achieved. Notably, Proposed+LDPC achieves equal or lower TER than LDPC across all E_b/N_0 values, with the most meaningful gain concentrated near the waterfall region around $E_b/N_0 = 1$ dB. Above the waterfall, both LDPC and Proposed+LDPC reach error-free transmission and the benefit of the LMM prior is diminished.

C. Quantitative Evaluation of Reconstructed Images

Fig. 3 illustrates the evaluation of the reconstructed images in terms of PSNR, LPIPS, and CLIP score under various E_b/N_0 conditions.

Proposed outperforms Uncoded in perceptual metrics throughout, with PSNR gain peaking at 9.5 dB at $E_b/N_0 = 3$ dB, although at very low E_b/N_0 LMM hallucinations cause a slight PSNR deficit while perceptual metrics remain superior. Proposed+LDPC shows non-monotone LPIPS behavior. Below the waterfall ($E_b/N_0 \leq -2$ dB) LDPC alone is better, but above it Proposed+LDPC transitions far more sharply to high fidelity.

The effectiveness of the proposed method in preserving semantic information is best exemplified by the CLIP score (Fig. 3c). While Uncoded degrades rapidly below 4 dB, both Proposed and Proposed+LDPC maintain scores above 0.8 at $E_b/N_0 = 1$ dB.

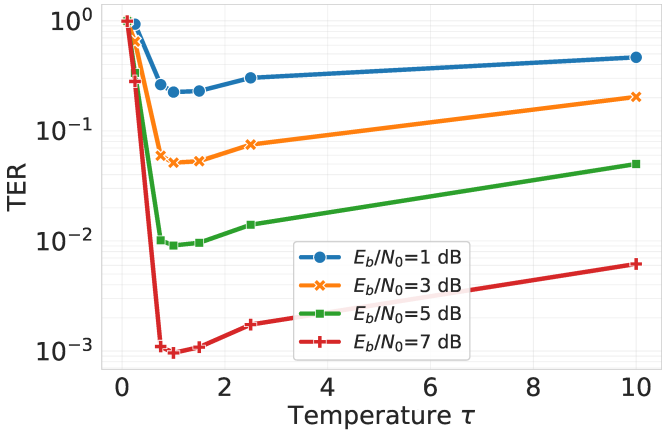


Fig. 4. Impact of temperature parameter τ on Token Error Rate (TER) with window size $W = 4096$.

D. Impact of Temperature Parameter

The temperature parameter τ in Eq. (12) controls how strongly the LMM prior influences the detection rule (14). Fig. 4 illustrates the TER for a range of τ values evaluated at $E_b/N_0 \in \{1, 3, 5, 7\}$ dB with window size $W = 4096$. For all E_b/N_0 conditions, $\tau = 1$ yields the lowest TER among all evaluated values. Extremely small values ($\tau = 0.1$) lead to TER above 0.99 across all evaluated E_b/N_0 values, as the overconfident LMM prior overrides the physical evidence and propagates early errors through the autoregressive chain. As τ increases, the LMM prior becomes more diffuse and the method gradually approaches the physical-layer likelihood-based baseline. These results confirm that the default setting $\tau = 1$ used throughout our experiments is well-calibrated, and that the proposed method is robust to moderate perturbations around this value.

E. Impact of Context Window Scaling and Performance Limits

Based on the observation that symbol LLRs play a dominant role in the joint detection process, we further examine the

impact of the context window size W. Table I summarizes the TER for various window sizes and E_b/N_0 levels.

TABLE I
IMPACT OF WINDOW SIZE W ON TER AT VARIOUS E_b/N_0 LEVELS

E_b/N_0 [dB]	$W = 10$	$W = 128$	$W = 512$	$W = 4096$
1	0.3488	0.2633	0.2365	0.2249
3	0.1238	0.0625	0.0535	0.0514
5	0.0280	0.0108	0.0094	0.0091
7	0.0035	0.0012	0.0010	0.0010

Table I reveals that increasing W consistently reduces TER, but with diminishing returns as W grows. At $E_\mathrm{b}/N_0 = 3$ dB, the TER decreases from 0.1238 ($W = 10$) to 0.0514 ($W = 4096$), a 2.4× reduction. However, the step from $W = 128$ to $W = 4096$ yields only a little gain. At higher E_b/N_0, all window sizes achieve near-zero TER, confirming that the LLR quality is the dominant factor once signal reliability is sufficient.

These results indicate that a moderate window size ($W \geq 128$) captures most of the available semantic context. The TER penalty for reducing W from 4096 to 128 is only 0.0111 at $E_\mathrm{b}/N_0 = 3$ dB, making $W = 128$–512 a practical operating point. This range significantly reduces KV cache overhead with minimal performance loss. With the image compressed to a 64×64 token grid, any W larger than 64 incorporates context from preceding rows. This explains why TER stays almost the same once W exceeds 64 tokens per row.

V. CONCLUSION AND FUTURE WORK

In this work, we proposed a joint semantic-channel token detection framework that incorporates an LMM prior into token detection through MAP estimation. By fusing physical-layer soft information (LLRs) and semantic context in a unified probabilistic rule, the receiver improves both token reliability and reconstruction quality compared with a physical-layer likelihood-based baseline. Experimental results show clear gains in TER, PSNR, LPIPS, and CLIP score, with especially strong improvements over the physical-layer likelihood-based baseline in the medium E_b/N_0 regime where PSNR gains of up to 9.5 dB are observed. The analysis also shows that performance is robust to temperature scaling around $\tau = 1.0$ and that larger context windows ($W > 128$) provide diminishing returns, making the approach practical for resource-constrained deployments.

This work suggests that future semantic communication systems should not treat LMMs merely as post-hoc error correctors, but rather as native components of the detection process itself. This perspective opens a new direction toward receiver-centric semantic communication architectures where communication reliability and semantic inference are jointly optimized.

However, since the current autoregressive inference entails inherent computational latency that scales linearly with the number of tokens N_t, reducing the inference latency remains a critical challenge for high-resolution real-time applications. Future work will focus on extending the framework to bidirectional semantic priors, such as BERT-style or non-autoregressive architectures, to enable parallelized token detection and significantly accelerate the decoding process.

ACKNOWLEDGMENT

This work was supported by the Japan Science and Technology Agency (JST) ASPIRE program (grant no. JPMJAP2326) and by the Mizuho Foundation for the Promotion of Sciences.

REFERENCES

[1] D. Gündüz *et al.*, "Beyond transmitting bits: Context, semantics, and task-oriented communications," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 5–41, 2023.
[2] E. Bourtsoulatze, D. B. Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2019, pp. 4774–4778.
[3] F. Wang, X. Chen, X. Deng, and S. Lin, "Deep joint source-channel coding for wireless image transmission: One-model-to-many," *IEEE Wireless Commun. Lett.*, vol. 14, no. 5, pp. 1501–1505, 2025.
[4] D. Huang, F. Gao, X. Tao, Q. Du, and J. Lu, "Toward semantic communications: Deep learning-based image semantic coding," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 55–71, 2023.
[5] C. Liang *et al.*, "Generative ai-driven semantic communication networks: Architecture, technologies, and applications," *IEEE Trans. Cogn. Commun. Netw.*, vol. 11, no. 1, pp. 27–47, 2025.
[6] Z. Zhao *et al.*, "Lrgd: Low-rank guided diffusion for robust image transmission in semantic communication," *IEEE Trans. Cogn. Commun. Netw.*, vol. 12, pp. 2439–2454, 2026.
[7] L. Qiao, M. B. Mashhadi, Z. Gao, R. Tafazolli, M. Bennis, and D. Niyato, "Token communications: A large model-driven framework for cross-modal context-aware semantic communications," *IEEE Wireless Commun.*, vol. 32, no. 5, pp. 80–88, 2025.
[8] K. Zhang, Z. Jin, Z. Cheng, M. Zeng, L. Qiao, and Z. Fei, "Tokcomuep: Semantic importance-matched unequal error protection for resilient image transmission," arXiv:2511.22859, 2025.
[9] L. Qiao, M. B. Mashhadi, Z. Gao, and D. Gündüz, "Token-domain multiple access: Exploiting semantic orthogonality for collision mitigation," in *Proc. IEEE INFOCOM Wkshps.*, 2025, pp. 1–6.
[10] R. Bommasani *et al.*, "On the opportunities and risks of foundation models," arXiv:2108.07258, 2021.
[11] T. Brown *et al.*, "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 1877–1901.
[12] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.
[13] J. Ainslie, J. Lee-Thorp, M. de Jong, Y. Zemlyanskiy, F. Lebrón, and S. Sanghai, "Gqa: Training generalized multi-query transformer models from multi-head checkpoints," in *Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, 2023, pp. 4895–4901.
[14] X. Wang *et al.*, "Emu3: Next-token prediction is all you need," arXiv:2409.18869, 2024.
[15] C. Zheng, L. T. Vuong, J. Cai, and D. Phung, "Movq: Modulating quantized vectors for high-fidelity image generation," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, 2022, pp. 23 412–23 425.
[16] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 4401–4410.
[17] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 586–595.
[18] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 139, 2021, pp. 8748–8763.
[19] 3GPP, "Multiplexing and channel coding (NR)," 3rd Generation Partnership Project (3GPP), Tech. Spec. TS 38.212, 2024, release 18.