Context-Aware Embedding Masking Based on Reinforcement Learning for Semantic Multiplexing

Ki-Ho Lee, Hyun-Ho Choi, *Senior Member, IEEE*, and Jung-Ryun Lee, *Senior Member, IEEE*

***Abstract*—In shared-embedding (SE) multiple access, U users share a single space and user-specific masks provide the attention-based separation that determines the multiplexing capacity. The original SE design uses state-agnostic masks. We instead learn a context-aware masking policy that maps (SNR, U) to a mask matrix through a low-rank actor trained by proximal policy optimization, rewarding per-user cosine similarity and mask orthogonality. On real BERT embeddings of AG News, it improves cosine similarity at every SNR and cuts the training-time mask-orthogonality penalty tenfold. Compared with a fixed-orthogonal scheme, the method shows that the gain originates from context-aware adaptation rather than orthogonality, while preserving recovery.**

***Index Terms*—Shared embedding, semantic communications, embedding masking, multiple access, context-aware reinforcement learning, proximal policy optimization.**

I. Introduction

SEMANTIC communications aim to convey task-relevant meaning rather than individual bits [1], [2], [3], and recent systems built on frozen pre-trained language models (PLMs) compress text into high-dimensional embeddings that are transmitted through deep joint source and channel coding [4], [5]. Supporting many users over shared resources is a central challenge of modern access networks, where artificial intelligence is increasingly used both to enable seamless and massive access [6] and to drive learning-based multi-dimensional resource scheduling [7], [8]. To serve many users within one resource block, the shared-embedding (SE) framework of [9] has emerged as a multi-user generalization of this paradigm. Unlike dedicated-embedding schemes that allocate disjoint portions of the embedding space to different users in an OFDMA-like fashion, SE lets U users share the same high-dimensional space, where each user's PLM embedding is projected into a shared space of dimension $d_{\mathrm{s}} = Kd_{\mathrm{b}}$, multiplied element-wise by a user-specific positional mask $\mathbf{m}_u \in \mathbb{R}^{d_{\mathrm{s}}}$, superimposed, and transmitted as a single block. The receiver demultiplexes the composite signal through user-indexed scaled-dot-product attention [10], exploiting the fact that the masks rotate each user into an almost-orthogonal direction within d_{s}. In this architecture, the masks $\{\mathbf{m}_u\}_{u=1}^{U}$ are the only mechanism distinguishing users within the shared space, and their pairwise correlation $|\cos(\mathbf{m}_u, \mathbf{m}_v)|$ directly sets the residual interference that the attention-based decoder must suppress. The mask design is therefore the central design variable of an SE system.

The original SE study [9] treats the masks as free trainable parameters and updates them jointly with the projection weights by stochastic gradient descent on the reconstruction loss. We refer to this procedure as *static masking*. It suffers from two limitations. First, the reconstruction loss encourages orthogonality only implicitly, which slows convergence and leaves the attained orthogonality sensitive to the random initialization. Second, a single mask bank must serve every channel realization, although the optimal masks depend on the signal-to-noise ratio (SNR) and the active user count U. Closing these gaps with supervised tools is non-trivial. A regressor from (SNR, U) to masks needs optimal-mask labels that have no closed form, while an end-to-end generator must backpropagate a per-user fidelity objective that is non-differentiable in the random Rayleigh realization, with an orthogonality regularizer competing against reconstruction. Reinforcement learning needs no labels and optimizes the deployed objective directly through a scalar reward. Indeed, even a fixed-orthogonal mask bank is outperformed at every SNR (§ IV-D), so adaptivity, not orthogonality alone, drives the gain.

Building on [9], our contributions are (i) a mask-learning framework based on deep reinforcement learning (DRL), in which a proximal policy optimization (PPO) [11] actor conditioned on (SNR, U) emits the mask matrix through a low-rank generator under a reward that jointly maximizes per-user cosine similarity and penalizes non-orthogonality $\mathcal{O}(\mathbf{M})$ [8], (ii) a real-data evaluation at the scale of bidirectional encoder representations from transformers (BERT) on 8,000 AG News embeddings ($d_{\mathrm{b}} = 768$, $d_{\mathrm{s}} = 3{,}072$), assessed by a task-oriented retrieval metric and against a fixed-orthogonal scheme, and (iii) a low-rank actor that drives the mask-orthogonality penalty an order of magnitude below static masking at negligible inference overhead, preserving the linear-in-K complexity of [9].[1]

Received 30 June 2026; accepted 28 July 2026. Date of publication 3 August 2026; date of current version 10 August 2026. This work was supported by the National Research Foundation of Korea (NRF) under Grant RS-2026-25480938. The associate editor coordinating the review of this article and approving it for publication was X. Gao. *(Corresponding authors: Hyun-Ho Choi; Jung-Ryun Lee.)*

Ki-Ho Lee is with the Department of Electronic Information Engineering, Anyang University, Anyang 14028, South Korea (e-mail: kiholee@anyang.ac.kr).

Hyun-Ho Choi is with the School of ICT, Robotics and Mechanical Engineering and the Research Center for Hyper-Connected Convergence Technology, Hankyong National University, Anseong 17579, South Korea (e-mail: hhchoi@hknu.ac.kr).

Jung-Ryun Lee is with the School of Electrical and Electronics Engineering, Department of Intelligent Energy and Industry Convergence, Chung-Ang University, Seoul 06974, South Korea (e-mail: jrlee@cau.ac.kr).

Digital Object Identifier 10.1109/LWC.2026.3718953

[1]The source code is available at https://github.com/KiHoLee/WCL.

II. System Model

We consider U uplink users sharing a common receiver. The text of user u is encoded by a frozen BERT model into a mean-pooled, centered, and ℓ_2-normalized embedding $\mathbf{b}_u \in \mathbb{R}^{d_\mathrm{b}}$ [12], [13]. A shared transmit projection f_tx maps this embedding into the expanded space as

$$\mathbf{e}_u = f_\mathrm{tx}(\mathbf{b}_u;\theta_\mathrm{tx}) \in \mathbb{R}^{d_\mathrm{s}}, \quad d_\mathrm{s} = K d_\mathrm{b}, \tag{1}$$

where K denotes the bandwidth-expansion factor. A user-specific mask $\mathbf{m}_u \in \mathbb{R}^{d_\mathrm{s}}$ is applied element-wise (with $\odot$ the Hadamard product), and the resulting signals are superimposed and power-normalized as

$$\mathbf{y}_\mathrm{tx} = \sqrt{d_\mathrm{s}}\,\boldsymbol{\xi}/\|\boldsymbol{\xi}\|_2, \quad \boldsymbol{\xi} = \sum_{u=1}^{U} \mathbf{e}_u \odot \mathbf{m}_u. \tag{2}$$

The scaling in (2) is applied per block and forces $\|\mathbf{y}_\mathrm{tx}\|_2^2 = d_\mathrm{s}$ exactly for every channel use, that is, unit power per dimension, regardless of the instantaneous codeword variance produced by the neural encoder. The power constraint is therefore instantaneous rather than statistical, since it is met for each message and each Rayleigh realization rather than only in expectation. As a result, the post-hoc amplitude scaling controls the transmitted instantaneous power. The signal experiences block Rayleigh fading of magnitude $|h|$ with $\mathbb{E}[|h|^2] = 1$ and additive white Gaussian noise (AWGN) $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma_n^2 \mathbf{I})$ with $\sigma_n^2 = 10^{-\mathrm{SNR_{dB}}/10}$, yielding $\mathbf{y}_\mathrm{rx} = |h|\mathbf{y}_\mathrm{tx} + \mathbf{n}$. The receiver forms U mask-projected candidates $\mathbf{c}_i = \mathbf{y}_\mathrm{rx} \odot \mathbf{m}_i$, and a user-indexed attention with learnable queries $\mathbf{q}_u \in \mathbb{R}^{d_\mathrm{s}}$ computes

$$a_{u,i} = \frac{\exp\big(\langle \bar{\mathbf{c}}_i, \bar{\mathbf{q}}_u \rangle / \sqrt{d_\mathrm{s}}\big)}{\sum_{j=1}^{U} \exp\big(\langle \bar{\mathbf{c}}_j, \bar{\mathbf{q}}_u \rangle / \sqrt{d_\mathrm{s}}\big)}, \tag{3}$$

where $\bar{\mathbf{c}}_i$ and $\bar{\mathbf{q}}_u$ denote the ℓ_2-normalized versions. The demultiplexed representation is $\hat{\mathbf{z}}_u = \sum_i a_{u,i}\mathbf{c}_i$, followed by the reverse projection $\hat{\mathbf{b}}_u = f_\mathrm{rx}(\hat{\mathbf{z}}_u;\theta_\mathrm{rx}) \in \mathbb{R}^{d_\mathrm{b}}$.

III. Proposed Context-Aware Embedding Masking

A. Mask Design as a Contextual-Bandit Policy

To overcome the two limitations of the static mask bank (channel-agnosticism and only-implicit orthogonality), we regard the mask matrix $\mathbf{M} \triangleq [\mathbf{m}_1; \ldots; \mathbf{m}_U] \in \mathbb{R}^{U \times d_\mathrm{s}}$ as the action of a policy conditioned on the wireless state

$$\mathbf{s} \triangleq [\mathrm{SNR}_\mathrm{norm}, U_\mathrm{norm}]^\top, \tag{4}$$

where $\mathrm{SNR}_\mathrm{norm} \in [0,1]$ and $U_\mathrm{norm} \in [0,1]$ are the range-normalized channel SNR and user count. This casts mask design as a single-step contextual bandit, since the channel and active user set are drawn independently per block, so the action does not affect the next state and the horizon is one. The reward for a single transmission combines per-user fidelity with explicit orthogonality pressure, and is given by

$$r(\mathbf{s},\mathbf{M}) = \frac{1}{U}\sum_{u=1}^{U} \mathrm{CosSim}(\hat{\mathbf{b}}_u, \mathbf{b}_u) - \beta \cdot \mathcal{O}(\mathbf{M}), \tag{5}$$

where $\mathrm{CosSim}(\mathbf{x},\mathbf{y}) \triangleq \mathbf{x}^\top\mathbf{y}/(\|\mathbf{x}\|_2\|\mathbf{y}\|_2)$ is the cosine similarity and $\|\cdot\|_F$ below is the Frobenius norm. The term

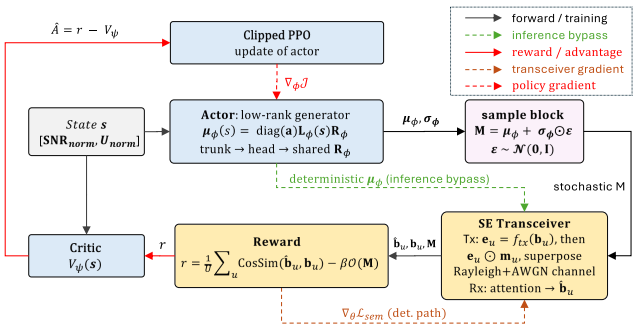


Fig. 1. Proposed context-aware mask policy. Solid and dashed arrows mark the training (stochastic) and inference (deterministic-mean) paths, respectively.

$\mathcal{O}(\mathbf{M}) = \|\widetilde{\mathbf{M}}\widetilde{\mathbf{M}}^\top - \mathbf{I}_U\|_F^2/U^2$ is the Gram-matrix deviation of the row-ℓ_2-normalized $\widetilde{\mathbf{M}}$ from the identity, and $\beta > 0$ balances orthogonality against fidelity. $\mathcal{O}(\mathbf{M})$ measures the average squared pairwise cosine similarity among user masks and vanishes when all masks are mutually orthogonal. The two terms are complementary rather than conflicting, since lowering inter-user correlation reduces the residual interference the decoder must suppress and thus raises the attainable fidelity, while any residual tension at low SNR or heavy overload is absorbed by conditioning on $\mathbf{s}$. The β-ablation in § IV-B confirms that the two improve together as β grows from 0 to 0.5. Because the horizon is one, we use PPO purely as a stable optimizer for the high-dimensional Gaussian action, not for sequential bootstrapping, so the discount factor, value bootstrapping, and generalized advantage estimation are absent.

B. Architecture Overview

Figure 1 illustrates the policy. The actor maps $\mathbf{s}$ to the Gaussian mean $\boldsymbol{\mu}_\phi(\mathbf{s})$ via a low-rank generator. The highlighted `sample` block realizes the action $\mathbf{M}$, drawing $\mathbf{M} \sim \pi_\phi(\cdot|\mathbf{s})$ in training by the reparameterization in (7), while at inference it is bypassed and $\boldsymbol{\mu}_\phi(\mathbf{s})$ is forwarded directly. The masks feed the transceiver, whose semantic and orthogonality metrics form the reward that, with the critic value, drives a clipped PPO update.

C. Low-Rank Mask Generator (Actor)

A naive flattening of the mask matrix would require the policy to output Ud_s real values (e.g., $4{\cdot}3{,}072 = 12{,}288$ at the operating point $U = 4$, $d_\mathrm{s} = 3{,}072$), which is prohibitive for PPO exploration. Instead, we express the mean of the policy $\boldsymbol{\mu}_\phi(\mathbf{s}) \in \mathbb{R}^{U \times d_\mathrm{s}}$ through the three-stage low-rank pipeline

$$\mathbf{g}(\mathbf{s}) = \mathrm{MLP}_\phi(\mathbf{s}) \in \mathbb{R}^{d_\mathrm{g}}, \tag{6a}$$

$$\mathbf{L}_\phi(\mathbf{s}) = \mathrm{reshape}\big(\mathbf{W}_L\,\mathbf{g}(\mathbf{s}) + \mathbf{b}_L\big) \in \mathbb{R}^{U \times r}, \tag{6b}$$

$$\boldsymbol{\mu}_\phi(\mathbf{s}) = \mathrm{diag}(\mathbf{a})\,\mathbf{L}_\phi(\mathbf{s})\,\mathbf{R}_\phi \in \mathbb{R}^{U \times d_\mathrm{s}}, \tag{6c}$$

where $\mathrm{MLP}_\phi : \mathbb{R}^2 \to \mathbb{R}^{d_\mathrm{g}}$ is a two-layer multilayer perceptron (MLP) with rectified linear unit (ReLU) activation ($d_\mathrm{g} = 256$) (the trunk), $\mathbf{W}_L, \mathbf{b}_L$ form the left-factor head, $\mathbf{R}_\phi \in \mathbb{R}^{r \times d_\mathrm{s}}$ is a shared learnable right factor, $r \ll d_\mathrm{s}$ is the latent rank (not to be confused with the scalar reward $r(\cdot,\cdot)$), and $\mathbf{a} \in \mathbb{R}^U$ absorbs

per-user scale. Here $\mathrm{reshape}:\mathbb{R}^{Ur}\to\mathbb{R}^{U\times r}$ is the parameter-free row-major reindexing $\mathbf{V}_{u,k}=v_{(u-1)r+k}$. Thus each mask row combines the shared basis $\mathbf{R}_\phi$ with per-user coefficients $\mathbf{L}_\phi(\mathbf{s})$, that is, $\boldsymbol{\mu}_\phi(\mathbf{s})_{u,:}=a_u\sum_k[\mathbf{L}_\phi(\mathbf{s})]_{u,k}[\mathbf{R}_\phi]_{k,:}$. The low-rank form is matched to the problem rather than a generic borrowing of low-rank adaptation (LoRA) [14], since the U near-orthogonal target rows live on a low-dimensional manifold, and collapsing the action from $Ud_\mathrm{s}=12{,}288$ to $Ur=256$ dimensions keeps PPO exploration tractable. With $r=64$ the mean has only 256 degrees of freedom yet realizes the near-orthogonal configurations that drive multiplexing gains.

The full policy is Gaussian $\pi_\phi(\mathbf{M}|\mathbf{s})=\mathcal{N}(\boldsymbol{\mu}_\phi(\mathbf{s}),\boldsymbol{\sigma}_\phi^2)$ with a state-independent diagonal $\boldsymbol{\sigma}_\phi$ (parameterized in log-space, initialized to e^{-1}), so early training is nearly deterministic. The action is realized as

$$\mathbf{M}=\begin{cases}\boldsymbol{\mu}_\phi(\mathbf{s})+\boldsymbol{\sigma}_\phi\odot\boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I}), & \text{training},\\ \boldsymbol{\mu}_\phi(\mathbf{s}), & \text{inference}.\end{cases}\tag{7}$$

The training branch realizes $\mathbf{M}\sim\pi_\phi(\cdot|\mathbf{s})$ via the reparameterization trick so that gradients flow through both $\boldsymbol{\mu}_\phi$ and $\boldsymbol{\sigma}_\phi$, whereas the inference branch uses the deterministic mean and bypasses the `sample` block.

Remark 1 (Model-agnostic operation): The policy acts only on the mask matrix $\mathbf{M}$ and on the scalar context $\mathbf{s}=(\mathrm{SNR},U)$, and the embedding $\mathbf{b}_u$ enters solely through the reward (5), never through the BERT weights or tokenizer. Hence any frozen encoder emitting fixed-dimensional embeddings (e.g., Sentence-BERT, a vision, or a speech backbone) can be substituted by changing only d_b and the embedding pool, so BERT/AG News is a concrete instance, not a requirement.

D. Critic and Clipped PPO Update

A three-layer MLP critic $V_\psi:\mathbb{R}^2\to\mathbb{R}$ with hidden width 128 estimates the expected reward under the current policy. During training we accumulate a transition buffer $\mathcal{B}=\{(\mathbf{s}^{(j)},\mathbf{M}^{(j)},r^{(j)})\}_{j=1}^{B}$ of size B, where each $\mathbf{s}^{(j)}$ is a context, $\mathbf{M}^{(j)}\sim\pi_{\phi_\mathrm{old}}(\cdot|\mathbf{s}^{(j)})$ is a mask drawn under the sampling policy ϕ_old, and $r^{(j)}=r(\mathbf{s}^{(j)},\mathbf{M}^{(j)})$ is the one-step reward. The clipped PPO surrogate is

$$\mathcal{J}(\phi)=\mathbb{E}_{j\sim\mathcal{B}}\Big[\min\Big(\rho_j(\phi)\hat{A}_j,\ \mathrm{clip}(\rho_j(\phi),1-\epsilon,1+\epsilon)\hat{A}_j\Big)\Big],\tag{8}$$

with likelihood ratio

$$\rho_j(\phi)=\frac{\pi_\phi(\mathbf{M}^{(j)}|\mathbf{s}^{(j)})}{\pi_{\phi_\mathrm{old}}(\mathbf{M}^{(j)}|\mathbf{s}^{(j)})},\tag{9}$$

one-step advantage $\hat{A}_j=r^{(j)}-V_\psi(\mathbf{s}^{(j)})$ (since the episode length is one, neither discount nor generalized advantage estimation is required), and clip operator $\mathrm{clip}(x,a,b)\triangleq\min(\max(x,a),b)$. The clip parameter $\epsilon=0.2$ bounds the per-sample change of the policy. A Gaussian entropy bonus $\mathcal{H}(\pi_\phi(\cdot|\mathbf{s}))=\frac{1}{2}\sum_{k=1}^{Ud_\mathrm{s}}\big(1+\log(2\pi\sigma_{\phi,k}^2)\big)$ is added as a regularizer with weight $\zeta=10^{-3}$, and the actor parameters ϕ are updated with learning rate η_ϕ by

$$\phi\leftarrow\phi+\eta_\phi\nabla_\phi\Big[\mathcal{J}(\phi)+\zeta\,\mathbb{E}_j\Big[\mathcal{H}(\pi_\phi(\cdot|\mathbf{s}^{(j)}))\Big]\Big].\tag{10}$$

TABLE I
SIMULATION PARAMETERS

Parameter	Value
BERT model	`bert-base-uncased` ($d_\mathrm{b}=768$)
Dataset	AG News (8,000 sentences)
d_s ($K=4$) / users U	3,072 / 1–6
Channel	Rayleigh + AWGN, block fading
Train / Eval SNR (dB)	$\{0,\dots,25\}$ / $\{0,\dots,30\}$
Weights β / λ	0.5 / 0.5
Actor hidden / rank r / critic	256 / 64 / 128
PPO clip ϵ / epochs / entropy	0.2 / 4 / 10^{-3}
Buffer size B	64
LR (actor / critic / trx)	3×10^{-4} / 10^{-3} / 10^{-3}
Epochs / steps per epoch	100 / 200
Seeds / eval. trials	6 / 200

The critic is fitted by mean-squared regression with learning rate η_ψ,

$$\psi\leftarrow\psi-\eta_\psi\nabla_\psi\sum_{j\in\mathcal{B}}\big(r^{(j)}-V_\psi(\mathbf{s}^{(j)})\big)^2.\tag{11}$$

After four optimizer epochs on a full buffer, we set $\phi_\mathrm{old}\leftarrow\phi$ and collect a new batch of transitions.

E. Joint Training of the Transceiver

The transceiver parameters $\{\theta_\mathrm{tx},\theta_\mathrm{rx},\{\mathbf{q}_u\}\}$ are differentiable and benefit from low-variance gradients, so a separate forward pass with the deterministic mean $\boldsymbol{\mu}_\phi(\mathbf{s})$ updates them by the mean-squared-error (MSE) plus cosine-similarity loss $\mathcal{L}_\mathrm{sem}$ (12), while the stochastic mask drives the PPO update. The two paths are detached so that the policy and transceiver receive consistent but decoupled signals. In summary, each iteration draws an SNR from the training set $\mathcal{T}$ to form $\mathbf{s}$, samples $\mathbf{M}\sim\pi_\phi(\cdot|\mathbf{s})$ and a user batch, evaluates the reward (5) and the value $V_\psi(\mathbf{s})$ through the Rayleigh channel, and updates the transceiver by (12). Once the buffer holds B transitions, ϕ and ψ are updated by (8) and regression.

F. Inference Complexity

At inference the scheme adds only a single low-rank actor pass in (6), where the trunk and head cost $O(d_\mathrm{g}(d_\mathrm{g}+Ur))$ and the right-factor product $O(Urd_\mathrm{s})$, dominated by $O(Urd_\mathrm{s})$ since $d_\mathrm{g}\ll d_\mathrm{s}$, on top of the transceiver's usual $O(Ud_\mathrm{b}d_\mathrm{s}+U^2d_\mathrm{s})$, for a per-block total of $O(U(r+d_\mathrm{b}+U)d_\mathrm{s})$. At $U=4$, $d_\mathrm{b}=768$, $d_\mathrm{s}=3{,}072$, $r=64$ the transceiver costs about 9.5 million multiply-accumulate operations (MMAC) per block while the actor adds only 0.2 MMAC, a negligible ($<3\%$) overhead that preserves the linear-in-K scaling of [9]. The parameterization also scales to larger d_s, since only the shared $\mathbf{R}_\phi$ grows with d_s while $\mathbf{L}_\phi(\mathbf{s})$ stays compact.

IV. NUMERICAL RESULTS

A. Simulation Setup

We adopt the setup of [9], where a frozen `bert-base-uncased` model ($d_\mathrm{b}=768$) produces mean-pooled, centered, and ℓ_2-normalized embeddings from 8,000 AG News headlines [15], with an expansion factor $K=4$ such that $d_\mathrm{s}=3{,}072$. The channel is block

TABLE II
TOP-1 RETRIEVAL ACCURACY (%) AT $U=4$ (6-SEED MEAN±STD, PAIRED CHANNELS)

Method	Top-1 retrieval Acc. (%)		
	0 dB	10 dB	20 dB
Fixed-Orth	91.2±0.3	97.5±0.1	97.6±0.1
Static, CE	**93.9**±0.4	97.4±0.1	97.5±0.1
Static, Sem	93.3±0.3	97.6±0.0	**97.7**±0.1
Proposed	92.7±0.3	**97.7**±0.1	**97.7**±0.1

Rayleigh, with the training SNR drawn uniformly from $\{0, 5, 10, 15, 20, 25\}$ dB and the evaluation SNR sweeping seven values in $[0, 30]$ dB, over 200 independent trials per point. Table I summarizes the key hyperparameters.

To attribute any improvement to the mask-learning strategy alone, we compare the proposed context-aware masking of § III with three alternatives that share the same transceiver, optimizer, and dataset. The first is the original static masking, whose masks $\{\mathbf{m}_u\}_{u=1}^{U}$ are free trainable parameters updated jointly with the transceiver on the MSE+CosSim objective

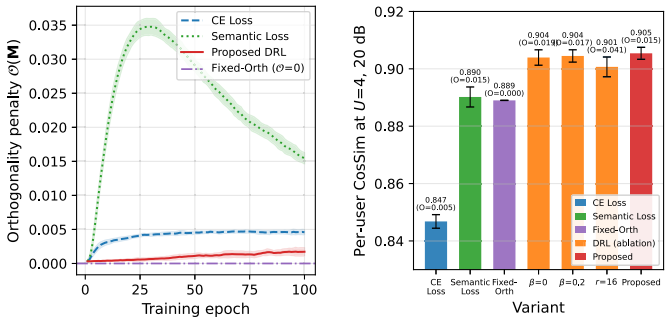
$$\mathcal{L}_{\text{sem}} = \frac{1}{U}\sum_{u}\big[(1-\lambda)\|\hat{\mathbf{b}}_u - \mathbf{b}_u\|_2^2 + \lambda(1-\text{CosSim}(\hat{\mathbf{b}}_u, \mathbf{b}_u))\big], \tag{12}$$

with $\lambda = 0.5$ balancing the MSE (norm) and cosine (direction) terms in the post-centered BERT space. The second alternative keeps the same free masks but, following [9], replaces the reconstruction objective with the C-way symbol-identity cross-entropy (CE)

$$\mathcal{L}_{\text{CE}} = -\sum_{u=1}^{U}\sum_{i=1}^{C} x_u^{(i)} \log \hat{p}_u^{(i)}, \;\; \hat{\mathbf{p}}_u = \text{softmax}\big(\tau \hat{\mathbf{b}}_u \mathbf{E}_{\text{pool}}^{\top}\big), \tag{13}$$

where $\mathbf{x}_u$ is the one-hot identity, $\mathbf{E}_{\text{pool}} \in \mathbb{R}^{C\times d_{\text{b}}}$ collects the $C = 8{,}000$ AG News embeddings, and $\tau = 16$ is a logit temperature, so that the receiver classifies each reconstructed embedding against the full 8,000-class pool as in [9]. Finally, to isolate the value of learning the masks, we include a fixed-orthogonal scheme whose masks are fixed to U mutually orthogonal rows of unit root-mean-square (RMS) value, taken from the QR factorization of a random $d_{\text{s}} \times U$ matrix, so that $\mathcal{O}(\mathbf{M}) = 0$ by construction, training only the transceiver $\{\theta_{\text{tx}}, \theta_{\text{rx}}, \{\mathbf{q}_u\}\}$ on (12). Sharing none of the learning machinery, it represents the best case for static orthogonality and isolates the benefit of adaptivity from orthogonality alone.

All methods share a matched budget of 100 epochs of 200 iterations. In the SNR and throughput curves the static-masking, cross-entropy, proposed, and DRL-ablation results at $U = 4$ are each averaged over six seeds, while the fixed-orthogonal scheme is shown as a single-seed reference. The retrieval evaluation in Table II averages all four schemes over six seeds. Here "epoch" and "step" count training iterations over independently drawn contexts, not time steps of a sequential trajectory, as in the single-step contextual bandit of § III-A.



(a) Orthogonality penalty vs. training epoch

(b) Ablation at $U = 4$, 20 dB

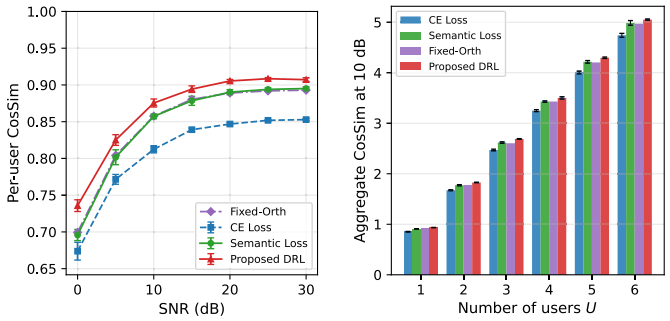
Fig. 2. (a) Training-time $\mathcal{O}(\mathbf{M})$ at $U = K = 4$, shown as $\pm 1\sigma$ over six seeds for static masking, DRL, and CE, with the fixed-orthogonal scheme ($\mathcal{O} = 0$) as a single-seed reference. (b) Ablation at $U = 4$, 20 dB, where each bar is labeled with CosSim and $\mathcal{O}(\mathbf{M})$ (parenthesized) as six-seed means with $\pm 1\sigma$, except the single-seed fixed-orthogonal bar.

B. Learning Dynamics and Ablation

Figure 2(a) traces $\mathcal{O}(\mathbf{M})$ at $U = 4$ during training ($\pm 1\sigma$ over six seeds). Both start near zero, since random high-dimensional masks are almost orthogonal, and then diverge. Static masking rises to $\mathcal{O} \approx 0.035$ near epoch 30 as the reconstruction objective correlates the masks, and relaxes only to 0.0164 ± 0.0009 by the end, whereas the DRL actor stays low and settles at 0.00172 ± 0.00053. This order-of-magnitude end gap, wider still at mid-training, arises because PPO penalizes $\mathcal{O}(\mathbf{M})$ through an explicit $\beta\mathcal{O}(\mathbf{M})$ term while static masking applies only indirect pressure, and the low-rank factorization (6c) further confines every mask to the shared basis $\mathbf{R}_\phi$. A third trace using the symbol-CE objective of [9] tightens the masks to $\mathcal{O} \approx 0.0046$, between MSE+CosSim and DRL, but at a semantic cost (CosSim ≈ 0.85 at 20 dB versus 0.905 for DRL). Figure 2(b) ablates the design choices at $U = 4$, 20 dB. Here CE gives the second-tightest geometry yet the lowest CosSim (0.847). Reducing β from $0.5 \to 0$ lowers CosSim ($0.9054 \to 0.9039$) and relaxes $\mathcal{O}$ ($0.015 \to 0.019$). Halving the rank to $r = 16$ is more harmful (CosSim 0.9007, $\mathcal{O} = 0.041$). Thus ($\beta = 0.5, r = 64$) attains the tightest Gram matrix among the DRL variants together with the highest CosSim of all schemes.

C. Mask Correlation and Performance

The end-of-training pairwise correlations $|\cos(\mathbf{m}_u, \mathbf{m}_v)|$ at $U = K = 4$ confirm the ordering, with static masking at 0.10 to 0.18 (mean 0.15), symbol-CE [9] at 0.03 to 0.11 (mean 0.08), and DRL at 0.02 to 0.07 (mean 0.05), so DRL attains a Gram matrix $\sim 1.6\times$ tighter than CE at no semantic cost. Figure 3(a) shows the 6-seed per-user CosSim versus SNR at $U = 4$, where DRL leads at every SNR with the gap shrinking from + 0.040 at 0 dB to + 0.012 at 30 dB and the low-SNR lead being largest because mask separability dominates under heavy noise. The CE scheme lies below static masking throughout (-0.010 to -0.044). Figure 3(b) shows the aggregate throughput $\sum_u \text{CosSim}(\hat{\mathbf{b}}_u, \mathbf{b}_u)$ at 10 dB for $U \in \{1, \ldots, 6\}$, where DRL leads at every load, with the gain over static masking tapering from + 3.05% at $U = 2$



(a) Per-user CosSim vs. SNR (b) Throughput vs. U

Fig. 3. (a) Per-user CosSim vs. SNR at $U = K = 4$ (static masking, DRL, and CE as 6-seed mean $\pm 1\sigma$, with the fixed-orthogonal scheme as a single-seed reference). (b) Aggregate throughput $\sum_u \mathrm{CosSim}$ vs. U at 10 dB, with 6-seed CE and a single-seed fixed-orthogonal bar.

to + 1.30% at $U = 6$ as the per-user budget $d_s/U = 512$ binds, while CE trails static masking by 4.9 to 6.2%.

D. Task-Oriented Evaluation and Fixed-Orthogonal Scheme

The fixed-orthogonal scheme is overlaid on Figs. 2–3. It attains $\mathcal{O}(\mathbf{M}) = 0$ exactly by construction, yet it is outperformed in per-user CosSim at every SNR by the learned policy (Fig. 3(a), e.g., + 0.036 at 0 dB), so perfect static orthogonality is not sufficient and the gain stems from context-aware co-adaptation. Because cosine similarity is a fidelity measure rather than a task-level one, we additionally evaluate the top-1 semantic retrieval accuracy Acc. Let c_u index the transmitted sentence of user u within the $C = 8{,}000$-sentence AG News pool $\mathbf{E}_{\mathrm{pool}}$ (13). The receiver returns the nearest neighbor in cosine similarity $\hat{c}_u = \arg\max_c \mathrm{CosSim}(\hat{\mathbf{b}}_u, [\mathbf{E}_{\mathrm{pool}}]_c)$, and $\mathrm{Acc} = \mathbb{E}[\mathbb{1}(\hat{c}_u = c_u)]$ is the fraction of symbols identified correctly (chance $1/C$), with $\mathbb{1}(\cdot)$ the indicator function. Its complement is the semantic symbol error rate $\mathrm{SER} = 1 - \mathrm{Acc}$, a symbol-level recoverability metric. Table II summarizes this metric at $U = 4$ over six seeds with paired channels. All methods saturate to within $\approx 0.3\%$ above 10 dB, so the proposed fidelity gains cost no symbol-level recoverability. The CE scheme is marginally best at 0 dB (93.9% versus 92.7%) yet worst in CosSim (Fig. 3(a)) and lowest in retrieval from 10 dB upward, where the proposed policy ties for the best, confirming that retrieval and fidelity are distinct axes and that only the proposed policy is strong on both. This crossover reflects objective alignment, as the CE loss (13) optimizes symbol identity directly and thus dominates at noise-limited 0 dB, whereas once retrieval saturates the proposed policy leads through tighter orthogonality and higher fidelity.

V. Conclusion

This letter replaces the gradient-trained mask bank of [9] (static masking) with a PPO-learned context-aware generator that emits the mask matrix through a compact low-rank factorization. On real BERT embeddings of AG News at a matched 100-epoch budget averaged over six seeds, it improves per-user CosSim at every SNR (up to + 0.040 at 0 dB), cuts the training-time mask-orthogonality penalty $\mathcal{O}(\mathbf{M})$ by an order of magnitude, and delivers up to + 3.05% higher aggregate throughput at 10 dB at negligible inference overhead. A task-oriented retrieval metric further shows that these fidelity gains incur no loss of symbol-level recoverability. Moreover, the fixed-orthogonal scheme, which is perfectly orthogonal yet context-agnostic, is outperformed at every SNR, confirming that the benefit stems from learned context adaptation rather than orthogonality alone. The gain arises from the explicit optimization of the orthogonality penalty $\mathcal{O}(\mathbf{M})$, the low-rank factorization (6c) that confines every mask to a common near-orthogonal subspace, and the context-conditioned mean $\boldsymbol{\mu}_\phi(\mathbf{s})$, so a single trained policy covers the entire SNR and user-count range with one forward pass per block, needing no per-condition retraining at deployment. The present study is limited to a block-Rayleigh channel, an i.i.d. user model, and a single dataset. Extending it to correlated channels, heterogeneous user priorities, cross-domain corpora, and relevance-aware state representations that share mask subspaces across semantically correlated users is left to future work.

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, Jan. 2023.

[2] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE Trans. Signal Process.*, vol. 69, pp. 2663–2675, 2021.

[3] E. Calvanese Strinati and S. Barbarossa, "6G networks: Beyond Shannon towards semantic and goal-oriented communications," *Comput. Netw.*, vol. 190, May 2021, Art. no. 107930.

[4] T. O'Shea and J. Hoydis, "An introduction to deep learning for the physical layer," *IEEE Trans. Cognit. Commun. Netw.*, vol. 3, no. 4, pp. 563–575, Dec. 2017.

[5] E. Bourtsoulatze, D. Burth Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Trans. Cognit. Commun. Netw.*, vol. 5, no. 3, pp. 567–579, Sep. 2019.

[6] Z. Lin, Z. Feng, K. Guo, A. Nauman, D. Niyato, and J. Wang, "AI-driven seamless and massive access in space-air-ground integrated networks," *IEEE Wireless Commun.*, vol. 32, no. 3, pp. 72–79, Jun. 2025.

[7] Z. Feng, Z. Lin, H. Wang, R. Ma, K. An, and Y. He, "Toward secure and reliable SAGIN: Learning-driven multi-dimensional resource scheduling for multi-RIS-assisted OTFS transmission," *IEEE J. Sel. Areas Commun.*, vol. 44, pp. 4848–4860, 2026.

[8] N. C. Luong et al., "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3133–3174, 4th Quart., 2019.

[9] K.-H. Lee, H.-H. Choi, and J.-R. Lee, "Transformer-based shared embedding for multiple access in semantic communications," *IEEE J. Sel. Areas Commun.*, vol. 44, pp. 2622–2637, 2026.

[10] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

[11] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, Hum. Lang. Technol., Minneapolis, MI, USA, vol. 1, Jun. 2019, pp. 4171–4186.

[13] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 3982–3992.

[14] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.

[15] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2015, pp. 649–657.