Transformer-Based Shared Embedding for Multiple Access in Semantic Communications

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

***Abstract*—This paper proposes a Transformer-based *shared embedding (SE)* scheme as an artificial intelligence (AI)-native multiple access for semantic communications. In SE, multiple users jointly occupy a high-dimensional embedding space and are separated by learnable user-specific positional masks together with attention-driven demultiplexing, so that inter-user interference is suppressed directly in the embedding domain. The resulting interference structure closely resembles interference alignment (IA), wherein residual multiuser energy is compressed into a low-rank subspace that the decoder can effectively attenuate. We develop closed-form expressions for signal power, effective signal-to-interference-plus-noise ratio (SINR), symbol error rate (SER), and per-user capacity, explicitly characterizing the roles of intra-user coherence and an alignment coefficient that quantifies attention-induced interference suppression. These formulas yield theoretical baselines that situate SE performance between a no-alignment lower bound and an ideal non-orthogonal multiple access (NOMA)-inspired upper bound and explain capacity gains in dense regimes via statistical multiplexing and alignment. Experimental evaluations are carried out under Rayleigh fading channels, and the results show that SE consistently outperforms the dedicated embedding (DE) scheme, which assigns each user an exclusive portion of the embedding space, across different channel conditions and user densities. The empirical SE curves match the theoretical predictions with fitted alignment factors and approach the NOMA-inspired bound as the embedding dimension increases, thereby showing that the embedding space can function as a new axis of multiple access complementary to time, frequency, and code domains. Finally, we outline adaptive and generalized training strategies that extend beyond symbol-level recovery to semantic-level classification and reconstruction tasks for next generation networks.**

***Index Terms*—Multiple access, semantic communications, interference management, deep learning, Transformer.**

I. Introduction

RECENT research has increasingly turned toward artificial intelligence (AI) techniques to dynamically optimize resource allocation, interference management, and channel modeling in wireless communication systems. In particular, AI-driven methods have shown promise in adapting to complex and time-varying environments, thereby enhancing system performance or spectrum utilization at the physical layer. Conventional approaches typically employ deep neural networks (DNNs) for channel modeling [1], convolutional neural networks (CNNs) for channel state information (CSI) compression [2], and recurrent neural networks (RNNs) for beamforming and resource allocation optimization in massive MIMO systems [3]. While these models have significantly advanced wireless performance, they often rely on static assumptions and require extensive retraining when deployed in fast-evolving or previously unseen channel conditions.

To overcome these challenges, recent efforts have explored generative AI (GAI) techniques, such as generative adversarial networks (GANs) and variational autoencoders (VAEs), which are capable of synthesizing realistic channel responses and estimating latent interference patterns [4], [5]. GANs exploit adversarial training to model complex propagation environments, while VAEs enable robust encoding of high-dimensional inputs into low-dimensional latent spaces, facilitating resilient decoding under uncertainty. These techniques have shown potential in enhancing physical layer reliability and security, especially under non-stationary conditions [6], [7].

As GAI-based wireless communication systems evolve, recent efforts have increasingly focused on enhancing GAI capabilities to model high-level abstractions and approximate semantic understanding. This evolution reflects a broader shift in wireless system design, moving away from the traditional goal of optimizing raw bit transmission and instead emphasizing the development of more intelligent and content-aware communication mechanisms. In this new paradigm, the objective is not merely to recover the transmitted bits but to preserve the intended meaning or task-specific informational content of the data [8], [9]. By prioritizing semantically important content and reducing unnecessary transmission redundancy, semantic communication enables systems to operate more efficiently, especially in bandwidth-limited or noisy environments [10].

Building on this perspective, a natural question arises: how can these systems efficiently encode and transmit semantic meaning in compact, interference-resilient forms? A key enabler of this capability is embedding, which learns high-level semantic representations that compactly encode user data while preserving meaning. The resulting embedding space [11] is a continuous high-dimensional latent domain in which raw user data (e.g., text tokens, sensor signals, or symbols) are projected into numerical vectors, with semantically similar inputs mapped to nearby points. This space supports compression and semantic preservation, and provides robust-

Received 15 May 2025; revised 13 October 2025; accepted 9 December 2025. Date of publication 12 December 2025; date of current version 20 February 2026. *(Corresponding authors: Hyun-Ho Choi; Jung-Ryun Lee.)*

Ki-Ho Lee is with the School of Electrical Engineering, Chung-Ang University, Seoul 06974, South Korea (e-mail: kiholee79@cau.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 Engineering and the 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/JSAC.2025.3643816

ness against noise and distortions. The embeddings within this space are typically generated through deep neural networks such as autoencoders [12], Transformers [13], or convolutional backbones [14], all of which serve as foundational blocks of today's large AI models (LAMs) and are designed to extract latent features aligned with the task objective [15]. In this context, AI models are trained end-to-end to project raw data, including images, text, or sensor signals, into a continuous latent space where semantically similar inputs are mapped to nearby vectors [16]. These embeddings not only serve as compressed transmission units but also enable critical operations such as interference suppression and semantic-level decoding. In particular, attention-based architectures like Transformers, which underpin state-of-the-art LAMs, are well suited for embedding generation because of their ability to model long-range dependencies and capture contextual meaning across complex data sequences [17], [18].

Complementing embedding-based representation, deep joint source–channel coding (JSCC) has emerged as a foundational component in semantic communication systems. By training the encoder and decoder jointly, JSCC learns continuous latent representations that are inherently resilient to fading and noise. This approach departs from traditional separation principles, enabling graceful degradation rather than abrupt cliff effects [19]. Recent advances have shown that semantic-aware JSCC can significantly improve image recovery [20], classification robustness, and even generalize to a variety of AI inference tasks [21]. Furthermore, Transformer-based LAMs, which already deliver state-of-the-art performance in channel prediction and traffic modeling [22], [23], are particularly suited to semantic tasks thanks to their powerful self-attention mechanisms, even under low signal-to-noise ratio (SNR) channel conditions [17]. Although deep JSCC has shown promising results in improving semantic robustness for individual users, its integration with multiple access remains largely unexplored.

Looking ahead to next-generation communication networks, scalable user multiplexing is expected to become increasingly critical, particularly in scenarios involving human-to-machine (H2M) and massive machine-type communication (mMTC). Conventional multiple access theory has largely focused on supporting a small number of users under asymptotic assumptions [19], [24]. However, these models do not directly extend to massive access scenarios, where thousands or even millions of devices must be supported simultaneously with stringent latency and reliability constraints [25]. In such environments, short-packet transmissions become necessary to reduce latency and receiver complexity [26], but they require a more refined information-theoretic framework that diverges from traditional Shannon-based metrics. Moreover, commonly adopted grant-based random access protocols may result in excessive scheduling delays and signaling overheads [27], [28].

In legacy systems such as 5G New Radio (NR), multiple access is typically achieved using orthogonal schemes like orthogonal frequency-division multiple access (OFDMA) and time-division multiple access (TDMA), or non-orthogonal schemes such as non-orthogonal multiple access (NOMA) [29]. While these approaches can be effective for moderate user densities, they rely on explicit grant signaling and rigid partitioning of time-frequency resources, which may lead to inefficiencies under dense or heterogeneous access demands. These limitations have motivated the search for more flexible and scalable multiple access mechanisms. A growing body of work has begun to explore embedding-aware protocols that relax rigid orthogonality. For instance, [30] proposes a hybrid NOMA framework where one user transmits semantically encoded signals while another transmits in a bit-wise fashion, improving spectral efficiency and offering robustness to interference. These directions signal a shift in how multiple users can be jointly supported in semantically meaningful and resource-efficient ways. However, it is noted that such schemes essentially realize multiplexing across heterogeneous transmission modes (semantic versus bit-based), rather than enabling pure multiplexing among semantic communication users in the embedding space.

Motivated by these insights, this paper explores a novel multiple access scheme called *shared embedding (SE)*. The core idea is to leverage high-dimensional embedding spaces for user multiplexing, where multiple users dynamically share a common embedding subspace and provide a new axis for multiple access, analogous to time, frequency, or code domains in conventional systems. Unlike conventional *dedicated embedding (DE)*, which can be understood as a resource-partitioning strategy where each user is assigned an exclusive portion of the embedding space, similar to allocating distinct time slots in TDMA or orthogonal subcarriers in OFDMA, thereby resembling the principle of orthogonal multiple access (OMA) as discussed in [31], SE allows flexible overlap among users, relying on AI-driven masking and representation learning to suppress inter-user interference. This dynamic allocation improves embedding reuse and spectral efficiency but introduces new analytical challenges in managing intra- and inter-user correlation. To address these challenges and assess the feasibility of the proposed approach, the main contributions of this paper are as follows:

- This work introduces a novel SE framework that enables scalable multiple access by allowing multiple users to share a Transformer-based embedding space, which stands in contrast to the rigid partitioning of DE schemes.
- The proposed study develops an AI-native encoder–decoder architecture that employs user-specific positional masking and attention-guided demultiplexing, and this design achieves efficient user separation without explicit orthogonalization.
- Theoretical analysis establishes closed-form expressions for key performance metrics, including signal-to-interference-plus-noise ratio (SINR), symbol error rate (SER), and capacity, and explicitly accounts for intra-user and inter-user correlations so that fundamental insights into embedding-domain multiplexing can be obtained.
- Extensive experiments validate the theoretical analysis and show consistent performance gains of SE over DE across user densities and SNR conditions under a fading environment, while also highlighting the potential of SE for integration with legacy multiple access schemes.

TABLE I
COMPARISON OF INTER-USER INTERFERENCE MINIMIZATION TECHNIQUES

Technique	Complexity	Scalability
Codebook-Based	Low	Limited by codes
Contrastive Learning	Medium	High, flexible
Interference Alignment (IA)	High	Moderate
Negative Correlation	High	Low

II. TECHNIQUES FOR MINIMIZING INTER-USER INTERFERENCE

Although correlation among users is often viewed as a source of interference, recent advances show that structured correlation design can be exploited to suppress it. This section outlines several representative strategies for minimizing inter-user interference in embedding-based systems.

One classical approach is to enforce orthogonality among user embeddings. This can be implemented by orthogonality regularization, which penalizes nonzero inner products among user-specific masks during training. Although effective, this approach is limited by the dimensionality of the embedding space, as the number of perfectly orthogonal vectors cannot exceed the dimensionality of the ambient space. Orthogonal spreading codes such as Walsh or Hadamard are examples used in traditional communication systems [32], [33].

A more flexible alternative is contrastive learning [34], which minimizes similarity (e.g., cosine similarity) between user embeddings through a contrastive loss. Although it does not enforce strict orthogonality, it encourages low inter-user correlation and scales better to large numbers of users.

Interference alignment (IA) offers another promising direction [35], [36]. Originally developed in the context of multiple-input multiple-output (MIMO) systems, IA aligns interference into low-dimensional subspaces, thereby freeing degrees of freedom for desired signals. Although more complex to realize in embedding-based domains, IA principles can still be applied through learned structures in neural attention.

Negative correlation is also a potential technique, where some user pairs are assigned deliberately anti-correlated embeddings to cancel their interference destructively. Although conceptually related to successive interference cancellation (SIC) [37], its scalability remains limited due to the difficulty of managing pairwise relations across many users. Table I summarizes the key trade-offs of these techniques.

These approaches have been investigated almost exclusively at the physical layer, where interference is suppressed through coding, beamforming, or linear precoding based on channel state information. Their direct extension to high-dimensional embedding spaces has not yet been systematically explored for semantic communications.

III. ATTENTION-BASED SHARED EMBEDDING FOR AI-NATIVE MULTIPLE ACCESS DESIGN

Inspired by contrastive learning and IA, the proposed SE framework achieves adaptive interference suppression not through explicit contrastive losses or manual subspace partitioning, but via Transformer-based attention masking that jointly learns to decorrelate user embeddings. This enables flexible and scalable multiple access through semantic-level coordination and representation learning, compared to traditional schemes that rely on rigid resource allocation.

Fig. 1 traces the end-to-end signal flow in the SE pipeline. The process begins with raw user-equipment embedding: every user converts its semantic message into a d-dimensional vector that occupies the same latent coordinates as the vectors of all other users. At this point, the representations overlap heavily, so the latent space offers little separation.

Separation is created by the next stage user-specific positional masking. A learnable mask $\mathbf{P}_u$ is applied element-wise to the raw embedding $\mathbf{E}_u$, rotating it into a user-unique subspace while preserving token-to-token coherence. Training pushes different masks toward near-orthogonality, so the masked embeddings of distinct users occupy almost orthogonal directions; in the figure this transition is annotated as a shift from "Low separation" to "High separation & alignment."

All masked embeddings are then summed and transmitted over the physical channel, resulting in a composite vector $\mathbf{Y}$ that is subject to both multiplicative fading and additive noise, as in a Rayleigh fading channel. The fading introduces random amplitude and phase distortions in addition to the noise, thereby spreading and rotating the composite embedding in the received domain. At the receiver, a multi-head Transformer decoder takes $\mathbf{Y}$ and, through its attention mechanism, locks each head onto the subspace associated with a user mask. In doing so, it isolates the desired signal, while aligned interference collapses into a low-rank component that subsequent layers can effectively suppress even under fading conditions. Finally, the decoder maps the cleaned embeddings back to the symbol domain, which produces per-user symbol streams with low error probability despite the combined effects of fading and noise.

Because the positional masks, the channel projection, and the decoder attention are trained jointly under a cross-entropy objective, the network learns to balance the antagonistic goals of maximizing intra-user coherence and minimizing inter-user correlation. The result is an AI-native multiple-access system in which many users can share a single high-dimensional embedding space without sacrificing decodability.

Fig. 2 illustrates a Transformer-based transmitter and receiver architecture designed to support the proposed SE framework. At the transmitter, user data sequences are first embedded and masked with user equipment (UE)-specific positional encodings, and the resulting user-specific embeddings are then multiplexed into a composite signal. The receiver employs user-specific attention strategies to demultiplex and reconstruct the data sequence of each user. The following subsections describe each component in greater detail, including the mechanisms for masking, attention-based separation, and embedding-driven demultiplexing.

A. Transmitter: Embedding Generation and Multiplexing

1) Input Representation and Embedding Generation: Let us consider U active users in the system, each transmitting a data sequence denoted as $\mathbf{X}_u = [x_{u,1}, x_{u,2}, \ldots, x_{u,T}]$, where T is the sequence length and $u \in \{1, \ldots, U\}$. To extract meaningful semantic representations from potentially diverse data

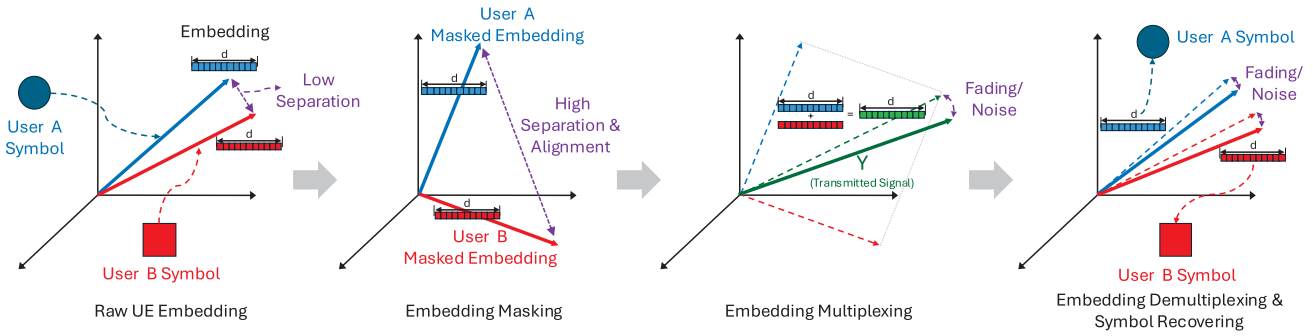

Fig. 1. System-level concept of the proposed SE framework.

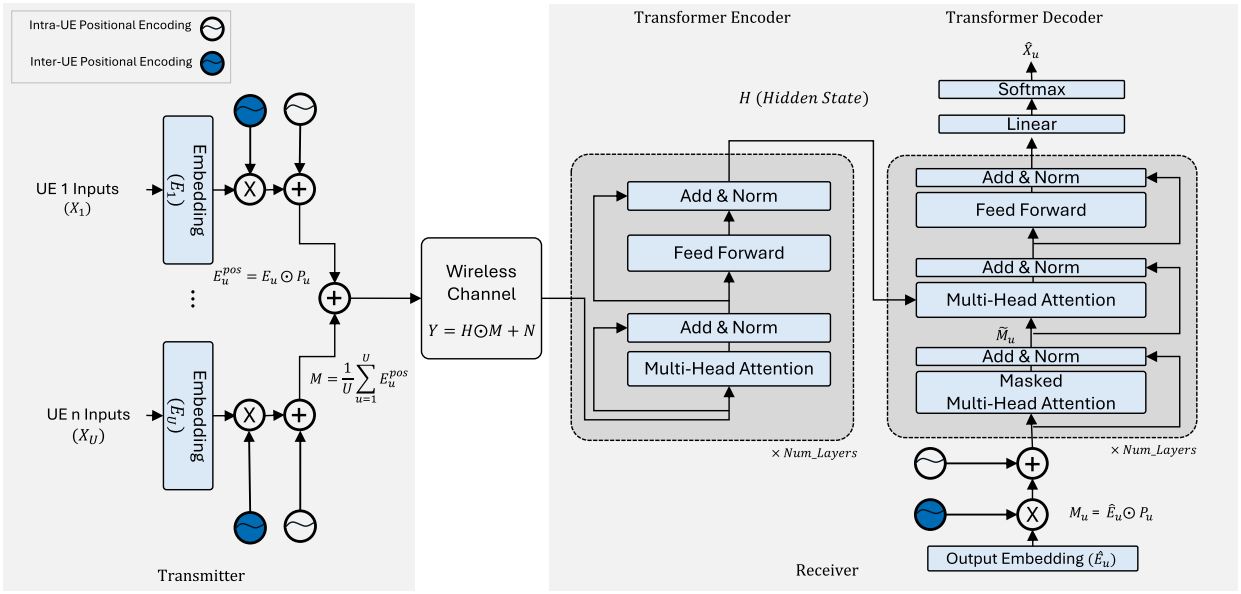

Fig. 2. Transformer-based multiple access architecture using UE-specific positional encoding.

sources (e.g., text tokens or symbol streams), each token $x_{u,t}$ is passed through a learnable embedding layer that produces a vector $\mathbf{E}_u \in \mathbb{R}^d$, where d is the embedding dimension. Here, $\mathbf{E}_u$ denotes the embedding vector corresponding to token $x_{u,t}$. This embedding layer can be either pre-trained for specific tasks such as text encoding or jointly optimized end-to-end for symbol-level communications.

2) Learnable UE Positional Encoding: Each user u is assigned a unique, learnable token-invariant positional encoding vector$\mathbf{P}_u \in \mathbb{R}^d$, which is initialized randomly and refined during training. These user-specific encodings serve as an additional "signature" that helps the receiver distinguish embeddings from different users within the shared space, with the entries of $\mathbf{P}_u$ acting as dimension-wise scaling factors. Masking values of $\mathbf{P}_u$ close to zero attenuate certain embedding dimensions, while larger masking values of $\mathbf{P}_u$ amplify them. During training, these values are optimized to increase coherence among embeddings of the same user while reducing similarity across different users, thereby shaping the embedding space so that the Transformer decoder can reliably separate user signals under diverse channel and user environments. To formalize this process, the element-wise combination of the user embedding and positional encoding for each token is written as

$$\mathbf{E}_u^{\text{pos}} = \mathbf{E}_u \odot \mathbf{P}_u, \tag{1}$$

which can equivalently be expressed in component form as$(\mathbf{E}_u^{\text{pos}})_i = E_{u,i} \cdot P_{u,i}$, for embedding dimension $i \in \{1, \ldots, d\}$. This operation effectively "masks the embedding of each user" with a unique positional encoding pattern, which is crucial to controlling the inter-user correlation and minimizing mutual interference.

Since the focus of this work is on user-level multiple access, the proposed framework primarily uses *inter-UE positional encoding* to distinguish between multiple users sharing the embedding space. Meanwhile, *intra-UE positional encoding*, which captures sequential or structural relationships between tokens *within* a user, is not explicitly considered in this work. However, we recognize that intra-user attention patterns are highly relevant for future extensions involving multi-modal user data (e.g., image-text fusion or sensor-gesture sequences). In such scenarios, intra-UE positional encoding could enable token-level attention modeling to enhance semantic representation and improve end-to-end reconstruction. This will be an important direction for future research.

3) Multiplexing of User Signals and Wireless Transmission: After positional masking, all user embeddings corresponding to a token are aggregated to form a multiplexed vector$\mathbf{M} = \sum_{u=1}^{U} \mathbf{E}_u^{\text{pos}}$, where $\mathbf{M} \in \mathbb{R}^d$ for each token. The i-th entry of the multiplexed vector is then given by

$$M_i = \sum_{u=1}^{U} W_{u,i}\, E_{u,i}, \tag{2}$$

where $W_{u,i}$ implicitly accounts for the effect of the positional mask $\mathbf{P}_u$ on embedding dimension i.

When transmitted over a Rayleigh fading channel, the multiplexed vector $\mathbf{M} \in \mathbb{R}^d$ for each token results in the received vector

$$\mathbf{Y} = \mathbf{H} \odot \mathbf{M} + \mathbf{N}, \quad \mathbf{Y}, \mathbf{H}, \mathbf{M}, \mathbf{N} \in \mathbb{R}^d, \tag{3}$$

where H_i denotes the element-wise Rayleigh fading coefficient for embedding dimension i, typically modeled as $H_i = \sqrt{Z_i^2 + W_i^2}/\sqrt{2}$ with $Z_i, W_i \sim \mathcal{N}(0,1)$. The additive Gaussian noise is $\mathbf{N} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Although the SE-based approach aggregates multiple users within a shared embedding space, it is important to note that transmission over the wireless medium can still coexist with legacy multiple access schemes such as OFDMA, TDMA, or single-carrier frequency-division multiple access (SC-FDMA). In practice, these conventional access techniques may be used to partition time-frequency or code resources among user groups or service types, while SE enables fine-grained signal multiplexing within each resource block. By introducing the embedding dimension as an additional domain for user separation, the system gains a scalable multiplexing capability that complements traditional resource axes. This hybrid approach improves both the scalability and overall resource granularity of the system, supporting a wide range of user densities and service requirements in future communication networks.

4) Complexity and Implementation Considerations: Compared to conventional orthogonal schemes, this multiplexing approach introduces two main computational considerations. First, the embedding dimension d must be large enough to support multiple users with a minimized correlation. Second, the positional encoding vectors $\mathbf{P}_u$ must be learned and updated according to the various channel conditions. Under Rayleigh fading, training accounts for multiplicative distortion caused by random channel gains in addition to additive noise, which ensures that the learned embeddings remain robust to the stochastic variability of wireless channels. Despite these additional steps, the overhead can be managed efficiently due to parallelized operations within modern deep learning frameworks. Furthermore, the system can be trained end-to-end, mapping raw user data to the transmitted signal, optimizing the entire pipeline via gradient-based methods for best performance across diverse channel conditions.

Importantly, the proposed Transformer-based user multiplexing framework is applicable to both uplink and downlink scenarios. In *downlink*, the base station aggregates user-specific embeddings into a composite signal and transmits it simultaneously to all users. In *uplink*, individual user devices transmit their respective encoded embeddings, which are naturally superimposed by the wireless channel and received as a composite signal at the base station. Under Rayleigh fading, the receiver sees a faded superposition $\mathbf{Y} = \mathbf{H} \odot \mathbf{M} + \mathbf{N}$, and the same demultiplexing architecture applies without modification by learning attention patterns that are invariant to the fading statistics. Therefore, in both cases, the multiplexed signal can be demultiplexed and decoded using the same Transformer-based receiver architecture. This implementation architecture is illustrated in the left portion of Fig. 2, where the shared embedding space is constructed by user-specific positional encoding at the transmitter.

B. Receiver: Demultiplexing and Reconstruction

Upon reception, the composite signal $\mathbf{Y}$, carrying superimposed embeddings from U users, is passed through the multi-head self-attention layers of a Transformer encoder to capture dependencies among all embedding dimensions, as illustrated in Fig. 3.

1) Input Preprocessing and Transformer Encoder: With Rayleigh fading, the encoder directly receives the faded superposition, i.e., the initial input is

$$\mathbf{H}^{(0)} = \mathbf{Y} = \mathbf{H} \odot \mathbf{M} + \mathbf{N}. \tag{4}$$

The encoder consists of L stacked layers, and the output of the l-th layer is denoted by $\mathbf{H}^{(l)}$. Each layer includes a multi-head self-attention (MHA) sublayer followed by a position-wise feed-forward network, both of which are wrapped with residual connections and layer normalization. Specifically, in the l-th layer, the input $\mathbf{H}^{(l-1)}$ is projected into queries, keys, and values as $\mathbf{Q}^{(l)} = \mathbf{H}^{(l-1)}\mathbf{W}_Q^{(l)}$, $\mathbf{K}^{(l)} = \mathbf{H}^{(l-1)}\mathbf{W}_K^{(l)}$, and $\mathbf{V}^{(l)} = \mathbf{H}^{(l-1)}\mathbf{W}_V^{(l)}$, where the projection matrices $\mathbf{W}_Q^{(l)}, \mathbf{W}_K^{(l)}, \mathbf{W}_V^{(l)} \in \mathbb{R}^{d \times d_k}$ are learnable parameters. The output of the layer is given by

$$\begin{aligned}\mathbf{H}^{(l)} = \text{LayerNorm}\Big(&\mathbf{H}^{(l-1)} \\ &+ \text{FeedForward}\Big(\text{LayerNorm}\big(\mathbf{H}^{(l-1)} \\ &+ \text{SelfAttn}(\mathbf{Q}^{(l)}, \mathbf{K}^{(l)}, \mathbf{V}^{(l)})\big)\Big)\Big).\end{aligned} \tag{5}$$

After processing through L such layers, the final encoder output is given by $\mathbf{H} = \mathbf{H}^{(L)}$.

2) Demultiplexing via Attention Mechanism: A key advantage of the SE approach is how the Transformer architecture naturally is well-suited for demultiplexing via attention-based separation. After encoding the received signal $\mathbf{Y}$ into a shared high-dimensional representation $\mathbf{H}$, the model enables each user to selectively retrieve its own information from this common representation using masked attention mechanisms, as further illustrated in the bottom path of Fig. 3.

Unlike traditional applications of Transformers, where attention is used to model intra-sequence dependencies (e.g., for next-token prediction in language models), the SE framework employs attention for inter-user signal separation within a shared sequence. Here, a single input sequence contains superimposed information from multiple users, and attention is repurposed to extract user-specific components rather than the semantic context. Despite this change in objective, the core Transformer structure remains unchanged; instead, the distinction lies in how queries are constructed and interpreted.

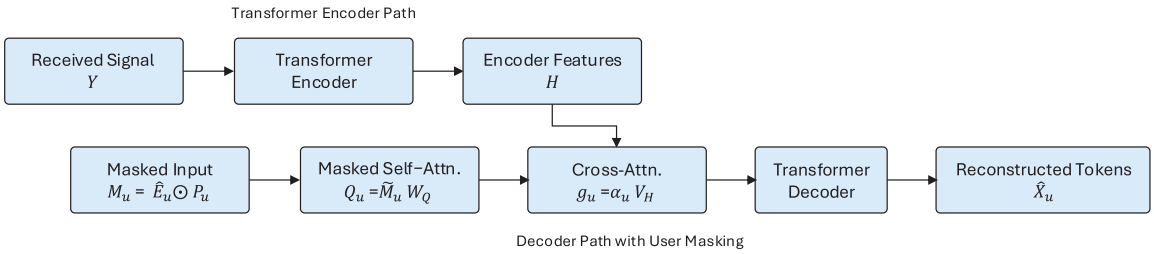

Fig. 3. Transformer-based multiuser demultiplexing and reconstruction procedure. The shared encoder output $\mathbf{H}$ is queried by user-specific masked embeddings to isolate each user's signal for final decoding.

For each user u, the decoder path begins by generating a masked input $\mathbf{M}_u = \hat{\mathbf{E}}_u \odot \mathbf{P}_u$, where $\hat{\mathbf{E}}_u$ is the estimated embedding of the user and $\mathbf{P}_u$ is the user-specific positional encoding reused from the transmitter. This masked embedding is then processed through a masked multi-head self-attention mechanism to yield a refined query representation $\tilde{\mathbf{M}}_u = \mathrm{MaskedMHA}(\mathbf{M}_u)$ which is shown in the decoder's lower attention block in Fig. 2. The processed embedding is further transformed into the user query $\mathbf{Q}_u = \tilde{\mathbf{M}}_u \mathbf{W}_Q$, which interacts with the encoder output $\mathbf{H}$, projected into key and value representations as $\mathbf{K}_H = \mathbf{H}\mathbf{W}_K$ and $\mathbf{V}_H = \mathbf{H}\mathbf{W}_V$, respectively.

The attention weights are computed by

$$\alpha_u = \mathrm{Softmax}\left(\frac{\mathbf{Q}_u \mathbf{K}_H^\top}{\sqrt{d_k}}\right), \tag{6}$$

guiding the model to selectively focus on user-specific segments of $\mathbf{H}$, thereby minimizing interference from other users.

3) Data Reconstruction With Transformer Decoder: Finally, each user's data is reconstructed by feeding the attended context $\mathbf{g}_u = \alpha_u \mathbf{V}_H$ into a Transformer decoder, yielding the decoded output

$$\hat{\mathbf{X}}_u = \mathrm{TransformerDecoder}(\mathbf{g}_u, \mathbf{H}), \tag{7}$$

where cross-attention layers further refine the user-specific representation. The input $\mathbf{g}_u$ serves as a user-adapted query embedding derived from the attention scores α_u, which selectively extract the relevant content of the user u from the encoder values $\mathbf{V}_H$. This vector provides a condensed representation of the user-specific context, guiding the decoder to reconstruct only the intended user's token sequence from the shared embedding space.

A final linear mapping (or multi-layer perceptron (MLP)) translates the decoder outputs to discrete token probabilities. The predicted output $\hat{\mathbf{X}}_u \in \mathbb{R}^{T \times C}$ represents the probabilities of the class for each time step t and symbol class C, from which the most likely token $\hat{x}_{u,t} = \arg\max \hat{\mathbf{X}}_{u,t}$ can be selected as the reconstructed symbol for user u. These predictions are compared with the original input symbols during training, forming the basis for end-to-end optimization.

Note: The $\mathrm{TransformerDecoder}(\cdot)$ block internally includes stacked self-attention, cross-attention, and feedforward layers, following the standard Transformer architecture.

4) Training Methodology and Considerations: The entire transmit–receive pipeline is differentiable, enabling end-to-end training. Typically, a task-specific loss function is used for optimization; for example, in symbol prediction, the cross-entropy (CE) loss is computed as

$$\mathcal{L} = \sum_{u=1}^{U}\sum_{t=1}^{T} \mathrm{CE}(\hat{x}_{u,t}, x_{u,t}) = -\sum_{u=1}^{U}\sum_{t=1}^{T}\sum_{i=1}^{C} x_{u,t}^{(i)} \log \hat{x}_{u,t}^{(i)}, \tag{8}$$

where C denotes the number of symbol classes, $x_{u,t}^{(i)}$ represents the one-hot encoded ground truth for user u, token t, and class i, and $\hat{x}_{u,t}^{(i)}$ is the predicted class probability [38]. Under the Rayleigh model, mini-batch generation samples $\mathbf{H}$ and $\mathbf{N}$ according to the desired fading regime and SNR. During backpropagation, gradients flow through both the transmitter (embedding and masking) and the receiver (Transformer encoder-decoder). This end-to-end optimization drives down inter-user correlation and interference while preserving intra-user coherence.

Overall, this AI-native multiple access framework leverages the Transformer's capacity to learn fine-grained relationships among tokens and users. With Rayleigh fading, it achieves robust demultiplexing through attention guided by embedding masking in a shared embedding space.

IV. Experimental Results

This section presents an empirical evaluation of the proposed SE framework, focusing on how the learned masking and attention mechanisms facilitate scalable multiuser communication within a shared embedding space. The experiments are designed to verify whether the system can suppress interference and preserve semantic coherence under varying signal conditions and user densities. We first examine how user-specific masking leads to decorrelated embeddings and gradually induces IA-like behavior during training. This is followed by a quantitative analysis of how these effects translate into SER performance. In parallel, we investigate the impact of Transformer model variants, including changes in depth and training strategy, to understand how architectural configuration influences the effectiveness of the SE framework. These evaluations provide a system-level perspective on how embedding design, model structure, and learning dynamics interact to support robust semantic multiple access.

A. Transformer-Based Training Setup

To investigate the impact of user-specific masking and token-level embeddings on interference mitigation, we employ a Transformer-based SE model introduced in Section III. The

TABLE II
LAM PARAMETERS FOR SE EMBEDDING ANALYSIS

Parameter	Value
AI Model Type	Separate per-SNR model, Unified training model
Transformer Depth (Enc/Dec)	1–4
Number of Users (N_{user})	8, 16
Embedding Dimension (d)	64–256
Token Sequence Length (T)	4
Vocabulary Size (V)	4
SNR Levels	0–25 dB
Channel Model	Rayleigh fading
Optimizer	AdamW [40]
Learning Rate (η)	1×10^{-4}
Epochs per Experiment (N_{epoch})	1–100
Batches per Epoch (N_{batch})	500–4000, 10,000
Loss Function ($\mathcal{L}$)	Cross-Entropy loss (CE)
Training Optimization	Gradient-based update: $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$
Parameter Initialization	PyTorch defaults, token/user embeddings $\sim \mathcal{N}(0,1)$
Evaluation Metric	SER, $\left\|\sum_{u=1}^{U} \mathbf{E}_u^{\text{pos}}\right\|^2$

system is designed to accommodate multiple users simultaneously sharing the embedding space, where interference is shaped and suppressed through learned user-specific masks and attention dynamics. During training, multiple users jointly occupy the embedding space to transmit short sequences of T tokens each, drawn from a discrete vocabulary of size V, over a Rayleigh fading channel. Synthetic symbol sequences are used to maintain control over interference patterns while ensuring reproducibility. The model is trained using the AdamW [39] optimizer with a learning rate $\eta = 1 \times 10^{-4}$, and optimized to minimize cross-entropy loss. θ, representing the trainable parameters of the Transformer model (e.g., attention weights, projection matrices, and embeddings), is updated via gradient descent according to $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$. Each training run consists of 500 to 4,000 or 10,000 batches per epoch and spans 1 to 100 epochs depending on the scenario. Unless otherwise noted, we rely on the default PyTorch initialization. Token embedding matrices are initialized from a zero-mean unit-variance normal distribution. These experimental settings are summarized in Table II.

To analyze the trade-off between model complexity and SER accuracy, and to gain architectural insights into the performance of SE-based multi-user communication, we explore two axes of variation that define a family of Transformer-based architectures within the SE framework. The first is the training scheme: in the per-SNR training, separate models are trained for each user count and SNR value, offering precise performance tuning but incurring significant computational cost. In contrast, the unified training uses a single model trained on a mixture of SNR conditions without explicit SNR tokens, requiring the Transformer to implicitly learn robust generalization across channels. For both of the per-SNR training and unified training, we divide the dataset into disjoint training and validation subsets. Both splits are generated using the same synthetic symbol generator, which guarantees consistent token distributions across sets. As a result, the model is not affected by distribution shift, and its generalization performance can be evaluated in a clean and unbiased manner under varying SNR and user configurations.

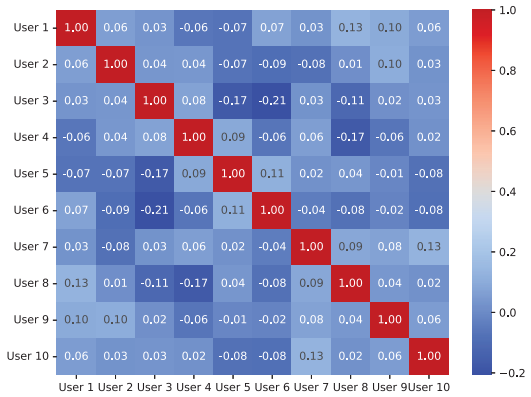

Fig. 4. Correlation heatmap of learned user embedding masks (inter-UE positional encodings) for 10 users at SNR = 20 dB. The average off-diagonal correlation = 0.0045.

The second axis involves architectural depth: by varying the number of encoder and decoder layers from two to four, we analyze how model complexity influences SER. This allows us to quantify the performance gain from deeper models, balanced against their increased computational cost.

These experimental settings are designed to explore key considerations for scalable SE-based multiple access, such as the trade-off between specialization and generalization, and the sensitivity of system performance to architectural configurations.

B. Correlation Heatmap of Inter-User Embedding Masks

To assess the degree of separability among user embeddings, we compute the correlation matrix of learned embedding masks, referred to as inter-user positional encodings, across different users. Here, we use $r_P^{(u,v)}$ to denote the *sample Pearson correlation coefficient* between the mask vectors of user u and user v, which is defined as follows [40]:

$$r_P^{(u,v)} = \text{corr}(\mathbf{P}_u, \mathbf{P}_v) = \frac{\sum_{i=1}^{d} (P_{u,i} - \bar{P}_u)(P_{v,i} - \bar{P}_v)}{\sqrt{\sum_{i=1}^{d} (P_{u,i} - \bar{P}_u)^2} \cdot \sqrt{\sum_{i=1}^{d} (P_{v,i} - \bar{P}_v)^2}}, \quad (9)$$

where $\bar{P}_u = \frac{1}{d}\sum_{i=1}^{d} P_{u,i}$ and $\bar{P}_v = \frac{1}{d}\sum_{i=1}^{d} P_{v,i}$ are the sample means of the positional encoding vectors.

Fig. 4 visualizes the pairwise correlations among the positional encodings of 10 users. The diagonal entries are all equal by definition, while the off-diagonal values reflect the degree of similarity between the masks of different users. The consistently low off-diagonal entries observed here indicate near-orthogonality between user-specific positional encodings, an essential condition for effective demultiplexing within the shared embedding space. Most of these coefficients cluster very close to zero; however, one can still observe a slight 'drift' of magnitude, which arises from stochastic gradient noise and IA. These inter-user correlations are not manually enforced, but rather emerge naturally during training as the model iteratively updates each $\mathbf{P}_u$ to minimize mutual interference while retaining intra-user coherence.

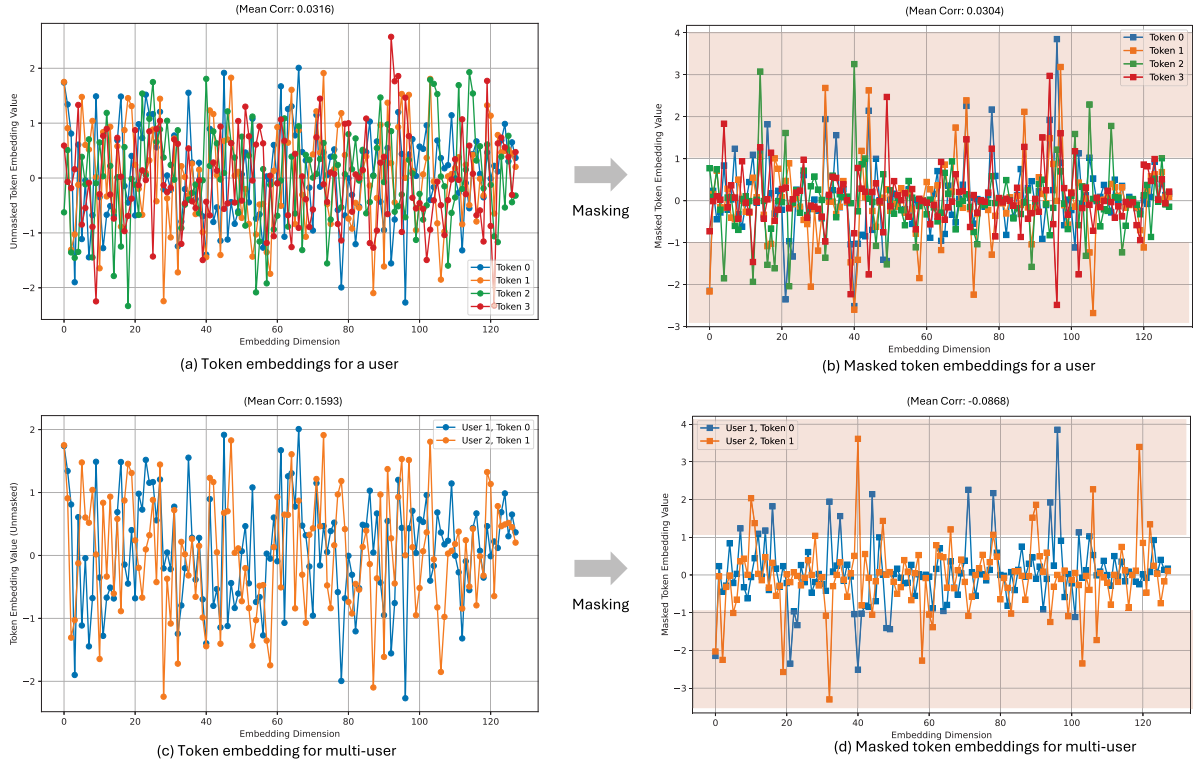

Fig. 5. Visualization of embedding masking behavior in the SE framework. (a) Raw token embeddings $\mathbf{E}_u$ for a single user show strong intra-user coherence. (b) Masked embeddings $\mathbf{E}_u^{\text{pos}}$ maintain this coherence with minimal distortion.(c) Raw token embeddings from multiple users exhibit weak inter-user separation.(d) The masked multiuser embeddings exhibit strong decorrelation across users, effectively suppressing interference.

C. Embedding Mask Visualization and Correlation Analysis

To understand how the proposed SE framework controls inter-user interference and preserves intra-user coherence, we visualize and quantify the impact of user-specific embedding in Fig. 5. As defined previously in (1), each user's masked embedding is obtained by element-wise multiplication between the token embedding matrix $\mathbf{E}_u \in \mathbb{R}^{T\times d}$ and the user-specific learnable positional encoding $\mathbf{P}_u \in \mathbb{R}^{1\times d}$, that is,

$$\mathbf{E}_u^{\text{pos}} = \mathbf{E}_u \odot \mathbf{P}_u. \tag{10}$$

Fig. 5 illustrates the effects of user-specific masking on embedding structure. Subplots (a) and (b) show the raw and masked token embeddings $\mathbf{E}_u$ and $\mathbf{E}_u^{\text{pos}}$ for a single user, respectively. Both embedding sets exhibit strong intra-user coherence, indicating that the Transformer has learned a consistent semantic structure across tokens. The masking operation preserves this internal alignment, as reflected in the nearly unchanged average intra-user correlation values of 0.0304 and 0.0316 before and after masking, respectively.

Subplots (c) and (d) compare the token embeddings of multiple users before and after masking. Prior to masking, the embeddings exhibit moderate inter-user overlap, with a correlation of 0.1593, indicating insufficient separation in the shared space. After applying user-specific positional masks, the inter-user correlation is reduced to -0.0868, demonstrating that the masking strategy effectively enhances user separation and mitigates interference.

Interestingly, beyond simple decorrelation, the masking process induces a form of *selective dimensional activation*, whereby the signal energy becomes concentrated in a small subset of dimensions that is unique to each user. These user-specific masks suppress irrelevant coordinates while amplifying those that best capture discriminative features, so the cumulative embedding of every user resides in a significantly lower-rank subspace. Because of this rank collapse, inter-user interference is confined to a common low-dimensional subspace, which is exactly the condition required for latent space IA, while the remaining orthogonal dimensions carry the desired signals.

This behavior closely parallels spatial-multiplexing precoding in network MIMO [41]: Just as virtual antennas steer user streams into (semi)orthogonal spatial lobes, the SE framework steers each user's embeddings into a distinct latent subspace. The result is a form of latent-space beamforming that allows simultaneous transmissions to coexist in the shared embedding space but remain readily separable at the decoder. In addition, although token-to-token correlations within a user remain low, indicating symbol-level independence, the masked embeddings still maintain consistent internal structure across each user's sequence. This preservation of intra-user coherence after applying user-specific positional masking is particularly important as it ensures that the semantic meaning embedded in each user's message is not disrupted during modulation. As a result, the decoder is able to consistently recover meaningful token representations, even when the embedding space is shared and affected by channel noise. This combination

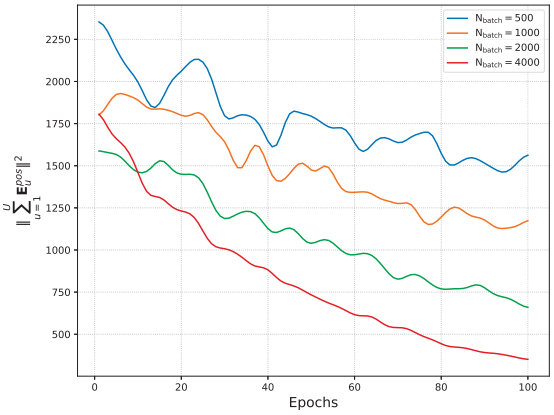

Fig. 6. Smoothed norm of the summed masked-embedding vectors, $\left\|\sum_u \mathbf{E}_u^{\text{pos}}\right\|^2$, measured over training epochs for varying batch sizes ($N_{\text{batch}} = 500$–4000), under SNR $= 15dB$, $U = 16$ and $d = 128$.

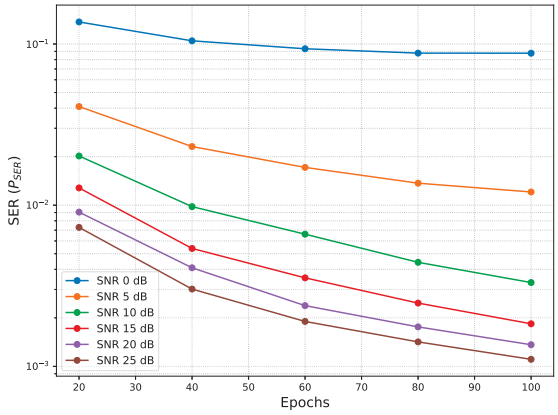

Fig. 7. P_{SER}.

of stable intra-user alignment, reduced inter-user correlation, and selective dimension modulation enables reliable multiuser demultiplexing.

Overall, the SE framework integrates sparsity, structural consistency, and latent-space separation to suppress interference in a manner functionally similar to interference-aware beamforming techniques employed in MIMO-based physical-layer communication systems.

D. Interference Suppression and SER

To evaluate how well the SE framework suppresses interference, we track the smoothed epoch-wise evolution of the total power of the masked embeddings under Rayleigh fading channel, $\left\|\sum_{u=1}^{U} \mathbf{E}_u^{\text{pos}}\right\|^2$, which is plotted in Fig. 6 for four training schedules that differ only in the number of batches per epoch. Here, $\|\cdot\|^2$ denotes the squared Euclidean norm measuring the total energy of the vector. This norm acts as a proxy for residual interference energy: the smaller its value, the more completely the user embeddings cancel when summed. All four curves exhibit a consistently decreasing trend, showing that the network progressively aligns user embeddings into mutually canceling directions. The larger N_B accelerates this decline because each epoch delivers more stochastic gradients, which accelerate the self-organization of the latent space. The behavior is analogous to *IA* in multiuser MIMO systems, where transmitters steer interference into a low-dimensional subspace that the receiver can null [42]. The same principle emerges automatically in a semantic latent space.

Next, we examine how this alignment translates into decoding accuracy by monitoring the SER, denoted by P_{SER}, across epochs and SNR values (0–25 dB), which is defined as

$$P_{\text{SER}} = \frac{\text{number of incorrectly detected tokens}}{\text{total number of transmitted tokens}}. \tag{11}$$

As Fig. 7 shows, P_{SER} decreases more rapidly as SNR increases, with noticeable improvements already at 10–15 dB and a sharp drop at 20–25 dB, where the error rate falls below 10^{-3} around 40 epochs. At lower SNR values (0–5 dB), the decrease is more gradual, reflecting the stronger impact of noise on the learning process. This trend arises because higher SNR values yield more reliable symbol observations at the decoder, which in turn allows the Transformer to more effectively extract gradient signals for learning attention patterns and refining embedding alignment. In particular, as the signal becomes cleaner, the model can more easily identify and reinforce meaningful token dependencies, leading to faster suppression of inter-user interference and improved convergence. At lower SNRs, the learning process is hindered by increased noise, which obscures semantic structure and slows the model's ability to organize the embedding space.

Observing the residual power and P_{SER} curves, we can see that the SE framework steadily achieves interference reduction throughout training, thereby enabling robust multiuser performance across a wide SNR range.

E. Impact of Transformer Model Variants

To investigate the impact of model configurations on performance, we examine the effect of the training strategy, either per-SNR or unified SNR-agnostic learning, and the Transformer depth defined by the number of encoder and decoder layers on the SER.

As illustrated in Fig. 8, Transformer depth has a significant effect on SER performance. Models with a greater number of encoder and decoder layers achieve consistently lower error rates across all SNR levels. In particular, the configuration with four encoder and decoder layers outperforms the shallower configurations with one or two layers, confirming the benefit of increased model capacity in learning robust token-level representations.

With regard to training scheme, models trained separately for each SNR condition exhibit better performance compared to their unified counterparts. Although unified training enables a single model to generalize across a wide range of SNRs, its performance exhibits degradation under noisy conditions. This is attributed to the increased difficulty in optimizing over a broader distribution of channel states.

These results show that both model depth and training scheme play a role in determining the trade-off between performance and complexity. Deeper Transformer configurations offer superior SER performance, albeit at the cost of

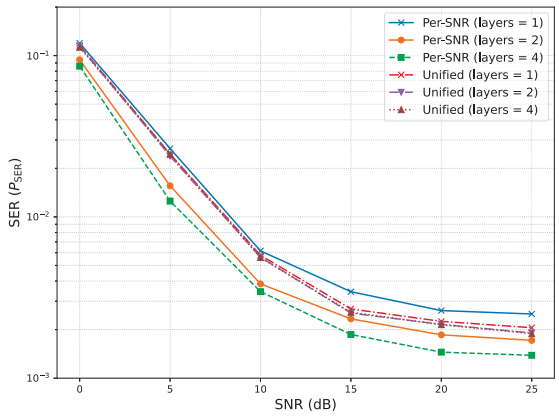

Fig. 8. SER performance comparison across SNR values under Rayleigh fading channels. The figure includes empirical SER results for both per-SNR training and unified SNR-agnostic training, using transformer configurations with encoder and decoder layers equal to 1, 2, and 4. All experiments were conducted under fixed conditions: $U = 16$, $d = 128$, and $N_{\text{batch}} = 10{,}000$.

increased computational burden. Similarly, per-SNR training delivers more stable convergence and error performance, but requires multiple models and training runs for deployment. Unified training, in contrast, offers improved scalability and efficiency, but is more sensitive to variation in training conditions. This sensitivity indicates that unified models generally require more careful optimization strategies, such as increased model depth, enhanced regularization, or context-aware/meta-learning schemes, to maintain robustness across a wide range of SNRs.

F. Experimental Results and Implications

The experimental results provide supporting evidence for the emergence of a conceptual chain from masking to embedding decorrelation to IA, as anticipated by the analytical framework. This progression is observed through three successive stages: mask design, embedding behavior, and interference suppression dynamics.

At the *masking level*, the heatmap shown in Fig. 4 reveals that the positional masks assigned to each user become increasingly decorrelated during training. Specifically, the pairwise correlations between the user masks approach zero, implying that the masks span nearly orthogonal subspaces. This mask orthogonality plays a key role in reducing inter-user interference at the embedding level.

At the *embedding level*, Fig. 5 shows that this decorrelation is reflected in the masked embeddings: different users' embeddings remain largely uncorrelated, while intra-user coherence, critical for semantic information preservation, is maintained. This suggests that attention-guided masking facilitates statistical separation among users in the embedding space.

At the *IA level*, the simultaneous decline in the norm of the aggregated masked embeddings and the SER, as seen in Fig. 6, indicates that the residual interference becomes increasingly structured in a way favorable to suppression. Although a direct rank measurement is not available, this trend suggests that the aggregate interference energy is concentrating in a more compressible or partially aligned subspace, which can be mitigated by linear projection. Although a precise characterization of the rank of the summed interference matrix is beyond the current experimental scope, the empirical results indicate a coherent progression. As the masks approach orthogonality, the embeddings of different users evolve toward negative correlation while preserving intra-user coherence; this separation in the embedding space fosters IA, which, in turn, suppresses residual interference at the receiver and yields consistent reductions in SER as shown in Fig. 7. End-to-end training implicitly realizes this progression within the SE framework by jointly optimizing masking and attention in the shared embedding space.

At the model-training level, the experimental results further reveal how specific Transformer configurations shape performance in SE-based communication in Fig. 8. The experimental results suggest that Transformer depth and training granularity jointly influence the trade-off between generalization, accuracy, and efficiency, and we adopt per-SNR training with four encoder and decoder layers as the default configuration in subsequent experiments to enhance accuracy (low SER across SNRs), stability (robust convergence and reproducibility), and model expressiveness (sufficient capacity to capture inter-user embedding interactions). Beyond this baseline, the results also indicate the need for more advanced training strategies that can reconcile the scalability of unified models with the accuracy of per-SNR training.

V. Numerical Analysis

In this section, we present a detailed theoretical analysis of the SE framework, beginning with a *correlation-based theoretical baseline* that highlights the interaction among user correlations, effective SINR, and capacity. This baseline is derived under idealized assumptions on intra-user and inter-user correlations, providing fundamental insights into how user embedding structures impact system performance.

While the above results are assuming fixed user and embedding correlations, extended training (e.g., with more epochs or batches) can lead to superior real-world performance that surpasses this baseline. Building upon this foundation, we introduce an additional alignment parameter to capture the degree of attention-based inter-user interference mitigation. This extended formulation enables us to analytically evaluate the benefits of implicit IA emerging from the SE framework and to compare its performance against the conventional DE scheme under varying correlation and alignment conditions.

A. Expected Signal and Interference Power in SE

Consider a system with U users, where each user transmits over d embedding dimensions. The combined transmitted signal at the i-th embedding dimension is expressed in equation (2) as:$M_i = \sum_{u=1}^{U} W_{u,i} E_{u,i}$,where $W_{u,i}$ represents the weight of the embedding mask applied to the user u at dimension i, and $E_{u,i}$ denotes the embedding value of the user u at dimension i.

The total received signal power across all embedding dimensions is formulated as

$$P_{\text{signal}} = \sum_{i=1}^{d} \left(\sum_{u=1}^{U} W_{u,i} E_{u,i} \right)^2. \tag{12}$$

To compute the expectation of P_{signal}, we expand the square term:

$$P_{\text{signal}} = \sum_{i=1}^{d}\sum_{u=1}^{U} W_{u,i}^2 E_{u,i}^2 + 2\sum_{i=1}^{d}\sum_{u<v} W_{u,i}W_{v,i}E_{u,i}E_{v,i}. \tag{13}$$

We impose the following statistical assumptions to facilitate the calculation of expectations.

- $\mathbb{E}[W_{u,i}^2] = \sigma_W^2$
- $\mathbb{E}[W_{u,i}W_{v,i}] = \rho_W\sigma_W^2$ for $u \neq v$, where ρ_W denotes the inter-user correlation coefficient among mask weights.
- $\mathbb{E}[E_{u,i}^2] = \sigma_E^2$
- $\mathbb{E}[E_{u,i}E_{v,i}] = \rho_E\sigma_E^2$ for $u \neq v$, where ρ_E denotes the inter-user correlation coefficient among embedding values.

The expectation of the first term simplifies to the following:

$$\mathbb{E}\left[\sum_{i=1}^{d}\sum_{u=1}^{U} W_{u,i}^2 E_{u,i}^2\right] = \sum_{i=1}^{d}\sum_{u=1}^{U}\sigma_W^2\sigma_E^2 = dU\sigma_W^2\sigma_E^2. \tag{14}$$

The second cross-term, which accounts for inter-user correlation, is computed as:

$$\mathbb{E}\left[2\sum_{i=1}^{d}\sum_{u<v} W_{u,i}W_{v,i}E_{u,i}E_{v,i}\right] = 2\sum_{i=1}^{d}\sum_{u<v}\rho_W\rho_E\,\sigma_W^2\sigma_E^2$$
$$= 2d\rho_W\rho_E\,\sigma_W^2\sigma_E^2\frac{U(U-1)}{2}$$
$$= d\rho_W\rho_E\,\sigma_W^2\sigma_E^2 U(U-1). \tag{15}$$

For notational convenience, we define the *effective inter-user correlation coefficient* as $\tilde{\rho} = \rho_W\rho_E$. Combining both terms, the total expected received signal power $\mathbb{E}[P_{\text{signal}}]$ becomes as follows:

$$\mathbb{E}[P_{\text{signal}}] = dU\sigma_W^2\sigma_E^2 + d\tilde{\rho}\,\sigma_W^2\sigma_E^2 U(U-1)$$
$$= d\sigma_W^2\sigma_E^2 U\left(1+\tilde{\rho}(U-1)\right). \tag{16}$$

This result delineates how the total received signal power is composed of the direct signal contributions from each user and the cross-term interference contributions arising from inter-user correlations. The expression in (16) was obtained by expanding the total signal power across all users and computing its expectation.

With this definition, the $P_{\text{signal}}^{\text{desired}}$ becomes

$$P_{\text{signal}}^{\text{desired}} = d\,\sigma_W^2\sigma_E^2\left(1+\rho_s(d-1)\right). \tag{17}$$

This formulation highlights that the d embedding dimensions of a single user behave analogously to a d-element antenna array: when positively correlated, they add coherently and yield an array-gain-like improvement in the effective SNR [43].

Next, we analyze the *interference power*, which arises from the $U-1$ users simultaneously sharing the embedding space with the target user. Each of these users contributes interference due to the overlap in their embedding vectors and the resulting multiplexed signal. This interference is not simply additive white noise; rather, it reflects *structured interference* influenced by correlations among user embeddings. Following the same expectation-expansion approach used earlier for the total and desired signal powers, but now applied to the $U-1$ interfering users, the aggregate interference power $P_{\text{interference}}$ can be expressed as

$$P_{\text{interference}} = d\,\sigma_W^2\sigma_E^2(U-1)\left(1+\rho_i(U-2)\right), \tag{18}$$

where the term $(U-1)$ represents the number of interfering users, excluding the target user, and ρ_i denotes the inter-user correlation coefficient among their effective embedding signals. For each interfering user, the contribution to the total interference power is proportional to $\sigma_W^2\sigma_E^2$, under the assumption that embedding dimensions are independent and have zero mean.

A crucial aspect is that the equicorrelation assumption for $U-1$ interfering users requires the covariance matrix to remain positive semidefinite, which imposes the range $-\frac{1}{U-2} \leq \rho_i \leq 1$. Thus, negative values of ρ_i are possible and represent partial interference suppression, while positive values indicate coherent interference growth. In the worst-case scenario where $\rho_i = 1$, the interference aligns coherently even under Rayleigh fading, scaling quadratically as $\sim (U-1)^2$ and leading to severe performance degradation due to overlap of embedding directions. In contrast, when $\rho_i = 0$, the users are uncorrelated, and interference grows only linearly with the number of users, allowing more scalable multiuser performance.

For the Rayleigh fading channel, the received signal includes multiplicative fading coefficients in addition to additive white Gaussian noise (AWGN). Assuming unit-variance Rayleigh fading, the average noise power remains $P_{\text{noise}} = d\sigma_N^2$, while the instantaneous signal and interference powers fluctuate with the fading realizations. Based on this fading-aware model, we aggregate the per-dimension signal, interference, and noise contributions (averaged over the fading realizations) to obtain the per-user effective SINR as follows. By combining the powers of the desired signal, inter-user interference, and background noise, the effective SINR for the SE scheme, denoted by $\gamma_{\text{eff}}^{\text{SE}}$, is derived as follows:

$$\gamma_{\text{eff}}^{\text{SE}} = \frac{d\sigma_W^2\sigma_E^2\left(1+(d-1)\rho_s\right)}{d\sigma_W^2\sigma_E^2(U-1)\left(1+(U-2)\rho_i\right)+d\sigma_N^2}$$
$$= \frac{\sigma_W^2\sigma_E^2\left(1+(d-1)\rho_s\right)}{\sigma_W^2\sigma_E^2(U-1)\left(1+(U-2)\rho_i\right)+\sigma_N^2}. \tag{19}$$

Here, ρ_s and ρ_i represent the intra-user and inter-user embedding correlation coefficients, respectively. The parameters σ_W^2, σ_E^2, and σ_N^2 denote the variances of the user-specific weight vector, token embedding, and additive noise, respectively. The term $\gamma_{\text{eff}}^{\text{SE}}$ captures how attention-driven alignment improves effective SINR by leveraging both intra-user coherence and inter-user interference suppression. To simplify the expression, we define the single-user SNR as $\text{SNR} = \frac{\sigma_W^2\sigma_E^2}{\sigma_N^2}$. This definition remains valid under unit-variance Rayleigh fading since the additive noise is AWGN and the fading-induced power fluctuations are accounted for in the averaged signal and interference terms. Accordingly, by substituting the normalized SNR form into the effective SINR expression, the final closed-form of $\gamma_{\text{eff}}^{\text{SE}}$ becomes

$$\gamma_{\text{eff}}^{\text{SE}} = \frac{\left(1+(d-1)\rho_s\right)\text{SNR}}{(U-1)(1+(U-2)\rho_i)\,\text{SNR}+1}, \tag{20}$$

which highlights the contributions of intra-user correlation ρ_s and inter-user alignment ρ_i to the effective SINR of the SE system. Motivated by the observed attention patterns in the Transformer (where user-specific masking compresses cross-user energy into a lower-dimensional subspace), we parameterize interference suppression directly via an alignment coefficient. We introduce the alignment coefficient $\rho_{\text{ia}} \in [0, 1]$, which quantifies the effectiveness of attention-driven IA in the embedding space. Conceptually, ρ_{ia} measures how well the learned positional masks and attention mechanisms compress inter-user interference into a low-dimensional subspace, thereby enabling its suppression. A larger value of ρ_{ia} indicates stronger alignment and more effective interference mitigation, while also reflecting the degree to which the Transformer's masking and attention mechanisms decorrelate user embeddings.

To formally capture this effect, we establish its relationship with the inter-user correlation coefficient ρ_i as follows:

$$\rho_i = -\frac{\rho_{\text{ia}}}{U-2}, \quad \text{for } U > 2, \tag{21}$$

where U is the number of users sharing the embedding space. This construction ensures that as $\rho_{\text{ia}} \to 1$, the inter-user correlation ρ_i approaches a strongly negative value, effectively driving inter-user embeddings toward orthogonality. In contrast, smaller values of ρ_{ia} correspond to weaker alignment and higher residual interference.

Substituting the alignment-based relation in (21) into the SINR in (20) yields a closed-form expression that depends directly on the alignment coefficient. Since $1 + (U-2)\rho_i = 1 + (U-2)\left(-\frac{\rho_{\text{ia}}}{U-2}\right) = 1 - \rho_{\text{ia}}$ for $U > 2$, the γ_{eff}^{SE} becomes

$$\gamma_{\text{eff}}^{\text{SE}} = \frac{(1 + (d-1)\rho_s)\,\text{SNR}}{(U-1)\,(1-\rho_{\text{ia}})\,\text{SNR} + 1}, \quad U > 2. \tag{22}$$

Here, a larger ρ_{ia} decreases the interference factor $(U-1)(1-\rho_{\text{ia}})$, thereby increasing the overall SINR. For completeness, when $U \leq 2$ the inter-user term vanishes and the denominator reduces to 1, recovering the interference-free form.

Overall, the joint characterization based on ρ_s and ρ_{ia} clarifies the respective roles of coherence and alignment in SE systems. The intra-user correlation ρ_s enhances the desired signal through coherent summation across the embedding dimensions, while stronger alignment (larger ρ_{ia}) directly suppresses multiuser interference through the factor $1 - \rho_{\text{ia}}$ in the denominator. As a result, the performance of SE systems depends on jointly maximizing ρ_s and ρ_{ia} by means of attention-driven masking and embedding-learning strategies. In addition, increasing the embedding dimension d amplifies the coherent gain, and when combined with high values of ρ_{ia}, provides improvements in SINR and capacity under favorable alignment conditions.

B. Theoretical and Empirical SER

1) Theoretical SER Under Rayleigh Fading: To contextualize the performance of the proposed SE scheme, we establish several analytical baselines. DE serves as the conventional orthogonal reference, while NOMA provides a modern upper bound under perfect interference cancellation. The SE formulation itself captures practical embedding-domain multiplexing with finite alignment factors, and the no-alignment case ($\rho_{ia} = 0$) represents a natural lower bound where inter-user interference accumulates independently. Notably, the comparison between SE and DE parallels the well-known distinction between OMA and NOMA discussed in heterogeneous semantic and bit communication frameworks [31]. In this analogy, DE corresponds to OMA, where each user is allocated an exclusive portion of the embedding space, whereas SE reflects a non-orthogonal philosophy similar to NOMA/Semi-NOMA, which enables overlapping access within the embedding domain. This perspective situates our SE versus DE comparison within established resource-domain baselines and thus provides a consistent theoretical context for evaluating embedding-domain multiple access.

For the DE under Rayleigh fading, where each user occupies a separate one-dimensional subspace, the average SER, denoted by $P_{\text{SER}}^{\text{DE}}$, for QPSK ($V = 4$) can be obtained from the corresponding BER expression as [44]

$$P_{\text{SER}}^{\text{DE}} = 1 - \left(\frac{1}{2}\left(1 + \sqrt{\frac{\text{SNR}}{1+\text{SNR}}}\right)\right)^2, \tag{23}$$

where SNR denotes the average symbol SNR.

In contrast, for the SE scheme, the effective SINR γ_{eff}^{SE} is given by equation (22), which captures the gain from both intra-user coherence (ρ_s) and inter-user IA (ρ_i). Accordingly, the average SER, denoted by $P_{\text{SER}}^{\text{SE}}$, under Rayleigh fading can be approximated by

$$P_{\text{SER}}^{\text{SE}} = 1 - \left(\frac{1}{2}\left(1 + \sqrt{\frac{\gamma_{\text{eff}}^{SE}}{1+\gamma_{\text{eff}}^{SE}}}\right)\right)^2, \tag{24}$$

where γ_{eff}^{SE} incorporates the effects of attention-driven interference suppression via shared embeddings.

For the lower bound case, corresponding to $\rho_{\text{ia}} = 0$ (or equivalently $\rho_i = 0$, i.e., no IA), the *effective SINR*, denoted by $\gamma_{\text{eff}}^{\text{LB}}$, reduces to

$$\gamma_{\text{eff}}^{\text{LB}} = \frac{(1 + (d-1)\rho_s)\,\text{SNR}}{(U-1)\,\text{SNR} + 1}. \tag{25}$$

Accordingly, the resulting average SER, denoted by $P_{\text{SER}}^{\text{LB}}$, under Rayleigh fading, becomes

$$P_{\text{SER}}^{\text{LB}} = 1 - \left(\frac{1}{2}\left(1 + \sqrt{\frac{\gamma_{\text{eff}}^{\text{LB}}}{1+\gamma_{\text{eff}}^{\text{LB}}}}\right)\right)^2, \tag{26}$$

which captures the error floor behavior caused by fully unaligned (independent) multiuser interference.

Finally, when the shared-embedding interference is assumed to be completely removed while the intra-user coherence gain (ρ_s) is still preserved, we define the corresponding effective SINR as

$$\gamma_{\text{eff}}^{\text{NOMA}} = (1 + (d-1)\rho_s)\,\text{SNR}, \tag{27}$$

where $\gamma_{\text{eff}}^{\text{NOMA}}$ denotes the effective SINR under the *ideal* assumption of perfect multi-user interference cancellation, analogous to the operation in NOMA. Accordingly, the average SER for this ideal case, denoted as $P_{\text{SER}}^{\text{NOMA}}$, is given by

$$P_{\text{SER}}^{\text{NOMA}} = 1 - \left(\frac{1}{2}\left(1 + \sqrt{\frac{\gamma_{\text{eff}}^{\text{NOMA}}}{1+\gamma_{\text{eff}}^{\text{NOMA}}}}\right)\right)^2, \tag{28}$$

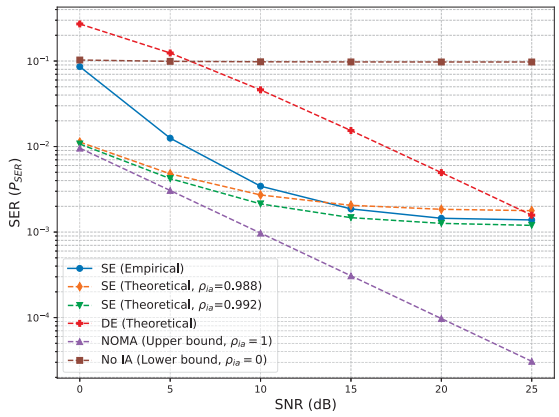

Fig. 9. SER versus SNR for the proposed SE with $U = 16$, $d = 128$, and QPSK modulation. The empirical results are compared with theoretical baselines for $\rho_s = 0.8$, including the no-alignment lower bound ($\rho_i = 0$), the NOMA upper bound ($\rho_{\text{ia}} = 1$), the DE baseline, as well as theoretical SE analysis with intermediate alignment factors ($\rho_{\text{ia}} = 0.988, 0.992$).

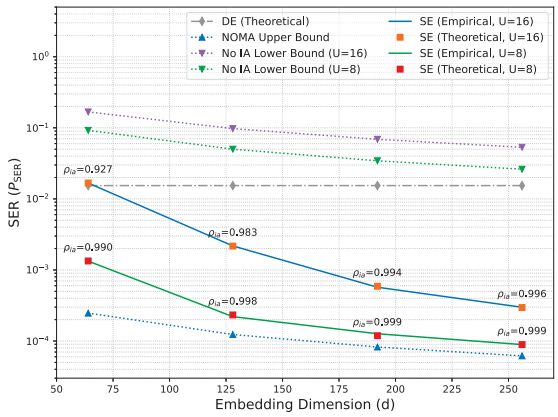

Fig. 10. SER of the proposed SE system versus embedding dimension for $U \in \{8, 16\}$ at SNR $= 15$ dB, assuming $\rho_s = 1$. The figure compares empirical results with per-dimension best-fit theoretical curves, and explicitly includes the DE baseline, the NOMA upper bound ($\rho_{\text{ia}} = 1$), and the user-dependent no-alignment lower bounds ($\rho_{\text{ia}} = 0$) for $U = 8$ and $U = 16$.

which serves as a natural *upper bound* on the achievable SER of the SE scheme. We refer to this as NOMA because, analogous to conventional non-orthogonal multiple access, it assumes perfect cancellation of inter-user interference. The key distinction is that, in our formulation, such cancellation is achieved via attention-driven alignment in the embedding domain, rather than through power-domain separation and SIC in the physical layer. It is noted that, in conventional NOMA, interference cancellation is limited to the signals of users with stronger power levels, whereas the proposed SE framework can suppress inter-user interference without such power-domain constraints by leveraging attention-driven alignment in the embedding space.

2) Empirical Validation and Performance Observations: Fig. 9 presents a comparison between the empirical SER performance and the theoretical predictions under Rayleigh fading. The empirical SE curves follow the theoretical baselines with $\rho_{\text{ia}} = 0.988$ and $\rho_{\text{ia}} = 0.992$, confirming that the Transformer effectively adapts its embedding masking to realize near-optimal alignment. The lower bound corresponding to the no-alignment case ($\rho_{ia} = 0$) exhibits an error floor due to the independent accumulation of multiuser interference, which scales with the number of users, whereas the reference case with $\rho_{\text{ia}} = 1$ serves as the ideal NOMA upper bound where all shared-embedding interference is fully canceled.

Between these two bounds, the SE scheme with practical alignment factors achieves notable performance gains, with empirical results consistently approaching the corresponding theoretical curves. Compared with the DE, the proposed SE achieves significantly lower SER across all SNR regimes, and the improvement margin becomes more pronounced at higher embedding dimensions and medium-to-high SNRs. These results show that the proposed SE effectively realizes attention-driven alignment, while maintaining scalability and robustness across the tested SNR range. At very low SNRs, however, a small gap remains between theory and experiment because noise dominates the learning process, and this dominance results in limited alignment precision and residual variability due to finite sample effects.

Fig. 10 presents the SER results under Rayleigh fading at SNR $= 15$ dB for both $U = 8$ and $U = 16$ users, assuming $\rho_s = 1$ for a theoretical reference. The DE and NOMA upper bounds are common to both user counts, while the no-alignment lower bounds differ due to the explicit dependence on the number of users. The empirical results closely match the theoretical predictions, with the alignment factor ρ_{ia} fitted separately for each embedding dimension.

As the embedding dimension d increases, the SER decreases rapidly because larger embedding spaces allow user embeddings to be more effectively aligned or separated, facilitating stronger attention-driven interference suppression. The fitted alignment factors approach unity as d grows, reflecting this improved alignment capability. For instance, in the case of $U = 16$, the fitted values rise from about $\rho_{\text{ia}} = 0.927$ at $d = 64$ to $\rho_{\text{ia}} = 0.996$ at $d = 256$, while for $U = 8$ the values are even higher, from about 0.990 to 0.999. This trend indicates that both higher-dimensional embeddings and lower user density contribute to stronger interference mitigation. Consequently, the SE curves progressively approach the NOMA upper bound while remaining below the user-dependent lower bounds.

In addition, comparison with the DE baseline reveals that SE consistently achieves superior performance, and the relative improvement becomes more pronounced as the embedding dimension increases. This advantage arises because SE can exploit alignment-driven coherence across embedding dimensions, whereas DE cannot. The close match between empirical and fitted theory further validates the effectiveness of embedding masking in approximating attention-driven alignment.

C. Theoretical Capacity Gains via Potential IA

To further assess the impact of embedding alignment on spectral efficiency, we analyze the achievable user capacity in terms of Shannon's information-theoretic framework. Specifically, we derive the per-user capacity of the SE scheme based on the effective SINR in (22) and compare it against the DE baseline, which provides a clean interference-free reference.

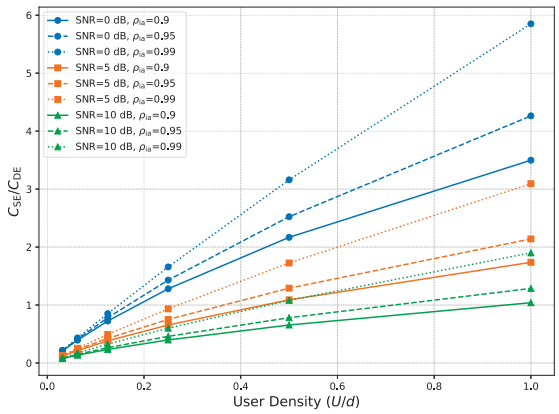

Fig. 11. Capacity ratio $C_{\mathrm{SE}}/C_{\mathrm{DE}}$ vs. user density U/d under varying IA levels and SNR conditions with embedding dimension $d = 128$.

The per-user capacity of the SE scheme is expressed as [45]:

$$C_{\mathrm{SE}} = \frac{1}{2}\log_2\left(1 + \frac{(1+(d-1)\rho_s)\cdot \mathrm{SNR}}{(U-1)(1-\rho_{\mathrm{ia}})\cdot \mathrm{SNR}+1}\right), \quad (29)$$

where ρ_s is the intra-user embedding correlation coefficient, and SNR is the base signal-to-noise ratio.

For comparison, the DE scheme assigns each user a disjoint $\frac{d}{U}$-dimensional subspace, eliminating inter-user interference. The DE per-user capacity is thus given by:

$$C_{\mathrm{DE}} = \frac{d}{2U}\log_2(1+\mathrm{SNR}). \quad (30)$$

The capacity ratio between SE and DE then becomes:

$$\frac{C_{\mathrm{SE}}}{C_{\mathrm{DE}}} = U\cdot\frac{\log_2\left(1+\frac{(1+(d-1)\rho_s)\cdot\mathrm{SNR}}{(U-1)(1-\rho_{\mathrm{ia}})\cdot\mathrm{SNR}+1}\right)}{d\log_2(1+\mathrm{SNR})}. \quad (31)$$

Fig. 11 illustrates how the SE-to-DE capacity ratio varies with the user density U/d, under different values of the IA factor $\rho_{\mathrm{ia}} = 0.9, 0.95$, and 0.99 (in the case of $U = 32$, $\rho_{\mathrm{i}} \approx -0.03, -0.0317$, and -0.033, respectively) and $\mathrm{SNR} = 0, 5, 10\,\mathrm{dB}$. The figure reveals that stronger alignment (that is, higher ρ_{ia}) consistently improves the relative performance of SE compared to DE, especially in high-density regimes where multiple users share the embedding space. In contrast, with the low-density regime ($U/d \ll 1$), interference suppression becomes less effective and the SE capacity approaches or falls below that of the DE scheme although SE achieves a significant SER gain over DE observed in Fig. 10. This analysis underscores the role of ρ_{ia} as a key indicator of attention-based alignment performance. Incorporating this alignment factor, the SE framework exhibits scalable capacity even in dense multiuser environments, outperforming conventional DE schemes under effective IA.

However, we note that ρ_{ia} is an abstract coefficient introduced to model the emergent effect of Transformer attention, rather than a directly measurable physical parameter. To achieve higher values of ρ_{ia} in practice, more extensive training, architectural refinements, and improved attention mechanisms are required. These directions lie beyond the present scope and remain important topics for future research.

D. Theoretical and Practical Implications

The theoretical SER analysis in this section clarifies how embedding correlations determine SE performance. The signal and interference power expressions show that intra-user coherence (ρ_s) enhances the desired signal through coherent combining across embedding dimensions, while inter-user correlation (ρ_i) governs interference growth with user density. These relationships yield the closed-form SINR in (20), which directly links embedding statistics to performance. To capture the effect of attention-driven suppression, we introduced the alignment coefficient ρ_{ia}, leading to the SINR form in (22). This formulation highlights that larger ρ_s strengthens signal gain, whereas higher ρ_{ia} reduces the effective interference factor $(U-1)(1-\rho_{\mathrm{ia}})$. Mapping this SINR to SER baselines places SE between the no-alignment lower bound and the NOMA (ideal cancellation) upper bound, thereby quantifying the role of alignment in reducing error floors. Capacity analysis further shows that SE can outperform DE when embedding alignment is strong and sufficient dimensions are available, especially in dense regimes where interference suppression is most critical. Importantly, unlike NOMA, which depends on SIC and favorable power-domain conditions, SE operates directly in the embedding domain without such constraints.

Overall, SE achieves its benefits through the joint contributions of intra-user coherence, attention-driven IA, embedding dimension, and statistical multiplexing, and these factors collectively explain the observed improvements in theoretical SINR, SER, and capacity.

VI. Conclusion

In this paper, we proposed the Transformer-based SE framework that leverages embedding structure, user correlation, and attention-driven masking to achieve scalable multiple access. Through combined theoretical analysis and empirical validation, we showed how these factors jointly determine key performance metrics such as SINR, SER, and capacity, and confirmed that SE consistently outperforms the DE scheme across diverse SNR conditions and system loads, naturally realizing IA and statistical multiplexing gains through end-to-end learning of shared embedding allocation.

In addition to these results, the study contributes methodological advances by providing closed-form analysis and an alignment-based interpretation of SE. In particular, we established performance bounds ranging from the no-alignment lower bound to the ideal NOMA-inspired upper bound, and systematically compared the SE performance by identifying the degree of attention-driven IA.

Reducing the complexity of deployment is another major priority. To this end, we will explore both model-agnostic meta-learning (MAML) [46], [47] and context-aware learning [48], [49]. MAML promises rapid adaptation of a pre-trained model to new user configurations and channel conditions with minimal fine-tuning, eliminating the need to train a separate model for every scenario. On the other hand, context-aware learning methods enable the network to adjust its behavior online using real-time signals such as user identity, embedding density, instantaneous SNR, or even the number of active users, thereby allowing a single shared model to flexibly operate across different multi-user settings.

Another important direction is enhancing robustness in harsh channels. We will incorporate denoising modules and embedding regularizers that preserve semantic integrity at low SNR, for example, by adding adaptive attention refinement and Transformer-stabilization techniques [50]. In addition, we will investigate techniques that ensure reliable data transmission under diverse user environments, where heterogeneous channel conditions and device capabilities such as different embedding dimensions coexist.

Equally critical is the need to broaden evaluation beyond symbol-level recovery. While the present study centers on symbol-level experiments, the proposed SE framework can be naturally extended to semantic classification tasks, where multiplexed embeddings are demultiplexed and decoded for user-specific predictions using objectives such as CE and task-level accuracy. For more demanding applications such as image reconstruction, however, symbol-level metrics are insufficient, and reconstruction-oriented losses such as mean squared error (MSE) must be employed. In this context, recent advances in Vision Transformer (ViT) and Swin Transformer-based semantic communication [51] have demonstrated that image-level fidelity can be effectively evaluated using peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM). Motivated by these studies, we plan to pursue ViT and Swin Transformer-based extensions of our SE framework for semantic data multiplexing, thereby enabling validation not only in terms of SER but also with respect to semantic fidelity and reconstruction quality. To further strengthen embedding separation and generalization across heterogeneous conditions, we will also investigate the integration of contrastive learning as an auxiliary objective in addition to the primary loss, which ensures that user-specific embeddings would remain well separated while preserving semantic fidelity.

Collectively, these research directions aim to generalize SE across a wide range of user channel conditions. By jointly addressing deployment complexity, robustness under fading, and semantic-level multiplexing with advanced learning strategies such as meta-learning, context-aware adaptation, and contrastive objectives, these efforts pave the way toward scalable and adaptive SE deployment. In doing so, they strengthen the applicability of SE not only in symbol-level evaluations but also in higher-level semantic scenarios, including image reconstruction and task-oriented communications, thereby laying the foundation for next-generation networks.

REFERENCES

[1] Z. Qin, H. Ye, G. Y. Li, and B. F. Juang, "Deep learning in physical layer communications," *IEEE Wireless Commun.*, vol. 26, no. 2, pp. 93–99, Apr. 2019.

[2] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, "Artificial neural networks-based machine learning for wireless networks: A tutorial," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3039–3071, 4th Quart., 2019.

[3] J. Johnston and X. Wang, "RNN beamforming optimizer for rate-splitting multiple access and cell-free massive MIMO," *IEEE Trans. Commun.*, vol. 73, no. 5, pp. 3579–3592, May 2025.

[4] A. Zappone, M. Di Renzo, and M. Debbah, "Wireless networks design in the era of deep learning: Model-based, AI-based, or both?," *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 7331–7376, Oct. 2019.

[5] K. Katsaros et al., "AI-native multi-access future networks—The REASON architecture," *IEEE Access*, vol. 12, pp. 178586–178622, 2024.

[6] F. Liu et al., "Integrated sensing and communications: Toward dual-functional wireless networks for 6G and beyond," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 6, pp. 1728–1767, Jun. 2022.

[7] N. Van Huynh et al., "Generative AI for physical layer communications: A survey," *IEEE Trans. Cognit. Commun. Netw.*, vol. 10, no. 3, pp. 706–723, Jun. 2024.

[8] Q. Lan et al., "What is semantic communication? A view on conveying meaning in the era of machine intelligence," *J. Commun. Inf. Netw.*, vol. 6, no. 4, pp. 336–352, Dec. 2021.

[9] 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.

[10] E. C. Strinati et al., "6G: The next frontier: From holographic messaging to artificial intelligence using subterahertz and visible light communication," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 42–50, Sep. 2019.

[11] J. Weston, F. Ratle, and R. Collobert, "Deep learning via semi-supervised embedding," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 1168–1175.

[12] X. Luo, B. Yin, Z. Chen, B. Xia, and J. Wang, "Autoencoder-based semantic communication systems with relay channels," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, May 2022, pp. 711–716.

[13] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inform. Process. Syst. (NIPS)*, 2017, pp. 5998–6008.

[14] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "End-to-end convolutional semantic embeddings," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5735–5744.

[15] E. Grassucci, Y. Mitsufuji, P. Zhang, and D. Comminiello, "Enhancing semantic communication with deep generative models: An overview," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 13021–13025.

[16] L. X. Nguyen, A. D. Raha, P. S. Aung, D. Niyato, Z. Han, and C. S. Hong, "A contemporary survey on semantic communications: Theory of mind, generative AI, and deep joint source-channel coding," 2025, *arXiv:2502.16468*.

[17] 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.

[18] H. Xie, Z. Qin, and G. Y. Li, "Semantic communication with memory," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 8, pp. 2658–2669, Aug. 2023.

[19] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed., Hoboken, NJ, USA: Wiley, 2006.

[20] Z. Lyu, G. Zhu, J. Xu, B. Ai, and S. Cui, "Semantic communications for image recovery and classification via deep joint source and channel coding," *IEEE Trans. Wireless Commun.*, vol. 23, no. 8, pp. 8388–8404, Aug. 2024.

[21] J. Xu, T.-Y. Tung, B. Ai, W. Chen, Y. Sun, and D. Gündüz, "Deep joint source-channel coding for semantic communications," *IEEE Commun. Mag.*, vol. 61, no. 11, pp. 42–48, Nov. 2023.

[22] B. Gu, J. Zhan, S. Gong, W. Liu, Z. Su, and M. Guizani, "A spatial–temporal transformer network for city-level cellular traffic analysis and prediction," *IEEE Trans. Wireless Commun.*, vol. 22, no. 12, pp. 9412–9423, Dec. 2023.

[23] T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin, "FedFormer: Frequency enhanced decomposed transformer for long-term series forecasting," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 27268–27286.

[24] S. Verdu, *Multiuser Detection*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[25] X. Chen, D. W. K. Ng, W. Yu, E. G. Larsson, N. Al-Dhahir, and R. Schober, "Massive access for 5G and beyond," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 615–637, Mar. 2021.

[26] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proc. IEEE*, vol. 104, no. 9, pp. 1711–1726, Sep. 2016.

[27] H. Zhou, Y. Deng, L. Feltrin, and A. Höglund, "Analyzing novel grant-based and grant-free access schemes for small data transmission," *IEEE Trans. Commun.*, vol. 70, no. 4, pp. 2805–2819, Apr. 2022.

[28] T. N. Weerasinghe, V. Casares-Giner, I. A. M. Balapuwaduge, and F. Y. Li, "Priority enabled grant-free access with dynamic slot allocation for heterogeneous mMTC traffic in 5G NR networks," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 3192–3205, May 2021.

[29] X. Chen, R. Jia, and D. W. K. Ng, "On the design of massive non-orthogonal multiple access with imperfect successive interference cancellation," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2539–2554, Mar. 2019.

[30] X. Mu and Y. Liu, "Exploiting semantic communication for non-orthogonal multiple access," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 8, pp. 2563–2576, Aug. 2023.
[31] X. Mu, Y. Liu, L. Guo, and N. Al-Dhahir, "Heterogeneous semantic and bit communications: A semi-NOMA scheme," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 155–169, Jan. 2023.
[32] R. L. Pickholtz, L. B. Milstein, and D. L. Schilling, "Spread spectrum for mobile communications," *IEEE Trans. Veh. Technol.*, vol. 40, no. 2, pp. 313–322, May 1991.
[33] A. Fukasawa, T. Sato, Y. Takizawa, T. Kato, N. Kawabe, and R. E. Fisher, "Wideband CDMA system for personal radio communications," *IEEE Commun. Mag.*, vol. 34, no. 10, pp. 116–123, Oct. 1996.
[34] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9726–9735.
[35] K. Gomadam, V. R. Cadambe, and S. A. Jafar, "A distributed numerical approach to interference alignment and applications to wireless interference networks," *IEEE Trans. Inf. Theory*, vol. 57, no. 6, pp. 3309–3322, Jun. 2011.
[36] V. R. Cadambe and S. A. Jafar, "Interference alignment and spatial degrees of freedom for the K user interference channel," in *Proc. IEEE Int. Conf. Commun.*, Nov. 2008, pp. 971–975.
[37] S. Sen, N. Santhapuri, R. R. Choudhury, and S. Nelakuditi, "Successive interference cancellation: Carving out MAC layer opportunities," *IEEE Trans. Mobile Comput.*, vol. 12, no. 2, pp. 346–357, Feb. 2013, doi: 10.1109/TMC.2012.17.
[38] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: https://www.deeplearningbook.org/
[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014, pp. 1–11.
[40] K. Pearson, "VII. Note on regression and inheritance in the case of two parents," *Proc. Roy. Soc. London*, vol. 58, nos. 347–352, pp. 240–242, 1895. [Online]. Available: https://royalsocietypublishing.org/doi/10.1098/rspl.1895.0041
[41] B. M. Hochwald, C. B. Peel, and A. L. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication—Part II: Perturbation," *IEEE Trans. Commun.*, vol. 53, no. 3, pp. 537–544, Mar. 2005.
[42] V. R. Cadambe and S. A. Jafar, "Interference alignment and degrees of freedom of the K-user interference channel," *IEEE Trans. Inf. Theory*, vol. 54, no. 8, pp. 3425–3441, Aug. 2008.
[43] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.
[44] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.
[45] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul. 1948.
[46] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, Aug. 2017, pp. 1126–1135.
[47] C.-F. Lin and H.-Y. Lee, "Master of puppets: Model-agnostic meta-learning via pre-trained parameters for natural language generation," in *Proc. NeurIPS Workshop Meta-Learning*, 2020, pp. 1–9.
[48] I. Mallioras, Z. D. Zaharis, P. I. Lazaridis, and S. Pantelopoulos, "A novel realistic approach of adaptive beamforming based on deep neural networks," *IEEE Trans. Antennas Propag.*, vol. 70, no. 10, pp. 8833–8848, Oct. 2022.
[49] R. K. Mahabadi, S. Ruder, M. Dehghani, and J. Henderson, "Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks," 2021, *arXiv:2106.04489*.
[50] Q. Zhou, R. Li, Z. Zhao, C. Peng, and H. Zhang, "Semantic communication with adaptive universal transformer," *IEEE Wireless Commun. Lett.*, vol. 11, no. 3, pp. 453–457, Mar. 2022.
[51] L. X. Nguyen et al., "Swin transformer-based dynamic semantic communication for multi-user with different computing capacity," *IEEE Trans. Veh. Technol.*, vol. 73, no. 6, pp. 8957–8972, Jun. 2024.

Ki-Ho Lee received the B.S., M.S., and Ph.D. degrees in electrical and electronic engineering from KAIST, Daejeon, South Korea, and the Executive M.B.A. degree from Seoul National University, Seoul, South Korea. He is currently a Research Professor with Chung-Ang University, Seoul. He joined KT in 2007, where he held various research and leadership positions until 2024. From 2018 to 2021, he was a Team Leader of the 5G Technology Planning Team. From 2022 to 2024, he led the Next-Generation NaaS Research Team, KT. His research interests include semantic communications, next-generation network architectures, multi-access edge computing, and AI-native network technologies. From 2023 to 2024, he actively contributed to technical standardization activities as a member of TTA TC11 (Mobile Communications Technology Committee) and the Chair of the Technical Standardization Subcommittee of the 5G MEC Forum.

Hyun-Ho Choi (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), South Korea, in 2001, 2003, and 2007, respectively. He is a Professor with the School of ICT, Robotics and Mechanical Engineering, Hankyong National University, South Korea. From February 2007 to February 2011, he was a Senior Engineer with the Communication Laboratory, Samsung Advanced Institute of Technology (SAIT), South Korea. Since March 2011, he has been a Professor with Hankyong National University. He was also a Visiting Researcher with TeleCIS Wireless Inc., CA, USA, in 2006; a Visiting Scholar with Stanford University, USA, in 2008; a Visiting Professor with UC Irvine, USA, in 2017; and a Visiting Professor with the University of Calgary, Canada, in 2024. His current research interests include mobile and wireless communications, semantic communications, wireless energy harvesting, machine learning, bio-inspired algorithms, and distributed optimization. He is a Life Member of KICS and KIICE.

Jung-Ryun Lee (Senior Member, IEEE) received the B.S. and M.S. degrees in mathematics from Seoul National University, in 1995 and 1997, respectively, and the Ph.D. degree in electrical and electronics engineering from Korea Advanced Institute of Science and Technology (KAIST) in 2006. From 1997 to 2005, he was a Chief Research Engineer with LG Electronics, South Korea. From 2006 to 2007, he was a full-time Lecturer of electronic engineering with the University of Incheon. Since 2008, he has been a Professor with the School of Electrical and Electronics Engineering, Chung-Ang University, South Korea. He was also a Visiting Researcher with UC Irvine, CA, USA, in 2016. His current research interests include low energy networks and algorithms, bio-inspired autonomous networks, semantic communication, space-air-ground integrated networks, and AI-native communication. He is a Life Member of KIISE and KICS.