Attention-Driven Interference Alignment for Multi-User Semantic Communications

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

***Abstract*—This paper presents an attention-driven interference alignment (IA) scheme for multi-user semantic communications that operates in the semantic embedding domain, unlike conventional IA in the baseband domain. To realize this, user-specific masking and attention-driven decoding are employed over a shared embedding space, which is mathematically modeled to characterize attention-based IA in the embedding domain and to enable implicit user separation without explicit orthogonalization. Built on a conventional cross-entropy (CE) loss, the proposed orthogonalization regularization further reinforces interference alignment by aligning user-specific semantic subspaces and suppressing inter-user interference. Experiments under Rayleigh fading channels show that the proposed approach outperforms both a resource-partitioned baseline with separate embeddings and a CE-only loss model without orthogonalization regularization, showing the effectiveness of an attention-driven IA solution for multi-user semantic communication.**

***Index Terms*—Interference alignment, semantic communication, transformer, attention.**

I. Introduction

Wireless communication systems have long relied on structured multi-user communication techniques such as time division multiple access (TDMA), frequency division multiple access (FDMA), and code division multiple access (CDMA) to support concurrent user transmissions [1], [2]. These schemes achieve user separation by allocating orthogonal resources in time, frequency, or code domains, thereby suppressing mutual interference. More recently, orthogonal frequency division multiplexing (OFDM) and non-orthogonal multiple access (NOMA) have been introduced to improve spectral efficiency by resource reuse through power-domain superposition and successive interference cancellation [3].

As wireless networks evolve toward 6G and beyond, these conventional symbol-level and channel-centric resource allocation methods are increasingly revealing fundamental limitations. Emerging applications such as autonomous vehicles, immersive media, and distributed AI require communication systems that deliver not just bits, but semantic meaning. This paradigm shift has given rise to *semantic communication*, in which the primary goal is to convey task-relevant content rather than raw data [4], [5]. By focusing on the contextual value of information, semantic communication improves bandwidth efficiency and reduces latency [6], particularly in scenarios constrained by power, time, or spectrum.

This evolution is further accelerated by the growing need to support massive connectivity in future networks, including machine-type communications (mMTC) and human-to-machine (H2M) interactions. Such applications demand ultra-reliable, low-latency communication with minimal signaling overhead, especially in short-packet transmission environments [7], [8]. While grant-free access schemes [9] have been proposed to reduce scheduling overhead, they still rely on explicit channel estimation and coordination, which limits their scalability in dense deployments.

To meet these requirements, recent research has increasingly explored deep learning models that encode multimodal inputs into high-dimensional *semantic embeddings*, such as adaptive Universal Transformers for sentence-wise semantic modulation [10] and Swin Transformer-based encoders tailored to user-specific capacities [11]. These embeddings abstract and compress source content for transmission and inference, allowing communication to take place in a latent semantic space. However, when multiple users project their semantic representations into a shared embedding space, a new challenge arises: the efficient management of this space to enable scalable and interference-aware multi-user communication.

This challenge has a conceptual connection to the principle of *interference alignment (IA)* [12], [13], a foundational multi-user information-theoretic technique that confines interference to structured subspaces, thereby maximizing the degrees of freedom available to desired signals. Inspired by this concept, we reinterpret user-level interference in semantic communication as a problem of *latent-space alignment*.

Unlike prior state-of-the-art IA methods, which operate exclusively at the physical layer using channel matrices and linear precoders, our approach extends the concept of IA to the semantic domain. In this work, we redesign IA for semantic communication: alignment is achieved not through beamforming or channel-domain signal processing, but through learned attention patterns that automatically adapt to each user's semantic representation. By operating directly in the embedding space, inter-user interference can be suppressed even when orthogonal resource allocation is unavailable, which is a scenario not addressed by existing IA methods.

To this end, we propose an attention-driven *soft IA* scheme that flexibly adapts through learned attention distributions, in contrast to strict or hard separation [14]. Instead of rigidly partitioning resources, attention-based decoding enables implicit user separation by dynamically focusing on semantically relevant features for each user. This design allows scalable and interference-aware access to shared semantic representations without explicit orthogonalization.

Our work is also situated within a broader trend of applying attention mechanisms to wireless communication tasks. Recent examples include attention-assisted multi-agent deep reinforcement learning for computation offloading in MEC-enabled IIoT [15], attention-based CSI feedback for massive MIMO [16], multi-attention architectures for RIS-assisted near-field beam training [17], and attention-enhanced small-sample wireless technology identification [18]. These studies demonstrate the potential of distributed and multi-head attention (MHA) for achieving user separation, resource optimization, and robust feature representation in complex wireless environments. However, the use of attention as a *soft IA mechanism* operating directly within the semantic embedding domain remains largely unexplored.

Received 2 June 2025; revised 16 September 2025 and 31 October 2025; accepted 27 November 2025. Date of publication 2 December 2025; date of current version 24 June 2026. This work was supported by Chung-Ang University Research Grants in 2026. The review of this article was coordinated by Prof. Geng Sun. *(Corresponding author: 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 Hankyong National University, Anseong 17579, South Korea (e-mail: hhchoi@hknu.ac.kr).

Jung-Ryun Lee is with the School of Electrical Engineering, Chung-Ang University, Seoul 06974, South Korea, and also with 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/TVT.2025.3639430

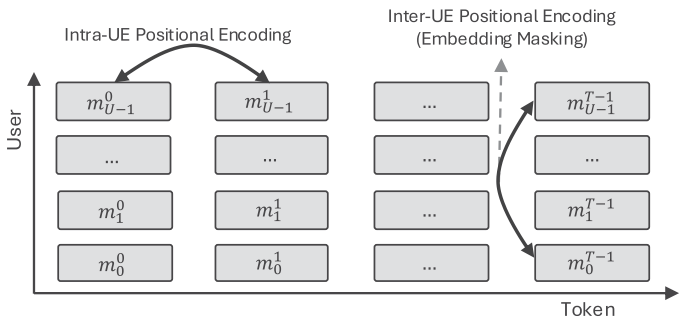

Fig. 1. Attention-based two-dimensional embedding layout.

To address this gap, we introduce a unified framework in which attention simultaneously serves as a semantic decoder and as an interference alignment engine. Unlike prior work that applies attention at the signal or channel level, our framework operates directly in the semantic embedding space. By integrating an alignment loss with user-specific masking, the learned attention distribution itself performs soft interference alignment, dynamically suppressing inter-user leakage in a shared semantic representation. This joint design explicitly tackles the open challenge of scalable multi-user access in shared semantic spaces, which has not been systematically studied in prior literature.

The main contributions of this paper are summarized as follows:

- A Transformer-based attention framework is proposed to achieve multi-user IA in shared semantic embedding spaces, enabling scalable and soft user separation without explicit orthogonalization.
- IA is introduced to guide attention distributions toward each user's intended semantic subspace, thereby enhancing attention-based IA during decoding.
- The proposed system is analytically and experimentally evaluated through multi-layer attention modeling, alignment metrics, and symbol error rate (SER) analysis under varying user densities and channel conditions.

II. Multi-User IA for Semantic Communications

A. Attention-Based Embedding Design and Architecture

In the proposed framework, instead of relying on explicit user separation or orthogonal resource allocation, attention is utilized as a learnable alignment mechanism to achieve implicit, interference-aware demultiplexing. To intuitively illustrate the proposed attention mechanisms for multiplexing, a two-dimensional diagram is presented in Fig. 1. This figure visualizes the simultaneous operation of token-wise positional encoding and user-wise embedding masking, which together enables scalable multi-user access within an embedding space. The horizontal axis represents the intra-user token sequence positions ranging from $t = 0$ to $t = T - 1$, where T denotes the length of each user's input sequence. Each user's semantic input is temporally ordered and injected with intra-user positional encoding. This positional information allows the Transformer to effectively capture contextual dependencies among tokens belonging to the same user.

The vertical axis, on the other hand, enumerates the different users $j \in \{0, 1, \ldots, U-1\}$, where U denotes the total number of users. Each row corresponds to a user-specific token sequence modulated by an embedding mask, ensuring that each user's tokens occupy a distinct subspace in the latent embedding domain. This dual encoding along temporal and user axes guarantees separability even when embeddings are fully superposed.

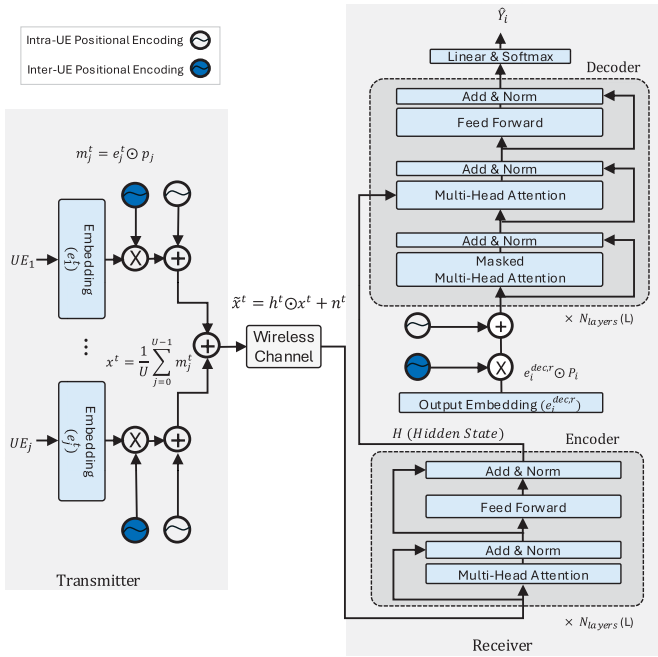

Fig. 2. Block-level architecture of the attention-based multiplexing scheme.

To clarify the system-level processing flow, Fig. 2 presents a block diagram illustrating the Transformer-based attention mechanism and user-specific masking operations. This figure depicts how masked embeddings from multiple users are injected, superposed, and transmitted over a shared wireless channel. At the receiver, the aggregated multi-user signal is distorted by both channel fading and additive noise before being processed by the Transformer encoder. This sequence is encoded into contextualized representations $\mathbf{H}$, which serve as a shared semantic memory. The Transformer decoder then uses masked MHA with user-specific queries to retrieve relevant semantic components to enable attention-based separation across users without user identifiers or dedicated decoders.

B. Mathematical Modeling of Attention-Driven IA

We present the mathematical formulation of the proposed model. The transmitted signal at each token position t is constructed as

$$\mathbf{x}^t = \sum_{j=0}^{U-1} \mathbf{m}_j^t, \tag{1}$$

where $\mathbf{m}_j^t = \mathbf{e}_j^t \odot \mathbf{p}_j$ denotes the masked embedding of user j at position t, obtained by elementwise multiplication of the semantic embedding $\mathbf{e}_j^t \in \mathbb{R}^d$ and the user-specific mask $\mathbf{p}_j \in \mathbb{R}^d$. This signal is transmitted through a Rayleigh fading channel with additive white Gaussian noise (AWGN) so that

$$\tilde{\mathbf{x}}^t = \mathbf{h}^t \odot \mathbf{x}^t + \mathbf{n}^t, \tag{2}$$

where $\mathbf{h}^t \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$ is the Rayleigh fading coefficient applied elementwise to the transmitted embedding and $\mathbf{n}^t \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ is zero-mean Gaussian noise with variance σ^2.

The received sequence $\{\tilde{\mathbf{x}}^t\}_{t=0}^{T-1}$ is then processed by a Transformer encoder as

$$\mathbf{H} = \text{Encoder}([\tilde{\mathbf{x}}^0, \ldots, \tilde{\mathbf{x}}^{T-1}]) \in \mathbb{R}^{T \times d}, \tag{3}$$

where $\mathbf{H}$ represents the contextualized hidden representation of the superposed input across all token positions. At decoding step r of R,

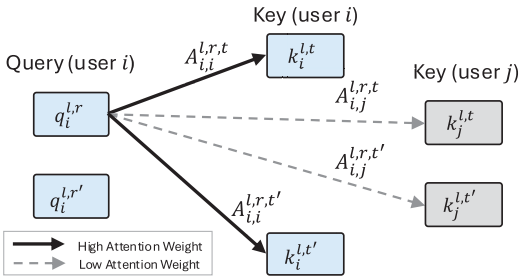

Fig. 3. Illustration of layer-wise attention weights from query tokens of user i to key tokens of user i and another user j across different token positions.

the query vector for user i is computed as

$$\mathbf{q}_i^{\ell,r} = \mathbf{W}_Q^{\ell}\left(\mathbf{e}_i^{\text{dec},r} \odot \mathbf{p}_i\right), \tag{4}$$

where $\mathbf{e}_i^{\text{dec},r} \in \mathbb{R}^d$ is the decoder embedding and $\mathbf{W}_Q^{\ell} \in \mathbb{R}^{d_k \times d}$ projects it to the query space. Each masked encoder embedding $\mathbf{m}_j^t$ is projected as

$$\mathbf{k}_j^{\ell,t} = \mathbf{W}_K^{\ell}\mathbf{m}_j^t, \tag{5}$$

explicitly linking keys to user j.

The attention score between $\mathbf{q}_i^{\ell,r}$ and $\mathbf{k}_j^{\ell,t}$ is given by $e_{i,j}^{\ell,r,t} = \langle \mathbf{q}_i^{\ell,r}, \mathbf{k}_j^{\ell,t} \rangle$, where $\langle \cdot,\cdot \rangle$ denotes the dot product in $\mathbb{R}^{d_k}$, and the attention weight is normalized as

$$A_{i,j}^{\ell,r,t} = \frac{\exp\left(e_{i,j}^{\ell,r,t}/\sqrt{d_k}\right)}{\sum_{j',t'}\exp\left(e_{i,j'}^{\ell,r,t'}/\sqrt{d_k}\right)}. \tag{6}$$

The learned attention behavior among multiple users is further illustrated in Fig. 3, which schematically visualizes how two query tokens $\mathbf{q}_i^{\ell,r}$ and $\mathbf{q}_i^{\ell,r'}$ from the same user i interact with key tokens $\mathbf{k}_i^{\ell,t}$, $\mathbf{k}_i^{\ell,t'}$ from user i and $\mathbf{k}_j^{\ell,t}$, $\mathbf{k}_j^{\ell,t'}$ from another user j across multiple token positions. The figure highlights the distribution of attention weights $A_{i,j}^{\ell,r,t}$ computed at decoding step r (and similarly for r') for different key positions t and t'. As shown in the figure, the queries $\mathbf{q}_i^{\ell,r}$ and $\mathbf{q}_i^{\ell,r'}$ focus their attention predominantly on key vectors $\mathbf{k}_i^{\ell,t}$ and $\mathbf{k}_i^{\ell,t'}$ that belong to the same user, while assigning much lower attention weights to key vectors $\mathbf{k}_j^{\ell,t}$ and $\mathbf{k}_j^{\ell,t'}$ associated with other users. This selective attention behavior illustrates the model's ability to confine attention within the intended semantic subspace through masking and learned alignment, leading to effective suppression of inter-user interference.

To track inter-user alignment, we compute the normalized user-level attention as $\hat{A}_{i,j}^{\ell} = \frac{\bar{A}_{i,j}^{\ell}}{\sum_{j'}\bar{A}_{i,j'}^{\ell}}$, where $\bar{A}_{i,j}^{\ell} = \frac{1}{RT}\sum_{r=1}^{R}\sum_{t=0}^{T-1} A_{i,j}^{\ell,r,t}$ is the average attention assigned by user i's decoder in layer ℓ to user j, aggregated over all decoding positions and token positions. The normalized attention matrix at layer ℓ is then defined as

$$\hat{A}^{\ell} = \begin{bmatrix} \hat{A}_{0,0}^{\ell} & \cdots & \hat{A}_{0,U-1}^{\ell} \\ \vdots & \ddots & \vdots \\ \hat{A}_{U-1,0}^{\ell} & \cdots & \hat{A}_{U-1,U-1}^{\ell} \end{bmatrix} \in \mathbb{R}^{U\times U}, \tag{7}$$

whose (i,j)-th entry corresponds to $\hat{A}_{i,j}^{\ell}$.

To capture the cumulative alignment pattern across the decoder depth, we compute the effective attention weight as the element-wise product of the per-layer normalized matrices:

$$A^{\text{eff}} = \hat{A}^{(1)} \odot \hat{A}^{(2)} \odot \cdots \odot \hat{A}^{(L)}, \tag{8}$$

where $\odot$ denotes the element-wise (Hadamard) product. This formulation reinforces user-specific attention through multi-layer accumulation, effectively enhancing consistent alignments and suppressing noisy or inconsistent ones. To maintain the probabilistic interpretation of attention, the final effective matrix is normalized row-wise as

$$\tilde{A}_{i,j}^{\text{eff}} = \frac{A_{i,j}^{\text{eff}}}{\sum_{j'} A_{i,j'}^{\text{eff}}}, \tag{9}$$

ensuring that each row of $\tilde{A}^{\text{eff}}$ sums to one. This normalized effective attention matrix, serving as a quantitative metric, highlights the persistent user-to-user alignment patterns that are preserved across decoder layers, thus facilitating clearer user separation.

Finally, the output sequence for user i is reconstructed by decoding the attended representation, denoted as

$$\hat{\mathbf{Y}}_i = \text{Decoder}_i([\mathbf{q}_i^1, \ldots, \mathbf{q}_i^R], \mathbf{H}), \tag{10}$$

where $\mathbf{q}_i^r = \mathbf{e}_{i,\text{dec}}^r \odot \mathbf{p}_i$ denotes the masked decoder input embedding at position r for user i. These vectors are projected to query representations across decoder layers as defined earlier, and jointly used to attend over the encoder output $\mathbf{H}$. This process highlights the central role of masking and attention in enabling scalable and interference-resilient multi-user communication within the shared embedding space.

C. Loss Function for Attention-Driven IA

To encourage interference-aware attention behavior, we define a regularization loss that penalizes semantic similarity between different users' embedding masks. Unlike conventional embedding-based learning frameworks that optimize only the cross-entropy objective, our formulation explicitly augments it with an interference-alignment regularization term, thereby strengthening the model's ability to suppress inter-user interference. The total loss function is formulated as

$$\mathcal{L}_{\text{IA}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{IA}}\mathcal{L}_{\text{orth}}, \tag{11}$$

where $\mathcal{L}_{\text{CE}}$ is the conventional cross-entropy (CE) loss for prediction accuracy, and $\mathcal{L}_{\text{orth}}$ is the newly proposed regularization term that promotes orthogonality between different users' masked embeddings. The hyperparameter $\lambda_{\text{IA}} \in \mathbb{R}_+$ controls the strength of this regularization.

The cross-entropy loss is defined as

$$\mathcal{L}_{\text{CE}} = -\sum_{i=0}^{U-1}\sum_{r=1}^{R}\log P\left(y_i^{r,\text{target}} \mid \mathbf{H}, \text{Decoder}_i\right), \tag{12}$$

where $y_i^{r,\text{target}}$ is the target token for user i and $P(\cdot)$ represents the softmax probability assigned by the user-specific decoder to the ground-truth token $y_i^{r,\text{target}}$ at decoding step r.

To define $\mathcal{L}_{\text{orth}}$, we reuse the user-specific embedding masks $\mathbf{p}_i \in \mathbb{R}^d$ introduced earlier. Each vector is L2-normalized as $\tilde{\mathbf{p}}_i = \mathbf{p}_i/\|\mathbf{p}_i\|$ for $i = 0, \ldots, U-1$, and these are stacked to form the normalized embedding matrix $\tilde{\mathbf{P}} \in \mathbb{R}^{U\times d}$. The cosine similarity matrix is then computed as $\mathbf{S} = \tilde{\mathbf{P}}\tilde{\mathbf{P}}^{\top} \in \mathbb{R}^{U\times U}$. The orthogonality loss penalizes any deviation from the identity matrix:

$$\mathcal{L}_{\text{orth}} = \|\mathbf{S} - \mathbf{I}\|_F^2, \tag{13}$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

Unlike $\mathcal{L}_{\text{CE}}$, which optimizes token prediction accuracy, $\mathcal{L}_{\text{orth}}$ explicitly promotes user-level semantic separation in the shared embedding space, allowing the attention mechanism to steer each user's focus toward its own semantic subspace. During backpropagation, $\mathcal{L}_{\text{CE}}$ generates attractive gradients that pull each prediction toward its ground-truth

TABLE I
SUMMARY OF KEY SIMULATION PARAMETERS

Parameter	Value / Range
Number of users (U)	8, 16
Embedding dimension (d)	128
Vocabulary size (V)	8
Sequence length (T)	8 tokens
Transformer layers (N_{layers})	2 – 6
Attention heads per layer	8
Training epochs	400
Batches per epoch (N_{batches})	10k – 30k sequences
Optimizer	AdamW, learning rate 10^{-4}
Channel model	Rayleigh fading
SNR (dB)	5 – 15
Loss function	$\mathcal{L}_{\text{IA}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{IA}}\mathcal{L}_{\text{orth}}$
Model parameters (θ)	All trainable parameters($\mathbf{e}, \mathbf{p}, \mathbf{W}$,...)
Training objective	$\theta^* = \arg\min_\theta \ \mathcal{L}_{\text{IA}}(\theta)$

token, whereas $\mathcal{L}_{\text{orth}}$ introduces repulsive gradient components between correlated user embeddings, effectively pushing them apart in the latent space and reinforcing user-specific attention alignment at each training step. By mitigating gradient coupling among users, this regularization further stabilizes the optimization trajectory and accelerates convergence toward an interference-minimized equilibrium.

III. Experimental Setup

The experiments are conducted using a Transformer-based encoder–decoder architecture with shared attention across users. Both the encoder and decoder consist of a variable number of Transformer layers, ranging from $N_{\text{layers}} = 2$ to 6, with 8 attention heads per layer. The embedding dimension is set to 128, and the vocabulary size is fixed at $V = 8$. Each user transmits a short sequence of $T = 8$ tokens.

The models are trained using the AdamW optimizer with a learning rate of 10^{-4}, over 400 training epochs. The number of training batches per epoch (N_{batches}) ranges from 10 k to 30 k, where each batch corresponds to a multi-user token sequence, depending on the experimental configuration. The training objective follows the total loss formulation $\mathcal{L}_{\text{IA}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{IA}}\mathcal{L}_{\text{orth}}$, as defined in (11). During training, a Rayleigh fading channel model is applied, with SNR values ranging from 5 to 15 dB.

At evaluation time, model performance is measured using the average symbol error rate (SER) and effective attention alignment. As a baseline for comparison, we consider an idealized setting where each user is assigned an orthogonal embedding vector. Under independent Rayleigh fading, the instantaneous signal-to-noise ratio γ follows an exponential distribution with mean $\overline{\gamma}$. To incorporate the dimensional gain from using a high-dimensional embedding space, the effective SNR is defined as

$$\overline{\gamma}_{\text{eff}} = \frac{d}{U \log_2 V}\,\overline{\gamma}, \tag{14}$$

where $U \log_2 V$ denotes the total number of dimensions required to represent the combined semantic content of all users. The baseline SER for the orthogonal embedding case can then be expressed in closed form [19] as $\text{SER}_{\text{baseline}} = \frac{1}{2}\left(1 - \sqrt{\frac{\overline{\gamma}_{\text{eff}}}{1+\overline{\gamma}_{\text{eff}}}}\right)$, which directly reflects the SNR scaling and fading characteristics of the reference channel.

Table I summarizes the key simulation parameters used throughout all experiments.

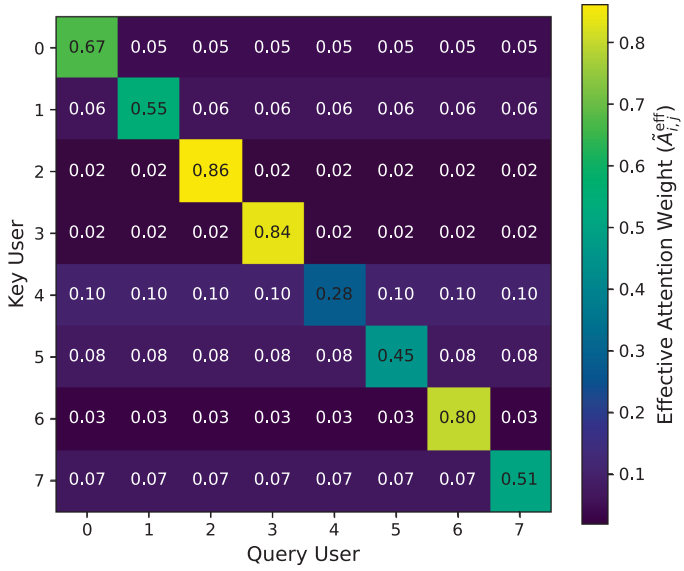

Fig. 4. Heatmap of the final effective attention matrix $\tilde{A}^{\text{eff}}$ obtained via element-wise accumulation across $L = 5$ decoder layers with fixed $U = 8$, $SNR = 15$ dB, $N_{\text{batches}} = 20$k, and $\lambda_{\text{IA}} = 0$ (i.e., $\mathcal{L}_{\text{IA}} = \mathcal{L}_{\text{CE}}$).

IV. Experimental Results and Analysis

A. Effective Multi-User IA Via Transformer Attention

We first visualize the final effective attention matrix $\tilde{A}^{\text{eff}}$ as a heatmap, obtained by element-wise accumulation of normalized per-layer attention matrices. As shown in Fig. 4, repeated attention across layers amplifies consistent intra-user attention while suppressing unstable or inter-user weights. The resulting diagonal dominance reflects reinforced intra-user alignment, where each user's decoder selectively concentrates on its own semantic region. This mechanism acts as a soft demultiplexing operation that naturally enhances user separation with depth.

We next investigate how the number of Transformer decoder layers affects overall attention alignment performance.

As observed, increasing the number of decoder layers (N_{layers}) generally strengthens intra-user attention alignment while reducing inter-user interference. This pattern is consistently visible across all training budgets, with particularly pronounced separation in the 10 k and 20 k batch cases. However, at 30 k batches, performance gains from depth begin to saturate or even slightly decline.

This non-monotonic trend suggests that while increasing depth and training budget synergistically improve semantic separation, excessive depth may lead to diminishing returns or instability in attention focus. Such effects align with recent findings that Transformer models exhibit early saturation in depth, and that excessively deep networks can exhibit degraded performance due to representational redundancy and rank deficiency in internal layers [20]. These results emphasize the need for careful balancing of architectural depth and training scale to maximize alignment and interference suppression in shared embedding spaces.

B. Impact of Attention-Driven IA Loss

Fig. 6 presents the SER performance under Rayleigh fading for CE-based IA ($\lambda_{\text{IA}} = 0$) and proposed IA with $\lambda_{\text{IA}} = 1.0$ under $U = 8$ and $U = 16$. For reference, baseline curves corresponding to conventional orthogonal embedding cases are included for each user configuration,

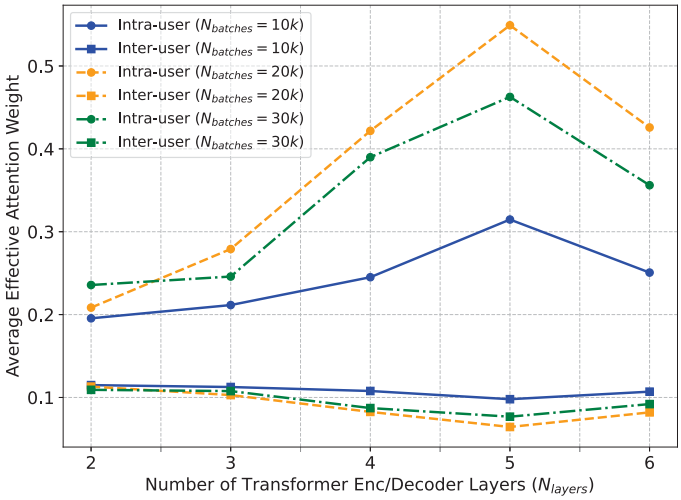

Fig. 5. Average cumulative effective attention weights versus Transformer depth N_{layers} for different training budgets with fixed $U = 8$, $SNR = 15$ dB, and $\lambda_{\text{IA}} = 0$.

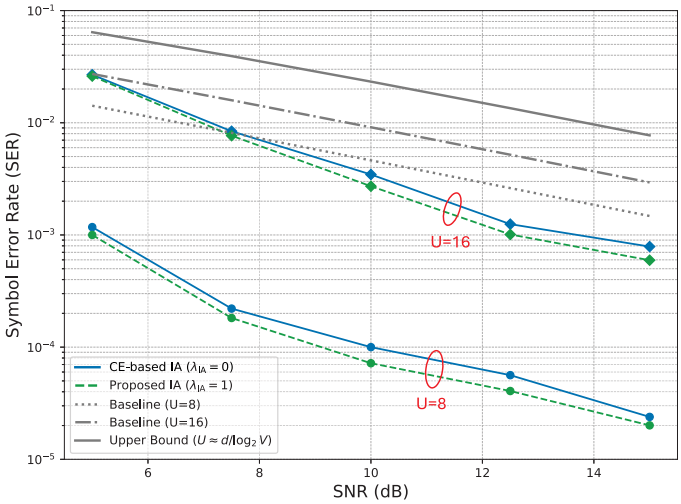

Fig. 6. SER versus SNR for CE-based IA ($\lambda_{\text{IA}} = 0$) and proposed IA with $\lambda_{\text{IA}} = 1.0$ under user configurations $U = 8$ and $U = 16$. $N_{\text{layers}} = 4$ and $N_{\text{batches}} = 20$ k.

as well as the *upper-bound*, defined as $U_{\text{ub}} = \frac{d}{\log_2 V} \approx 43$, representing the maximum user capacity without embedding gain.

Across most SNR values, both CE-based IA and proposed IA-regularized models outperform the corresponding baselines, showing the benefit of learned embeddings over idealized orthogonal allocation. The performance gap becomes more pronounced when proposed IA regularization is applied. In particular, $\lambda_{\text{IA}} = 1.0$ achieves the largest improvement for $U = 8$ in the high-SNR regime, where the SER decreases from 10^{-3} (Baseline, $U = 8$) to 10^{-5} at 15 dB, achieving approximately a two-orders-of-magnitude reduction.

In the baselines of the conventional orthogonal embedding case, a smaller U directly yields a higher effective SNR per user. In the CE-based IA and proposed IA-regularized models, part of the gain can be attributed to an *embedding dimension gain*, which improves separation capability in the embedding space. Beyond this dimensional benefit, IA regularization provides an additional gain by explicitly reducing inter-user correlation in the shared embedding space, enabling the attention mechanism to more effectively suppress interference and preserve semantic fidelity.

At very low SNR, the performance of the CE-based IA and proposed IA-regularized models becomes closer to that of the baseline, particularly for $U = 16$. This effect is mainly due to reduced learning effectiveness under severe channel conditions, where poor channel quality hampers training accuracy. Nevertheless, even the least favorable baseline corresponding to the lower bound configuration is consistently outperformed by the proposed models.

C. Computational Complexity Analysis

In practical deployment, the dominant factor that affects latency is the inference-time computational cost rather than the training-time overhead. Therefore, in this subsection we analyze the inference complexity of the proposed framework. The primary contributors to computational cost are the MHA blocks and the feed-forward sublayers in the Transformer. For a Transformer with N_{layers} decoder layers, embedding dimension d, sequence length T, and U concurrent users, the computational cost per attention head is dominated by the matrix multiplications in the attention mechanism, given by $\mathcal{O}_{\text{attn}} = \mathcal{O}(T^2\, d)$, since the attention score computation involves a $T \times T$ matrix for each head. With h heads and N_{layers} layers, the total cost scales as

$$\mathcal{O}_{\text{MHA}} = N_{\text{layers}} \cdot h \cdot \mathcal{O}(T^2 d). \tag{15}$$

In the multi-user embedding setting, input sequences are combined in a shared embedding space before decoding, allowing the attention mechanism to operate once over the aggregated representation. As a result, the attention complexity remains $\mathcal{O}(T^2 d)$, independent of the number of users U. User-specific masking is applied as a lightweight element-wise operation with complexity $\mathcal{O}(UTd)$, which is asymptotically negligible and does not affect the overall attention cost. This eliminates the need for per-user attention modules and enables efficient and scalable interference suppression. The feed-forward sublayers have complexity

$$\mathcal{O}_{\text{FFN}} = N_{\text{layers}} \cdot \mathcal{O}(Td^2), \tag{16}$$

which typically dominates when $d \gg T$. Combining these terms, the overall complexity of the proposed framework is

$$\mathcal{O}_{\text{total}} = N_{\text{layers}} \left[h \cdot (T^2 d) + (Td^2)\right], \tag{17}$$

which scales linearly with N_{layers} and h, and quadratically with the sequence length T and d. The cost grows with the number of users U only in the masking and loss computation stages, both of which incur minimal overhead. Note that increasing the embedding dimension d enhances the effective SNR in (14) but also raises computational complexity, revealing a trade-off between performance and cost. Moreover, larger vocabularies in practical multi-modal scenarios, beyond the current $V = 8$ setting applicable to classification-level applications, will require even higher d, calling for efficient complexity reduction strategies in future work.

V. Conclusion

This paper presented a new approach for achieving multi-user IA in semantic communication systems through a Transformer-based attention mechanism. By leveraging user-specific masking, the proposed method enables user separation without explicit orthogonalization. This reinterprets attention not only as a decoding mechanism, but as a semantic-domain resource allocator capable of aligning inter-user interference within a shared embedding space. A mathematical

framework and experimental results were presented to characterize the attention-driven demultiplexing process, including user-wise masking, multi-layer attention aggregation, and an interference alignment loss that reinforces semantic focus.

Future work includes exploring dynamic attention control and adaptive IA under varying conditions, investigating variants such as multi-query attention for scalability and semantic multicast/broadcast, and extending with larger architectural parameters such as embedding dimensions, vocabulary size, and sequence length to validate in more realistic scenarios.

REFERENCES

[1] T. S. Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed. Upper Saddle River, NJ, USA: Prentice Hall, 2002.

[2] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 388–404, Mar. 2000.

[3] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5 G nonorthogonal multiple-access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, Aug. 2016.

[4] T. M. Getu, G. Kaddoum, and M. Bennis, "Semantic communication: A survey on research landscape, challenges, and future directions," *Proc. IEEE*, vol. 112, no. 11, pp. 1649–1685, Nov. 2024.

[5] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE Trans. Signal Process.*, vol. 69, pp. 2663–2674, 2021.

[6] 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.

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

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

[9] 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 5 G NR networks," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 3192–3205, May 2021.

[10] 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.

[11] 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. 8957–8971, Jun. 2024.

[12] 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.

[13] 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.

[14] A. Vaswani, N. Shazeer, and N. Parmar, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30.

[15] C. Ling, K. Peng, S. Wang, X. Xu, and V. C. M. Leung, "A multi-agent DRL-based computation offloading and resource allocation method with attention mechanism in MEC-enabled IIoT," *IEEE Trans. Serv. Comput.*, vol. 17, no. 6, pp. 3037–3051, Nov./Dec. 2024.

[16] R. Li, T. Liu, J. Lv, W. Yang, and J. Ma, "An efficient attention mechanism based neural network for CSI feedback in massive MIMO system," in *Proc. 10th Int. Conf. Comput. Commun.*, Dec. 2024, pp. 1199–1203.

[17] Q. Zhou, J. Zhao, K. Cai, Y. Gong, and Y. Zhu, "Multi-attention mechanism for beam training in RIS-assisted near-field communications," in *Proc. 2025 IEEE Wireless Commun. Netw. Conf. Workshops*, Mar. 2025, pp. 1–6.

[18] Y. Shen, P. Liu, C. Zhu, W. Xun, B. Xu, and H. Zhao, "Small sample wireless technology identification enhanced by cyclic shift and attention mechanism," in *Proc. 6th Int. Conf. Robot., Intell. Control Artif. Intell.*, Dec. 2024, pp. 1225–1230.

[19] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[20] J. Petty, S. Steenkiste, I. Dasgupta, F. Sha, D. Garrette, and T. Linzen, "The impact of depth on compositional generalization in transformer language models," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, Jun. 2024, pp. 7239–7252.