Transformer-Based Contrastive Learning for Multi-User Semantic Communications

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

***Abstract*—This letter proposes multi-user semantic communication (MUSC), a Transformer-based framework that enables scalable multi-user transmission through contrastive embedding multiplexing. User-specific positional masking allows multiple users to share a common encoder while preserving semantic separability. A contrastive learning objective (InfoNCE) is jointly optimized with cross-entropy reconstruction loss to enhance inter-user discrimination under fading channels. Simulation results across diverse signal-to-noise ratio conditions and user counts demonstrate consistent symbol error rate reduction compared with cross-entropy-only and dedicated embedding baselines, confirming that contrastive embedding-space coordination provides an effective and practical foundation for multi-user semantic communication.**

***Index Terms*—Semantic communication, multi-user communication, multiplexing, contrastive learning, Transformer.**

I. Introduction

SEMANTIC communication conveys task-relevant information rather than raw bits, enabling efficient bandwidth utilization for next-generation wireless systems [1], [2]. Transformer architectures have emerged as a powerful foundation for this paradigm owing to their ability to capture long-range contextual dependencies through scaled dot-product attention [3]. The deep learning enabled semantic communication (DeepSC) framework [4] demonstrated that Transformers can effectively extract and preserve semantic meaning during transmission over noisy channels.

Most existing semantic communication systems, however, are designed for single-user scenarios. In practical 6G environments such as edge intelligence and the Internet of Things (IoT), multiple devices must simultaneously share a common medium [5]. When multiple devices project their semantic representations into a shared embedding space, efficient multiplexing with minimal inter-user interference becomes essential. Task-oriented multi-user semantic communication has been explored using multiple-input multiple-output (MIMO) spatial multiplexing [6], but its scalability is constrained by the available spatial degrees of freedom. Dedicated channel allocation [7] eliminates inter-user interference at the cost of spectral efficiency, as the required resources grow proportionally with the number of users. Semi-non-orthogonal multiple access (NOMA) [8] improves spectral efficiency but requires successive interference cancellation, adding receiver complexity and potentially limiting robustness under imperfect channel estimation.

A shared embedding framework was introduced in [9], where multiple users multiplex their semantic representations via user-specific masking and shared encoding. However, that approach placed both the encoder and decoder at the receiver side, deviating from the standard transmitter–receiver semantic communication architecture [10]. Moreover, it relied solely on architectural separation without explicitly enforcing inter-user discriminability at the representation level.

This letter proposes *multi-user semantic communication (MUSC)*, a framework that incorporates *contrastive embedding multiplexing*. The key idea is to extend interference coordination from the physical domain to the semantic embedding space by explicitly enforcing inter-user separability at the representation level. A contrastive learning objective [11]–[13] is jointly optimized with the cross-entropy (CE) loss to enhance inter-user discrimination while preserving intra-user semantic consistency under practical fading channels. The contrastive term acts as a soft orthogonalization mechanism in the latent space, reducing inter-user overlap without requiring dedicated physical resources or additional antennas.

The contributions are threefold: (i) a Transformer-based MUSC framework that separates the encoder and decoder into the transmitter and receiver sides, compliant with the standard semantic communication architecture [10], and enables scalable multiple access in a shared embedding space through user-specific positional masking; (ii) a contrastive embedding multiplexing mechanism that jointly optimizes InfoNCE and CE objectives to achieve user-level semantic separation under practical fading channels; and (iii) comprehensive experiments that verify the effectiveness, scalability, and computational efficiency of the proposed framework across diverse SNR conditions and user counts.

II. System Model and Architecture

A. Transformer-Based MUSC Architecture

Consider a system with U users indexed by $u \in \{1, \ldots, U\}$. Each user generates a token embedding matrix $\mathbf{E}_u \in \mathbb{R}^{T \times d}$, where T denotes the token length and d the embedding dimension. Throughout this letter, bold uppercase and lowercase letters denote matrices and vectors, respectively.

Fig. 1 illustrates the end-to-end architecture. Unlike [9], which placed both the encoder and decoder at the receiver,

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

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

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

Corresponding authors: Hyun-Ho Choi; Jung-Ryun Lee.

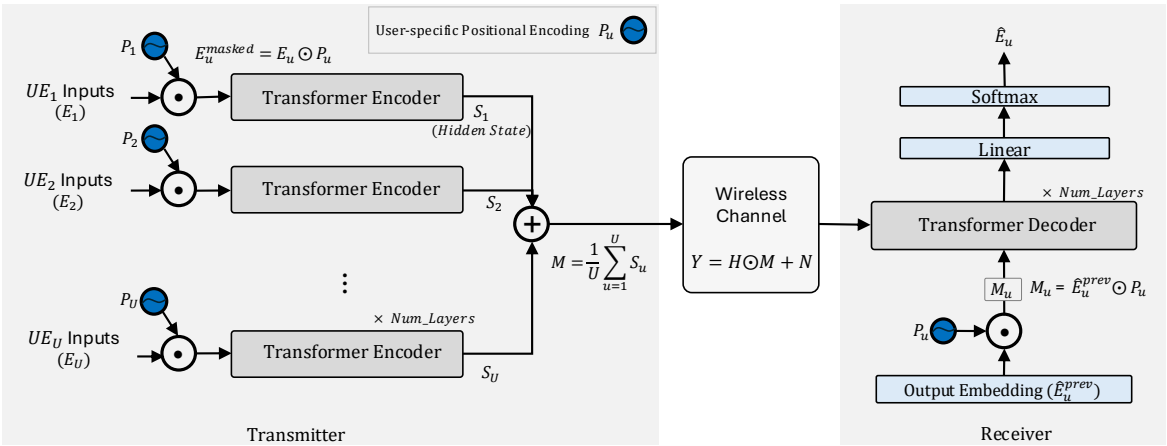

Fig. 1. End-to-end MUSC architecture. User-specific masking selects embedding coordinates via $\mathbf{P}_u$; a shared Transformer encoder produces per-user representations $\mathbf{S}_u$; mean aggregation forms a multiplexed sequence $\mathbf{M}$; the Rayleigh channel applies fading $\mathbf{H}$ and noise $\mathbf{N}$; and each user reconstructs $\hat{\mathbf{E}}_u$ through mask-reused cross-attention decoding.

the proposed MUSC framework separates them into the transmitter and receiver sides, conforming to the standard semantic communication paradigm [10]. At the transmitter, each user applies a user-specific positional mask $\mathbf{P}_u \in \mathbb{R}^d$ that deterministically selects and weights embedding coordinates:

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

where $\odot$ denotes the Hadamard (elementwise) product. The mask $\mathbf{P}_u$ is applied identically across all T tokens: entries close to zero suppress coordinates not allocated to user u, while larger entries emphasize user-specific coordinates. Through this mechanism, each user effectively occupies a distinct subspace of the shared encoder, realizing a soft partitioning of the embedding space that mitigates inter-user interference without relying on orthogonal physical resources.

All users share a single Transformer encoder with tied parameters. The masked sequence is encoded as

$$\mathbf{S}_u = \text{Encoder}\big(\mathbf{E}_u^{\text{masked}}\big), \tag{2}$$

where $\mathbf{S}_u \in \mathbb{R}^{T\times d}$. Each encoder block comprises multi-head self-attention [3], residual connections with layer normalization, and a position-wise feed-forward network. Per-user encoded representations are then aggregated into a single transmitted sequence via elementwise averaging:

$$\mathbf{M} = \frac{1}{U}\sum_{u=1}^{U} \mathbf{S}_u, \tag{3}$$

which maintains constant transmission bandwidth regardless of U, since the dimensionality of $\mathbf{M}$ does not grow with the number of users. This aggregation also preserves gradient flow to all users through the shared encoder during backpropagation and avoids additional signaling overhead at the physical layer.

The multiplexed sequence is transmitted over a memoryless Rayleigh fading channel [14] with additive white Gaussian noise (AWGN):

$$\mathbf{Y} = \mathbf{H} \odot \mathbf{M} + \mathbf{N}, \tag{4}$$

where $\mathbf{H} \in \mathbb{R}^{T\times d}$ contains independent Rayleigh-distributed fading gains with scale parameter $\sigma_h = 1/\sqrt{2}$, yielding unit average channel power (i.e., $\mathbb{E}[|[\mathbf{H}]_{t,k}|^2] = 1$), and $\mathbf{N} \in \mathbb{R}^{T\times d}$ is AWGN with variance σ_n^2 set according to the target signal-to-noise ratio (SNR).

At the receiver, user u forms a masked query $\mathbf{M}_u = \hat{\mathbf{E}}_u^{\text{prev}} \odot \mathbf{P}_u$ using a learnable prior and the same mask $\mathbf{P}_u$, then performs cross-attention decoding:

$$\hat{\mathbf{E}}_u = \text{Decoder}(\mathbf{M}_u, \mathbf{Y}), \tag{5}$$

where the masked query serves as the query input while the received sequence $\mathbf{Y}$ provides both key and value. By reusing the same mask $\mathbf{P}_u$ at both the encoder and decoder, the framework ensures transmitter–receiver consistency and focuses decoding attention on user-u coordinates, enabling per-user reconstruction without physical-layer orthogonalization. Since only the mask $\mathbf{P}_u$ varies across users while all weights are shared, the framework supports an arbitrary number of simultaneous users without user-specific model instances.

B. Learning Objective

The training objective addresses two complementary goals: accurate reconstruction of each user's semantic content and inter-user separability in the shared embedding space. The CE loss supervises token-level prediction accuracy:

$$\mathcal{L}_{\text{CE}} = -\sum_{u=1}^{U}\sum_{t=1}^{T}\sum_{v=1}^{V} y_u^{(t,v)} \log \hat{p}_u^{(t,v)}, \tag{6}$$

where $\hat{p}_u^{(t,v)}$ is the predicted probability of token v at position t for user u, $y_u^{(t,v)}$ is the one-hot ground-truth indicator, and V is the vocabulary size.

While the CE loss ensures reconstruction fidelity, it does not explicitly constrain the geometric arrangement of different users' representations. To enforce inter-user separability, the normalized temperature-scaled cross-entropy (InfoNCE) loss [11], [12] is introduced. Let $\mathbf{z}_u^a$ and $\mathbf{z}_u^b$ denote ℓ_2-normalized projections of the mean-pooled encoder output of user u from two independent channel realizations. These representations are obtained by passing $\bar{\mathbf{s}}_u = \frac{1}{T}\sum_{t=1}^{T}[\mathbf{S}_u]_t \in \mathbb{R}^d$ through a learnable two-layer MLP projection head followed

TABLE I
SIMULATION PARAMETERS

Parameter	Value
Number of users (U)	4, 8, 16
Embedding dimension (d)	128
Encoder / Decoder layers	4 / 2
Vocabulary size (V)	4 (QPSK-equivalent)
SNR range (dB)	0–25 (5 dB steps)
Channel model	Rayleigh fading
Epochs / Training batches	100 / 10,000
Evaluation batches	100,000
Optimizer / Learning rate	AdamW / 10^{-4}
Contrastive weight ($\lambda_{\text{contrast}}$)	10^{-2} (optimized)

by ℓ_2-normalization, so that each projected vector lies on the unit hypersphere. The contrastive loss is defined as

$$\mathcal{L}_{\text{NCE}} = -\frac{1}{U}\sum_{u=1}^{U}\log\frac{\exp\big(\text{sim}(\mathbf{z}_u^a,\mathbf{z}_u^b)/\tau\big)}{\sum_{v=1}^{U}\exp(\text{sim}(\mathbf{z}_u^a,\mathbf{z}_v^b)/\tau)}, \tag{7}$$

where $\text{sim}(\mathbf{a},\mathbf{b}) = \mathbf{a}^\top\mathbf{b}/(\|\mathbf{a}\|\|\mathbf{b}\|)$ is the cosine similarity and $\tau > 0$ is a temperature parameter controlling the sharpness of the similarity distribution. The numerator encourages intra-user embedding cohesion, while the denominator pushes apart representations of different users. The overall training objective combines both terms as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{contrast}}\mathcal{L}_{\text{NCE}}, \tag{8}$$

where $\lambda_{\text{contrast}} \geq 0$ balances reconstruction fidelity against inter-user separability. Setting $\lambda_{\text{contrast}} = 0$ recovers the pure CE baseline, while increasing $\lambda_{\text{contrast}}$ places greater emphasis on embedding-space discrimination.

III. Simulation Results

A. Experimental Setup

Three schemes are compared: (1) *CE baseline* trained with $\mathcal{L}_{\text{CE}}$ alone; (2) *CE + Contrastive (proposed)* trained with $\mathcal{L}_{\text{total}}$; and (3) *dedicated embedding (DE)*, an interference-free reference in which each user occupies a distinct orthogonal subspace. The DE symbol error rate (SER) under flat Rayleigh fading with QPSK-equivalent symbols is given by [15]

$$P_{\text{SER}} = 1 - \left(\frac{1}{2}\left(1+\sqrt{\frac{\text{SNR}}{1+\text{SNR}}}\right)\right)^2. \tag{9}$$

Table I summarizes the simulation parameters. All models share identical encoder/decoder architectures, embedding dimensions, and hyperparameters to ensure a strictly controlled comparison. Training proceeds for 100 epochs with 10,000 batches per epoch under independent Rayleigh fading realizations, and the empirical SER is estimated over 100,000 independent Monte Carlo evaluation batches.

B. Effect of Contrastive Weight

Fig. 2 shows the SER as a function of $\lambda_{\text{contrast}}$ for $U = 8$ users across four representative SNR conditions. At every SNR level, the proposed model consistently outperforms the CE baseline (dashed lines), confirming that the contrastive term enhances inter-user separability. The SER exhibits a non-monotone dependence on $\lambda_{\text{contrast}}$: performance improves as $\lambda_{\text{contrast}}$ increases from 10^{-3} to 10^{-2}, achieving the best trade-off between reconstruction fidelity and inter-user separation. Beyond this point, excessive contrastive regularization distorts the embedding geometry and degrades reconstruction quality. This confirms that an intermediate value strikes the best balance between inter-user separation and embedding-structure preservation. Accordingly, $\lambda_{\text{contrast}} = 10^{-2}$ is adopted for all subsequent experiments.

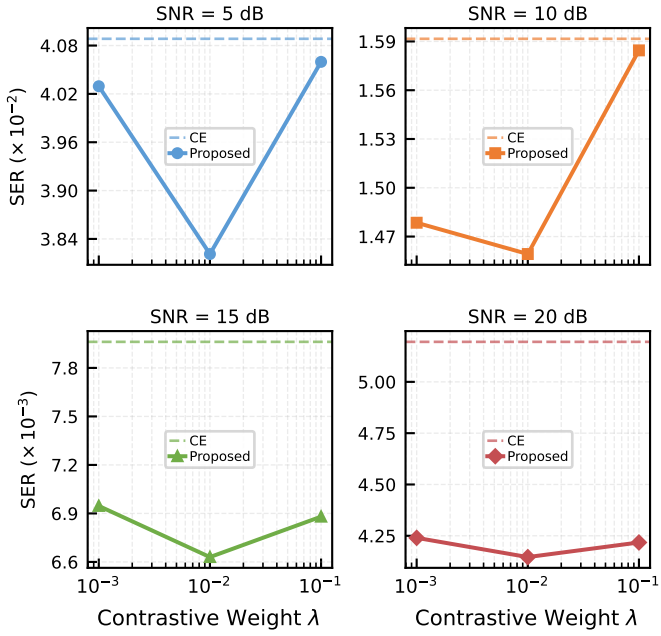

Fig. 2. SER versus $\lambda_{\text{contrast}}$ for $U = 8$ and $d = 128$ at SNRs of 5, 10, 15, and 20 dB. Dashed lines indicate the CE baseline.

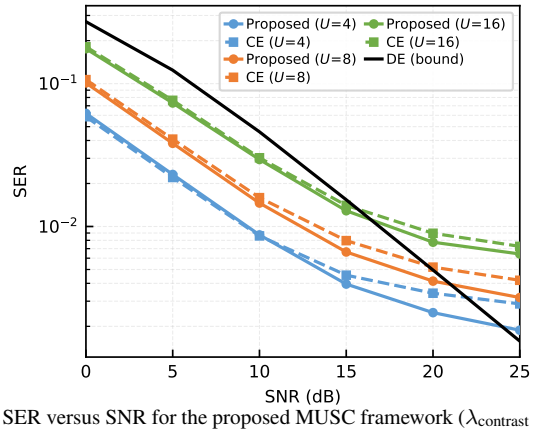

Fig. 3. SER versus SNR for the proposed MUSC framework ($\lambda_{\text{contrast}} = 0.01$) with 4, 8, and 16 users ($d = 128$). The DE curve represents the theoretical lower bound under a non-shared embedding scheme.

C. SER Performance Comparison

Fig. 3 presents the SNR-wise SER for the CE baseline and the proposed CE + Contrastive scheme under 4-, 8-, and 16-user configurations, alongside the theoretical DE bound in (9). Across all user counts and the entire SNR range, the proposed scheme achieves consistently lower SER, confirming that contrastive regularization yields generalizable gains beyond

masking-based separation alone. The improvement is most pronounced at $U = 4$, where the contrastive loss operates on fewer negative pairs and achieves cleaner user-specific subspace assignments. As U increases to 8 and 16, the gap narrows due to denser inter-user interference under a fixed embedding dimension, suggesting that adaptive masking or increased embedding capacity could provide further gains in dense deployments.

At high SNR, all models converge toward the DE bound, indicating that residual inter-user interference, rather than channel noise, becomes the dominant performance bottleneck. In this regime, the system performance is primarily governed by the model's ability to discriminate among user embeddings in the shared latent space. The proposed model achieves closer proximity to the DE bound than the CE baseline across all user configurations, further confirming that contrastive regularization effectively suppresses residual inter-user interference at the embedding level. Notably, these gains are achieved without modifying the encoder–decoder architecture or the physical-layer scheme, demonstrating that learning-based embedding coordination alone yields substantial improvements.

D. Complexity Analysis

Both configurations employ identical Transformer encoder and decoder architectures, so the dominant computational cost is shared. Each Transformer layer involves multi-head attention with complexity $\mathcal{O}(T^2 d)$ and feed-forward projections with complexity $\mathcal{O}(Td^2)$. Since all users share the same encoder and decoder weights, the per-iteration cost scales linearly with the number of users:

$$\mathcal{O}_{\text{shared}} = \mathcal{O}\big(U \cdot L \cdot (T^2 d + T d^2)\big), \tag{10}$$

where L denotes the total number of encoder and decoder layers. The InfoNCE loss (7) additionally requires pairwise cosine similarity computation across all U users, incurring

$$\mathcal{O}_{\text{NCE}} = \mathcal{O}(U^2 d). \tag{11}$$

This quadratic dependence on U arises only during loss evaluation, while the forward and backward passes remain identical to those of the CE baseline. The relative overhead ratio is

$$\frac{\mathcal{O}_{\text{NCE}}}{\mathcal{O}_{\text{shared}}} \approx \frac{U}{L \cdot T \cdot d}. \tag{12}$$

Under our settings ($U = 16$, $L = 6$, $T = 32$, $d = 128$), this ratio evaluates to approximately 6.5×10^{-4}, confirming that the contrastive overhead is negligible. Moreover, this ratio diminishes as either T or d increases, making the proposed approach increasingly practical for large-scale deployments.

IV. Conclusion

This letter proposed MUSC, a Transformer-based multi-user semantic communication framework that achieves scalable embedding-space multiplexing through contrastive learning. The framework employs user-specific positional masking to partition the shared embedding space without dedicated encoders or orthogonal physical resources, and jointly optimizes cross-entropy reconstruction with an InfoNCE-based contrastive objective to enforce inter-user separability. By combining architectural separation through masking with learning-based separation through contrastive regularization, MUSC achieves synergistic improvements in both semantic fidelity and interference suppression.

Simulation results demonstrated consistent SER reduction across all SNR levels and user configurations, with the performance gain most pronounced at lower user counts and at high SNR where inter-user interference dominates. The contrastive overhead was confirmed to remain below 10^{-3} relative to the overall Transformer complexity, verifying practical feasibility for deployment in resource-constrained wireless environments.

Future work includes adaptive masking strategies that dynamically allocate embedding subspaces based on channel state information, extension of the semantic source model to richer modalities such as natural language and images, and investigation of contrastive regularization under frequency-selective or time-varying channel models to further validate the generality of the proposed approach.

References

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

[2] D. Gündüz, Z. Qin, I. Estella Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong, and C.-B. Chae, "Beyond transmitting bits: Context, semantics, and task-oriented communications," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 5–41, 2023.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

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

[5] X. Chen, D. W. K. Ng, W. Yu, E. G. Larsson, N. Al-Dhahir, and R. Schober, "Massive access for 5G and beyond," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 3, pp. 615–637, 2021.

[6] H. Xie, Z. Qin, X. Tao, and K. B. Letaief, "Task-oriented multi-user semantic communications," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 9, pp. 2584–2597, 2022.

[7] R. Fantacci and B. Picano, "Multi-user semantic communications system with spectrum scarcity," *Journal of Communications and Information Networks*, vol. 7, no. 4, pp. 375–382, Dec. 2022.

[8] X. Mu, Y. Liu, L. Guo, and N. Al-Dhahir, "Heterogeneous semantic and bit communications: A semi-NOMA scheme," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 155–169, 2023.

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

[10] Q. Zhou, R. Li, Z. Zhao, C. Peng, and H. Zhang, "Semantic communication with adaptive universal transformer," *IEEE Wireless Communications Letters*, vol. 11, no. 3, pp. 453–457, 2022.

[11] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[12] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, vol. 119. PMLR, 2020, pp. 1597–1607.

[13] S. Tang, Q. Yang, L. Fan, X. Lei, A. Nallanathan, and G. K. Karagiannidis, "Contrastive learning-based semantic communications," *IEEE Transactions on Communications*, vol. 72, no. 10, pp. 6328–6343, Oct. 2024.

[14] A. Goldsmith, *Wireless Communications*. Cambridge University Press, 2005.

[15] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.