Multi-Modal Multi-Task Semantic Communication: A Distributed Information Bottleneck Perspective

Yujie Zhou, Yiwei Liao, Cheng Peng, Rulong Wang, Yong Xiao, Yingyu Li, Guangming Shi

***Abstract*—Semantic communication (SemCom) shifts the focus from data transmission to meaning delivery, enabling efficient and intelligent communication. Existing AI-based coding schemes for multi-modal multi-task SemCom often require transmitters with full-modal data to participate in all receivers' tasks, which leads to redundant transmissions and conflicts with the physical limits of channel capacity and computational capability. In this paper, we propose PoM^2-DIB, a novel framework that extends the distributed information bottleneck (DIB) theory to address this problem. Unlike the typical DIB, this framework introduces modality selection as an additional key design variable, enabling a more flexible tradeoff between communication rate and inference quality. This extension selects only the most relevant modalities for task participation, adhering to the physical constraints, while following efficient DIB-based coding. To optimize selection and coding end-to-end, we relax modality selection into a probabilistic form, allowing the use of score function estimation with common randomness to enable optimizable coordinated decisions across distributed devices. Experimental results on public datasets verify that PoM^2-DIB achieves high inference quality compared to full-participation baselines in various tasks under physical limits.**

I. Introduction

THE superiority of multiple data modalities over a single modality lies in the fact that it can unlock deeper insights about entities from multiple perspectives [2], thereby enabling the handling of complex intelligent tasks [3]. This increases the popularity of multi-modal data sources: Nowadays, advanced intelligent applications tend to leverage such informative data to improve reasoning quality, e.g., autonomous driving with vision and radar. With the development of such applications at the network edge, the demand for multi-modal, multi-task and cross-device coordination is increasing. Centralized processing is unrealistic because of its significant communication overhead and data security concerns. However, efficient communication across distributed multi-modal sources to achieve coordination is challenging. Compared to single-modal data, the large volume of multi-modal data far exceeds the capacity of typical communication paradigms, especially since this type of data is usually used for multiple downstream tasks, the complexity of communication is further exacerbated. Therefore, a new unified and efficient communication paradigm is needed to address such complex communication tasks.

The task-oriented semantic communication (SemCom) is a promising solution. It focuses on transporting and delivering the task-specific meaning of messages, instead of transmitting the raw data [4], highly reducing communication overhead and transmission latency as well as enabling customized intelligent services [5], [6]. Despite its great potential to achieve efficient multi-modal multi-task communication, one pivotal challenge impeding its widespread application is the lack of efficient coding schemes capable of capturing and reconstructing task-specific semantics from distributed observable data [7].

Unlike traditional communication systems, which primarily aim at redundancy reduction and accurate bit-stream recovery, SemCom codecs must accurately capture task-relevant correlations at a low communication cost. Note that semantics inherently vary with tasks, these codecs need to adapt dynamically according to different task objectives. This challenge is further complicated by the multi-modal data scenario, where transmitters observe heterogeneous modalities exhibiting potentially complex and task-dependent correlations.

To systematically tackle these complexities, this paper studies the codec design problem for multi-modal multi-task SemCom systems, where distributed transmitters sense distinct data modalities. We investigate this problem from the information-theoretic perspective of the distributed information bottleneck (DIB) theory [8], which forms a theoretically optimal tradeoff between compressing input data into low-complexity, i.e., low-rate, representations and preserving task-relevant information, a.k.a. rate-relevance tradeoff [8] among multiple distributed sources. It is a theoretical extension of the classical information bottleneck (IB) theory [9] to balance communication costs and task-specific semantic recovery rigorously.

Existing DIB coding frameworks often consider degenerated single-task scenarios, where all devices are involved and each transmitter only has a single modality. A non-trivial gap exists in extending DIB to multi-modal multi-task cases. Real-world communication systems face physical constraints, esp. limited channel and computational resources, making it impractical for each device to communicate simultaneously with all others in a full-device manner (traditional DIB) for heavy multi-modal multi-task inference. On the other hand, high data redundancy is implied in multi-modal sources, i.e., distinct modalities may convey redundant information for a specific task, e.g., visual and thermal data both indicating object recognition. The above two observations derive a key requirement of the multi-modal

A preliminary version of this paper was presented at IEEE International Conference on Communications (ICC), Montreal, Canada, June 2025 [1].

Yujie Zhou, Yiwei Liao, Cheng Peng, Rulong Wang and Yong Xiao are with the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China. Yong Xiao is also with Peng Cheng Laboratory, Shenzhen, China, and Pazhou Laboratory (Huangpu), Guangzhou, China (e-mail: {zhouyujie2357, liao_yiwei, m202372990, rulongwang, yongxiao}@hust.edu.cn).

Yingyu Li is with the School of Mechanical Engineering and Electronic Information, China University of Geosciences, Wuhan 430074, China (e-mail: liyingyu29@cug.edu.cn).

Guangming Shi is with the Peng Cheng Laboratory, Shenzhen 518055, China (e-mail: gmshi@xidian.edu.cn).

This version is supplementary material.

multi-task setting: limiting the number of modality-level communication links to remove redundancy among modalities and meet physical resource constraints. Specifically, receivers must strategically choose a proper subset of transmitters that most effectively support their specific tasks. Meanwhile, transmitters that observe multiple raw data modalities must properly select the most (task-specific) informative modalities to transmit to each requesting receiver. An extended theoretical framework is necessary to incorporate such a two-sided selection into the rate-relevance tradeoff to meet the requirement. We construct such a framework by generalizing DIB to multiple modalities for each transmitter and performing multiple tasks on different receivers. We refer to this form as the multi-modal multi-task distributed information bottleneck (M^2-DIB) problem.

Although this M^2-DIB successfully generalizes the classical DIB to multi-modal multi-task SemCom scenarios, solving this problem remains computationally and theoretically challenging due to its combinatorial nature and the interdependence of distributed selection decisions, esp., selecting optimal subsets of modalities requires joint optimization alongside continuous semantic codec parameters, highly complicating the computational landscape.

To tackle these challenges, we propose a computable version of M^2-DIB, namely *probabilistic multi-modal multi-task DIB (PoM^2-DIB)*. In PoM^2-DIB, the combinatorial complexity of discrete selection decisions is mitigated by transforming selection policies into differentiable probabilistic forms. Leveraging score function estimation and incorporating common randomness [10] among distributed devices, our solution achieves coordinated yet decentralized decision-making, compatible with the distributed networking. In addition, we employ variational approximation techniques to provide computationally tractable bounds for conditional entropy terms, avoiding the need for explicit knowledge of complex underlying data distributions. Computational efficiency is further enhanced by new empirical approximations of high-dimensional mutual information (MI).

In sum, our work makes the following contributions:

- We propose the PoM^2-DIB framework, which generalizes DIB theory by introducing probabilistic modality selection as a new degree of freedom in the DIB-induced rate-relevance tradeoff, enabling efficient resource-constrained multi-modal multi-task semantic communication.
- PoM^2-DIB is established as a computationally tractable solution for M^2-DIB, involving a cooperative probabilistic selection with common randomness to enable efficient end-to-end optimization of semantic codecs and modality selection policies.
- We analyze theoretical properties of PoM^2-DIB, proving the equivalence between the primal deterministic M^2-DIB and the relaxed probabilistic form. More importantly, we prove that even when relaxing all the resource constraints to allow full participation, the optimal solution may still exclude certain modalities due to redundancy, demonstrating that PoM^2-DIB achieves a more flexible tradeoff than the standard DIB.
- Extensive experiments are conducted on publicly available multi-modal multi-task datasets to demonstrate the effectiveness of PoM^2-DIB: it performs as well as the full participation baselines under hard physical constraints. The convergence, esp. the selection policy, is also studied.

II. Related Work

A. Multi-Modal Multi-Task Semantic Communication

Recently, research on semantic communication has shifted its focus to the recovery of task-relevant implicit information at the receiving end rather than the restoration of the original data. In this field, most existing works [11]–[13] tackle single-transmitter-to-receiver single-task cases. Some existing works consider more realistic scenarios involving multiple transmitting devices, e.g., work [14] investigates the distributed semantic coding through the distributed information bottleneck (DIB) coding framework [8] for collaborative feature extraction to handle a single task.

For practical scenarios involving multi-modality and multi-tasking, SemCom systems must be able to accomplish a set of tasks concurrently. Early efforts [15]–[17] explore multi-modal multi-tasking in a non-cooperative pattern, where every task operates independently on its own multi-modal dataset. While recent studies [18], [19] investigate the joint/cooperative multi-tasking across multiple distributed devices. Although all the early efforts [15]–[19] try to construct practical multi-modal multi-task SemCom systems, they always adopt an unrealistic assumption: all transmitters with multiple local data modalities are involved in all receivers' tasks. None of them considers the hard constraint induced by the limit of channel capacity and computational capability, i.e., in a finite time frame, all devices can only simultaneously communicate with a finite number of modalities of a finite number of other devices. Besides, since most solutions [15]–[17] are built on DL techniques to enhance inference quality without rate concerns, they lack information-theoretic guarantees to achieve the optimal tradeoff between communication rate and inference quality. In contrast, PoM^2-DIB is built based on DIB theory, which deals with this issue and rigorously models multi-modal multi-task scenarios.

B. Distributed Information Bottleneck

The Tishby's information bottleneck (IB) is first proposed in [9]. It is grounded in RD theory [20]. Unlike the data reconstruction emphasized by the RD-based lossy compression, IB focuses on recovering the hidden meaning or semantics in the data under a rate limit. In which the meaning is introduced as a relevant hidden variable according to the raw data. It extracts the relevant information of targets (meaning) from raw signals to yield representations that are minimally informative about raw signals and maximally informative about targets, in which the informativeness is measured by MI. The current work [21] indicates that IB is essentially a remote source coding problem in which the distortion is measured under logarithmic.

Slepian-Wolf's theorem [22] indicates that distributed coding can achieve the joint coding rate of correlated memoryless sources. Work [8] extends IB theory to distributed scenarios as DIB, which involve multiple transmitters and a receiver. This extension is rigorous, which follows the classical result [22], [23] to characterize the optimal rate-relevance tradeoff region.

IB-based coding can be implemented efficiently. The paper [24] establishes the deep variational IB (VIB), which utilizes variational techniques and deep learning-based parameterization schemes to achieve IB. By applying the powerful gradient backpropagation in the learning paradigm, the feasibility of IB has been greatly enhanced. This also illustrates that leveraging DIB to establish multi-modal multi-task SemCom systems is promising: it is theoretically grounded in information theory and can also be efficiently implemented by DL.

III. System Model and Problem Formulation

A. Notations

We first clarify the main notations used in this paper. Calligraphic letters (e.g., $\mathcal{X}$) denote sets or functionals, uppercase letters (e.g., X) represent random variables or constants, and lowercase letters (e.g., x) represent elements of sets. Boldface letters (e.g., $\boldsymbol{x}$ or $\boldsymbol{X}$) emphasize vector structures. We denote the cardinality of a set $\mathcal{X}$ by $|\mathcal{X}|$. The distribution induced by a random variable X is denoted by P_X, and its probability density function (pdf) or probability mass function (pmf) by p_X. The Kullback-Leibler (KL) divergence between two distributions P and Q is given by $D_{\mathrm{KL}}(P\|Q) \triangleq \mathbb{E}_P[\log \frac{dP}{dQ}]$ where $P \ll Q$. The mutual information (MI) between random variables X and Z is defined as $I(X;Z) \triangleq D_{\mathrm{KL}}(P_{XZ}\|P_X P_Z)$. The notation $[N]$ abbreviates the index set $\{1,...,N\}$.

B. System Model

We consider a multi-modal multi-task semantic communication system illustrated in Fig. 1(a), consisting of multiple distributed transmitters and receivers. Let $\mathcal{K} := \{1,...,K\}$ denote the set of transmitters and $\mathcal{T} := \{1,...,T\}$ represent the set of receivers. All devices are synchronized, that is, their observed data share a common time frame.

Each transmitter $k \in \mathcal{K}$ observes multi-modal data $\boldsymbol{X}_k$, consisting of $m(k)$ distinct modalities, formally represented as $\boldsymbol{X}_k := (X_k^1,...,X_k^{m(k)})$. Even when transmitters sense similar modalities, the geographical distribution could result in diverse perspectives and measurements, which leads to $\boldsymbol{X}_k \neq \boldsymbol{X}_{k'}$ for any $k \neq k' \in \mathcal{K}$. This "$\neq$" allows $\boldsymbol{X}_k$ and $\boldsymbol{X}_{k'}$ to be ranged on different value spaces, overlapping or not. Therefore, each transmitter's data is unique, though potentially correlated, and follows a joint distribution $P_{\boldsymbol{X}_1\cdots\boldsymbol{X}_K}$.

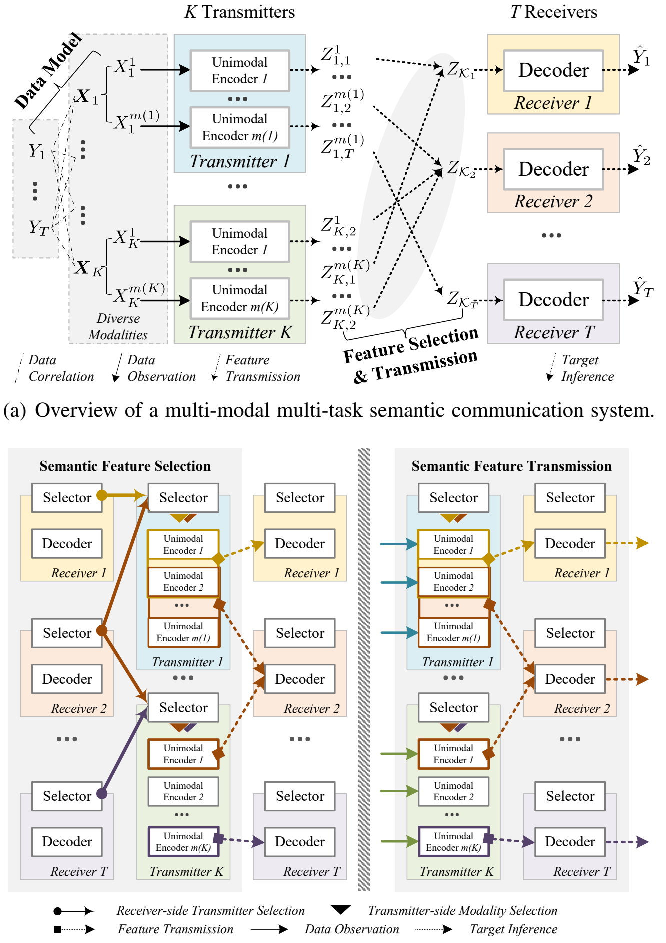
Each receiver $t \in \mathcal{T}$ aims to execute its respective semantic inference task, e.g., decision-making or pattern recognition. Formally, receiver t intends to optimally infer its task-specific semantic target Y_t using available transmitter data. The optimal inference can be formally described as a conditional distribution $P_{Y_t|\boldsymbol{X}_1\cdots\boldsymbol{X}_K}$ for such t.

C. Semantic Feature Selection and Transmission

Modality selection is first determined, followed by communication with DIB-based coding, as shown in Fig. 1(b).

1) Transmitter Side: Each transmitter independently encodes local raw data in a unimodal manner. Specifically, given a modality $m \in [m(k)]$ and a particular task t, the transmitter k generates a semantically compressed feature:

$$Z_{k,t}^m = f_k^m(X_k^m, t), \tag{1}$$



(a) Overview of a multi-modal multi-task semantic communication system.

(b) Feature selection and transmission mechanism: 1) receiver-side selection → 2) transmitter-side selection → 3) multi-modal multi-task communication

Fig. 1: A multi-modal multi-task semantic communication system and details about feature selection and transmission.

where $f_k^m : \mathcal{X}_k^m \times \mathcal{T} \to \mathcal{Z}_k^m$ represents the unimodal encoder for the m-th modality at transmitter k. Due to limited channel capacity, this encoding should yield task-specific compressed features, effectively reducing task-irrelevant redundancy. For each given t, f_k^m induces a t-specific conditional $P_{Z_{k,t}^m|X_k^m}$.

Due to limited computing resources, transmitter k can only simultaneously process E_k-times task-specific encoding in a finite time frame. In this case, a transmitter-side selector $\upsilon_k : \mathcal{T} \to 2^{[m(k)]}$ is introduced, which maps receivers to a subset of available modalities. Formally, the limit is described as

$$|\mathcal{T}_k| \leq E_k, \text{ where } \mathcal{T}_k := \textstyle\bigsqcup_{t \in \hat{\mathcal{T}}_k} \upsilon_k(t) \tag{2}$$

with $\hat{\mathcal{T}}_k \subseteq \mathcal{T}$ denoting the set of receivers that transmitter k supports for and $\bigsqcup$ standing for disjoint union.

2) Receiver Side: Similarly, each receiver t must select a subset of transmitters due to physical constraints. The receiver-side selector $\upsilon_t : \{\mathcal{K}\} \to \mathcal{K}_t \in 2^{\mathcal{K}}$ determines a subset $\mathcal{K}_t$ of all available transmitters $\mathcal{K}$, i.e., $\mathcal{K}_t = \upsilon_t(\mathcal{K})$ (whose input is always the complete index set), subject to:

$$1 \leq |\mathcal{K}_t| \leq E_t \leq K. \tag{3}$$

Also, in terms of all $\{\mathcal{K}_t\}_t$, we can formally define $\hat{\mathcal{T}}_k$ at k:

$$\hat{\mathcal{T}}_k := \{t : \text{ this specific } k \in \mathcal{K}_t, \forall t \in \mathcal{T}\}. \tag{4}$$

Receiver t receives compressed semantic features $Z_{k,t} = \{Z_{k,t}^m : m \in \upsilon_k(t)\}$ from all selected transmitters $k \in \mathcal{K}_t =$

$\upsilon_t(\mathcal{K})$, forming a fused representation $Z_{\mathcal{K}_t} := \{Z_{k,t}\}_{k \in \mathcal{K}_t}$. It attempts to recover its target Y_t via a decoder $g_t : \mathcal{Z}_{\mathcal{K}_t} \to \mathcal{Y}_t$,

$$\hat{Y}_t = g_t(Z_{\mathcal{K}_t}). \tag{5}$$

A Markov chain $Y_t \leftrightarrow X_k^m \leftrightarrow Z_{k,t}^m \leftrightarrow \hat{Y}_t$ holds for any $k \in \mathcal{K}$, $t \in \mathcal{T}$, and $m \in [m(k)]$ in this system [8].

The main problem of this SemCom system is to recover the true Y_t as precisely as possible via $\hat{Y}_t$ for every t with a low rate under the above communication link limitations. This problem is fundamentally different from the traditional source coding problems, focusing on directly compressing raw data at the transmitters to fit the capacity of communication channels.

D. Distributed Information Bottleneck for Feature Coding

To rigorously formulate the semantic communication and coding processes, we utilize the distributed information bottleneck (DIB) theory [8]. Unlike typical learning-based SemCom that mainly focuses on approximating the ideal $P_{Y_t|\boldsymbol{X}_1 \ldots \boldsymbol{X}_K}$, the IB theory aims to find the optimal way to compress the information rate from a random variable X that preserves the relevance about a target Y [9]. DIB extends the classical IB framework with information-theoretic tools to cases involving multiple encoders. A standard DIB consists of K encoders and a single decoder for recovering the target signal Y, where each encoder converts its local observation X_k to a low-dimensional and low-rate feature Z_k. With a Lagrangian multiplier $\beta \geq 0$, we expect encoders to satisfy the following

$$\begin{aligned} &\min\nolimits_{\{p(z_k|x_k)\}_k} \Big\{ \mathcal{L}_{\text{DIB}}[\{p(z_k|x_k)\}_k] \\ &\quad := H(Y|Z_{\mathcal{K}}) + \beta \sum_{k \in \mathcal{K}} (H(Y|Z_k) + I(X_k; Z_k)) \Big\}, \end{aligned} \tag{6}$$

with $Z_{\mathcal{K}} = \{Z_1, ..., Z_K\}$. The first term represents reconstruction quality, while the second term balances compression and relevance. Particularly, the recovery quality of $\hat{Y}_t$ w.r.t. Y_t is measured by conditional entropy (CE), which is equivalent to the IB-induced MI when the data distribution is invariant [8].

E. Problem Formulation

We now present the multi-modal multi-task DIB (M^2-DIB) problem. Our objective is to find selectors $\{\upsilon_t\}_{t \in \mathcal{T}}$, $\{\upsilon_k\}_{k \in \mathcal{K}}$, and encoders $\{f_k^m\}_{m \in [m(k)], k \in \mathcal{K}}$ jointly to achieve the optimal multi-modal multi-task SemCom under physical constraints:

$$\begin{aligned} &\min_{\substack{\{\upsilon_t\}_{t \in \mathcal{T}}, \{\upsilon_k\}_{k \in \mathcal{K}}, \\ \{f_k^m\}_{m \in [m(k)], k \in \mathcal{K}}}} \Bigg\{ \mathcal{L}_{\text{M}^2\text{-DIB}}[\{\upsilon_t\}_{t \in \mathcal{T}}, \{\upsilon_k\}_{k \in \mathcal{K}}, \{f_k^m\}_{m \in [m(k)], k \in \mathcal{K}}] \\ &:= \sum_{t \in \mathcal{T}} \Bigg\{ \underbrace{H(Y_t|Z_{\mathcal{K}_t})}_{\substack{\text{multi-modal} \\ \text{relevance}}} + \beta \sum_{\substack{k \in \upsilon_t(\mathcal{K}), \\ m \in \upsilon_k(t)}} \underbrace{H(Y_t|Z_{k,t}^m)}_{\substack{\text{single-modal} \\ \text{relevance}}} + \underbrace{I(X_k^m; Z_{k,t}^m)}_{\substack{\text{single-modal} \\ \text{rate}}} \Bigg\} \Bigg\} \\ &\qquad \text{s.t.} \begin{cases} 1 \leq |\mathcal{K}_t| \leq E_t, \forall t \in \mathcal{T}, \\ |\mathcal{T}_k| \leq E_k, \forall k \in \mathcal{K}. \end{cases} \end{aligned} \tag{$\mathbf{P}_0$}$$

This formulation unifies selection with coding as optimizable variables, extending (6) to multi-modal multi-task cases. Since it is based on DIB, the modality-fused $Z_{\mathcal{K}_t}$ is low-dimensional and low-complexity, and can be depicted with a low information rate [8] at a given level of semantics recovering quality.

This problem ($\mathbf{P}_0$) is a multivariable optimization problem and is generally difficult to solve. Two major challenges that hinder the address of ($\mathbf{P}_0$) include: (i) Finding optimal selectors that are most beneficial for solving tasks is quite challenging due to the decentralization and combinatorial complexity. Esp., selections ($\mathcal{K}_t$ and $\mathcal{T}_k$) are highly interdependent and mutually restrictive through (2)-(4). (ii) Directly computing information terms and ideal decoders $(g_t)_t$ is practically infeasible, since the analytical joint distributions are unknown in most realistic scenarios, and only accessible via finite empirical samples. Addressing these challenges demands a computationally feasible yet theoretically rigorous solution.

IV. Probabilistic Feature Selection

The first challenge corresponds to the design of a selection policy. Previous works [1], [14] query a value table based on pre-defined metrics to find the subset that has the highest value as their selection. This type of approach must be sub-optimal because the true optimal subset should be induced through the original objective ($\mathbf{P}_0$) itself inherently.

Based on this observation, we directly optimize the selection policy towards minimizing the objective functional $\mathcal{L}_{\text{M}^2\text{-DIB}}[\cdot]$, which is, however, extremely intractable.

To tackle this issue, we transform the deterministic selection design into a variational approximation problem in probability. This form of selection policy is directly optimizable using the score function estimation detailed in Section VI.

A. Characterization of Selection in Probability

Each selector υ_t at receiver-side yields the selected set $\mathcal{K}_t$. It maps a set to a set. To facilitate subsequent calculations, we process these set objects as characteristic vectors. We define a vector corresponding to $\mathcal{K}_t$, i.e.,

$$\hat{\boldsymbol{a}}_t := (a_{k,t})_{k=1}^K \in (\{0,1\}^K, \|\cdot\|_1), \tag{7}$$

with $a_{k,t} = 1$ if $k \in \mathcal{K}_t$ else 0 w.r.t. the t-th receiver. This $\hat{\boldsymbol{a}}_t$ represents all transmitters selected by the receiver t.

Then, we define all transmitter-side selectors υ_k as a single vectorized $\upsilon_{\mathcal{K}} : \{0,1\}^K \to \{0,1\}^{j_K}$ with $j_k = \sum_{j=1}^k m(j)$ if $k \in [K]$ else 0, which satisfies

$$\boldsymbol{a}_t = \upsilon_{\mathcal{K}}(\hat{\boldsymbol{a}}_t) := ((a_{k,t}^m)_{m=1}^{m(k)})_{k=1}^K, \tag{8}$$

where $a_{k,t}^m = 1$ if $m \in \upsilon_k(t)$ and $k \in \mathcal{K}_t$ else 0. $\boldsymbol{a}_t$ represents all the modalities sent to the receiver t. Let $\boldsymbol{a}_{\mathcal{T}} = \sum_{t=1}^T \boldsymbol{a}_t$, the total modality usage across all the receivers. We can rewrite constraints in ($\mathbf{P}_0$) as

$$1 \leq |\mathcal{K}_t| \leq E_t, \forall t \in \mathcal{T} \Leftrightarrow 1 \leq \|\hat{\boldsymbol{a}}_t\| \leq E_t, \forall t \in \mathcal{T}, \tag{9}$$

$$|\mathcal{T}_k| \leq E_k, \forall k \in \mathcal{K} \Leftrightarrow \|\boldsymbol{a}_{\mathcal{T}}^{(k)}\| \leq E_k, \forall k \in \mathcal{K}, \tag{10}$$

where $\boldsymbol{a}^{(k)}$ represents a sub-vector of $\boldsymbol{a}$ from the $j_{k-1} + 1$-th dimension to the j_k-th. Finally, we concatenate all $\boldsymbol{a}_t$ to have

$$\boldsymbol{a} := \oplus_{t=1}^T \boldsymbol{a}_t \in (\{0,1\}^{j_K T}, \|\cdot\|_1), \tag{11}$$

This $\boldsymbol{a}$ represents all the modalities sent to all receivers. We call $\boldsymbol{a}$ the equivalent expression of all selectors. *We now aim to find the optimal $\boldsymbol{a}^\star$, which is equivalent to finding the overall optimal selection $\{\upsilon_t\}_{t \in \mathcal{T}}, \{\upsilon_k\}_{k \in \mathcal{K}}$.*

Note that this selection is prior to the transmission of low-dimensional features. We can *randomize the full $\boldsymbol{a}$ as its random variable counterpart $\boldsymbol{A}$* following an initial distribution $P_{\boldsymbol{A}}$. In so doing, our goal transforms to finding the optimal distribution $P^{\star}_{\boldsymbol{A}}$ that assigns probability mass close to one to that optimal $\boldsymbol{a}^{\star}$. This transformation has several advantages. Since task and data distributions are arbitrary, the uniqueness of $\boldsymbol{a}^{\star}$ can not be guaranteed. For example, let $\boldsymbol{X}_1, ..., \boldsymbol{X}_K$ be iid (independent and identically distributed) at $k \in \mathcal{K}$, then any $\boldsymbol{a}$ satisfying (9) and (10) is optimal. In this case, it is unfaithful for selectors to return a deterministic $\boldsymbol{a}$, while returning a randomized $\boldsymbol{A}$ can capture all the optimal vector $\boldsymbol{a} \in \text{supp}(\boldsymbol{A})$, where $\text{supp}(\boldsymbol{A})$ is the support of $\boldsymbol{A}$. More importantly, instead of directly solving the intractable primal ($\mathbf{P}_0$), this randomized version can be directly addressed via gradient-based methods.

B. Cooperative Policy Design and Objective Transformation

The selectors of distributed devices are also distributed. It is necessary to find an effective and efficient approach to generate the global $P_{\boldsymbol{A}}$ by coordinating all distributed selectors. Note that all selectors are synchronized. We properly establish a design that mainly leverages a type of prior information, common randomness (CR) [10], which corresponds to a shared source of randomness among devices, from this synchronization.

We expect that each selector can locally decide its selection through a shared (random) observation, e.g., a common time frame. To be specific, we introduce a common random variable $U \sim P_U$ deployed on all devices, which maps $\Omega \to \{0,1\}^R$. CR is described by an outcome $\omega \in \Omega$. Let ω be observed by all the selectors and U be independent of data.

That is, each $\hat{\boldsymbol{A}}_t$ behaves following $P_{\hat{\boldsymbol{A}}_t|U=U(\omega)}$ given ω. Also, each $\boldsymbol{A}_t$ follows $P_{\boldsymbol{A}_t|\hat{\boldsymbol{A}}_t U=U(\omega)}$ given ω and $P_{\boldsymbol{A}_t|\hat{\boldsymbol{A}}_t U} = \prod_{k=1}^{K} P_{\boldsymbol{A}_t^{(k)}|A_{k,t}U}$. In this case, $P_{\hat{\boldsymbol{A}}_t|U}$ is fully induced by υ_t, and $\{P_{\boldsymbol{A}_t^{(k)}|A_{k,t}U}\}_t$ are fully induced by υ_k. Gathering all the local decisions, we have the global $P_{\boldsymbol{A}}$:

Definition 1. Policy $P_{\boldsymbol{A}}$ is cooperative if there exists a variable U projecting CR $\omega \in \Omega$ to $u := U(\omega)$ in a finite set $\{0,1\}^R$ with $R \leq j_K T$ such that

$$\begin{aligned} P_{\boldsymbol{A}\hat{\boldsymbol{A}}U} &= P_U \textstyle\prod_{t=1}^{T} P_{\hat{\boldsymbol{A}}_t|U} P_{\boldsymbol{A}_t|\hat{\boldsymbol{A}}_t U} \\ &= P_U \textstyle\prod_{t=1}^{T} P_{\hat{\boldsymbol{A}}_t|U} \prod_{k=1}^{K} P_{\boldsymbol{A}_t^{(k)}|A_{k,t}U}, \end{aligned} \tag{12}$$

where $\hat{\boldsymbol{A}} = (\hat{\boldsymbol{A}}_t)_{t=1}^{T}$ and $\hat{\boldsymbol{A}}_t = (A_{k,t})_{k=1}^{K}$. All such $P_{\boldsymbol{A}}$ that satisfy (9) and (10) constitute to a set

$$\begin{aligned} \mathcal{P}_{\boldsymbol{A}} := \Big\{ & P_{\boldsymbol{A}} : P_{\boldsymbol{A}\hat{\boldsymbol{A}}U} = P_U \textstyle\prod_{t=1}^{T} P_{\hat{\boldsymbol{A}}_t|U} \prod_{k=1}^{K} P_{\boldsymbol{A}_t^{(k)}|A_{k,t}U}, \\ & R \leq j_K T, \text{supp}(\boldsymbol{A}) \subseteq \big\{\boldsymbol{a} : \boldsymbol{a} \in \{0,1\}^{j_K T} \text{s.t. (9) and (10)}\big\}\Big\}. \end{aligned} \tag{13}$$

Remark 1. The cardinality constraint does not hurt the expressiveness of U for the joint $P_{\boldsymbol{A}}$, since the support of U covers the set $\{0,1\}^{j_K T}$, including all values of $\boldsymbol{a}$ at $R = j_K T$.

Remark 2. CR generally exists in our considered system due to synchronization. In the pioneer DIB [8, Theorem 1], this CR has been introduced and used as a time-sharing variable. In the rest, we do not draw any distinction between the CR ω and its induced value $u = U(\omega)$.

Remark 3. How does U help cooperative selection: U provides additional information for selectors to assist them in making decisions: it acts as an additional input for either υ_t or υ_k by Definition 1. Take υ_k as an example. It is redefined on $\mathcal{T} \times \mathcal{U}$ rather than $\mathcal{T}$. It decides the selection by involving the shared information brought by $U \in \mathcal{U}$, rather than just the received t-index. This common random source can provide good support for cross-device collaboration.

In the preceding, we transform selectors to be randomized to yield their respective conditionals as random policies. We now denote $\boldsymbol{\upsilon} := (\upsilon_k)_{k=1}^{K} \oplus (\upsilon_t)_{t=1}^{T}$ as the primal deterministic version of all the selectors. Then, we can set its randomized version to $\boldsymbol{\upsilon}_{\text{p}}$ distinctively, which decides $\boldsymbol{A}$ with $P_{\boldsymbol{A}} \in \mathcal{P}_{\boldsymbol{A}}$.

We extend the original M^2-DIB objective ($\mathbf{P}_0$) to the probabilistic form, termed probabilistic M^2-DIB (PoM^2-DIB):

$$\min_{\boldsymbol{\upsilon}_{\text{p}}, \boldsymbol{f} : P_{\boldsymbol{A}} \in \mathcal{P}_{\boldsymbol{A}}} \mathcal{L}_{\text{PoM}^2\text{-DIB}}[\boldsymbol{\upsilon}_{\text{p}}, \boldsymbol{f}] := \mathbb{E}_{P_{\boldsymbol{A}}}\big[\mathcal{L}_{\text{M}^2\text{-DIB}}[\boldsymbol{\upsilon}_{\text{p}}, \boldsymbol{f}]\big], \quad (\mathbf{P}_1)$$

in which $\boldsymbol{f} := ((f_k^m)_{m=1}^{m(k)})_{k=1}^{K}$ represents the overall unimodal encoding functions, and $\mathcal{L}_{\text{M}^2\text{-DIB}}[\boldsymbol{\upsilon}_{\text{p}}, \boldsymbol{f}] := \sum_{t=1}^{T} H(Y_t|\boldsymbol{A}_t \circ \boldsymbol{Z}_t) + \beta\langle \boldsymbol{A}_t, ((H(Y_t|Z_{k,t}^m) + I(X_k^m; Z_{k,t}^m))_{m=1}^{m(k)})_{k=1}^{K}\rangle$. In which, $\boldsymbol{Z}_t := ((z_{k,t}^m)_{m=1}^{m(k)})_{k=1}^{K}$, $\circ$ stands for Hadamard product. When we set $\boldsymbol{\upsilon}_{\text{p}}$ to $\boldsymbol{\upsilon}$, i.e., replacing $\boldsymbol{A}$ with $\boldsymbol{a}$, this problem recovers ($\mathbf{P}_0$) with a concise description, ensuring consistency between probabilistic and deterministic treatments.

Note that $P_{\boldsymbol{A}}$ is cooperative. We do a decomposition as

Theorem 1. *The probabilistic $\mathcal{L}_{\text{PoM}^2\text{-DIB}}[\boldsymbol{\upsilon}_{\text{p}}, \boldsymbol{f}]$ can be rewritten into a cooperative form with $P_{\boldsymbol{A}} \in \mathcal{P}_{\boldsymbol{A}}$, i.e.,*

$$\begin{aligned} \mathcal{L}_{\text{PoM}^2\text{-DIB}}[\boldsymbol{\upsilon}_{\text{p}}, \boldsymbol{f}] = \mathbb{E}_{P_U}\Big[\sum_{t=1}^{T} \mathbb{E}_{P_{\boldsymbol{A}_t|U}}\big[H(Y_t|\boldsymbol{A}_t \circ \boldsymbol{Z}_t) \\ + \beta\langle \boldsymbol{A}_t, ((H(Y_t|Z_{k,t}^m) + I(X_k^m; Z_{k,t}^m))_{m=1}^{m(k)})_{k=1}^{K}\rangle\big]\Big], \end{aligned} \tag{14}$$

where $\mathbb{E}_{P_{\boldsymbol{A}_t|U}}[\cdot] = \mathbb{E}_{P_{\hat{\boldsymbol{A}}_t|U}}\Big[\mathbb{E}_{\prod_{k=1}^{K} P_{\boldsymbol{A}_t^{(k)}|A_{k,t}U}}[\cdot]\Big]$.

Proof. For details, please see Appendix A. □

Remark 4. Theorem 1 shows that each receiver t can individually process its task by observing a common randomness U when we apply the CR-based coordination. *All selectors know each other's decisions when this common U is observed.* The decomposition is useful to ensure the computational feasibility of decision-making in the distributed coordination paradigm.

V. Variational Feature Coding and Probability Parameterization

For the second challenge, we introduce variational decoders and a computable bound for $\mathcal{L}_{\text{PoM}^2\text{-DIB}}[\cdot]$. Then, we parameterize all optimizable distributions as neural networks.

A. Variational Decoding

The t-th ideal decoder g_t perceives its respective modality-fused representation $Z_{\mathcal{K}_t}$ (equivalently, the vectorized $\boldsymbol{A}_t \circ \boldsymbol{Z}_t$), to return the best prediction for Y_t based on this representation. This g_t, i.e., $P_{Y_t|\boldsymbol{A}_t \circ \boldsymbol{Z}_t}$, and unimodal versions $\{P_{Y_t|Z_{k,t}^m}\}_{k \in \mathcal{K}}$, can be determined through a Markov chain [14, Appendix A] if encoders $\boldsymbol{f}$ and selectors $\boldsymbol{\upsilon}_{\text{p}}$ are given. But directly calculating such terms via high-dimensional integrals is difficult [8], [14].

We introduce variational distributions, $\{Q_{Y_t|\boldsymbol{A}_t \circ \boldsymbol{Z}_t}\}_{t \in \mathcal{T}}$ and $\{Q_{Y_t|Z_{k,t}^m}\}_{k \in \mathcal{K}, t \in \mathcal{T}}$, to approximate these terms similar to the approach [24]. With variational probabilities, useful bounds of information terms in $\mathcal{L}_{\text{PoM}^2\text{-DIB}}[\cdot]$ can be derived:

Proposition 1. *The following bounds hold given any k and t,*

(i) An upper bound holds for $H(Y_t|Z_{k,t}^m)$:

$$H(Y_t|Z_{k,t}^m) \leq \mathbb{E}_{P_{Z_{k,t}^m}}[H(P_{Y_t|Z_{k,t}^m}, Q_{Y_t|Z_{k,t}^m})], \tag{15}$$

where $H(P_{Y|Z}, Q_{Y|Z}) := -\int p(y|z)\log q(y|z)dy$.

(ii) For given $\boldsymbol{A}_t$, an upper bound holds for $H(Y_t|\boldsymbol{A}_t \circ \boldsymbol{Z}_t)$:

$$H(Y_t|\boldsymbol{A}_t \circ \boldsymbol{Z}_t) \leq \mathbb{E}_{P_{\boldsymbol{Z}_t|\boldsymbol{A}_t}}[H(P_{Y_t|\boldsymbol{A}_t\circ\boldsymbol{Z}_t}, Q_{Y_t|\boldsymbol{A}_t\circ\boldsymbol{Z}_t})]. \tag{16}$$

Proof. For details, please see Appendix A. □

Based on Theorem 1 and Proposition 1, we establish a computable bound $\mathcal{L}_{\text{PoM}^2\text{-DVIB}}[\cdot]$ for $\mathcal{L}_{\text{PoM}^2\text{-DIB}}[\cdot]$ in the following:

$$\begin{aligned}
&\min_{\boldsymbol{\upsilon}_{\text{p}},\boldsymbol{f},\boldsymbol{g},\boldsymbol{g}_{\text{loc}}:P_{\boldsymbol{A}}\in\mathcal{P}_{\boldsymbol{A}}} \mathcal{L}_{\text{PoM}^2\text{-DVIB}}[\boldsymbol{\upsilon}_{\text{p}}, \boldsymbol{f}, \boldsymbol{g}, \boldsymbol{g}_{\text{loc}}] \\
&:= \mathbb{E}_{P_U}\Big[\sum_{t=1}^{T}\mathbb{E}_{P_{\boldsymbol{A}_t|U}}\big[\mathbb{E}_{P_{\boldsymbol{Z}_t|\boldsymbol{A}_t}}[H(P_{Y_t|\boldsymbol{A}_t\circ\boldsymbol{Z}_t}, Q_{Y_t|\boldsymbol{A}_t\circ\boldsymbol{Z}_t}) \\
&\quad + \beta\big\langle \boldsymbol{A}_t, ((\mathbb{E}_{P_{Z_{k,t}^m}}[H(P_{Y_t|Z_{k,t}^m}, Q_{Y_t|Z_{k,t}^m})] \\
&\quad + I(X_k^m; Z_{k,t}^m))_{m=1}^{m(k)})_{k=1}^{K}\big\rangle\big]\Big] \geq \mathcal{L}_{\text{PoM}^2\text{-DIB}}[\boldsymbol{\upsilon}_{\text{p}}, \boldsymbol{f}],
\end{aligned} \tag{$\mathbf{P}_2$}$$

where $\boldsymbol{g} := (g_t)_{t\in\mathcal{T}}$ denotes the global variational decoder and $\boldsymbol{g}_{\text{loc}} := (\boldsymbol{g}_{\text{loc},t})_{t\in\mathcal{T}}$ with any t-th local $\boldsymbol{g}_{\text{loc},t} := ((g_{k,t}^m)_{m=1}^{m(k)})_{k=1}^{K}$ is the overall variational unimodal decoder. Every defined $g_{k,t}^m$ induces the unimodal variational $Q_{Y_t|Z_{k,t}^m}$. In the remainder of this paper, we use $\boldsymbol{g}$ to represent the variational multi-modal decoder by default.

B. Probability Parameterization

Parameterization techniques enable optimizing tractable parameters in a typical finite-dimensional $\mathbb{R}$-linear space, rather than processing the function itself, which is often difficult. In this part, we parameterize selectors $\boldsymbol{\upsilon}_{\text{p}}$, encoders $\boldsymbol{f}$, decoders $\boldsymbol{g}$ and local decoders $\boldsymbol{g}_{\text{loc}}$ as $\boldsymbol{\upsilon}_{\text{p};\boldsymbol{\theta}}$, $\boldsymbol{f}_{\boldsymbol{\psi}}$, $\boldsymbol{g}_{\boldsymbol{\phi}}$ and $\boldsymbol{g}_{\text{loc};\boldsymbol{\varphi}}$, respectively, by deep neural networks (DNNs). Let $NN(\cdot)$ be a DNN. Then, we detail the implementation of selection and coding modules.

1) Selection: All the selectors must cooperatively generate an $\boldsymbol{A}$ from the $\boldsymbol{\upsilon}_{\text{p}}$-induced $P_{\boldsymbol{\theta};\boldsymbol{A}|U}$. We now detail the structure of this $P_{\boldsymbol{\theta};\boldsymbol{A}|U}$ following Definition 1. $\boldsymbol{\theta} := (\theta_k)_{k=1}^{K} \oplus (\theta_t)_{t=1}^{T}$. Each k-th transmitter-side selector is presented by υ_{θ_k}, which induces $\{P_{\theta_k;\boldsymbol{A}_t^{(k)}|A_{k,t}U}\}_{t=1}^{T}$. Each t-th receiver-side selector is υ_{θ_t}, which induces $P_{\theta_t;\hat{\boldsymbol{A}}_t|U}$. In which, the CR U is priorly defined. It is noticed that all such parameterized distributions for selections are essentially discrete.

Take P_{θ_t} as an example. When we directly implement p_{θ_t}, i.e., the conditional probability mass of P_{θ_t}, by $\sigma(NN(\cdot;\theta_t))$, where σ denotes softmax, the dimension of this DNN's output is exponentially large according to the number of all possible $\hat{\boldsymbol{a}}_t$. This holds if for every possible realization $\hat{\boldsymbol{a}}_t$, we use an output's dimension to store its mass, i.e., we leverage a simple categorical distribution to represent P_{θ_t}. Realizing p_{θ_t} by this approach requires an output dimension on the order of $\mathcal{O}(2^K)$. This high space complexity causes this simple DNN modeling method to fail at large-scale device communication.

To fully tackle this issue in selection parameterization, we apply a type of structural decomposition. We carefully design a DNN for p_{θ_t} in the following process. We implement a DNN $NN(\cdot;\theta_t)$ whose input is $u \sim U$ and whose output is a non-normalized $\mathbb{R}$-vector $\pi_{\hat{\boldsymbol{A}}_t}$, which is $E_t + K$-dimensional. This $\pi_{\hat{\boldsymbol{A}}_t}$ can be divided into two vectors: the first E_t-dimensional part $\pi_{\hat{\boldsymbol{A}}_t,\text{num}}$ and the second K-dimensional part $\pi_{\hat{\boldsymbol{A}}_t,\text{prob}}$. The number of selected transmitters $\|\hat{\boldsymbol{A}}_t\|$ is decided by the mass $\sigma(\pi_{\hat{\boldsymbol{A}}_t,\text{num}})$. After given $\|\hat{\boldsymbol{A}}_t\|$, we can generate $\hat{\boldsymbol{A}}_t$ following a probability corresponding to a joint distribution that can yield a $\|\hat{\boldsymbol{A}}_t\|$-times without-replacement sampling from $\sigma(\pi_{\hat{\boldsymbol{A}}_t,\text{prob}})$. We write this joint distribution as $\sigma^{\times\|\hat{\boldsymbol{A}}_t\|}(\pi_{\hat{\boldsymbol{A}}_t,\text{prob}})$. For any $\hat{\boldsymbol{a}}_t \sim \hat{\boldsymbol{A}}_t$, its mass is exactly a product of $\sigma(\pi_{\hat{\boldsymbol{A}}_t,\text{num}})(\hat{\boldsymbol{a}}_t)$ and $\sigma^{\times\|\hat{\boldsymbol{A}}_t\|}(\pi_{\hat{\boldsymbol{A}}_t,\text{prob}})(\hat{\boldsymbol{a}}_t)$ according to the product rule.

This construction can be analogized and understood as that we construct a "point process" by using a DNN. The number of points (transmitters) is first determined, then their positions (indices) are decided. P_{θ_t} can be regarded as a joint distribution of all points. This means that we introduce a structural random prior to P_{θ_t} to reduce the space complexity $\mathcal{O}(2^K)$ to $\mathcal{O}(K)$, where an exponential compression is achieved.

In the preceding, P_{θ_t} is used to show how to use DNN to establish each t-th receiver-side selector. For the transmitter-side, we also perform a similar treatment with a $E_k + m(k)$-dimensional vector. We omit this redundant part.

2) Coding: Encoder: $\boldsymbol{\psi} := ((\psi_k^m)_{m=1}^{m(k)})_{k=1}^{K}$. Each m-th unimodal encoder at its corresponding transmitter k is denoted by $f_{\psi^m}(\cdot, t)$, which induces $P_{\psi_k^m;Z_{k,t}^m|X_k^m}$ for a task t. Following [8], [14], we define the encoder to be a multivariate Gaussian encoder. Specifically, we use a DNN $NN(\cdot,\cdot;\psi_k^m)$ to generate $\mathcal{N}(z_{k,t}^m|\mu_{k,t}^m, \Sigma_{k,t}^m)$, where the DNN's input is (x_k^m, t) and the output is $(\mu_{k,t}^m, \Sigma_{k,t}^m)$. μ is the mean and Σ is the covariance, which is diagonal. We also apply the reparameterization trick to draw $z_{k,t}^m$, i.e., $z_{k,t}^m = \mu_{k,t}^m + \sqrt{\Sigma_{k,t}^m\mathbf{1}}\circ\epsilon$ with $\epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, to allow gradient backpropagation [24]. Decoder: $\boldsymbol{\phi} := (\phi_t)_{t=1}^{T}$. Each t-th decoder is denoted by g_{ϕ_t}, inducing the variational $Q_{\phi_t;Y_t|\boldsymbol{A}_t\circ\boldsymbol{Z}_t}$. $\boldsymbol{\varphi} := (\varphi_t)_{t=1}^{T}$. Each m-th unimodal decoder at the receiver t is $g_{k,t;\varphi_t}^m$, which induces $Q_{\varphi_t;Y_t|Z_{k,t}^m}$. This $g_{k,t;\varphi_t}^m$ can be also described as $g_{\varphi_t}(\cdot;k,m)$. Note that any unimodal decoder at t shares a common φ_t for efficient computation.

VI. Joint Optimization of Selection and Coding

In this section, we first establish $\mathcal{L}_{\text{PoM}^2\text{-DVIB}}[\cdot]$, the empirical objective for the preceding parameterized coding and selection modules with finite samples. Then, we minimize $\hat{\mathcal{L}}_{\text{PoM}^2\text{-DVIB}}(\cdot)$ with gradient-based methods.

A. Empirical Objective and Gradient-based Optimization

For any task $t \in \mathcal{T}$, all available labeled data are depicted as $\mathcal{D}_t := \{(\boldsymbol{x}_{1,t}^i, ..., \boldsymbol{x}_{K,t}^i, y_t^i)\}_{i=1}^{N}$, where all samples are iid and $\boldsymbol{x}_{k,t}^i := (x_{k,t}^{m,i})_{m\in[m(k)]}$. Any raw data $(\boldsymbol{x}_{1,t}^i, ..., \boldsymbol{x}_{K,t}^i)$ of $\mathcal{D}_t$ can differ from any $(\boldsymbol{x}_{1,t'}^i, ..., \boldsymbol{x}_{K,t'}^i)$ of $\mathcal{D}_{t'}$ when $t \neq t'$, but they do follow a same joint distribution $P_{\boldsymbol{X}_1\ldots\boldsymbol{X}_K}$. In which, we add a subscript t at each $\boldsymbol{x}_{k,t}^i$ to emphasize that the sample belongs to the dataset specific to a task t. For simplicity, we assume $|\mathcal{D}_t| = N$ for any $t \in \mathcal{T}$. Then, $\mathcal{L}_{\text{PoM}^2\text{-DVIB}}[\cdot]$ can be estimated by these labeled data $\mathcal{D}_1, ..., \mathcal{D}_t$ empirically.

We estimate three key terms appeared in $\mathcal{L}_{\text{PoM}^2\text{-DVIB}}[\cdot]$ as follows. At any task t, for any $k \in \mathcal{K}_t$, with $1 \leq i, j \leq N$, we directly estimate the MI term $I(X_k^m; Z_{k,t}^m)$:

$$I(X_k^m; Z_{k,t}^m) = \mathbb{E}_{P_{X_k^m Z_{k,t}^m}}\left[\log \frac{p_{\psi_k^m}(z_{k,t}^m | x_k^m)}{\mathbb{E}_{P_{X_k}}[p_{\psi_k^m}(z_{k,t}^m | x_k^m)]}\right] \approx \frac{1}{N}\sum_{i=1}^{N} \log \frac{p_{\psi_k^m}(z_{k,t}^{m,i} | x_{k,t}^{m,i})}{\frac{1}{N}\sum_{j=1}^{N} p_{\psi_k^m}(z_{k,t}^{m,i} | x_{k,t}^{m,j})}. \tag{17}$$

We also estimate

$$\mathbb{E}_{P_{Z_{k,t}^m}}[H(P_{Y_t|Z_{k,t}^m}, Q_{Y_t|Z_{k,t}^m})] \approx -\frac{1}{N}\sum_{i=1}^{N} \log q_{\varphi_t}(y_t^i | z_{k,t}^{m,i}; k, m). \tag{18}$$

Given the t-th $\boldsymbol{A}_t$, we have

$$\mathbb{E}_{P_{\boldsymbol{Z}_t|\boldsymbol{A}_t}}[H(P_{Y_t|\boldsymbol{A}_t \circ \boldsymbol{Z}_t}, Q_{Y_t|\boldsymbol{A}_t \circ \boldsymbol{Z}_t})] \approx -\frac{1}{N}\sum_{i=1}^{N} \log q_{\phi_t}(y_t^i | \boldsymbol{A}_t \circ \boldsymbol{z}_t^i). \tag{19}$$

Since samples in $\mathcal{D}_t$ are iid, our empirical approximations, i.e., (17)-(19), are asymptotically unbiased. In particular, *the crucial MI term $I(X_k^m; Z_{k,t}^m)$ is successfully estimated without variational or contrastive log-ratio bounds [25] since the parameterized conditional $p(z|x)$ is explicitly given in this case.* We can approximate the marginal $p(z) = \mathbb{E}_{x \sim p(x)}[p(z|x)] \approx \frac{1}{N}\sum_{i=1}^{N} p(z|x^i)$ with $\mathcal{D}_t$.

Combine (17)-(19) to derive the empirical version of ($\mathbf{P}_2$),

$$\begin{aligned} &\min_{\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\phi}, \boldsymbol{\varphi}: P_{\boldsymbol{A}} \in \mathcal{P}_{\boldsymbol{A}}} \hat{\mathcal{L}}_{\text{PoM}^2\text{-DVIB}}(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\phi}, \boldsymbol{\varphi}) \\ &:= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{P_U}\left[\sum_{t=1}^{T} \mathbb{E}_{P_{\theta_t, (\theta_k)_k; \boldsymbol{A}_t|U}}\left[\left\{\hat{\mathcal{L}}_t^i(\boldsymbol{A}_t) := \right.\right.\right. \\ &- \log q_{\phi_t}(y_t^i | \boldsymbol{A}_t \circ \boldsymbol{z}_t^i) + \beta \left\langle \boldsymbol{A}_t, \left(- \log q_{\varphi_t}(y_t^i | z_{k,t}^{m,i}; k, m) \right.\right. \\ &\left.\left.\left.\left.\left. + \log \frac{p_{\psi_k^m}(z_{k,t}^{m,i} | x_{k,t}^{m,i})}{\frac{1}{N}\sum_{j=1}^{N} p_{\psi_k^m}(z_{k,t}^{m,i} | x_{k,t}^{m,j})} \right)_{k=1}^{K} \right\rangle \right\}\right]\right], \end{aligned} \tag{$\mathbf{P}_3$}$$

where $\mathbb{E}_{P_{\theta_t, (\theta_k)_k; \boldsymbol{A}_t|U}}[\cdot] = \mathbb{E}_{P_{\theta_t; \hat{\boldsymbol{A}}_t|U}}\left[\mathbb{E}_{\prod_{k=1}^{K} P_{\theta_k; \boldsymbol{A}_t^{(k)}|A_{k,t}U}}[\cdot]\right]$. By tackling this empirical problem ($\mathbf{P}_3$), we can approximate the solution of the primal ($\mathbf{P}_0$).

It is challenging to directly backpropagate the gradients of all parameters in addressing ($\mathbf{P}_3$) using conventional gradient descent methods, even when reparameterization techniques are used to design encoders. This difficulty arises because of the peculiarities of $\{P_{\theta_t, (\theta_k)_k; \boldsymbol{A}_t|U}\}_t$, i.e., they are discrete and can not directly achieve a differentiable propagation of gradients due to sampling. We elaborate on the update of parameters in the following.

1) The Update of $\boldsymbol{\theta}$: Note that $\hat{\mathcal{L}}_{\text{PoM}^2\text{-DVIB}}(\cdot)$ is essentially a function of $\boldsymbol{\theta}$. Ignoring parameters $\boldsymbol{\psi}, \boldsymbol{\phi}, \boldsymbol{\varphi}$, we have

$$\hat{\mathcal{L}}_{\text{PoM}^2\text{-DVIB}}(\boldsymbol{\theta}) = \mathbb{E}_{P_U}\Big[\sum_{t=1}^{T} \mathbb{E}_{P_{\theta_t, (\theta_k)_k; \boldsymbol{A}_t|U}}\big[\hat{\mathcal{L}}_t(\boldsymbol{A}_t)\big]\Big], \tag{20}$$

where $\hat{\mathcal{L}}_t(\boldsymbol{A}_t) := \frac{1}{N}\sum_{i=1}^{N} \hat{\mathcal{L}}_t^i(\boldsymbol{A}_t)$. Then, for any component θ_t, we can transform its gradient into a score function, namely the gradient of a log-likelihood, $\nabla_{\theta_t} \log p_{\theta_t, (\theta_k)_k}(\boldsymbol{A}_t|U)$, i.e.,

$$\begin{aligned} \nabla_{\theta_t} \hat{\mathcal{L}}_{\text{PoM}^2\text{-DVIB}}(\boldsymbol{\theta}) &= \mathbb{E}_{P_U}\Big[\sum_{\boldsymbol{a}_t} \nabla_{\theta_t} p_{\theta_t, (\theta_k)_k}(\boldsymbol{a}_t|u) \hat{\mathcal{L}}_t(\boldsymbol{a}_t)\Big] \\ &= \mathbb{E}_{P_U}\Big[\sum_{\boldsymbol{a}_t} p_{\theta_t, (\theta_k)_k}(\boldsymbol{a}_t|u) \nabla_{\theta_t} \log p_{\theta_t, (\theta_k)_k}(\boldsymbol{a}_t|u) \hat{\mathcal{L}}_t(\boldsymbol{a}_t)\Big] \\ &= \mathbb{E}_{P_U}\Big[\mathbb{E}_{P_{\theta_t, (\theta_k)_k; \boldsymbol{A}_t|U}}\big[\nabla_{\theta_t} \log p_{\theta_t, (\theta_k)_k}(\boldsymbol{a}_t|u) \hat{\mathcal{L}}_t(\boldsymbol{a}_t)\big]\Big] \\ &\approx \underbrace{\nabla_{\theta_t} \frac{1}{N}\sum_{i=1}^{N} \log p_{\theta_t, (\theta_k)_k}(\boldsymbol{a}_t^i | u^i) \hat{\mathcal{L}}_t^i(\boldsymbol{a}_t^i)}_{\text{empirical estimation of log-likelihood}}, \end{aligned} \tag{21}$$

where the last approximation follows the Gibbs sampling, i.e., for every sample i, we assign realizations of U and $\boldsymbol{A}$ as u^i and $\boldsymbol{a}^i$. Estimation (21) is unbiased and is also called *policy gradient* following classical policy optimization [26]. We can also estimate $\nabla_{\theta_k} \hat{\mathcal{L}}_{\text{PoM}^2\text{-DVIB}}(\boldsymbol{\theta})$ for any k in the same way.

2) The Update of Other Parameters: Unlike the above complex transformations required to obtain computable gradients of $\boldsymbol{\theta}$, the reparameterization trick [24] introduced in Section V-B enables the gradients of parameters $\boldsymbol{\psi}, \boldsymbol{\phi}, \boldsymbol{\varphi}$ being directly calculated via backpropagation [27].

B. Training and Inference Procedure

The *training procedure* corresponding to PoM^2-DIB is summarized in Algorithm 1.

In Step 4, the synchronizing signal is adopted to derive a CR sample ω. Then, through an identical auxiliary variable U, a common u is obtained at each receiver.

In Step 6, every receiver in parallel selects its cooperative transmitters based on $u = U(\omega)$ with $\mathcal{P}_{\boldsymbol{A}}$ satisfying (9) and (10). General cases have been discussed in the above. Directly replacing (20) with (22) makes Algorithm 1 compatible with general cases.

In Step 9, each selected transmitter in parallel decides data modalities provided for the receiver based on u.

In Step 11, each selected modality m corresponding to t samples $x_{k,t}^m$ from the dataset $\mathcal{D}_{k,t}^m$.

In Step 16, unimodal local decoders of task t are instantiated as a single DNN, i.e., $q_{\varphi_t}(\cdot|\cdot; \cdot, \cdot)$, deployed at receiver t.

In Steps 17-19, a mini-batch of data samples can be generated by repeating Steps 4-17 B times. Then, $\hat{\mathcal{L}}_{\text{PoM}^2\text{-DVIB}}$ is calculated by these B samples instead of full N samples.

In Step 20, parameters of selectors, encoders, and decoders are updated. The backpropagation begins at each receiver and executes in parallel.

The *inference procedure* is summarized in Algorithm 2.

In Steps 2, u is sampled from U only once. In this case, a fixed selection characterized by $\boldsymbol{a}$ can be derived via Step 3. This operation maintains a stable communication link between two parties (receivers and their selected transmitters). It does not hurt the optimality if the converged $P_{\boldsymbol{\theta}; \boldsymbol{A}|U}$ is optimal. For more details, see the proof of Theorem 2.

Algorithm 2 does not involve local decoders, which are only used as regularization in Algorithm 1 to support the derivations of optimized encoders, decoders, and selectors [8].

VII. Further Analysis and Discussion

A. The Feasibility of Policy Update under Hard Constraints

In the preceding, we usually assume the variational distribution of selectors being in $\mathcal{P}_{\boldsymbol{A}}$ by default. But the parameterized

Algorithm 1: Training Procedure of PoM^2-DIB

```
Input: datasets $\{\mathcal{D}_t\}_{t\in\mathcal{T}}$, batch size $B$.
Output: optimized parameters $\boldsymbol{\theta}$, $\boldsymbol{\psi}$, $\boldsymbol{\phi}$, and $\boldsymbol{\varphi}$.
initialize $\boldsymbol{\upsilon}_{\mathrm{p};\boldsymbol{\theta}}, \boldsymbol{f}_{\boldsymbol{\psi}}, \boldsymbol{g}_{\boldsymbol{\phi}}, \boldsymbol{g}_{\mathrm{loc};\boldsymbol{\varphi}}$;
synchronize all participating devices;
while not converged do
    yield CR $u \sim P_U$ from synchronizing signals;
    for each receiver $t \in \mathcal{T}$ in parallel do
        randomly select $\hat{\boldsymbol{a}}_t \sim P_{\theta_t;\hat{\boldsymbol{A}}_t|U=u}$;
        request low-rate representations from
          transmitters characterized by $\hat{\boldsymbol{a}}_t$;
        for each transmitter $k|_{a_{k,t}=1}$ in parallel do
            randomly select $\boldsymbol{a}_t^{(k)} \sim P_{\theta_k;\boldsymbol{A}_t^{(k)}|A_{k,t}U}$;
            for each modality $m|_{a_{k,t}^m=1}$ in parallel do
                synchronously sample $x_{k,t}^m$ from
                  $\mathcal{D}_{k,t}^m := \{x_{k,t}^{m,i}\}_{i=1}^N$ of $\mathcal{D}_t$;
                encode $z_{k,t}^m$ given $x_{k,t}^m$ as
                  $z_{k,t}^m = f_{\psi_k^m}(x_{k,t}^m, t) = \mu_{k,t}^m + \sqrt{\Sigma_{k,t}^m \mathbf{1}} \circ \epsilon$,
                  where $\epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$;
            send all unimodal features $z_{k,t}^m|_{a_{k,t}^m=1}$ to the
              receiver $t$;
        receive all features as $\boldsymbol{a}_t \circ \boldsymbol{z}_t$;
        infer the recover $\hat{y}_t$ of the true $y_t$ from
          $\hat{y}_t \sim q_{\phi_t}(y_t|\boldsymbol{a}_t \circ \boldsymbol{z}_t)$;
        each $m|_{a_{k,t}^m=1}$-th unimodal decoder infers the
          local recover $\hat{y}_{k,t}^m$ of the true $y_t$ from
          $\hat{y}_{k,t}^m \sim q_{\varphi_t}(y_t|z_{k,t}^m; k, m)$;
    collect all above data as a sample;
    if $B$-batch samples are collected then
        compute $\hat{\mathcal{L}}_{\mathrm{PoM^2\text{-}DVIB}}$ with this mini-batch;
        update parameters $\boldsymbol{\theta}$, $\boldsymbol{\psi}$, $\boldsymbol{\phi}$, and $\boldsymbol{\varphi}$ through
          techniques introduced in Section VI-A;
```

$P_{\boldsymbol{\theta};\boldsymbol{A}} \notin \mathcal{P}_{\boldsymbol{A}}$ usually holds, since DNN-based distributions are randomly initialized. In this general case, supp$(\boldsymbol{A})$ always fails to meet constraints (9) and (10) in training, e.g, it is possible for any transmitter k to have more than E_k times task inference requests through its parameterized transmitter-side selector. To tackle this issue, we add a random choice at transmitter-side that if the transmitter k receives $E_k' > E_k$ times requests from receivers, then it uniformly samples E_k requests from these E_k' ones to execute. After doing so, to avoid a singular case where some receivers are unable to receive the unimodal feature from any of their selected transmitters, we provide a penalty term in the loss function. For more engineering details, please refer to our online code.

For the cases, we now define an intermediate variable $\boldsymbol{A}' \sim P_{\boldsymbol{\theta};\boldsymbol{A}'}$, which is a copy of the primal $\boldsymbol{A} \sim P_{\boldsymbol{\theta};\boldsymbol{A}} \notin \mathcal{P}_{\boldsymbol{A}}$ with $\boldsymbol{A}' = (\boldsymbol{A}_t')_{t=1}^T$. The integration of the above two mechanisms, random choice and penalty, can generate a stochastic mapping with a kernel $P_{\boldsymbol{A}|\boldsymbol{A}'}$, which is a naturally non-parameterizable conditional without exact forms, where the output $\boldsymbol{A}$ is regular that satisfies (9) and (10). This means that we construct a new $\widetilde{P}_{\boldsymbol{\theta};\boldsymbol{A}} = \mathbb{E}_{P_{\boldsymbol{\theta};\boldsymbol{A}'}}[P_{\boldsymbol{A}|\boldsymbol{A}'}] \in \mathcal{P}_{\boldsymbol{A}}$ from the two mechanisms.

This result demonstrates that it is feasible to update a parameterized distribution without explicitly involving (9) and (10), since all the constraints can be externalized into an additional probability kernel. We are always able to replace $P_{\boldsymbol{\theta};\boldsymbol{A}}$ with its regular version $\widetilde{P}_{\boldsymbol{\theta};\boldsymbol{A}} = \mathbb{E}_{P_U}[\prod_{t=1}^T \mathbb{E}_{P_{\theta_t,(\theta_k)_k;\boldsymbol{A}_t'|U}}[P_{\boldsymbol{A}_t|\boldsymbol{A}_t'}]]$ where $U \leftrightarrow \boldsymbol{A}_t' \leftrightarrow \boldsymbol{A}_t$ for any $t \in \mathcal{T}$.

In this case, we make a minor modification to the update of $\boldsymbol{\theta}$, i.e., we rewrite (20) as

$$\hat{\mathcal{L}}_{\mathrm{PoM^2\text{-}DVIB}}(\boldsymbol{\theta}) = \mathbb{E}_{P_U}[\textstyle\sum_{t=1}^T \mathbb{E}_{P_{\boldsymbol{A}_t|\boldsymbol{A}_t'}}[\mathbb{E}_{P_{\theta_t,(\theta_k)_k;\boldsymbol{A}_t'|U}}[\hat{\mathcal{L}}_t(\boldsymbol{A}_t)]]]. \tag{22}$$

The unknown term $\mathbb{E}_{P_{\boldsymbol{A}_t|\boldsymbol{A}_t'}}[\cdot]$ can be estimated with empirical approximation via Gibbs sampling. This is also the reason why we use policy gradient, which works well when the workflow involves an unknown and non-parameterizable $P_{\boldsymbol{A}|\boldsymbol{A}'}$, where the cases, nevertheless, cannot be tackled by traditional reparameterization, just as we do for other parameters.

B. The Primal ($\mathbf{P}_0$): Optimal Rate-Relevance Tradeoff

Suppose a Markov chain hold, i.e., for any $k, k' \in \mathcal{K}$ and $t \in \mathcal{T}$, $X_k^m \leftrightarrow Y_t \leftrightarrow X_{k'}^{m'}$. Then, we can claim that for any task t and its corresponding modalities across transmitters, the encoding functions of the primal ($\mathbf{P}_0$) achieve the lowest sum communication cost at a given level β of relevance between the true Y_t and the recovered $\hat{Y}_t$ under a logarithmic loss. This is proved from the property of DIB [8, Proposition 2].

However, this property may not always hold [8, Remark 4], especially in the considered multi-modal multi-task scenario, where each raw unimodal data may rely on multiple tasks. It is exactly because the optimum under strict conditions is not usually achieved that we introduce a new degree of freedom, i.e., the task-aware selection of modality to control communication links to further remove data redundancy among modalities to reduce communication costs. This elaborate control allows a more flexible feature transmission than DIB with a lower rate. We will empirically verify this statement in the next section.

C. Relations among Objectives

We summarize the relations among optimization objectives of ($\mathbf{P}_0$)-($\mathbf{P}_3$) in the following. This guarantees our probabilistic relaxation converges to the primal deterministic optimum.

Theorem 2. *For objectives $\mathcal{L}_{\mathrm{M^2\text{-}DIB}}$, $\mathcal{L}_{\mathrm{PoM^2\text{-}DIB}}$, $\mathcal{L}_{\mathrm{PoM^2\text{-}DVIB}}$, and $\hat{\mathcal{L}}_{\mathrm{PoM^2\text{-}DVIB}}$ with a unified $\beta \geq 0$ at any feasible solution of $\hat{\mathcal{L}}_{\mathrm{PoM^2\text{-}DVIB}}$, i.e., $\boldsymbol{\upsilon}_{\mathrm{p};\boldsymbol{\theta}}, \boldsymbol{f}_{\boldsymbol{\psi}}, \boldsymbol{g}_{\boldsymbol{\phi}}, \boldsymbol{g}_{\mathrm{loc};\boldsymbol{\varphi}}$, the following holds*

$$\begin{aligned}\lim_{N\to\infty} \hat{\mathcal{L}}_{\mathrm{PoM^2\text{-}DVIB}}(\boldsymbol{\theta}, \psi, \phi, \boldsymbol{\varphi}) &\overset{\text{a.e.}}{\to} \mathcal{L}_{\mathrm{PoM^2\text{-}DVIB}}[\boldsymbol{\upsilon}_{\mathrm{p};\boldsymbol{\theta}}, \boldsymbol{f}_{\boldsymbol{\psi}}, \boldsymbol{g}_{\boldsymbol{\phi}}, \boldsymbol{g}_{\mathrm{loc};\boldsymbol{\varphi}}] \\ &\geq \mathcal{L}_{\mathrm{PoM^2\text{-}DIB}}[\boldsymbol{\upsilon}_{\mathrm{p};\boldsymbol{\theta}}, \boldsymbol{f}_{\boldsymbol{\psi}}].\end{aligned} \tag{23}$$

Besides, their minimums satisfy

$$\begin{aligned}\min_{\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\phi},\boldsymbol{\varphi}} \lim_{N\to\infty} \hat{\mathcal{L}}_{\mathrm{PoM^2\text{-}DVIB}} &\overset{\text{a.e.}}{\to} \min_{\boldsymbol{\upsilon}_{\mathrm{p};\boldsymbol{\theta}}, \boldsymbol{f}_{\boldsymbol{\psi}}, \boldsymbol{g}_{\boldsymbol{\phi}}, \boldsymbol{g}_{\mathrm{loc};\boldsymbol{\varphi}}} \mathcal{L}_{\mathrm{PoM^2\text{-}DVIB}} \\ &\geq \min_{\boldsymbol{\upsilon}_{\mathrm{p}}, \boldsymbol{f}} \mathcal{L}_{\mathrm{PoM^2\text{-}DIB}} = \min_{\boldsymbol{\upsilon}, \boldsymbol{f}} \mathcal{L}_{\mathrm{M^2\text{-}DIB}}.\end{aligned} \tag{24}$$

Proof. For details, please see Appendix B. □

Remark 5. This theorem ensures that our transformation and approximation minimally hurt the achievement of the minimizer of the primal ($\mathbf{P}_0$), while enhancing its tractability. The crucial last equality in (24) strongly supports this point.

Algorithm 2: Inference Procedure of PoM^2-DIB

```
Input: trained encoders, decoders, and selectors.
Output: predictions $\{\hat{y}_t\}_{t\in\mathcal{T}}$.
synchronize all participating devices;
yield CR $u \sim P_U$ from synchronizing signals;
decide $\boldsymbol{a}$ from $P_{\boldsymbol{\theta};\boldsymbol{A}|U=u}$ induced by all selectors.
while true do
    for each receiver $t \in \mathcal{T}$ in parallel do
        request low-dimensional representations from
          transmitters characterized by $\hat{\boldsymbol{a}}_t$;
        for each transmitter $k|_{a_{k,t}=1}$ in parallel do
            locally observe its raw data $\boldsymbol{x}_k$ with
              synchronization;
            for each modality $m|_{a^m_{k,t}=1}$ in parallel do
                encode $z^m_{k,t}$ given $x^m_k$ of $\boldsymbol{x}_k$;
            send all unimodal features $z^m_{k,t}|_{a^m_{k,t}=1}$ to the
              receiver $t$;
        receive all features as $\boldsymbol{a}_t \circ \boldsymbol{z}_t$;
        infer the recover $\hat{y}_t$ of the true $y_t$ from
          $\hat{y}_t \sim q_{\phi_t}(y_t|\boldsymbol{a}_t \circ \boldsymbol{z}_t)$;
```

D. Selection for A Flexible Rate-Relevance Tradeoff

We investigate the property for optimal selection in this part. We indicate that there exists optimal selections at the performance limits of all receivers and transmitters, i.e., $|\mathcal{K}_t| = E_t$ for all receivers $t \in \mathcal{T}$ and $|\mathcal{T}_k| = E_k$ for all transmitters $k \in \mathcal{K}$ in a degenerate case that $\beta \to 0$. Formally,

Proposition 2. *There always exists a selection $\boldsymbol{v}^\star$ that is optimal corresponding to ($\boldsymbol{P}_0$) with $|\mathcal{K}^\star_t| = E_t$ for any $t \in \mathcal{T}$ and $|\mathcal{T}_k| = E_k$ for any $k \in \mathcal{K}$ at $\beta \to 0$.*

Proof. For details, please see Appendix C. □

Corollary 1. *There always exists a selection policy $\boldsymbol{v}^\star_{\rm p}$ that is optimal corresponding to ($\boldsymbol{P}_1$) with $\|\hat{\boldsymbol{a}}^\star_t\| = E_t$ for any $t \in \mathcal{T}$ and $\|\boldsymbol{a}^{(k)}_{\mathcal{T}}\| = E_k$ for any $k \in \mathcal{K}$ at $\beta \to 0$.*

Proof. It directly follows from (24) and Proposition 2. □

Remark 6. In the case $\beta \to 0$, communication costs essentially captured by MI are ignored. That is, evaluating relevance-rate tradeoff from ($\mathbf{P}_0$) degenerates to evaluating task relevance between transmitters' modalities and receivers only. This results that the more modalities are provided for a receiver, the better the performance evaluated by ($\mathbf{P}_0$), because communication is free. *In the general case $\beta > 0$, Proposition 2 and Corollary 1 do not hold.* It indicates that the optimal selection does not necessarily exist on the boundary; it may also be located within the interior since the non-degenerate tradeoff works. This also illustrates that PoM^2-DIB is more flexible than DIB since the selection participates in the tradeoff. An interesting extreme case is that even if we relax constraints, i.e., communication links and computation capabilities, E_k and E_t, to infinity, the optimal participation may not be the full participation either. We show the existence of this case in Fig. 5-6, Appendix D.

Remark 7. If hard constraints (9) and (10) are explicitly given, we can find $\boldsymbol{v}^\star_{\rm p}$ through our carefully constructed PoM^2-DIB. But sometimes in practical design, we only have a "tendency" that, given inference quality, we hope that all devices have as few connections as possible with others. In this case, we can leverage the general multiplier rule [28] to transform (9) and (10) as additional penalties to the optimization objective, i.e., $\sum_k \|\boldsymbol{a}^{(k)}_{\mathcal{T}}\| + \sum_t \|\hat{\boldsymbol{a}}_t\|$, with a multiplier γ to form a three-way tradeoff among rate, relevance, and communication link. For details, please see Fig. 7-8, Appendix D.

VIII. Performance Evaluation

A. Experimental Setup

1) Dataset: Two multi-modal multi-task datasets are used to evaluate our PoM^2-DIB: the simple *HandWritten*, a.k.a. *AV-MNIST* dataset [29], and the recent *MM-Fi* dataset [30].

HandWritten Dataset: This dataset is established by aligning the MNIST handwritten digit images with the FSDD spoken digit audio samples. MNIST contains 70000 28×28 gray-scale images of digits (0-9). FSDD contains 3000 audio recordings of spoken digits (0-9) from different speakers, sampled at 8000 Hz. We pair each MNIST image with an FSDD audio sample of the same digit label in (0-9). Due to the data size mismatch between FSDD and MNIST, we introduce data augmentation (e.g., Gaussian noise, random gain, time/frequency masking) to the audio to ensure a one-to-one correspondence. Each entry in this dataset consists of a normalized MNIST image and a Mel-spectrogram representation of an augmented FSDD audio sample with the identical label.

MM-Fi Dataset: It is a public multi-modal non-intrusive 4D human dataset with 27 daily or rehabilitation action categories, scales up to 152986 entries, and synchronizes 5 non-intrusive sensing modalities including RGB-D frames, point cloud from mmWave Radar and LiDAR, and WiFi channel state information (CSI) data.

2) Modality & Task: For HandWritten, we develop a SemCom system involving 3 transmitters and 3 receivers. Every transmitter has 3 modalities while each receiver has its specific task. We customize 3 types of modalities for experiments on HandWritten. Type-A is the magnitude of the 2D Fast Fourier Transform (FFT) of the cropped image, where the primal gray-scale image is cropped to a 14×14 patch of its center. Type-B is the Mel-spectrogram representation of the audio with 16 Mel bands, using an FFT window of 2048 and a hop length of 256. Type-C is the vectorized random noise. The 1-st transmitter's observed data is $\boldsymbol{X}_1 := (X^C, X^C, X^A)$, the 2-nd one's data is $\boldsymbol{X}_2 := (X^C, X^A, X^B)$, and the 3-rd $\boldsymbol{X}_3 := (X^A, X^B, X^C)$. We write these vectors into a matrix form, i.e.,

$$\begin{bmatrix} \boldsymbol{X}_3 \\ \boldsymbol{X}_2 \\ \boldsymbol{X}_1 \end{bmatrix} = \begin{bmatrix} X^A, X^B, X^C \\ X^C, X^A, X^B \\ X^C, X^C, X^A \end{bmatrix} \tag{25}$$

where we artificially add redundancy and repetitive modalities to control and track the behavior of the optimal selection. Also, we customize 3 types of tasks for different receivers. The first receiver needs to distinguish the parity of handwritten digits, which is a classification task with 2-classes. The second needs to identify whether digits have a ring structure (0, 4, 6, 8, 9 or not) with 6-classes. The third needs to accomplish the typical 0-9 recognition with 10-classes.

For MM-Fi, we also develop a SemCom system involving 4 transmitters and 3 receivers. Every transmitter has 4 modalities while each receiver has its specific task. We customize 8 types of modalities (A-H), where A-G correspond respectively to RGB, view-1 of infra-red, view-2 of infra-red, depth, LiDAR, mmWave, and WiFi-CSI, all of which are inherent entries of the dataset. In addition, H is the vectorized random noise. We have a similar matrix as (25) to represent local observations,

$$\begin{bmatrix} \boldsymbol{X}_4 \\ \boldsymbol{X}_3 \\ \boldsymbol{X}_2 \\ \boldsymbol{X}_1 \end{bmatrix} = \begin{bmatrix} X^B, X^C, X^F, X^H \\ X^G, X^E, X^F, X^H \\ X^D, X^E, X^F, X^H \\ X^A, X^B, X^C, X^H \end{bmatrix} \tag{26}$$

We naturally have 3 types of tasks from MM-Fi's data annotation. The first receiver needs to predict 3D human pose key points with a form in $\mathbb{R}^{17\times 3}$, which is regarded as a regression task. The second needs to categorize 27 daily or rehabilitation actions corresponding to these data entries (27 classes). Data are collected from 4 different scenes. The third one needs to identify from which scenario the data comes (4 classes).

3) Model: We realize a simple feed-forward network (FFN) as a unified base model for encoders, decoders, and selectors on both datasets. This model consists of two residual blocks. The first block projects the input to 512 dimensions, and the second to 256 dimensions, each followed by a ReLU activation and layer normalization. The final layer is linear with the target output dimension.

4) Baseline: We compare the proposed PoM^2-DIB against the following representative baselines:

- Typical DL-based SemCom (DLSC) [15]–[17]: DLSC is designed to minimize learning risk without rate concerns.
- TADIB & VDDIB [1], [14]: Typical DIB-based solutions for multi-modal multi-task cases with full participation.
- DIB with random selection (RS-DIB) [8]: A typical DIB-based solution with random participation.

DLSC applies deterministic encoding and full participation. It is used to compare and verify the effectiveness of IB-based methods in rate reduction. The rest two are leveraged to verify the effectiveness of the newly introduced key variable, namely selection, and they both apply the typical DIB objective (6) to multiple tasks. All baselines are realized by the base model.

5) Metric: Relevance, characterized by *negative cross entropy (N-CE)*, is set to a common metric for all tasks, which is the logarithmic loss [21]. *Top-1 Accuracy* is set to the task-specific metric for classification, while *mean per joint position error (MPJPE)* and *procrustes analysis-MPJPE (PA-MPJPE)* are also set to the task-specific metrics for the MM-Fi's human pose estimation. CE terms are easy to estimate in classification but difficult in regression. In this regard, we adopt mean square error (MSE) instead of CE for regression, since the two, MSE and CE, are equivalent under the Gaussian assumption [31] to some extent. Sum-rate, namely, the total rate, is set to measure the communication overhead besides (9) and (10).

6) Default Training Hyper-Parameter: Training rounds are set to 2000 epochs, with a batch size of 20. The learning rate is set to 1e-4 for coding, while the learning rate is set to 5e-5 for selection, i.e., half of the coding. The multiplier β is set to 1e-3. E_t in (9) is set to 2 for all $t \in \mathcal{T}$. E_k in (10) is set to 4 for all $k \in \mathcal{K}$. The above two constraints are valid since for both datasets, E_t is up to 3, for HandWritten, E_k is up to 9, and for MM-Fi, it is up to 12. Moreover, the output dimension of encoders is set to 24 for HandWritten and 48 for MM-Fi. CR is initialized in prior as a 24-dimensional normal distribution for HandWritten and a 48-dimensional one for MM-Fi.

TABLE I: Overall and task-specific performance on (a) HandWritten and (b) MM-Fi datasets (95% confidence intervals). All metrics include sum-rate (nat), N-CE (nat), Top-1 Acc (%), and MPJPE/PA-MPJPE (mm). Besides, t1, t2, and t3 denote tasks 1, 2, and 3, respectively.

	metric/method	PoM^2-DIB	RS-DIB	TADIB	DLSC
	under limits	✓	✓	✗	✗
(a)	sum-rate	166.78±0.96	141.76±1.08	190.11±1.18	788.43±1.89
	N-CE	-0.39±0.05	-0.70±0.03	-0.30±0.02	-0.10±0.01
	t1:Top-1Acc	98.60±1.89	87.20±2.49	98.29±1.29	98.79±1.05
	t2:Top-1Acc	99.39±2.17	86.20±3.95	98.39±1.06	98.39±1.19
	t3:Top-1Acc	96.50±1.59	92.60±1.70	98.60±1.46	98.60±1.20
	under limits	✓	✓	✗	✗
(b)	sum-rate	311.43±2.34	67.78±2.12	447.32±1.91	1479.38±153.43
	N-CE	-0.17±0.02	-2.58±0.01	-0.19±0.02	-0.13±0.01
	t1:MPJPE	107.53±3.12	137.72±4.80	120.66±5.27	99.21±5.86
	t1:PA-∼	58.35±1.16	94.90±0.89	63.60±1.04	56.29±2.47
	t2:Top-1Acc	100.00±0.00	16.01±15.93	100.00±0.00	100.00±0.00
	t3:Top-1Acc	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00

7) Platform: We conduct our experiments on a workstation with an NVIDIA GeForce RTX 5090 GPU and an Intel(R) Core(TM) i9-14900KF CPU. The experimental code is implemented using Python 3.9.23, CUDA 13.0, and PyTorch 2.2.2, which is available at https://github.com/SamuChamp/pom2dib.

B. Numerical Results

1) Performance Comparison: In Table I, we compare the overall and task-specific performance of PoM^2-DIB and baselines on both datasets. The first two, i.e., PoM^2-DIB and RS-DIB, are realistic, satisfying resource constraint conditions (9) and (10). While the last two, i.e., TADIB and DLSC, do not meet the requirements since they assume full participation.

The inference quality of PoM^2-DIB, both in terms of the global metric N-CE and task-specific metrics (Top-1 Acc and MPJPE/PA-MPJPE), can be compared with that of TADIB and DLSC. Moreover, while meeting (9) and (10), its communication overhead is actually lower than that of both, demonstrating significant superiority. Compared to the second RS-DIB with random selection, PoM^2-DIB shows the substantial advantage of the deployment of an optimizable selection policy. Simple random selection always lets its corresponding coding scheme converge to a sub-optimal point or even collapse, e.g., at task 2 of MM-Fi. While in PoM^2-DIB, each task performs well.

Finally, in Table I, we observe that the traditional DL-based SemCom, represented by DLSC, consumes large communication resources due to the lack of rate control, but only achieves analogous results to IB solutions, PoM^2-DIB and TADIB, in inference quality. The sum-rate of DLSC is characterized by the sum of the entropy of low-dimensional outputs of encoders, since it applies deterministic encoders typically rather than the stochastic one of IB-based coding solutions, whose estimation is based on NPEET.

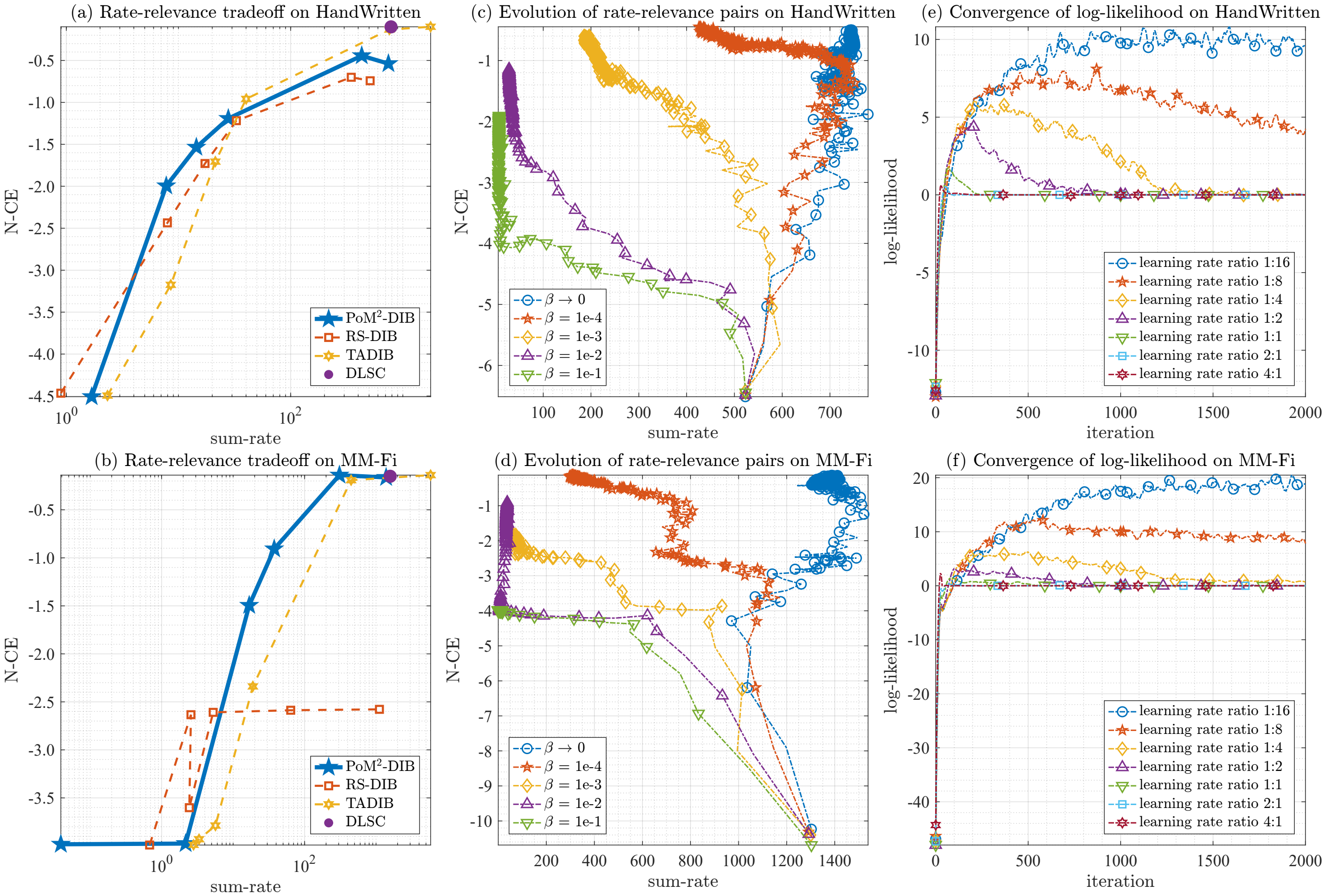


Fig. 2: Effect of hyper-parameter. (a)-(b) Rate-relevance tradeoff. (c)-(d) Training dynamics on information plane. (e)-(f) Effect of learning rate ratio (learning rate for selection to learning rate for coding, e.g., the default ratio is 1:2) on the convergence of log-likelihood (21).

TABLE II: Overall performance of PoM^2-DIB varying with learning rate ratio from 1:16 to 4:1 on (a) HandWritten and (b) MM-Fi datasets.

	metric/ratio	1:16	1:8	1:4	1:2	1:1	2:1	4:1
(a)	sum-rate	174.43	187.97	183.25	165.08	140.42	142.47	143.13
	N-CE	-0.72	-0.67	-0.61	-0.45	-0.92	-0.96	-0.95
(b)	sum-rate	93.92	72.71	47.60	49.31	52.50	53.12	76.75
	N-CE	-0.40	-0.26	-1.81	-1.88	-1.98	-1.87	-0.31

2) Effect of Hyper-Parameter: Effects of hyper-parameters are summarized in Fig. 2.

Rate-Relevance Tradeoff: In Fig. 2(a) and (b), we explore rate-relevance curves on the two datasets. DLSC does not involve a rate control variable like β. It is thereby represented as a fixed point in the two sub-figures. Other methods apply a range of β from 0 to 1 to control their communication rates. Then, we collect the converged value of the "sum-rate"-"N-CE" pair at each sampled $\beta \in [0, 1]$ to roughly depict the tradeoff curve between the two. It is observed that PoM^2-DIB achieves the ideal rate-relevance curve characterized by TADIB under hard limits by introducing a new degree of freedom, i.e., selection. In MM-Fi, its performance is even better than that of TADIB (the further to the upper left, the better), which fully matches our discussion in Section VII-B. While RS-DIB, which lacks an update mechanism of selection, performs much worse. It does not even work on the complex MM-Fi, corresponding to its crash on task 2 as shown in Table I. This contrast strongly supports the effectiveness of the cooperative selection policy presented in Section IV-B.

Effect of β: Fig. 2(c) and (d) sketch the training trajectories of PoM^2-DIB on the information plane varying with the value of β. Each point on any trajectory corresponds to the value of "sum-rate"-"N-CE" pair at an iteration step. It means that each trajectory corresponds to one complete training dynamics. It is observed that each trajectory finally converges to a stationary point. The tail of the trajectory holds denser markers than the head, which directly reflects the convergence dynamics. Also, we find that these trajectories converge to different points, in which β explicitly controls the update direction. Specifically, β decreases from 1e-1 to 0. The trajectories for these different β values exhibit distinctly different behaviors. In (c), when $\beta \to 0$, the value of rate increases monotonically. While β = 1e-4 or 1e-3, the information flow exhibits the classical behavior, i.e., the rate value increases then decreases while the relevance monotonically increases, which is previously indicated in [8], [32]. If we proceed to pose a greater rate penalty by increasing β, then the rate value of the trajectory will degenerate into a monotonically decreasing state. In Fig. 2(d), we also observe a similar phenomenon but with more irregular oscillations. It shows that different values of β correspond to different types of training dynamics. The phenomena are basically consistent with the pioneering work [8]. This further illustrates that our extended IB solution, i.e., PoM^2-DIB, involving selection as a

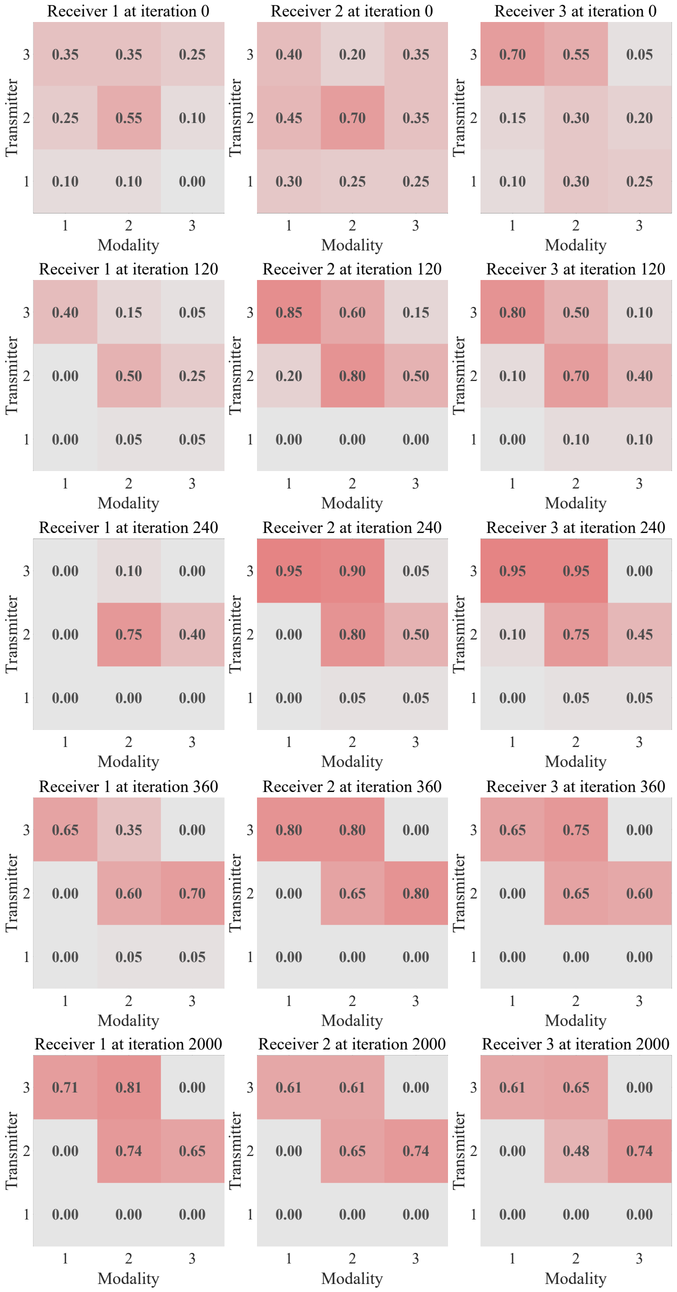


Fig. 3: Convergence of randomized selection $\boldsymbol{A}$ on HandWritten. Corresponding rate: 165.08; N-CE: -0.45. Total selection number (converged): 8.

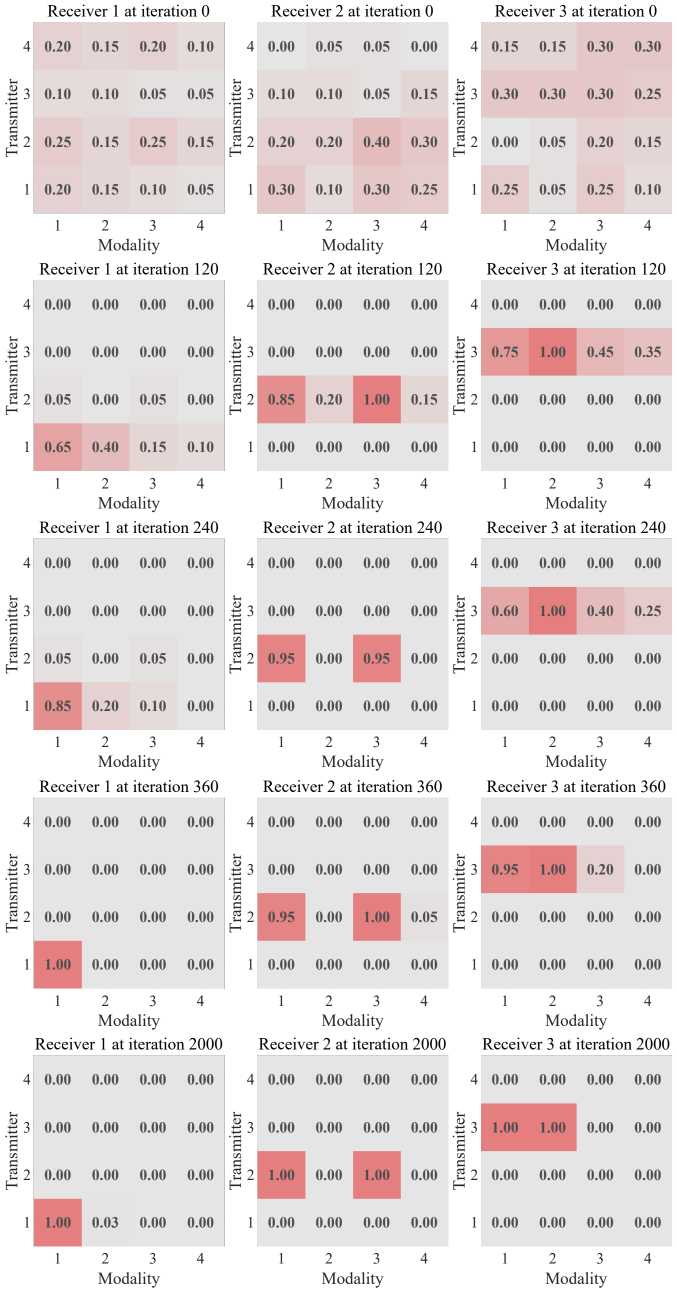


Fig. 4: Convergence of randomized selection $\boldsymbol{A}$ on MM-Fi. Corresponding rate: 78.17; N-CE: -0.35. Total selection number (converged): 5.

new variable to tackle hard limits, still fully inherits the good properties of IB, achieving effective rate-relevance tradeoff.

Effect of Learning Rate Ratio: In the rest of this section, we elaborate on the behaviors of the optimizable selection policy. In Fig. 2(e)-(f), we detail the convergence of the log-likelihood (21), controlled by the learning rate ratio. It is observed that all curves start from a low negative value, then grow to a positive value, and finally drop to zero if the ratio is sufficiently large, or converge to a certain positive value. The rapid growth in the early stage of training shows the effectiveness of policy gradient. It shows that the update of the cooperative selection policy matches the overall optimization goal. The converged selection policies also exhibit different characteristics with the differences in ratios. If the log-likelihood converges to zero, a deterministic policy $P_{\boldsymbol{A}|U}$ conditioned on CR is obtained. This is directly derived from (21) to let $\log p(\boldsymbol{a}|u) = 0$. This also demonstrates that if this term converges to a positive value, we have a stochastic version with CR. Both types of policies work well as shown in Table II. Additionally, we illustrate that there exists an optimal ratio, which is 1:2 on HandWritten and 1:8 on MM-Fi. We also find that using random selection at a ratio of 4:1 can achieve nearly optimal inference quality on MM-Fi. This reveals the highly nonlinear influence of the learning rate ratio on optimality. In which, the value of N-CE always converges to around -0.3 or -1.9 while the value of sum-rate always converges to around 80 or 50 on MM-Fi, which means that the selection and its corresponding coding jump between the optimal solution and the suboptimal one due to different learning rate ratios. The convergence of the log-likelihood only partially indicates selection convergence, as this log-likelihood also approaches zero when the loss value vanishes for every type of selection. Secondly, we present the convergence of the selection and the favorable properties it exhibits.

3) Characterization of Optimizable Randomized Selection: The convergence of the optimizable selection is represented in Figs. 3 and 4. The two are under hard limits with $E_t = 2$ and $E_k = 4$. In Figs. 3 and 4, these color blocks has the following meaning: Each row, presenting an iteration round, maintains 3 chunks. Each chunk represents all available modalities from transmitters for a specific task of a receiver. The value of color coding for a modality corresponds to the marginal probability mass of selecting that modality for the specific task. Hence, the value range is from 0 to 1. With the definition of $\boldsymbol{A}$ in Section IV-A, we can equivalently consider each row as corresponding to a "mixed" selection $\bar{\boldsymbol{a}}$ in the expectation sense, i.e., $\bar{\boldsymbol{a}} := \mathbb{E}[\boldsymbol{A}]$. Starting from a random initial distribution, $\bar{\boldsymbol{a}}$ gradually converges to a stationary distribution. Although we point out in the preceding that $P_{\boldsymbol{A}|U}$ could converge to a deterministic policy at the default ratio 1:2, due to the inherent randomness of U in training, this leads to the non-degenerate marginal $P_{\boldsymbol{A}}$. As shown in Fig. 3, the converged selection is mixed, namely, every possible selection has a mass lower than 1. Also, we can easily measure the effectiveness of the converged selection on HandWritten. Recall the setting in (25), i.e.,

$$\begin{bmatrix} {\color{red}X^A}, {\color{red}X^B}, X^C \\ X^C, {\color{red}X^A}, {\color{red}X^B} \\ X^C, X^C, {\color{red}X^A} \end{bmatrix} \tag{27}$$

where X^A corresponds to image information, X^B corresponds to audio information, and finally X^C corresponds to independent noises. We can easily confirm that the effective modalities are X^A and X^B. This exactly corresponds to the convergent selection expectation, i.e., the last row of Fig. 3, which drops all noise terms X^C. This is in line with our expectations and once again strongly illustrates the effectiveness of the selection optimization. In Fig. 4, a similar phenomenon is observed, but at this time, $\boldsymbol{A}$ is deterministic and degenerate: the probability mass of choosing these modalities equals 1. Besides, it is worth noticing that in every iteration, the expected selection number (the sum of type masses) is 8 for HandWritten, which exactly falls at the boundary of our limits. In contrast, the expected value of MM-Fi is less than 8 (converging to 5), which reveals the high data redundancy of the real dataset MM-Fi.

IX. Conclusion

This paper proposes the PoM^2-DIB framework for multimodal multi-task SemCom, which introduces a new scheme to jointly optimize modality selection alongside semantic coding based on DIB theory. By transforming discrete selection into a probabilistic form and leveraging common randomness for cross-device coordination, this distributed framework achieves a flexible modality-aware rate-relevance tradeoff. It can reduce the overhead of unnecessary transmission and computation of redundant data significantly, while maintaining high inference quality under physical limits. Besides theoretical guarantees, extensive experimental results also illustrate the effectiveness of PoM^2-DIB across different tasks and datasets, especially in resource-constrained multi-modal multi-task scenarios.

References

[1] C. Peng, Y. Zhou, R. Wang, Y. Xiao, Y. Li, and G. Shi, "Multi-user information bottleneck for semantic-aware communication," in *ICC 2025 - IEEE International Conference on Communications*, 2025.

[2] Z. Lu, "A theory of multimodal learning," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[3] G. K. Walia, M. Kumar, and S. S. Gill, "Ai-empowered fog/edge resource management for iot applications: A comprehensive review, research challenges, and future perspectives," *IEEE Communications Surveys & Tutorials*, vol. 26, no. 1, pp. 619–669, 2024.

[4] G. Shi, Y. Xiao, Y. Li, and X. Xie, "From semantic communication to semantic-aware networking: Model, architecture, and open problems," *IEEE Communications Magazine*, vol. 59, no. 8, pp. 44–50, 2021.

[5] Y. Xiao *et al.*, "Reasoning over the air: A reasoning-based implicit semantic-aware communication framework," *IEEE Trans. Wireless Commun.*, vol. 23, no. 4, Apr. 2024.

[6] Y. Xiao, Z. Sun, G. Shi, and D. Niyato, "Imitation learning-based implicit semantic-aware communication networks: Multi-layer representation and collaborative reasoning," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 3, pp. 639–658, 2022.

[7] D. Gündüz, M. A. Wigger, T.-Y. Tung, P. Zhang, and Y. Xiao, "Joint source–channel coding: Fundamentals and recent progress in practical designs," *Proc. of the IEEE*, early access, Nov. 2024, doi:10.1109/JPROC.2024.3477331.

[8] I. E. Aguerri and A. Zaidi, "Distributed variational representation learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 1, pp. 120–138, 2019.

[9] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[10] P. W. Cuff *et al.*, "Coordination capacity," *IEEE Transactions on Information Theory*, vol. 56, no. 9, pp. 4181–4206, 2010.

[11] J. Shao, Y. Mao, and J. Zhang, "Learning task-oriented communication for edge inference: An information bottleneck approach," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 1, pp. 197–211, 2021.

[12] S. Xie, S. Ma *et al.*, "Robust information bottleneck for task-oriented communication with digital modulation," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 8, pp. 2577–2591, 2023.

[13] H. Li, J. Shao, H. He, S. Song, J. Zhang, and K. B. Letaief, "Tackling distribution shifts in task-oriented communication with information bottleneck," *arXiv preprint arXiv:2405.09514*, 2024.

[14] J. Shao, Y. Mao, and J. Zhang, "Task-oriented communication for multidevice cooperative edge inference," *IEEE Transactions on Wireless Communications*, vol. 22, no. 1, pp. 73–87, 2022.

[15] 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.

[16] G. He, S. Cui, Y. Dai, and T. Jiang, "Learning task-oriented channel allocation for multi-agent communication," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 11, pp. 12 016–12 029, 2022.

[17] G. Zhang, Q. Hu, Z. Qin, Y. Cai, G. Yu, and X. Tao, "A unified multi-task semantic communication system for multimodal data," *IEEE Transactions on Communications*, vol. 72, no. 7, pp. 4101–4116, 2024.

[18] M. Gong *et al.*, "Compression before fusion: Broadcast semantic communication system for heterogeneous tasks," *IEEE Transactions on Wireless Communications*, vol. 23, no. 12, pp. 19 428–19 443, 2024.

[19] A. H. Razlighi, M. H. Tillmann, E. Beck, C. Bockelmann, and A. Dekorsy, "Cooperative and collaborative multi-task semantic communication for distributed sources," *arXiv preprint arXiv:2411.02150*, 2024.

[20] Z. Goldfeld and Y. Polyanskiy, "The information bottleneck problem and its applications in machine learning," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 19–38, 2020.

[21] A. Zaidi, I. Estella-Aguerri, and S. Shamai, "On the information bottleneck problems: Models, connections, applications and information theoretic views," *Entropy*, vol. 22, no. 2, p. 151, 2020.

[22] D. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Transactions on Information Theory*, vol. 19, no. 4, pp. 471–480, 1973.

[23] T. A. Courtade and T. Weissman, "Multiterminal source coding under logarithmic loss," *IEEE Transactions on Information Theory*, vol. 60, no. 1, pp. 740–761, 2014.

[24] A. A. Alemi *et al.*, "Deep variational information bottleneck," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, 2017*. OpenReview.net, 2017.

[25] P. Cheng, W. Hao, S. Dai, J. Liu, Z. Gan, and L. Carin, "Club: A contrastive log-ratio upper bound of mutual information," in *International conference on machine learning*. PMLR, 2020, pp. 1779–1788.

[26] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, “Policy gradient methods for reinforcement learning with function approximation,” *Advances in neural information processing systems*, vol. 12, 1999.

[27] D. E. Rumelhart *et al.*, “Learning representations by back-propagating errors,” *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[28] F. Clarke, *Functional analysis, calculus of variations and optimal control*. Springer, 2013, vol. 264.

[29] V. Vielzeuf, A. Lechervy, S. Pateux, and F. Jurie, “Centralnet: a multilayer approach for multimodal fusion,” in *Proceedings of the European conference on computer vision (ECCV) workshops*, 2018, pp. 0–0.

[30] J. Yang *et al.*, “Mm-fi: Multi-modal non-intrusive 4d human dataset for versatile wireless sensing,” in *NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[31] S. Yu, X. Yu, S. Løkse, R. Jenssen, and J. C. Príncipe, “Cauchy-schwarz divergence information bottleneck for regression,” in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[32] N. Tishby and N. Zaslavsky, “Deep learning and the information bottleneck principle,” in *2015 ieee information theory workshop (itw)*. Ieee, 2015, pp. 1–5.

Appendix A
Proof of Theorem 1 and Proposition 1

Proof of Theorem 1. By introducing an auxiliary U matching Definition 1, we directly decompose $\mathcal{L}_{\text{PoM}^2\text{-DIB}}[\boldsymbol{v}_{\text{p}}, \boldsymbol{f}]$ as

$$\begin{aligned}&\mathbb{E}_{P_{\boldsymbol{A}}}\left[\mathcal{L}_{\text{M}^2\text{-DIB}}[\boldsymbol{v}_{\text{p}}, \boldsymbol{f}]\right] = \mathbb{E}_{P_U}\left[\mathbb{E}_{P_{\boldsymbol{A}|U}}\left[\mathcal{L}_{\text{M}^2\text{-DIB}}[\boldsymbol{v}_{\text{p}}, \boldsymbol{f}]\right]\right] \\ &= \mathbb{E}_{P_U}\left[\sum_{t=1}^{T}\mathbb{E}_{P_{\boldsymbol{A}_t|U}}\left[I(\boldsymbol{A}_t \circ \boldsymbol{Z}_t; Y_t) + \beta\langle \boldsymbol{A}_t, \boldsymbol{L}_{\text{IB},t}\rangle\right]\right],\end{aligned} \tag{28}$$

which is our desired result. □

Proof of Proposition 1. For (i), we have

$$\begin{aligned}I(Z;Y) &\geq \int dydzp(y,z)\log\frac{q(y|z)}{p(y)} \\ &= \int dydzp(y,z)\log q(y|z) - \int dyp(y)\log p(y) \\ &= H(Y) - \mathbb{E}_{z\sim P_Z}[H(P_{Y|Z}, Q_{Y|Z})],\end{aligned} \tag{29}$$

where the first inequality holds due to the non-negativity of $D_{\text{KL}}(\cdot\|\cdot)$, i.e.,

$$\int p(z)p(y|z)\log\frac{p(y|z)}{q(y|z)}dzdy = \mathbb{E}_{P_Z}[D_{\text{KL}}(P_{Y|Z}\|Q_{Y|Z})] \geq 0. \tag{30}$$

Note that $H(Y|Z) = H(Y) - I(Y;Z)$. We obtain the result stated in (i) by replacing Z with $Z_{k,t}^m$ and Y with Y_t.

For (ii), following the preceding inequality (29), we replace Z with $\boldsymbol{A}_t \circ \boldsymbol{Z}_t$ and Y with y_t to yield

$$H(Y_t|\boldsymbol{A}_t \circ \boldsymbol{Z}_t) \leq \mathbb{E}_{P_{\boldsymbol{A}_t\circ\boldsymbol{Z}_t}}[H(P_{Y_t|\boldsymbol{A}_t\circ\boldsymbol{Z}_t}, Q_{Y_t|\boldsymbol{A}_t\circ\boldsymbol{Z}_t})]. \tag{31}$$

We first define a function $h(\boldsymbol{A}_t, \boldsymbol{Z}_t) := \boldsymbol{A}_t \circ \boldsymbol{Z}_t$. Then, by taking the law of the unconscious statistician, we identically rewrite $\mathbb{E}_{P_{h(\boldsymbol{a}_t,\boldsymbol{Z}_t)}}[f(h(\boldsymbol{a}_t,\cdot))] = \mathbb{E}_{P_{\boldsymbol{Z}_t|\boldsymbol{A}_t=\boldsymbol{a}_t}}[f(h(\boldsymbol{a}_t,\cdot))]$ for any measurable f. Since arbitrary $\boldsymbol{a}_t \sim \boldsymbol{A}_t$ is admitted, we can finally obtain the result stated in (ii). □

Appendix B
Proof of Theorem 2

We first give a technical lemma, then we prove Theorem 2.

Lemma 1. *If $P_{\boldsymbol{A}}$ is (degenerated) Dirac, then $P_{\boldsymbol{A}} \in \mathcal{P}_{\boldsymbol{A}}$.*

Proof of Theorem 2. In (23), the almost everywhere convergence holds due to the strong law of large numbers, and the inequality holds due to Proposition 1. In (24), the inequality can be derived analogously to [8, Lemma 1].

In the remainder of the proof, we show that the last equality in (24) always holds. Let the optimal encoders $\boldsymbol{f}^\star$ of each realization $\boldsymbol{a}$ be known as $\boldsymbol{f}^\star(\boldsymbol{a})$. We then regard objectives $\mathcal{L}_{\text{PoM}^2\text{-DIB}}$ and $\mathcal{L}_{\text{M}^2\text{-DIB}}$ as functions w.r.t. $\boldsymbol{a}$ only, since $\boldsymbol{f}^\star(\cdot)$ of $\boldsymbol{a}$ are identical w.r.t. the two objectives under a same data distribution.

We first prove $\min\mathcal{L}_{\text{PoM}^2\text{-DIB}} \leq \min\mathcal{L}_{\text{M}^2\text{-DIB}}$ and then prove $\min\mathcal{L}_{\text{PoM}^2\text{-DIB}} \geq \min\mathcal{L}_{\text{M}^2\text{-DIB}}$. The minimizer of $\mathcal{L}_{\text{M}^2\text{-DIB}}$ can induce its corresponding optimal $\{\mathcal{K}_t^\star\}_{t\in\mathcal{T}}$, which is a deterministic selection and yields its characterization $\boldsymbol{a}^\star$ from $\boldsymbol{v}^\star$. We can construct a Dirac distribution at $\boldsymbol{a}^\star$, denoted as $\widetilde{P}_{\boldsymbol{A}}^\star$. We claim $\widetilde{P}_{\boldsymbol{A}}^\star \in \mathcal{P}_{\boldsymbol{A}}$ due to Lemma 1. This shows that $\mathcal{L}_{\text{M}^2\text{-DIB}}$ is a special case of $\mathcal{L}_{\text{PoM}^2\text{-DIB}}$, where $P_{\boldsymbol{A}}$ should be given as a Dirac distribution. Since $P_{\boldsymbol{A}}^\star$ of $\mathcal{L}_{\text{PoM}^2\text{-DIB}}$ is optimal over $\mathcal{P}_{\boldsymbol{A}}$ and $\widetilde{P}_{\boldsymbol{A}}^\star$ may not be identical to $P_{\boldsymbol{A}}^\star$, we claim $\min\mathcal{L}_{\text{PoM}^2\text{-DIB}} \leq \min\mathcal{L}_{\text{M}^2\text{-DIB}}$. We then prove $\min\mathcal{L}_{\text{PoM}^2\text{-DIB}} \geq \min\mathcal{L}_{\text{M}^2\text{-DIB}}$. We explicitly write

$$\begin{aligned}\min\mathcal{L}_{\text{M}^2\text{-DIB}} &= \mathbb{E}_{\widetilde{P}_{\boldsymbol{A}}^\star}[\mathcal{L}_{\text{M}^2\text{-DIB}}[\boldsymbol{A}, \boldsymbol{f}^\star(\boldsymbol{A})]] \\ &= \mathcal{L}_{\text{M}^2\text{-DIB}}[\boldsymbol{a}^\star, \boldsymbol{f}^\star(\boldsymbol{a}^\star)]\end{aligned} \tag{32}$$

by reformulating ($\mathbf{P}_0$), where we characterize the impact of $\boldsymbol{v}$ as its characterization $\boldsymbol{a}$. Also, with a certain $1 \geq p \geq 0$,

$$\begin{aligned}&\min\mathcal{L}_{\text{PoM}^2\text{-DIB}} = \mathbb{E}_{P_{\boldsymbol{A}}^\star}[\mathcal{L}_{\text{M}^2\text{-DIB}}[\boldsymbol{A}, \boldsymbol{f}^\star(\boldsymbol{A})]] \\ &= p\mathcal{L}_{\text{M}^2\text{-DIB}}[\boldsymbol{a}^\star, \boldsymbol{f}^\star(\boldsymbol{a}^\star)] + (1-p)\mathcal{L}_{\text{M}^2\text{-DIB}}[\boldsymbol{a}', \boldsymbol{f}^\star(\boldsymbol{a}')]\end{aligned} \tag{33}$$

as ($\mathbf{P}_1$), where for simplicity, we consider a case that $P_{\boldsymbol{A}}^\star$ with its support at $\boldsymbol{a}^\star$ and $\boldsymbol{a}'$ only. Then the following holds

$$\begin{aligned}&\min\mathcal{L}_{\text{PoM}^2\text{-DIB}} - \min\mathcal{L}_{\text{M}^2\text{-DIB}} \\ &= p\mathcal{L}_{\text{M}^2\text{-DIB}}[\boldsymbol{a}^\star, \boldsymbol{f}^\star(\boldsymbol{a}^\star)] + (1-p)\mathcal{L}_{\text{M}^2\text{-DIB}}[\boldsymbol{a}', \boldsymbol{f}^\star(\boldsymbol{a}')] \\ &\quad - \mathcal{L}_{\text{M}^2\text{-DIB}}[\boldsymbol{a}^\star, \boldsymbol{f}^\star(\boldsymbol{a}^\star)] \\ &= (1-p)(\mathcal{L}_{\text{M}^2\text{-DIB}}[\boldsymbol{a}', \boldsymbol{f}^\star(\boldsymbol{a}')] - \mathcal{L}_{\text{M}^2\text{-DIB}}[\boldsymbol{a}^\star, \boldsymbol{f}^\star(\boldsymbol{a}^\star)]) \\ &= (1-p)(\mathcal{L}_{\text{M}^2\text{-DIB}}[\boldsymbol{a}', \boldsymbol{f}^\star(\boldsymbol{a}')] - \min\mathcal{L}_{\text{M}^2\text{-DIB}}) \\ &\geq 0.\end{aligned} \tag{34}$$

This also holds for general supports that have finite counts. It follows that $\min\mathcal{L}_{\text{PoM}^2\text{-DIB}} \geq \min\mathcal{L}_{\text{M}^2\text{-DIB}}$. □

Proof of Lemma 1. Let random variables V, W follow a joint distribution P_{VW}, which is Dirac at (v_0, w_0), i.e., its probability mass function satisfies

$$p_{VW}(v,w) = \begin{cases}1 \text{ if } (v,w) = (v_0,w_0), \\ 0 \text{ else.}\end{cases} \tag{35}$$

In this case, we can directly have $p_{VW}(v,w) = p_V(v)p_W(w)$, which implies the independence between V and W. Extending the above conclusion to the Dirac $P_{\boldsymbol{A}}$, it follows that $P_{\boldsymbol{A}} = \prod_{t=1}^{T}\prod_{k=1}^{K}P_{\boldsymbol{A}_t^{\lfloor k\rfloor}}$, which is in $\mathcal{P}_{\boldsymbol{A}}$ with an arbitrary U. □

Appendix C
Proof of Proposition 2

Proof. This proof accounts for the receiver-side selection for simplicity and it can be straightforwardly applied to the case of the selection at the transmitter-side.

We demonstrate that at $\beta \to 0$ if one selection is optimal and in which the number of transmitters selected by a certain receiver is less than its performance boundary, i.e., $|\mathcal{K}_t| < E_t$, then there must exist an extension of it, which is also optimal and satisfies $|\mathcal{K}_t| = E_t$ for all receivers $\mathcal{T}$.

We choose a selection $\{\mathcal{K}_t\}_{t\in\mathcal{T}}$ with $|\mathcal{K}_{t'}| < E_{t'}$ at some t' and $|\mathcal{K}_t| = E_t$ for any $t \in \mathcal{T} : t \neq t'$. All t' construct a subset of $\mathcal{T}$ as $\mathcal{T}'$. Note that the transmitter-side selection is not assigned, therefore it is arbitrary but should satisfy (10). We now write the minimum of the $\mathcal{L}_{\text{M}^2\text{-DIB}}$ objective with this fixed receiver-side selection in the following

$$
\begin{aligned}
&\min_{\boldsymbol{f}} \mathcal{L}_{\text{M}^2\text{-DIB}}|_{\{\mathcal{K}_t\}_{t\in\mathcal{T}}} \\
&= \min_{\boldsymbol{f}} \left\{ \mathcal{L}_{\text{M}^2\text{-DIB}}|_{\mathcal{K}_{t'}} + \mathcal{L}_{\text{M}^2\text{-DIB}}|_{\{\mathcal{K}_t\}_{t\in\mathcal{T}/t'}} \right\} \\
&= \min_{\boldsymbol{f}} \Big\{ H(Y_{t'}|Z_{\mathcal{K}_{t'}}) + \beta \sum_{k\in\mathcal{K}_{t'}, m\in v_k(t')} \big(H(Y_{t'}|Z^m_{k,t'}) + I(X^m_k; Z^m_{k,t'})\big) \\
&\qquad\qquad + \mathcal{L}_{\text{M}^2\text{-DIB}}|_{\{\mathcal{K}_t\}_{t\in\mathcal{T}/t'}} \Big\} \\
&= \min_{\boldsymbol{f}} \left\{ H(Y_{t'}|Z_{\mathcal{K}_{t'}}) + \mathcal{L}_{\text{M}^2\text{-DIB}}|_{\{\mathcal{K}_t\}_{t\in\mathcal{T}/t'}} \right\},
\end{aligned}
\tag{36}
$$

where $\boldsymbol{f} := ((f^m_k)^{m(k)}_{m=1})^K_{k=1}$. Note that for a k, t-pair that is not compatible with the current selection $\{\mathcal{K}_t\}_{t\in\mathcal{T}}$, the value of $f^m_k(\cdot, t)$ is not defined and can be arbitrary. We now add a new k, t-pair for t' to extend $\mathcal{K}_{t'}$ as $\widetilde{\mathcal{K}}_{t'} := \mathcal{K}_{t'} \cup k'$. The new k' does exist due to $|\mathcal{K}_{t'}| < E_{t'}$. We also construct a encoder for this pair as $f^{m'}_{k'}(\cdot, t') \equiv C$ for any m', where C is a constant (vector). It means that any unimodal encoder of k' induces a constant mapping. In so doing, $H(Y_{t'}|Z_{\mathcal{K}_{t'}}) = H(Y_{t'}|Z_{\widetilde{\mathcal{K}}_{t'}})$. We now set the minimizer of (36) to be $\boldsymbol{f}'$ and set the extended version of $\boldsymbol{f}'$ with some constant $f^{m'}_{k'}$ to be $\tilde{\boldsymbol{f}}'$, then we have

$$
\begin{aligned}
&\min_{\boldsymbol{f}} \mathcal{L}_{\text{M}^2\text{-DIB}}|_{\{\mathcal{K}_t\}_{t\in\mathcal{T}}} \\
&= \left\{ \mathcal{L}_{\text{M}^2\text{-DIB}}|_{\mathcal{K}_{t'}} + \mathcal{L}_{\text{M}^2\text{-DIB}}|_{\{\mathcal{K}_t\}_{t\in\mathcal{T}/t'}} \right\}\Big|_{\boldsymbol{f}'} \\
&= \left\{ \mathcal{L}_{\text{M}^2\text{-DIB}}|_{\mathcal{K}_{t'}} + \mathcal{L}_{\text{M}^2\text{-DIB}}|_{\{\mathcal{K}_t\}_{t\in\mathcal{T}/t'}} \right\}\Big|_{\tilde{\boldsymbol{f}}'} \\
&\geq \min_{\boldsymbol{f}} \mathcal{L}_{\text{M}^2\text{-DIB}}|_{\{\widetilde{\mathcal{K}}_{t'}\}\cup\{\mathcal{K}_t\}_{t\in\mathcal{T}/t'}}.
\end{aligned}
\tag{37}
$$

If $\{\mathcal{K}_t\}_{t\in\mathcal{T}}$ is optimal, then $\{\widetilde{\mathcal{K}}_{t'}\} \cup \{\mathcal{K}_t\}_{t\in\mathcal{T}/t'}$ is also optimal, i.e.,

$$
\begin{aligned}
\min_{\boldsymbol{f},\boldsymbol{v}} \mathcal{L}_{\text{M}^2\text{-DIB}} &= \min_{\boldsymbol{f}} \mathcal{L}_{\text{M}^2\text{-DIB}}|_{\{\mathcal{K}_t\}_{t\in\mathcal{T}}} \\
&\geq \min_{\boldsymbol{f}} \mathcal{L}_{\text{M}^2\text{-DIB}}|_{\{\widetilde{\mathcal{K}}_{t'}\}\cup\{\mathcal{K}_t\}_{t\in\mathcal{T}/t'}} \\
&\geq \min_{\boldsymbol{f},\boldsymbol{v}} \mathcal{L}_{\text{M}^2\text{-DIB}}.
\end{aligned}
\tag{38}
$$

We repeat the above construction for all $t' \in \mathcal{T}'$ until the updated subset $\mathcal{T}' = \varnothing$. This derives the result. □

Appendix D
Additional Experiments to Verify Remarks 6 and 7

We now verify our statement in Remark 6, which indicates that at $\beta > 0$, even when there exist no physical constraints, the optimal selection does not degrade into full participation. This claim is empirically demonstrated as illustrated in Fig. 5-6. In this case, a few modalities that are beneficial for task inference are selected, all the useless noise modalities are not included. In contrast, if we set $\beta \to 0$ as illustrated in Fig. 6, free communication and unlimited computing power make the convergence selection approach full participation.

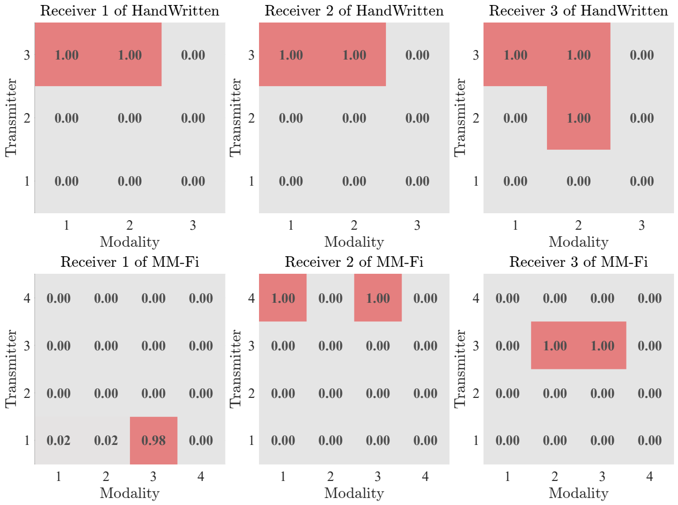


Fig. 5: Converged $\boldsymbol{A}$ with no limits at β=1e-3. For HandWritten, rate: 172.27; N-CE: -0.15. For MM-Fi, rate: 78.17; N-CE: -0.36.

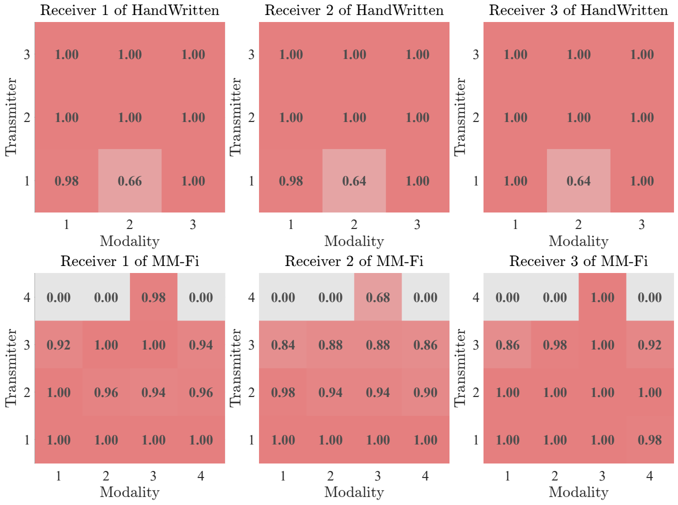


Fig. 6: Converged $\boldsymbol{A}$ with no limits at $\beta \to 0$. For HandWritten, rate: 1690.59; N-CE: -0.08. For MM-Fi, rate: 4592.59; N-CE: -0.14.

Fig. 7 and 8 allow full participation and adopt the sparse prior discussed in Remark 7. They depict the convergence of the optimizable selection, albeit with soft resource constraints. It can be verified from the expected values of the first line, i.e., the initial stage, that the HandWritten expected value is 8.4 and the MM-Fi expected value is 14.35, both of which are greater than the 8 we set. However, the modal selection that eventually converges to is highly sparse with the HandWritten expected value 3.44 and the MM-Fi expected value 3.15, which is due to the application of the sparse regularization with $\gamma = 0.1$. This indicates that by introducing the additional multiplier γ, the rate-relevance tradeoff under modality-aware constraints can be successfully extended to a three-way tradeoff of rate-relevance-selection, thereby allowing for efficient and simple SemCom networking.

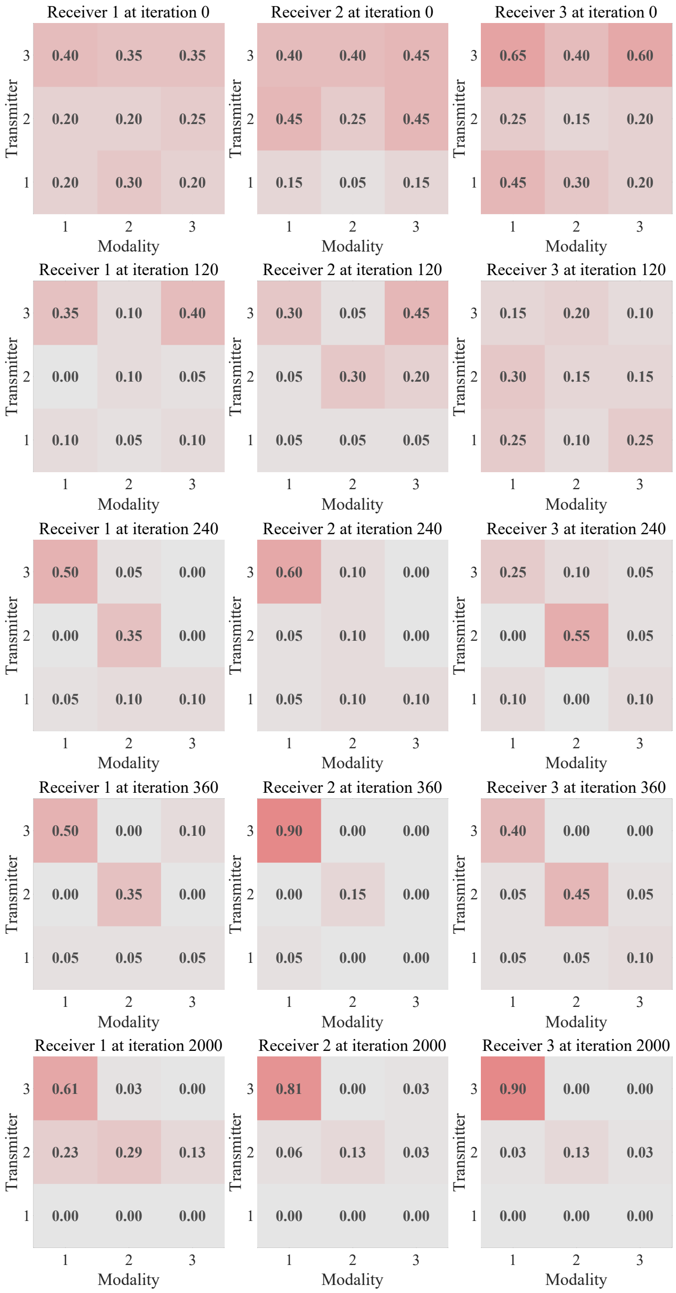


Fig. 7: Convergence of $\boldsymbol{A}$ with the sparse prior on HandWritten. Corresponding rate: 135.71; N-CE: -0.70. Total selection number (converged): 3.44.

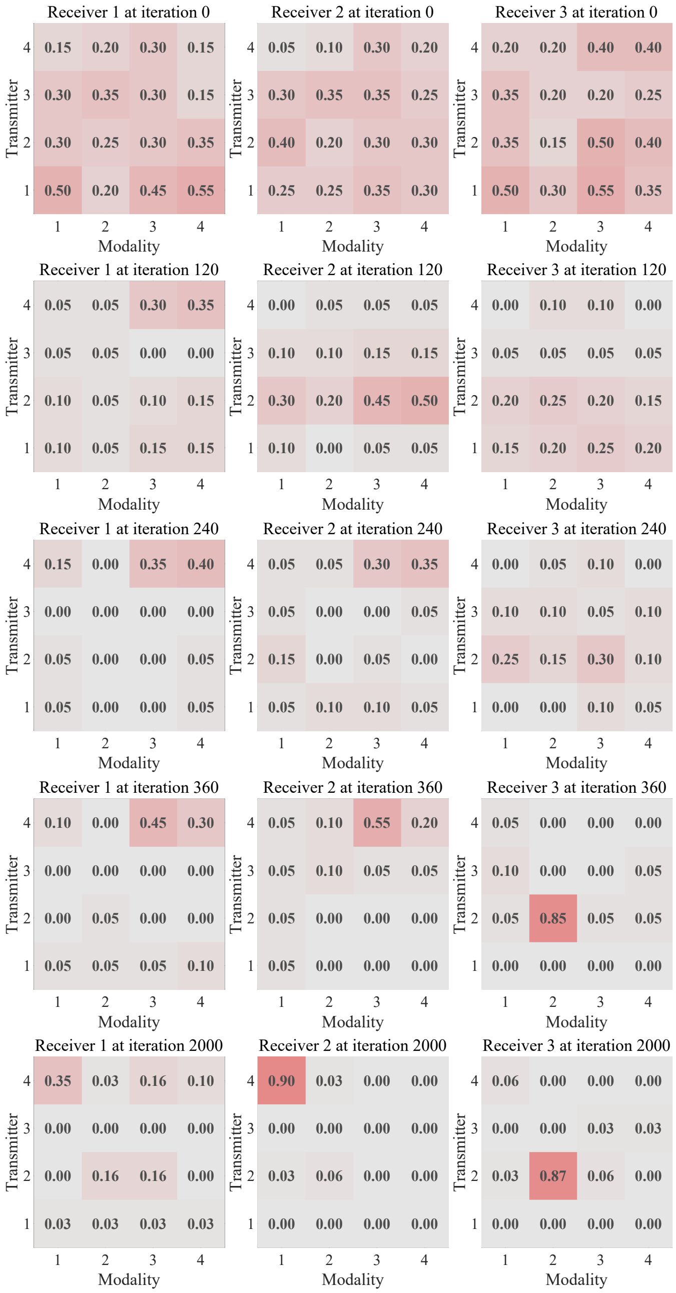


Fig. 8: Convergence of $\boldsymbol{A}$ with the sparse prior on MM-Fi. Corresponding rate: 75.99; N-CE: -0.35. Total selection number (converged): 3.15.