Large AI Model for Multimodal Integrated Sensing and Communication

Yubo Peng, Luping Xiang, Kun Yang, and Jienan Chen

Abstract

Multimodal integrated sensing and communication (ISAC) exploits heterogeneous modalities to enhance perception accuracy, communication robustness, and environmental adaptability, becoming a key enabler of the Internet of Everything (IoE). Nevertheless, existing multimodal ISAC systems remain constrained by heterogeneous data characteristics, dynamic modality availability, and the limited adaptability of current fusion strategies. To overcome these limitations, we propose a LAM-enabled multimodal ISAC (LAM-MSAC) framework. First, modality-specific feature encoders are introduced to provide native compatibility for diverse sensing data. Second, to cope with dynamically changing modality combinations, we design a Mixture-of-Experts (MoE) fusion module in which multiple experts process different modal combinations. Finally, the MoE structure activates only a subset of experts for each inference, substantially reducing computational cost without sacrificing model capacity. A case study shows that LAM-MSAC achieves over 90% beam prediction accuracy while significantly lowering computation compared with maintaining multiple single- or multi-modality models. Potential research directions are further discussed to advance the integration of LAMs into multimodal ISAC.

Introduction

Background

Multimodal integrated sensing and communication (ISAC) has recently emerged as a key enabler of the Internet of Everything (IoE) [1]. Despite its promise, existing ISAC systems still face several fundamental challenges. Traditional single-modality solutions remain highly sensitive to occlusions, illumination fluctuations, and environmental dynamics, leading to degraded sensing accuracy and unstable communication performance. Although multimodal ISAC incorporates heterogeneous sources such as RF, radar, vision, LiDAR, and GPS to enhance robustness, practical fusion is hindered by substantial modality heterogeneity in structure, resolution, and sampling rate, as well as inherent spatio-temporal asynchrony. Moreover, most fusion pipelines rely on fixed, manually designed modality combinations, rendering them insufficiently adaptive to the dynamic, large-scale IoE environments characteristic of autonomous mobility, smart city infrastructure, emergency monitoring, and disaster-response applications.

Additionally, existing multimodal ISAC frameworks predominantly depend on task-specific traditional AI models, which suffer from limited generalization, weak cross-domain transferability, and inadequate capacity to capture cross-modal dependencies. Vision-based models often overfit localized single-modality cues and fail under low visibility or unfamiliar scenes. Radar/RF-based models, constrained by fixed channel assumptions and narrow training distributions, lose robustness in dynamic, occluded, or non-stationary environments.

The rapid development of large AI models (LAMs) introduces new opportunities for advancing intelligent sensing and communication [2]. Powered by Transformer attention and Mixture-of-Experts (MoE) architectures [3], LAMs provide strong feature abstraction, transferability, and continual learning capabilities. Their ability to capture long-range dependencies and process heterogeneous modalities aligns naturally with the complexity of multimodal ISAC. For instance, under severely degraded visual conditions, a LAM can infer plausible environmental states from rich semantic priors; when radar signals become unreliable, it can perform cross-modal completion by projecting sparse radar or channel state information (CSI) observations into a unified semantic space, thereby improving robustness and reliability [4].

Contributions

To address the limitations of multimodal ISAC, we propose a LAM-enabled multimodal ISAC (LAM-MSAC) framework that acts as a generalized backbone for downstream tasks. The primary contributions are as follows:

1. *Challenges for LAM-MSAC:* We comprehensively examine the key challenges in integrating LAMs with multimodal ISAC, including: (i) insufficient native support for 6G-relevant modalities such as RF signals, LiDAR point clouds, and infrared sensing; (ii) the need for adaptation to time-varying and spatially heterogeneous modality combinations; and (iii) the tension between the high computational overhead of LAMs and the stringent low-latency and low-complexity requirements of ISAC.

Yubo Peng, Luping Xiang (corresponding author), and Kun Yang are with the State Key Laboratory of Novel Software Technology, Nanjing University, Nanjing 210008, China, and also with the Institute of Intelligent Networks and Communications (NINE) and the School of Intelligent Software and Engineering, Nanjing University, Suzhou Campus, Suzhou 215163, China; Jienan Chen is with the National Key Laboratory of Wireless Communications, University of Electronic Science and Technology of China, Chengdu, Sichuan 611731, China.

These insights inform the design of LAMs that are both versatile and resource-efficient.

2. *Design for LAM-MSAC:* We survey publicly available datasets for multimodal ISAC training and analyze representative LAM architectures, highlighting their advantages and limitations. Building on this analysis, we propose a LAM-MSAC architecture that accepts free-form modality inputs and generates unified multimodal features with high fidelity. This generalized backbone allows task-specific output heads to be lightweight and easily fine-tuned, thereby reducing computational cost while preserving adaptability to diverse downstream applications.
3. *Case Validation for LAM-MSAC:* We conduct a case study to validate the proposed framework. Results show that LAM-MSAC supports arbitrary input combinations from three or more modalities. In a representative beam prediction task, it achieves higher prediction accuracy with lower cumulative computational cost compared to ensembles of separate single-, dual-, and tri-modality models.

Challenges in LAMs for Multimodal ISAC

To design an effective LAM-MSAC framework, as illustrated in Table 1, this section provides a comprehensive clarification of the critical challenges in integrating LAMs into multimodal ISAC systems.

First, multimodal heterogeneity in ISAC systems has been shown to significantly degrade performance [5], [6], [7]. For instance, mismatched sampling rates or feature distributions between radar and camera streams can reduce object detection accuracy, obviously. Distribution shifts in RF or LiDAR sensing under varying environments also severely affect sensing reliability and communication robustness. Existing multimodal LAMs, although effective for conventional modalities such as text, images, audio, and video, are fundamentally incompatible with these heterogeneous sensing modalities, which differ widely in physical characteristics, sampling structures, and spatio–temporal resolutions. These limitations highlight the necessity of designing LAMs tailored for multimodal ISAC, capable of modeling diverse sensing modalities within a unified representation space and mitigating cross-modal inconsistencies.

Second, real-world ISAC deployments are characterized by spatially and temporally varying sensing configurations. For example, some regions may only deploy cameras and mmWave radar, whereas others rely primarily on LiDAR; similarly, daytime sensing favors cameras while nighttime operations require infrared sensors. LAMs designed for multimodal ISAC must therefore support flexible and adaptive handling of dynamically changing modality combinations, as opposed to the static input configurations assumed in most current multimodal foundation models.

Finally, 6G applications impose strict requirements on latency, energy efficiency, and computational complexity. While LAMs excel in representation learning and generalization, their massive parameter counts and high inference complexity result in prohibitive resource demands. This makes them unsuitable for latency-critical communication or edge computing scenarios without further optimization. Hence, developing lightweight, resource-efficient, and adaptive model architectures, while preserving the performance benefits of LAMs, remains a critical research challenge.

Collectively, these challenges underscore the importance of constructing representative multimodal datasets, designing adaptable model architectures, and optimizing inference for constrained environments. These considerations directly inform the design and implementation of the proposed LAM-MSAC framework, as presented in the subsequent section.

LAMs-Enabled Multimodal ISAC Framework

To address the identified challenges, we propose the LAM-MSAC framework. Specifically, we first investigate existing multimodal datasets that can serve as training data for the framework. We then analyze mainstream LAM architectures to provide design guidance for LAM-MSAC. Building upon these insights, we present the detailed framework design, in which a multimodal encoder is introduced to flexibly accommodate diverse input modalities, thereby overcoming the lack of native support for 6G-relevant data. Moreover, by integrating cross-attention and MoE-based balancing mechanism, solving the challenges of adapting to time-varying, heterogeneous modality combinations while reconciling the heavy overhead of LAMs with stringent latency and complexity constraints.

Datasets for LAM-MSAC

High-quality datasets form the foundation of LAM training, as they directly affect model accuracy, generalization, and reliability. Current multimodal ISAC datasets can be broadly classified into real-world datasets and simulation-based datasets. Several representative examples are described as follows:

Challenge	Reason Analysis	LAM Mitigation
Modality heterogeneity	6G sensing modalities (RF, LiDAR, radar, IR) differ significantly from conventional RGB/audio/text.	Unified semantic embedding space enables cross-modal alignment.
Dynamic modality availability	Sensor combinations vary across time and space.	Modular and adaptive encoders enable flexible selection of modalities.
High computational cost	Massive parameters conflict with 6G latency and resource constraints.	Lightweight adapters and compression reduce inference complexity.
Data scarcity	Multimodal ISAC datasets are limited and difficult to construct.	Pretraining and synthetic data reduce dataset dependence.
Domain shift & misalignment	Sensors differ in resolution, noise level, and sampling domain.	Robust pretrained semantics improve domain generalization.

TABLE 1. Challenges of Multimodal ISAC and LAM-based Solutions.

1) Raymobtime: A simulation-based dataset tailored for vehicle-to-infrastructure (V2I) communications, particularly for 5G/mmWave multiple-input multiple-output (MIMO) channel modeling. It is generated using Simulator for Urban Mobility (SUMO) for traffic simulation, Blender for 3D rendering, and Remcom Wireless InSite for ray-tracing. The dataset provides over 6,000 synchronized samples, including camera images, LiDAR point clouds, GPS coordinates, and channel data under both LoS and NLoS conditions, making it a valuable benchmark for vision- and LiDAR-assisted beam prediction tasks [8].

2) ViWi: A large-scale synthetic framework that combines wireless communication data with visual sensing for mmWave beam prediction. Generated with Blender and Wireless InSite, it covers diverse urban environments populated with vehicles and pedestrians. The dataset offers RGB images, depth maps, LiDAR data, channel state information, and beam indices, providing more than 400,000 samples across training, validation, and test sets. Its open-source nature has made it one of the most widely adopted multimodal benchmarks for ISAC research [9].

3) FLASH/e-FLASH: FLASH is a real-world dataset collected from autonomous vehicles equipped with RGB cameras, 16-channel LiDAR, and GPS, with measured mmWave received signal strength as the RF label. Data are recorded in urban canyon environments under both LoS and NLoS conditions, yielding 31,923 synchronized multimodal samples (~20 GB) that capture realistic vehicular dynamics and prop-agation effects [10]. Its extension, e-FLASH, scales up both the number of scenarios and the sensing modalities, further supporting research on mmWave vehicular communication under realistic conditions [11].

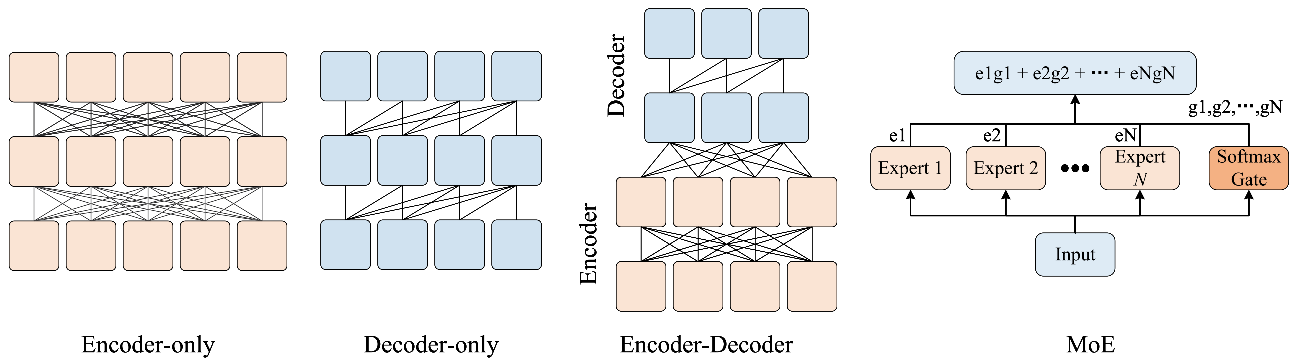
4) DeepSense 6G: Developed under the 6G research initiative, DeepSense 6G is a comprehensive real-world dataset that combines communication data (CSI, beam indices, Signal-to-Noise Ratio (SNR)) with sensing data (RGB images, LiDAR, GPS, IMU) across 30+ scenarios. It spans diverse times of day and environments, including highways, intersections, and dense urban areas, offering high realism and diversity at the cost of substantial data collection and labeling efforts [12].

5) M3SC: A high-fidelity simulation-based dataset designed for multimodal ISAC in vehicular networks. Using AirSim, WaveFarer, and Wireless InSite, it generates aligned RGB images, depth maps, LiDAR point clouds, radar signals, and massive MIMO channel data under dynamic crossroad scenarios with multiple vehicles, variable weather, and time-of-day conditions. This dataset supports a broad range of joint sensing–communication research tasks [13].

Collectively, these datasets provide rich multimodal priors that are essential for pretraining LAM-MSAC, thereby enhancing its ability to generalize effectively to diverse downstream sensing and communication tasks.

Architectures for LAM-MSAC

While datasets provide the foundation for training large models, the architectural design is a critical factor equally as important. Inappropriate design choices may lead to excessive computational costs, limited generalization, or difficulties in scaling to real-world applications. As shown in Fig. 1, the mainstream architectures for LAMs can be



(a)

	Advantages	Limitations	Representative Models
Transformer	Long-range dependency modeling Generalizable across modalities Mature training ecosystem	High computation and memory cost Inefficient for long sequences Resource-intensive training/inference	Encoder-only: BERT Decoder-only: GPT series, LLaMA Encoder–decoder: T5, BART
Mixture-of-Experts	Efficient scaling via subset of experts Lower forward-pass cost Flexible across tasks	Training stability issues Complex expert routing Requires careful load balancing	Switch Transformer GShard Mixtral, DeepSeek-MoE

(b)

FIGURE 1. The mainstream architectures of the LAM. a) The structure illustration. b) The characteristic summarizarion.

broadly categorized into two groups: Transformer and MoE [3] architectures, as detailed below:

1) Transformer: The mainstream development of foundation models has been largely driven by Transformer-based architectures, which represent the earliest and most influential paradigm. As shown in Fig. 1(a), these architectures can be categorized into three main types: encoder-only models, such as BERT [14], which are primarily used for understanding tasks through bidirectional contextual encoding; decoder-only models, such as GPT series and LLaMA, which excel in autoregressive text generation by predicting tokens sequentially; and encoder–decoder models, such as T5 and BART, which adopt a sequence-to-sequence structure well-suited for tasks like translation and summarization. Transformer-based architectures are highly effective in capturing long-range dependencies and have demonstrated strong generalization across diverse tasks. However, they are also characterized by high computational complexity, as the quadratic cost of self-attention scales poorly with sequence length, leading to significant memory and efficiency bottlenecks in large-scale deployments.

2) MoE: To mitigate these limitations, MoE architectures have emerged as a promising alternative. As shown in Fig. 1(a), MoE extends the Transformer framework by incorporating a large pool of expert subnetworks, where only a small subset of experts is activated for each input through a routing mechanism. This design significantly reduces the computational cost per forward pass while allowing the overall parameter size to scale far beyond dense Transformer counterparts. Compared to standard Transformer-based models, MoE architectures offer enhanced flexibility and efficiency, as they balance the benefits of massive model capacity with manageable training and inference costs.

Fig. 1(b) summarizes the two architectures. MoE models are increasingly regarded as a natural evolution of Transformer-based architectures, as they effectively address the scalability and computational limitations of dense Transformers.

Design for LAM-MSAC

Based on the above investigation of multimodal ISAC datasets and the preliminary analysis of LAM architectures, we present the design scheme of LAM-MSAC to address the challenges identified in the section "Challenges in LAMs for Multimodal ISAC". The scheme primarily consists of two components:

1) Modal Feature Extraction: To tackle the challenge of insufficient native support for 6G-relevant modalities, such as RGB images, power measurements, point clouds, and RF signals, we design a set of modality-specific feature encoders. As illustrated in Fig. 2(a), these encoders are tailored to the unique characteristics of the representative modalities introduced in the section "Datasets for LAM-MSAC," which are described as follows:

1. For the visual modality, we employ a ResNet-based image encoder, where the final fully connected layer is replaced by a projection layer to extract normalized feature embeddings of fixed dimensionality. This design leverages the proven ability of ResNet to capture rich semantic and structural information in RGB images, while the residual connections ensure robust training and representation learning. The projection layer further aligns the visual embeddings with those from other modalities, ensuring effective multimodal fusion.
2. For one-dimensional communication data such as power measurements, we adopt a feed-forward network (FFN) with multiple dense layers and ReLU activations, followed by a projection layer. Since power measurements are essentially low-dimensional scalar sequences without strong spatial

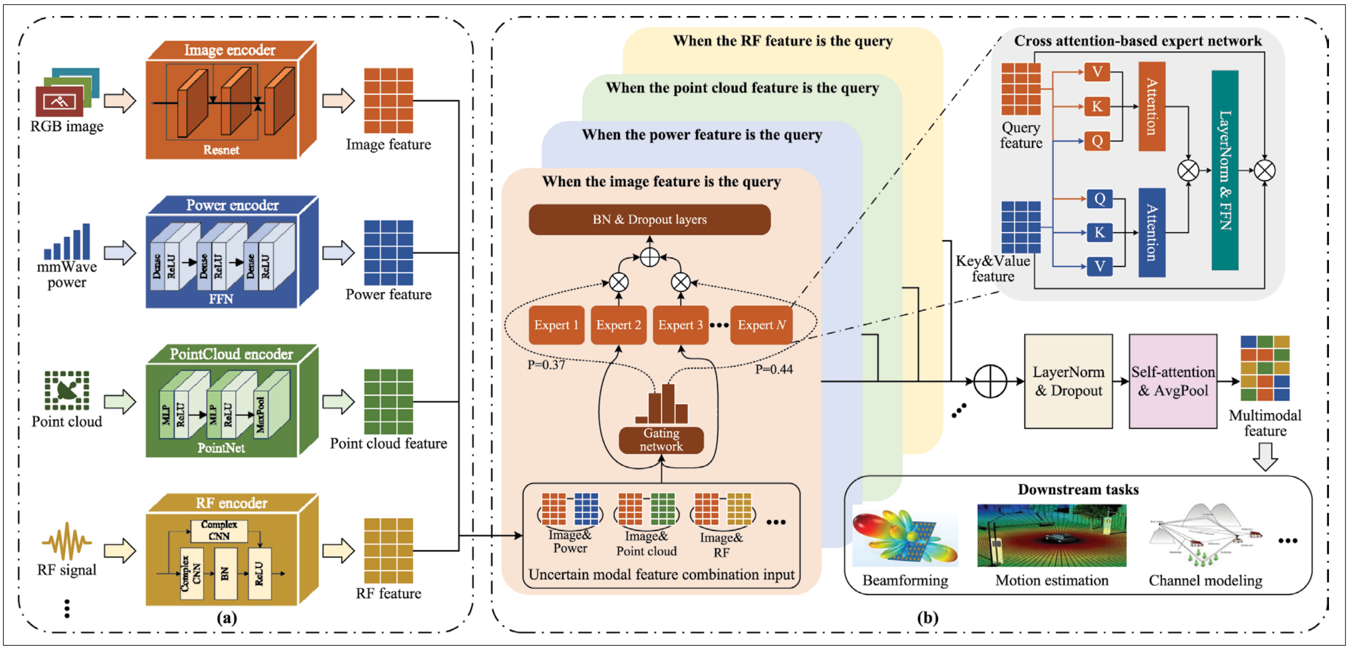


FIGURE 2. The illustration of the proposed LAM-MSAC framework. a) Modal feature extraction. b) MoE-based multimodal fusion.

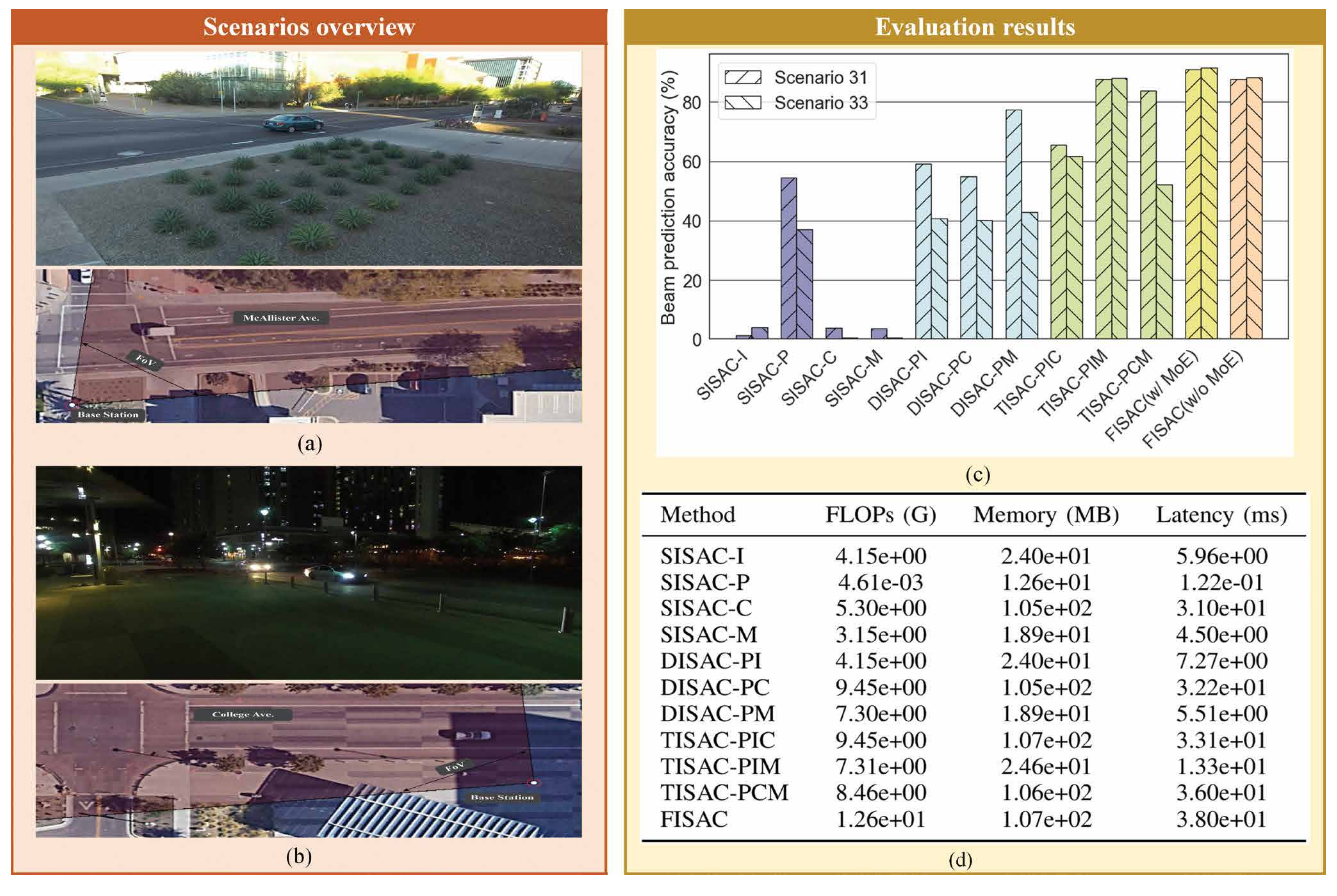
dependencies, a lightweight feed-forward architecture is sufficient to model their nonlinear patterns.

3. For geometric data in the form of point clouds, we design a point net as the encoder. In this net, each point's low-dimensional spatial coordinates are processed through a multi-layer perceptron (MLP), and point-level features are aggregated using max pooling. This architecture explicitly accounts for the irregular and unordered nature of point sets, with MLPs learning local geometric descriptors and max pooling ensuring permutation invariance while capturing global shape information.
4. For RF signals, which are inherently complex-valued, we construct an encoder that integrates complex residual blocks combining complex-valued convolution neural network (CNN), batch normalization (BN), and nonlinear activation, with skip connections to preserve signal integrity. By directly operating in the complex domain, this encoder retains critical amplitude and phase information that conventional real-valued networks would fail to capture. The residual structure further enhances stability and preserves physically meaningful signal characteristics, ensuring that the extracted features remain discriminative and consistent with the underlying RF properties.

Overall, the multimodal encoder aims to flexibly adapt to various input modalities, thereby enabling the extraction and alignment of multimodal data features. This paves the way for subsequent multimodal fusion.

2) MoE-Based Multimodal Fusion: Based on the previous architecture analysis, MoE architectures achieve significantly lower computational cost while maintaining high model capacity and flexibility, through activating only a subset of experts for each input. This property aligns well with the design goals of LAM-MSAC, which requires adaptive handling of heterogeneous multimodal inputs and efficient processing under constrained resources. Therefore, we design a cross-modal fusion module based on MoE, achieving effective integration of heterogeneous modality-specific features.

As illustrated in Fig. 2(b), the module contains a set of parallel experts and a learnable gating network responsible for routing. Each expert is implemented as a transformer block consisting of a cross-attention layer followed by a feed-forward subnetwork. In the cross-attention layer, query embeddings from one modality attend to key-value pairs from another, enabling explicit modeling of inter-modal dependencies. Residual connections and layer normalization are applied after each sublayer to ensure training stability. The gating network takes as input the concatenated multimodal features and outputs a probability



(a)

(b)

(c)

Method	FLOPs (G)	Memory (MB)	Latency (ms)
SISAC-I	4.15e+00	2.40e+01	5.96e+00
SISAC-P	4.61e-03	1.26e+01	1.22e-01
SISAC-C	5.30e+00	1.05e+02	3.10e+01
SISAC-M	3.15e+00	1.89e+01	4.50e+00
DISAC-PI	4.15e+00	2.40e+01	7.27e+00
DISAC-PC	9.45e+00	1.05e+02	3.22e+01
DISAC-PM	7.30e+00	1.89e+01	5.51e+00
TISAC-PIC	9.45e+00	1.07e+02	3.31e+01
TISAC-PIM	7.31e+00	2.46e+01	1.33e+01
TISAC-PCM	8.46e+00	1.06e+02	3.60e+01
FISAC	1.26e+01	1.07e+02	3.80e+01

(d)

FIGURE 3. Scenario overview and evaluation results. a) Scenario 31: A city street in the daytime. b) Scenario 33: A city street at night. c) Beam prediction results under different modal combinations. d) Computing performance analysison NVIDIA Jetson Orin NX.

distribution over all experts. A top- k routing strategy is adopted, where only the k experts with the highest gating scores are activated. Their outputs are aggregated using the normalized gating weights, forming the fused representation for that layer. After expert aggregation, the fused features pass through normalization, dropout regularization, and a final multi-head self-attention pooling layer across modalities, producing the unified representation used by downstream ISAC tasks. To avoid expert collapse and ensure balanced utilization, we incorporate a load-balancing regularizer based on the Kullback-Leibler divergence [15] between gating probabilities and a uniform distribution.

Overall, cross-attention facilitates fine-grained interaction between heterogeneous embeddings, the MoE structure introduces modular specialization for different modality combinations, and the balancing mechanism ensures that model capacity is fully utilized. As a result, LAM-MSAC ensures adaptability to dynamic and heterogeneous modality combinations and balances the heavy computational overhead with ISAC's strict latency and complexity constraints.

Case Study

To evaluate the effectiveness of the proposed LAM-MSAC framework, we conduct a case study that demonstrates its ability to accommodate various combinations of multimodal inputs.

Experimental Settings

1) Data Settings: To evaluate the proposed LAM-MSAC framework under realistic multimodal ISAC conditions, we conduct experiments on Scenario 31 (daytime street) and Scenario 33 (nighttime street) of the DeepSense 6G dataset, as illustrated in Fig. 3(a) and (b). The raw data are first partitioned according to scenario identifiers, where each scenario corresponds to a specific combination of sensing modalities. Each partition is mapped to an independent multimodal dataset class, and a unified dataloader is built to support batched, shuffled, and parallel sampling across scenarios. To ensure balanced training over heterogeneous modalities, in each iteration, mini-batches are drawn from different scenario-specific loaders either uniformly or in proportion to dataset sizes.

Additionally, we apply (i) resizing, normalization, and standardized data augmentation for RGB images, (ii) min-max normalization and NaN cleaning for RF power sequences, (iii) point-cloud centering, unit-sphere scaling, and point-number padding for LiDAR data, and (iv) complex-value normalization and sequence reshaping for radar IQ inputs. These operations follow common practice in multimodal sensing literature and ensure consistent numerical ranges across modalities.

2) Model Settings: Table 2 summarizes the overall network architecture of LAM-MSAC, which supports four potential sensing modalities. All modality-specific encoders are initialized using Xavier uniform initialization, while the parameters in the MoE-based fusion module adopt Kaiming initialization to stabilize expert selection. Layer normalization parameters are initialized to identity settings.

We conduct hyperparameter tuning through grid search on a held-out validation subset, exploring learning rates in {1×10^{-3}, 5×10^{-4}, 2×10^{-4}}, weight decay in {10^{-5}, 10^{-6}}, and different expert activation choices (top-1 vs. top-k). The final configuration uses the AdamW optimizer with learning rate 2×10^{-4}, weight decay 10^{-6}, and a StepLR scheduler that decays the learning rate by a factor of 0.1 every 50 epochs.

3) Baseline Settings: We investigate four modality configurations: SISAC, DISAC, TISAC, and FISAC.

1. SISAC includes four single-modality settings: image (I), power (P), point cloud (C), and mmWave (M).
2. DISAC fuses the power modality with another modality: PI, PC, and PM.
3. TISAC incorporates three modalities: PIC, PIM, and PCM.
4. FISAC corresponds to full-modality fusion. To further analyze the effectiveness of the MoE architecture, we consider two variants of FISAC:
   - FISAC (w/ MoE): the proposed LAM-MSAC framework.
   - FISAC (w/o MoE): a baseline full-modality model with the MoE module removed while keeping the backbone unchanged.

4) Metric Settings: This experimental design enables systematic comparison across modality combinations and model variants. Beam prediction accuracy is used as the evaluation metric, formulated as a multi-class classification task.

Evaluation Results

1) Performance Analysis Across Modalities: Fig. 3(c) demonstrates consistent performance improvements as the number of modalities increases from SISAC to FISAC under both scenarios. In Scenario 31 (daytime street), single-modality results reveal clear differences across sensing types: SISAC-I and SISAC-P provide stronger baselines, whereas SISAC-C and SISAC-M yield significantly lower accuracy. DISAC significantly improves performance, where power information acts as a strong complementary modality. TISAC further enhances accuracy, with TISAC-PIM and TISAC-PCM achieving the highest gains, highlighting the critical role of mmWave when combined with other modalities. Notably, in Scenario 33 (nighttime street), the accuracy of

	Module	Layer Name	Activation
Feature Extraction	Image Encoder	ResNet-50	ReLU
	Power Encoder	3×Dense	ReLU
	PointCloud Encoder	2×Dense + MaxPool	ReLU
	mmWave Encoder	3×Complex ResBlock (CNN+BN)	ReLU
Feature Fusion	Experts	4×Cross-Attn + FFN + Norm	ReLU
	MoE Fusion	Top- k gating	None
Task Heads	Beam Predictor	1×Dense	Softmax
	Blockage Detector	1×Dense	Sigmoid

TABLE 2. Network structure of the proposed LAM-MSAC.

SISAC-I decreases significantly due to low-visibility conditions, but the integration with power, point cloud, and mmWave in DISAC and TISAC effectively compensates for this weakness. Additionally, FISAC delivers the best prediction accuracy in both scenarios, and FISAC (w/ MoE) outperforms FISAC (w/o MoE) across both scenarios. We speculate that this improvement stems from the adaptive nature of the MoE module, where the gating network selects the most suitable subset of expert networks for each input sample. In contrast, FISAC (w/o MoE) treats all samples uniformly, leading to suboptimal processing for certain samples and thus lower accuracy.

2) Computational Cost Analysis: We further evaluated the computational implications of using different modality combinations by measuring FLOPs, GPU memory usage, and inference latency on an NVIDIA Jetson Orin NX. As shown in Fig. 3(d), the results show a clear trade-off between model complexity and beam prediction accuracy. Single-modality models exhibit the lowest computational cost but also the poorest accuracy, especially for vision-only and point cloud–only settings. Introducing additional modalities increases the consumption of computational resources, for example, from 4 GFLOPs in dual-modality ISAC to 7 GFLOPs in tri-modality configurations, but the consistent accuracy improves. The full-modality ISAC model incurs the highest computational load, yet achieves the best performance across all scenes. This illustrates a clear Pareto trend: as more modalities are incorporated, computational cost increases, but the accuracy gain is also significant.

Moreover, among all combinations, "power + mmWave" offers the best efficiency: it achieves a notable accuracy gain while requiring only moderate resource cost, outper-forming other dual-modality settings with similar or even lower computational budgets. In contrast, tri-modality and full-modality fusion yield the highest accuracy but incur substantially higher cost. Thus, dual-modality configurations (e.g., power + mmWave) offer a balanced option for resource-constrained platforms, while full-modality fusion is preferable when accuracy is prioritized and sufficient hardware resources are available.

Open Issues and Future Directions

Despite the promising results of LAM-enabled multimodal ISAC, several open challenges remain to be addressed in future research.

Distributed Multimodal ISAC

This article assumes that all sensing modalities are co-located; however, in practical deployments, sensors such as cameras, radars, and power meters are often spatially distributed. This distributed setting poses challenges, including synchronization among heterogeneous devices, communication bandwidth and reliability constraints, and the difficulty of reconstructing global features from incomplete or noisy local observations. Future research may explore lightweight distributed feature extraction and fusion schemes to reduce communication overhead, federated learning and edge intelligence to enable collaborative training without raw data sharing, and graph-based or consensus-driven algorithms to integrate dispersed multimodal information under network constraints.

Data Privacy and Security

Multimodal ISAC systems rely on data collected from end devices, such as visual and location information, which may contain sensitive user privacy. Direct data sharing is often infeasible due to legal and ethical concerns, making privacy-preserving learning a key challenge. Potential research directions include federated or split learning frameworks, differential privacy mechanisms to mitigate leakage risks, and secure multiparty computation techniques that allow collaborative ISAC model training while protecting raw user data.

Theoretical Analysis of AI-Based Multimodal ISAC

Although AI models provide intelligent solutions for multimodal ISAC, they inevitably introduce significant computational complexity, particularly with LAMs. This raises fundamental questions regarding the trade-offs between communication, sensing, and computation. Extending conventional ISAC theory to a new communication-sensing-computation paradigm is therefore essential. Future work should aim to establish theoretical frameworks for resource allocation, performance bounds, and system scalability when AI-driven models are integrated into ISAC.

Comprehensive Multimodal ISAC Datasets

Current multimodal ISAC datasets have made notable progress in terms of the number and diversity of modalities. However, they largely lack task-specific labeled data. Most existing datasets provide only simplistic labels, such as the optimal beam index, which severely limits the development and training of robust AI models for multimodal ISAC. This gap results in a significant barrier to advancing AI-driven ISAC systems capable of handling complex sensing and communication tasks. Future direction could focus on constructing comprehensive datasets that include rich, task-specific annotations for multiple modalities, covering diverse scenarios such as target detection, localization, and environmental understanding.

Conclusion

This article has presented a comprehensive study on the integration of LAMs into multimodal ISAC. We first identified the key challenges for LAM-enabled ISAC, including limited modality support, the need for dynamic multimodal adaptability, and the tension between large-scale model complexity and limited computation resource constraints. To address these issues, we proposed the LAM-MSAC framework, a MoE-based architecture that can flexibly accommodate arbitrary modality combinations and generate reliable fused representations for downstream tasks. Furthermore, a case study validated its effectiveness, demonstrating that it not only achieves higher accuracy through multimodal integration but also ensures computational efficiency compared to training and maintaining multiple modality-specific models. Finally, we highlighted several open issues and suggested future research directions, including distributed multimodal deployment, dataset standardization, and efficient training strategies. These advances will be crucial for realizing the full potential of LAMs in next-generation ISAC systems and supporting large-scale, intelligent, and resilient IoE applications.

Acknowledgment

This work was supported in part by Jiangsu Major Project on Fundamental Research under Grant BK20243059, in part by the Natural Science Foundation of China under Grant 62531008 and Grant 62301122, in part by the Gusu Innovation Project under Grant ZXL2024360, in part by the High-Tech District of Suzhou City under Grant RC2025001, in part by Xiaomi Young Scholar Award, and in part by Nanjing University–China Mobile Communications Group Company Ltd. Joint Institute.

References

[1] Z. Wei et al., "Integrated sensing and communication enabled multiple base stations cooperative sensing towards 6G," *IEEE Netw.*, vol. 38, no. 4, pp. 207–215, Jul. 2024.

[2] Y. Peng et al., "SIMAC: A semantic-driven integrated multimodal sensing and communication framework," *IEEE J. Sel. Areas Commun.*, vol. 44, pp. 673–688, 2026.

[3] Y. Zhou et al., "Mixture-of-experts with expert choice routing," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35 et al., Eds., Red Hook, NY, USA: Curran Associates, 2022, pp. 7103–7114. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/

[4] F. Jiang et al., "Large language model enhanced multi-agent systems for 6G communications," *IEEE Wireless Commun.*, vol. 31, no. 6, pp. 48–55, Dec. 2024.

[5] F. Jiang et al., "Large AI model empowered multimodal semantic communications," *IEEE Commun. Mag.*, vol. 63, no. 1, pp. 76–82, Jan. 2025.

[6] X. Cheng et al., "Intelligent multi-modal sensing-communication integration: Synesthesia of machines," *IEEE Commun. Surv. Tut.*, vol. 26, no. 1, pp. 258–301, 1st Quart., 2024.

[7] N. Chen et al., "Multimodal heterogeneous data sensing and communication integration for CIoT," *IEEE Trans. Consum. Electron.*, vol. 71, no. 3, pp. 7454–7472, Aug. 2025.

[8] A. Klautau et al., "5G MIMO data for machine learning: Application to beam-selection using deep learning," in *Proc. Inf. Theory Appl. Workshop (ITA)*, Feb. 2018, pp. 1–9.

[9] M. Alrabeiah et al., "ViWi: A deep learning dataset framework for vision-aided wireless communications," in *Proc. IEEE 91st Veh. Technol. Conf. (VTC-Spring)*, May 2020, pp. 1–5.

[10] B. Salehi et al., "FLASH: Federated learning for automated selection of high-band mmWave sectors," in *Proc. IEEE Conf. Comput. Commun.*, London, U.K., May 2022, pp. 1719–1728.

[11] J. Gu et al., "Multimodality in mmWave MIMO beam selection using deep learning: Datasets and challenges," *IEEE Commun. Mag.*, vol. 60, no. 11, pp. 36–41, Nov. 2022.

[12] A. Alkhateeb et al., "DeepSense 6G: A large-scale real-world multi-modal sensing and communication dataset," *IEEE Commun. Mag.*, vol. 61, no. 9, pp. 122–128, Sep. 2023.

[13] X. Cheng et al., "M3SC: A generic dataset for mixed multi-modal (MMM) sensing and communication integration," *China Commun.*, vol. 20, no. 11, pp. 13–29, Nov. 2023.

[14] J. Ni et al., "Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models," 2021, *arXiv:2108.08877*.

[15] S. Ji et al., "Kullback–Leibler divergence metric learning," *IEEE Trans. Cybern.*, vol. 52, no. 4, pp. 2047–2058, Apr. 2022.

Biographies

Yubo Peng (Graduate Student Member, IEEE) (ybpeng@smail.nju.edu.cn) is currently pursuing the Ph.D. degree with the School of Intelligent Software and Engineering, Nanjing University.

Luping Xiang (Senior Member, IEEE) (luping.xiang@nju.edu.cn) is currently an Assistant Professor with the State Key Laboratory of Novel Software Technology, School of Intelligent Software and Engineering, Nanjing University, China.

Kun Yang (Fellow, IEEE) (kunyang@nju.edu.cn) is currently the Director of the Institute of Intelligent Networks and Communications (NINE) and the Chair Professor with the State Key Laboratory of Novel Software Technology, School of Intelligent Software and Engineering, Nanjing University, China.

Jienan Chen (Senior Member, IEEE) (Jesson.chen@outlook.com) is currently a Professor with the National Key Laboratory of Science and Technology on Communications, University of Electronic Science and Technology of China, China.