A Study on Multimodal Large Models for Wireless Channels

Qianyi Qian

HD Shanghai School, Shanghai, China

michael.qian1119@gmail.com

Abstract. The rapid evolution of Fifth-Generation (5G) and Sixth-Generation (6G) communication technologies, coupled with complex and dynamic wireless propagation environments and diverse application scenarios, has imposed increasingly stringent demands on channel modeling. Traditional approaches, which predominantly rely on limited measurements or a restricted set of environmental parameters, often fail to comprehensively capture channel variations caused by the superposition of factors in real world such as diverse building shapes and terrain. This inherent limitation constrains the generalization capability and prediction accuracy of such models. In response, Multi-Modal Large Models (MMLMs) have emerged as a prominent research focus in wireless channel modeling nowadays. By integrating multi-source environmental data—such as visual, radar, and sensor information, MMLMs can more comprehensively characterize propagation environments, uncover correlations between different modalities, and enhance modeling accuracy and generalization as well. Hence, this paper will systematically review the fundamental methodologies and practical challenges of traditional wireless channel modeling. It analyzes the advantages of MMLMs in terms of input information richness, multi-perspective environmental comprehension, spatiotemporal sequence modeling, and generalization in complex scenarios, while also comparing their similarities and differences with conventional approaches. Furthermore, the paper synthesizes recent advancements in the application of MMLMs to channel modeling, evaluates their practical utility and developmental prospects, and discusses key implementation challenges along with potential future research directions.

Keywords: Wireless channels, multimodal model, communication technology.

1. Introduction

The proliferation of Fifth-Generation (5G) mobile communication technology has given rise to complex and dynamic wireless propagation environments alongside diverse application scenarios. Simultaneously, the rapid advancement of Sixth-Generation (6G) technology and the Industrial Internet of Things (IIoT) has increased the demand for a greater number of devices in industrial settings. As a result, the highly dynamic and complex nature of IIoT environments poses new challenges to traditional channel modeling approaches [1]. Since the wireless channel serves as the foundation of wireless communications; it is crucial to accurately understand signal propagation mechanisms and develop precise channel modeling techniques [2]. However, traditional methods—such as deterministic models and stochastic models (e.g., the Geometry-Based Stochastic Model, GBSM)—exhibit inherent limitations in such contexts.

For instance, deterministic models like ray tracing incur high computational costs, while stochastic models such as GBSM offer only limited stochastic representations of channels, making it difficult to predict channel characteristics in unknown environments [1]. Moreover, both traditional stochastic and deterministic channel models struggle to balance prediction accuracy and generalization capability [3]. Oversimplifying environmental information reduces accuracy, whereas modeling all scatterers in detail leads to excessive complexity [2]. Additionally, conventional channel modeling largely depends on limited data measurements and a small set of environmental parameters. This makes it challenging to comprehensively, accurately, and precisely capture channel variations caused by real-world obstacles such as houses, buildings, terrain, and weather. Consequently, the generalization ability and prediction accuracy of these models are severely constrained.

By considering these challenges, and driven by recent rapid progress in multimodal artificial intelligence (AI) and deep learning (DL) based large-scale models, multimodal large models

(MMLMs) for wireless communication channels have gradually emerged as a new research focus in the field of channel modeling [1, 3]. As mentioned in abstract MMLMs offer advantages in input information richness, multi-perspective environmental understanding, spatiotemporal sequence modeling, and generalization in complex scenarios. These capabilities enable a more comprehensive characterization of wireless propagation environments and facilitate the capture of correlations across different modalities [4-5]. Therefore, MMLM-assisted wireless channel modeling can enhance both the accuracy and generalization capability of channel models. This paper aims to review fundamental approaches in traditional channel modeling, highlight the challenges they currently face, and analyze the advantages of MMLMs over conventional methods in terms of input information diversity, multi-perspective environmental comprehension, spatiotemporal sequence modeling, and performance in complex scenarios. The similarities and differences between the two paradigms are also compared [2, 6]. Furthermore, this study anticipates major difficulties in practical implementation and suggests future research directions, thereby providing clearer pathways for subsequent researchers [5, 7].

After reviewing relevant survey literature, this paper will adopt a systematic approach by categorizing and synthesizing different aspects of the field. This structure will not only provide readers with a clear research landscape but also help identify gaps in current studies and uncover potential future research opportunities [1-3]. Accordingly, for the first time, this paper will systematically classify existing work along three dimensions: information sources, model architecture, and modality fusion strategies. This framework will be used to elaborate on recent progress in MMLMs for wireless channel modeling, discuss the practical utility and development prospects of such models, and anticipate major challenges in real-world deployment. Please refer to the following figure (Fig.1) for details.

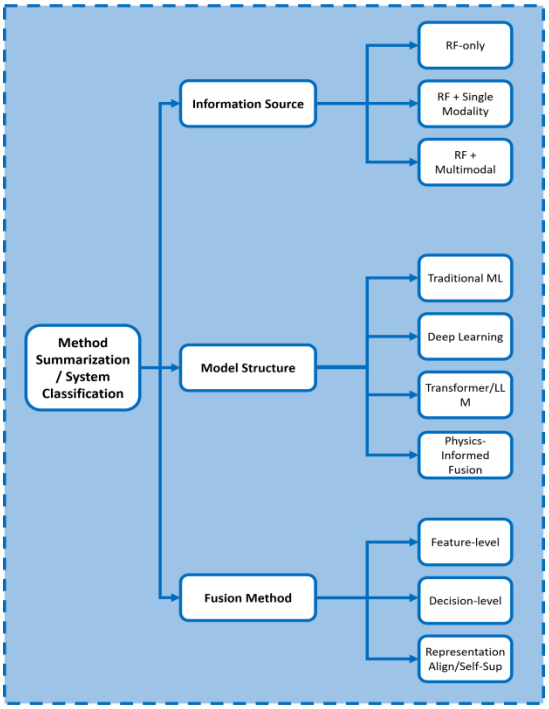

Figure 1. The three major classification modules and their corresponding small branches. (Picture credit: Original)

A massive quantity of literature and representative studies indicate that the evolution of wireless communication technologies, particularly in the era of 6G signals, has intensified the demand for higher communication rates and improved reliability. In this context, AI-based channel information prediction offers significant potential for 6G communication systems. This paper presents relatively earlier compared to other papers, highlighting the limitations of traditional channel prediction methods in high-mobility, high-density 6G environments, emphasizing the comparative advantages of AI-based techniques. By leveraging big data and deep learning (DL), AI-based approaches

adaptively learn underlying channel variation patterns, thereby achieving superior prediction performance [1, 3]. In introducing AI-based channel models, the authors describe Multiple-Input Multiple-Output (MIMO) systems and present a general mathematical expression for channel information prediction in such systems: using N historical channel states to predict M future channel states. Through clear comparative analysis, the paper demonstrates that AI's powerful self-adaptive and autonomous learning capabilities can deeply excavate channel patterns, leverage large volumes of historical data, improve prediction accuracy, and enable proactive system optimization—thereby enhancing spectral efficiency and resource allocation [8-9]. However, even with the integration of MMLMs into wireless channel prediction, the quality and quantity of input data remain critical. Acquiring large volumes of high-quality channel data is extremely challenging; issues such as data loss, noise, and imbalanced distributions can adversely affect model performance [3]. Moreover, compared to traditional channel modeling and prediction methods, MMLMs involve significantly higher complexity and computational demands [4, 7]. Thus, reducing computational complexity while maintaining accuracy remains a major technical challenge.

Similar studies echo this perspective, noting that 6G communication systems not only require high data rates and reliability but also place extreme demands on precise positioning, environmental sensing, and efficient communication link establishment. In other words, there is a strong dependency on a comprehensive understanding of the surrounding environment [10]. The authors focus on Multi-Modal Intelligent Channel Modeling (MMICM) to explore the mapping relationship between communication and sensing [1, 5].

This article systematically structures the field of AI-based channel modeling along three key axes: information sources (RF, single-modal, multi-modal), model architectures (from traditional ML to physics-informed large models), and fusion strategies. We provide a comparative analysis of these paradigms, critically assessing their performance-complexity trade-offs, and conclude by identifying research gaps and future directions for 6G.

2. Classify by Information Source

2.1. Communication Signals and Artificial Intelligence

Regarding current wireless channel modeling methods, we can categorize them based on the information source. Common classifications include communication signals + AI, communication signals + single perception modality, and communication signals + multimodal perception. Communication signals are a necessary component of multimodal large model-assisted channel modeling, while the remaining mainstream classifications consist of AI, single, or multimodal perception.

For communication signals + AI, the essence lies in utilizing only radio frequency measurement data, such as CSI, CIR, and I/Q signals, for data-driven modeling [3]. MIMO-GAN, as a typical example of channel modeling for communication signals + AI, is one of the tools based on Generative Adversarial Networks (GANs) to learn the auxiliary channel modeling generated by MIMO channel distribution [8]. GANs are limited to single input and single output. Traditional GANs generally only accept a single random physical noise input value and subsequently generate a single channel prediction result. Consequently, the generator tends to learn to generate only a few modeling samples, leading to a lack of diversity, which can result in mode collapse and difficulty in adapting to more modeling scenarios. MIMO-GAN, which learns MIMO channel distribution based on GANs, is easier to obtain and can capture complex statistical patterns compared to the single GAN. MIMO greatly enhances the channel capacity and reliability by using multiple antennas to simultaneously transmit and receive signals (multiple-input multiple-output, MIMO).

However, to design and test these advanced MIMO systems, it is first necessary to have a channel model that can accurately simulate their complex propagation characteristics (simulating the propagation properties of radio waves in real cities and indoors, such as multipath transmission and fading during transmission). Just as it is difficult to model complex channels using GANs alone,

modeling channels using only MIMO modeling methods also faces many challenges. Therefore, the goal of MIMO-GAN is to utilize AI models in the form of adversarial networks to repeatedly learn from real channel inputs, thereby generating highly realistic data that matches the real-world distribution of MIMO channel data. This allows for the construction of a more accurate channel modeling model [8]. Similar ANN-based MIMO Channel Playback utilizes artificial neural networks to learn and capture the complex characteristics of new-line-of-sight wireless channels, simulating and reproducing (Playback) the real behavior of the channel through digital twins. Its high efficiency, low cost, and high fidelity make ANN-based MIMO Channel Playback widely used, but its lack of characterization of environmental sets and semantic features limits its generalization ability. For example, ANN-based MIMO Channel Playback models trained in indoor scenarios are difficult to directly apply to new environments such as subway stations [9]. Therefore, to overcome these limitations, the core of research in communication signals + AI has shifted to exploring more advanced models to tap into the potential of radio frequency data itself [3]. Generative modeling, represented by MIMO-GAN, is one of the important directions [8]. Its modeling methods can be mainly divided into the following categories (generative modeling, measurement data fitting, and physical prior integration): Generative modeling, as the name suggests, treats wireless channel modeling as a generation problem. It is applicable to the simulation and testing of MIMO systems, utilizing Generative Adversarial Networks (GANs) to learn the statistical distribution of Multiple-Input Multiple-Output (MIMO) channels. Therefore, it can directly generate samples that are consistent with real channels, without the need for explicit parameterized models. Experimental results of MIMO-GAN (ICC 2022) reveal that the channels generated by the generative modeling approach are highly consistent with the 3GPP standard model in terms of power, delay, and spatial correlation, achieving an impressive average error of 3.57ns in delay.

However, the limitation of MIMO-GAN lies in its dependence on training with large amounts of data and also its limited generalization ability for complex scenarios as well [8]. Measurement data fitting methods primarily utilize Artificial Neural Networks (ANNs) to fit and replay measured 5G MIMO channels, enabling better prediction of channel coefficients even in missing scenarios [9]. Meanwhile, ANN-based MIMO Channel Playback (JSAC 2020) is the first to demonstrate that ANNs can capture the complex characteristics of real channels, thus achieving channel "replay". However, the ANN model's approach to replay measured 5G MIMO channels is limited in scale, making it difficult to adapt to large model scenarios [9]. Physical prior models are of profound significance for enhancing the physical consistency of channel modeling [2, 6]. The essence of their work is to embed physical laws into deep learning, allowing the deep learning model to "understand" the correlation between physical formulas and observed input-output values, thus avoiding the black box effect. It utilizes neural networks to enhance nonlinear fitting capabilities while ensuring physical consistency, significantly improving generalization and interpretability in indoor channel prediction tasks. Due to the embedding of path loss formulas, the physical prior integration method is highly suitable for complex propagation scenarios where physical laws are clear but difficult to analyze. However, due to its complex design, requiring redefinition and constraints for different environments, the design difficulty of the physical prior integration method is extremely high [2, 6].

2.2. Communication Signals and Single Perception Model

As mentioned, although the "communication signal + AI" method has made some progress, its performance ceiling is limited by the insufficient information dimension of radio frequency (RF) measurement data, especially in characterizing specific environmental features, which has inherent limitations [3]. To overcome this bottleneck, researchers have begun to explore the integration of communication signals with other perception modalities, giving rise to the classification of "communication signal + single perception modality" [10-12]. This method aims to compensate for the lack of RF information by introducing new information sources such as conventional images, point clouds, and semantic segmentation results. Due to the simultaneous operation of communication signals and single perception modalities, cross-validation using complementary information can

connect two independent information sources to recognize the same perception target, thereby reducing error rates and improving accuracy [10-12]. At the same time, the combination of communication signals and single perception models can enhance the robustness of the system in complex environments. For example, in environments with minimal changes in lighting (such as at night, strong light, or even smoke or fog), visual sensors used to provide point cloud maps or indoor structure maps may have reduced accuracy or even fail, while RF signals can still work, providing a safeguard for different extreme environments and effectively enhancing the overall robustness of the system [11]. Conversely, when RF signals are severely interfered with or blocked (such as in nuclear power plants, rooms with many floors or walls), another perception modality sensor can serve as the main source of information to ensure continuity [10, 12]. It is worth noting that even though a single modality can significantly improve prediction accuracy in specific environments, it still has numerous limitations, such as being susceptible to lighting effects and having local limitations such as limited point cloud coverage [10-12].

To overcome the limitations of environmental scale, researchers have explored methods using satellite remote sensing images for large-scale path loss prediction [10, 12-13]. For example, the principle of satellite remote sensing image-assisted modeling proposed by Ates et al. is to use satellite panchromatic images as input to a convolutional neural network (CNN) to predict the path loss index and shadow fading in the area. This assisted modeling approach demonstrates for the first time that visual information can directly assist in large-scale channel parameter modeling, and through effective research, it has been shown that this form of prediction error is significantly reduced compared to traditional empirical modeling methods. However, the limitations of the applicable scenarios still restrict the popularity of satellite panchromatic image-assisted modeling. It can only be used for path loss modeling in macro-cellular and geographic coverage areas. This limited application scenario is similar to the signal availability issues caused by insufficient base station coverage during the early stages of 5G penetration. Similarly, satellite images lack 3D structural information, so different lighting and resolutions can easily affect the accuracy of modeling [10, 12-13].

Complementary to satellite images at the macro scale, semantic segmentation provides an alternative approach for modeling complex indoor environments. This method incorporates semantic segmentation results of walls, furniture, and other indoor scenes as additional inputs, embedding them within the prediction channel parameters of PINN. By integrating computer vision results, it enhances the model's perception of the environmental geometric structure. Experimental results also show that in indoor propagation scenarios, its prediction accuracy is significantly higher than that of models using only RF inputs. However, there are very few prompt words suitable for semantic segmentation, making it only applicable to complex indoor scenes such as office buildings and factories, but not suitable for outdoor environments. Additionally, relying too heavily on high-quality segmentation labels leads to high data acquisition costs, which is also a negative factor from an economic perspective. Recognition accuracy is also a major pain point for semantic segmentation. For example, Apple's assistant Siri only uses semantic segmentation to attempt to guess the questions users are asking, but it struggles to make efficient choices when faced with multi-source inputs. For instance, "Please help me check the weather of the friend I last spoke with," is a "complex task" that pure semantic segmentation finds difficult to understand, requiring the opening of the contact list, call history, and weather information simultaneously. Similarly, at the auxiliary channel modeling level, when users input complex multi-source structural information, the accuracy of semantic segmentation still needs to be guaranteed.

On the other hand, compared to providing semantic labels, point cloud images can provide more precise geometric structure information. This auxiliary modeling approach utilizes LiDAR point clouds to provide the locations of scatterers in the environment, commonly used for modeling small-scale channel characteristics. By introducing multi-path characteristics closely related to the physical environment into the channel model, it more realistically reconstructs the propagation effects in NLOS scenarios. In scenarios such as vehicle networking, drones, and complex urban environments,

point cloud images can better reflect the three-dimensional positioning of scatterers, effectively providing accurate location information for sensing or channel modeling. However, due to the high cost of point cloud acquisition and limited coverage, the popularization of point cloud-assisted modeling still requires some time to mature [14].

2.3. Communication Signals and Multi-sensory Modalities

However, a single perception modality often only provides information on a certain dimension of the environment (such as satellite images providing macro-topography, point clouds providing geometric structure, and semantic segmentation providing object categories), and its capabilities still have boundaries [4-5, 14]. Therefore, can multiple perception modalities be synergistically utilized to form a more comprehensive and robust understanding of the environment? This leads to scholars' research on "communication signals + multimodal perception".

This method aims to construct a more superior channel model by fusing communication signals with various data such as images, depth maps, point clouds, maps, and GPS. Among them, many wireless channel models such as CSI-CLIP, MFEF-Net IIoT, and dual-camera prediction models belong to this category [4-5, 14]. The essence of CSI-CLIP is to transfer the CLIP model from the visual language domain to the RF language domain. Its core capability is to learn images and text, describe, and understand the associations in the same semantic state. This enables zero-shot image classification and image-text retrieval capabilities. CSI is extracted from wireless signals and represents physical layer data that characterizes environmental changes, such as the impact of object movement on wireless network signals, which is captured by CSI [4]. Another approach is to use MFEF-Net Network, which belongs to the category of communication signals + multimodal perception. As the name suggests, it extracts features by simultaneously processing and fusing data from multiple different modalities such as RGB images, point clouds, and GPS, and uses an attention mechanism for weighted prediction to obtain more robust and accurate perception results.

Notably, the multi-attention mechanism fusion is a noteworthy point. Essentially, it uses a cross-attention module to allow one of the multimodal perception data, such as RGB feature maps, to "ask" another perception data, such as infrared feature maps, which areas deserve attention, or vice versa. This not only enables adaptive weighting and complementarity between features but also greatly improves the accuracy of communication signals [5].

The core principle of the IIoT dual-camera prediction model is to use two cameras of different modalities to produce high-precision dimensional measurements and three-dimensional positioning through cross-observation, allowing data to be truly "seen" and observed, providing real reference value for various physical input noises in subsequent channel modeling [14]. For instance, the dual-camera configuration of RGB+Depth or RGB+RGB is the most common. It involves recognizing the target and subsequently measuring the distance to derive a high-precision 3D model, which is widely used in measurements prior to indoor channel modeling. On the other hand, simulating real human eyes by calculating parallax prediction and depth information to reconstruct a 3D scene is more commonly employed in fields such as AGV vehicle navigation rather than channel modeling. Despite the extensive fields involved in this method, its high demand for bandwidth and lightweight optimization of the model remain challenges that need to be addressed, laying a solid foundation for further expansion in future fields. Later scholars need to consider the integrity of video data transmission, balancing transmission stability, optimization, and transmission quality to find the right point [14].

Multimodal large models have more extensive and comprehensive input data compared to the former [5, 8]. For example, multimodal self-supervised modeling uses CSI and CIR as natural multimodal inputs, employing self-supervised contrastive learning to train a general channel foundation model. This technology is the first multimodal channel foundation model, utilizing cross-modal consistency learning. This approach reduces the error rate by 22% in localization tasks and improves accuracy by 1% in book management tasks. Furthermore, it facilitates cross-task and cross-scenario channel modeling and prediction. However, due to the high complexity of data input, the

training cost is very high, requiring large-scale data to support multimodal self-supervised modeling [4].

A similar multimodal fusion network (MFEF-Net) achieves modal weighting through an attention mechanism after fusing RGB color images, point clouds, and GPS positioning data. The introduction of an adaptive attention mechanism is the main reason for the improvement in accuracy and robustness compared to single-modality prediction. It is primarily suitable for path loss prediction in urban vehicle networking scenarios. However, the need for complex multimodal synchronous acquisition platforms has become a limiting factor in the advancement of MFEF-Net-assisted channel modeling [5].

The IIoT dual-camera prediction model mentioned earlier is a typical representative method of industrial IoT multimodal prediction models [14]. This method utilizes dual-end image information to assist channel modeling in industrial scenarios for the first time, combining synchronous images from the transmitting and receiving ends and physical parameters of transceivers to jointly predict signal power and RMS DS using CNN and MLP. Similarly, its experimental results still demonstrate that its prediction error is significantly lower than traditional methods. However, the layout of dual-end cameras and the synchronization of data between the transmitting and receiving ends still pose engineering challenges [14].

The large model and beam prediction, known as BeamLLM, utilizes video frames captured by base station cameras to extract user location features through object detection and then convert them into input prompts for LLMs to predict beam directions. It is the first time that a pretrained LLM has been used for beam prediction tasks in visual sequences. Experimental results show that BeamLLM still achieves significantly better performance than traditional LSTM modeling methods under few-sample conditions. It is commonly used for rapid beam alignment in millimeter-wave and terahertz high-frequency communications. However, in terms of model complexity, high complexity leads to higher inference latency, making it unsuitable for direct deployment on terminal devices [7].

To sum up, the wireless channel modeling method has experienced an evolution process from internal mining to external integration [1-4]. The "communication signal+ai" method, which played a key role in the early stage but has exposed the lack of generalization, has indeed deeply excavated the statistical laws of RF data itself, but the lack of environmental information has led to the limitation of generalization ability, which has prompted scholars to explore new channel modeling methods [1, 3]. Therefore, the "communication signal+single sensing mode" method significantly improves the ability to depict environmental characteristics by introducing external multi-source sensing sources such as images and point clouds [10-12]. But inevitably, each mode has its own competent application scenarios and limitations at different levels [10-12]. Therefore, the search for new modeling methods has not stopped. The latest "communication signal+multimodal sensing" research shows that integrating a variety of complementary sensing information is an effective way to achieve high-precision, high robustness and cross scene channel modeling [4, 7, 14]. Looking forward to the future, scholars can take this as the exploration goal, combine multimodal input with physical prior and large model base, and design a set of efficient reasoning mechanism, so as to build the next generation of portable and scalable channel foundation model [2, 4, 7].

3. Classification According to Model Structure

Different large models belong to different model structure classifications. The recently popular AI models such as DeepSeek and ChatGPT belong to typical deep learning models (DL), while earlier traditional machine learning models such as KNN, SVM and decision tree also lay a good foundation for the field of channel modeling. They were often used in simple prediction and channel classification based on measurement data in the early stage. As the cornerstone of the early indoor positioning system, the traditional machine learning model uses KNN and other algorithms to match the position according to the received signal strength indication (RSSI) or channel state information (CSI) of each observation point as the "fingerprint". At the same time, the traditional machine learning

model has been very mature in the field of simple channel classification, such as identifying LOS and NLOS scenarios. Because the algorithm is fully transparent in training, its interpretability is relatively clear, and due to the less resources required for training and reasoning, the computational overhead is small, so it is suitable for deployment on devices with low on-line availability of resources. And when the input data is limited, the model can still be used for machine training and effective work.

With the advancement of the times, in the field of auxiliary infinite channel modeling, similar to the DL model just mentioned has gradually played a vital role in this field [3, 8-9]. The deep learning model mainly consists of convolutional neural network (CNN), large-scale input-output (MIMO) hybrid precoding, CSI wireless sensing, recurrent neural network (RNN), long-term short-term memory network and generative countermeasure network (GAN) [8-9]. The main operation mode of helping wireless channel modeling can be regarded as identifying the nonlinear relationship of similar signal changes such as radio waves or raindrops and high-dimensional characteristics such as indoor layout from CSI. In addition, due to the addition of recurrent neural network (RNN) and long-term short-term memory network, the channel modeled with the aid of deep learning model even contains the characteristics of channel changing with time, which not only reflects the strong generalization ability, but also can be used to predict the dynamic changes of channel capacity and quality. Therefore, the deep learning model has been widely used in path loss prediction and channel generation [8-9].

Secondly, transformer with large model and physical prior combination model are also relatively advanced large models to assist wireless channel modeling in recent years [2, 4, 6-7]. The former mainly uses the cross-modal attention mechanism to realize the interaction of multiple features, which is similar to the multimodal perception mentioned above. It mainly relies on wireless signals such as CSI, RSSI and i/q data, as well as the visual information captured by cameras such as images/point clouds, 3D point clouds and other spatial information to observe geographic information such as maps and building layout [4, 7]. So as to establish the relationship between these heterogeneous multi-source data, rather than just connecting into a long vector input to the model, which is what the traditional machine learning model lacks. The attention mechanism indicates that the system can associate the input value accepted at present with the memory network. For example, the pronoun "it" may be understood as "apple" mentioned earlier (under a specific conversation scenario). Similar attention mechanisms at the channel modeling level will improve the overall generalization ability of the system, "remember" the previous "conversation" scenario, and continuously update and iterate in the next similar situation and environment, extending this mechanism to different modes [4, 7].

The physical prior combination model aims to improve generalization and ensure physical consistency. If the observed actual phenomenon is contrary to the previously proposed physical laws, the interpretability of this situation will become extremely difficult [2, 6]. These basic physical laws mainly include the attenuation of signal strength with distance (path loss), the physical properties of reflected, diffracted and scattered waves, and the superposition effect of inphase, antiphase waves such as superposition. If these basic physical principles are violated, it will be physical inconsistency, which is not only considered absolutely not allowed in safety critical areas such as autonomous vehicles, but also will sharply increase the difficulty of phenomenon explicability, and even lead to unexplainable results. In addition, the combination is also to better get rid of the limitations brought by the traditional data-driven model, because the deep learning convolutional neural network and other algorithms are easy to imitate the statistical fields in the training data end-to-end, so the physical inconsistency may be reflected in the plates not covered by the training. In addition, relying on a large amount of data to learn these laws that can be simply described by the laws of Physics (such as the data loss mentioned above) leads to data inefficiency. As a solution, previous scholars in this field tried to embed the laws of propagation physics into the framework of deep learning. PINN (Physics informed neural networks) compared the difference between the predicted value of the model and the true measured value by combing the constraint loss function, so as to ensure that the model learned from the actual data. At the same time, the physical inconsistency can also be avoided by substituting the predicted value predicted by the model into the currently known physical equation and calculating the degree to which it meets the physical equation. Therefore, the physical prior model ensures the

explicability of certain phenomena, improves the generalization, and ensures the consistency of physics [2, 6].

4. Classification According to Mode Fusion Mode

Fusion more intuitively endows different models with the ability of cross modal reasoning, rather than telling users what has changed through input data and not explaining why, as with low interpretative modes. For example, MFEF-net is a good example of feature and fusion. It performs multi-source physical input value splicing or transmembrane attention interaction in the feature extraction stage, so as to allocate weights for different modal features [5].

In the decision level fusion, each mode is predicted separately, and then weighted or voted at the decision level. This module is mostly used for post-processing in the system deployment phase, which is flexible, but its fusion depth is limited.

Representation alignment and sub supervised learning are another mode fusion method. Its essence is to map unreasonable modes into the shared representation space through sub supervision or comparative learning. For example, CSI-clip often uses the consistency of the dual observation means of CSI and CIR for comparative learning. Its essence is to learn image recognition and CSI action recognition by inputting a large amount of annotation free knowledge. The pre-similarity of the same object or radio wave will be very high, ensuring the accuracy and accuracy of information transmission [4].

Through the analysis from the three dimensions of information source, model structure and fusion mode, we know that the frontier research is gradually moving from single RF data to multimodal fusion, that is, only RF to the combination of single-mode communication and then to multimodal auxiliary wireless communication channel modeling [3-5]. The structural level of similar models has gradually changed from the original traditional model to the deep learning model, and then to the transformer/LLM and other more advanced models. Ultimately, in order to ensure the physical consistency, it has turned to the direction of the physical prior combination model [2, 8-9]. A chronological order (from top to bottom) mind map is shown below.

Finally, through the analysis of fusion methods, this paper elaborates three different fusion dimension methods: feature fusion, decision fusion and representation alignment [4-6]. It is not limited to the large-scale model fusion at the feature level, but turns to the decision fusion at different levels and later at the decision-making level. Finally, the characterization alignment mentioned above achieves the accuracy of identification by comparing and learning massive parameters, which greatly improves the possibility of assisting wireless channel modeling [4]. Therefore, in future research, we need to build a general wireless channel basic model with multimodal input, physical consistency and University reasoning ability, so as to improve the existing wireless communication quality [2, 4, 7].

5. Analysis and Comparison

In summary, this paper will compare the advantages and disadvantages of different wireless channel modeling methods through analyzing the modeling accuracy, generalization ability, computational complexity and real-time of different models, data dependent on the difficulty of acquisition and the depth of fusion by combination of a clear mind map and comparison below (Fig.2).

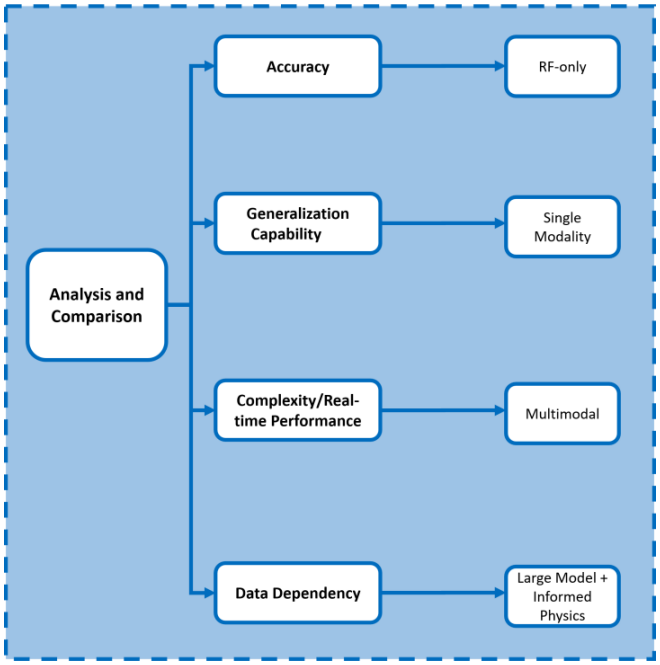

Figure 2. Characteristics of different models assisting wireless channel modeling. (Picture credit: Original)

The RF only modeling method mentioned at the beginning has high-cost performance in terms of the ease of data acquisition, so its computational overhead is very low, and it is suitable for rapid modeling and simulation. Even though this modeling method has been developed more mature, it still lacks the support of environmental information, resulting in poor generalization ability, which makes it difficult to adapt to future 6G complex scenes [1, 3].

Uni-modal method can significantly improve the prediction accuracy in specific scenarios such as satellite image/semantic segmentation. Compared with the RF only modeling method, the single-mode method combines more spatial or suggestive information and has higher prediction accuracy and accuracy. However, the single-mode method is too easy to be limited by the modal defects such as the image under strong light or darkness [10-12].

Multi-modal modeling method combines a large number of multi-source information, can more comprehensively characterize the propagation environment, and performs best in accuracy and robustness [4-5, 14]. On the contrary, due to the extremely high cost of collecting multiple information (such as point cloud image shooting and RGB camera collection) and the large scale of the model itself, there are still challenges at the real-time level [4-5, 14].

Table 1. Summary Comparison of Different Methods

Method Category	Accuracy	Generalization	Complexity/ Real-time	Data Dependency	Fusion Method	Advantages	Limitations	Accuracy
RF-only	★★★	★★	★★★★★	★★★★★	None	Easy data acquisition, low computation cost	Lacks env info, poor generalization	RF-only
Single Modality	★★★★	★★★	★★★★	★★★★	Feature/Decision-level	Significant accuracy boost in specific scenarios	Relies on single modality, obvious limitations	Single Modality
Multimodal	★★★★★	★★★★	★★★	★★	Feature/Align/ Self-Sup	Best accuracy and robustness	High data acquisition & alignment difficulty	Multimodal
Large Model + Physics	★★★★★	★★★★★	★★	★★	Representation Align + Physical Constraints	Strong generalization, good interpretability	High computation/deployment cost, lack of standards	Large Model + Physics

Because of the embedding of the basic laws of physics, the large model and the physical bright model have the characteristics of high accuracy and high interpretability, and can also take into account the generalization and even have the ability of cross task migration [2, 4, 6-7]. These comparisons are shown on the Table 1.

6. Conclusion and Critical Assessment of the Future

Table 2. Comparison and Classification of Channel Modeling Methods

Classification Dimension	Sub-category	Representative Work	Technical Indicators	Advantages	Limitations
Information Source	Comm Signal Only (RF-only)	MIMO-GAN (ICC 2022); ANN-based MIMO Playback (JSAC 2020)	Learns statistical data/generates distribution from CSI/CIR	Low data complexity, applicable	Cannot use env semantics, poor generalization
	Comm Signal - Single Modality	Satellite Image Path Loss (IEEE Access 2019/2020); PINN-Semantic Seg (2024)	RF - Image Seg/Cloud Input → CNN/MLP Prediction	Accuracy improvement in range	Single modality limitations large
	Comm Signal - Multimodal	CSI-CLIP (2024); MFEF-Net (RGB+PC+GPS); IoT Dual-cam (2025)	Multi-freq data → Transformer Self-supervised Analysis	Rich representation, strong robustness	High acquisition difficulty, high complexity
Model Architecture	Traditional ML	SVM, KNN, Decision Tree	Simple feature learning & classification	Easy implementation, interpretability	Limited performance
	Deep Learning	CNN, RNN, GAN	Nonlinear feature extraction & generative modeling	High fitting capability	High data dependency
	Transformer/LLM	LLM/WM (2025); BeamLLM (2025)	Cross-modal generation & large-scale prediction	Generalization, transferability	High inference cost
	Physics-Informed	PINN-based Model (Zhu 2024)	Combine propagation model + SDL	Ensures consistency, enhances generalization	Complex design
Fusion Method	Feature-level Fusion	MFE+Net	Concatenation/Attention Weighting	Flexible fusion	Difficult concept alignment
	Decision-level Fusion	Multimodal Independent Prediction + Weighting	Post-processing fusion	Simple implementation	Insufficient information utilization
	Decision-level Fusion	CSI-CLP	Contrastive Learning Shared Representation Space	Strong cross-modal adaptability	High training cost

With the large-scale popularity of 5G today, people's yearning and exploration for 6G are also gradually rising. Therefore, leading-edge scholars are slowly evolving from the original RF only to multimodal to meet the needs of environmental perception in 6G scenarios [1, 3]. However, the current multimodal research relies too much on the experimental environment and lacks real-scale network authentication. Therefore, the multimodal large-scale model combined with a priori physical model to assist wireless channel modeling still needs to solve some essential problems [4, 7]. From the previous paragraphs, we fully understand that the large multimodal model significantly improves the generalization and accuracy of wireless channel modeling [4-5, 14]. But the high price and cost is the huge consumption of data and computing. Therefore, in future research, we need to find a balance between performance and overhead, solve the cost problem and retain the advantages of multimodal large model for auxiliary channel modeling [4, 7]. At the same time, in terms of interpretability and credibility, physical prior is one of the best effective means to ensure these two aspects. However, at present, most of the work in this area is still in the "shallow integration", that is,

it is possible that there is still the possibility of physical inconsistency or lack of depth in the case of implanting some basic physical rules. Therefore, in the future, it is necessary to achieve the unity of theory and model through theoretical party lessons that are more in-depth and more convenient for multimodal large-scale model understanding [2, 6]. In addition, in future research, researchers are also worth exploring edge cloud collaborative reasoning, knowledge distillation, multimodal data compression and its application to ensure effective and lightweight data storage [2, 6].

To sum up, by Table 2, we understand that mature RF only technology is still valuable in simple scenarios [3]. However, compared with the RF detection assisted channel modeling method, the single-mode method significantly improves the accuracy of modeling in specific application scenarios after fusing a small part of image information. However, it is difficult to adapt to complex environments or extreme scenarios such as high light exposure [10-12]. Therefore, the multimodal and large model methods have been verified to be reasonable in many papers, ensuring the ability to show ultra-high intensity in complex environments while achieving high generalization, accurate prediction and high accuracy [4-5, 14]. However, due to the high cost and difficulty in acquisition, multi-modal and large model assisted channel modeling methods still need time to precipitate. After solving these problems, they can be slowly accepted by the market and successfully popularized in mass production [4, 7]. It can be seen that the future development direction should be the integration paradigm of multi-modal and large-scale models embedded in the physical prior model and University reasoning [2, 4, 6-7].

References

[1] Chen Huang, Ruisi He, Bo Ai, Andreas F. Molisch, Buon Kiong Lau, Katsuyuki Haneda, Bo Liu, Cheng-Xiang Wang, Mi Yang, Claude Oestges, Zhangdui Zhong. IEEE Transactions on Antennas and Propagation, Artificial Intelligence Enabled Radio Propagation for Communications—Part II: Scenario Identification and Channel Modeling, 2022.

[2] Ethan Zhu, Haijian Sun, Mingyue Ji. Physics-Informed Generalizable Wireless Channel Modeling with Segmentation and Deep Learning Fundamentals, Methodologies, and Challenges, 2024.

[3] Mi Yang, Ruisi He, Bo Ai, Chen Huang, Chenlong Wang, Yuxin Zhang, Zhangdui Zhong. IEEE Communications Magazine. AI-Enabled Data-Driven Channel Modeling for Future Communications, 2024.

[4] Jun Jiang, Wenjun Yu, Yunfan Li, Yuan Gao, Shugong Xu. ICMLCN. MIMO Wireless Channel Foundation Model via CIR-CSI Consistency, 2025.

[5] Kai Wang, Li Yu, Jianhua Zhang, Yixuan Tian, Eryu Guo, Guangyi Liu. China Mobile Research Institution. Multi-Modal Environmental Sensing Based Path Loss Prediction for V2I Communications, 2024.

[6] Ethan Zhu, Haijian Sun, and Mingyue Ji. arXiv: 2401.01288v1 [cs.IT] 2 Jan 2024. Physics-Informed_Generalizable_Wireless_Channel_Modeling_with_Segmentation_and_Deep_Learning_Fundamentals_Methodologies_and_Challenges, 2024.

[7] Can Zheng, Jiguang He, Guofa Cai, Zitong Yu, Chung G. Kang. arXiv: 2503.10432v2 [cs.LG] 27 Jun 2025. BeamLLM: Vision-Empowered mmWave Beam Prediction with Large Language Models, 2025.

[8] Tribhuvanesh Orekondy, Arash Behboodi, Joseph B. Soriaga. IEEE International Conference on Communications. MIMO-GAN: Generative MIMO Channel Modeling, 2022.

[9] Xiongwen Zhao, Fei Du, Suiyan Geng, Zihao Fu, Zhongyu Wang, Yu Zhang, Zhenyu Zhou, Lei Zhang, Liuqing Yang. IEEE Journal on Selected Areas in Communications. Playback of 5G and Beyond Measured MIMO Channels by an ANN-Based Modeling and Simulation Framework, 2020.

[10] Chenlong Wang, Bo Ai, Ruisi He, Mi Yang, Shun Zhou, Long Yu, Yuxin Zhang, Zhicheng Qiu, Zhangdui Zhong, Jianhua Fan. IEEE Transactions on Machine Learning in Communications and Networking. Channel Path Loss Prediction Using Satellite Images: A Deep Learning Approach, 2024.

[11] Sreelal M. Mana, Kerolos G. K. Gabra, Sepideh M. Kouhini, Malte Hinrichs, Dominic Schulz, Peter Hellwig, Anagnostis Paraskevopoulos, Ronald Freund, Volker Jungnickel. IEEE Access. LIDAR-Assisted Channel Modelling for LiFi in Realistic Indoor Scenarios, 2022.

[12] Hasan F. Ates, Syed Muhammad Hashir, Tonçer Baykas, Bahadır K. Gunturk. IEEE Access. Path Loss Exponent and Shadowing Factor Prediction from Satellite Images Using Deep Learning, 2019.

[13] Omar Ahmadien, Hasan F. Ates, Tuncer Baykas, and Bahadır K. Gunturk. IEEE Access. Predicting Path Loss Distribution of an Area from Satellite Images Using Deep Learning, 2020.

[14] Shudong Zhou, Yu Liu, Rui Wang, Ziheng Li, Zhichao Xin, Jie Huang, and Ji Bian. IEEE Internet of Things Journal. A Multimodal Predictive Channel Model Based on Dual-Camera Images for IIoT Communications, 2025.