GOAL-ORIENTED SEMANTIC COMMUNICATION AND NETWORKING

Streamlined Transmission: A Semantic-Aware XR Deployment Framework Enhanced by Generative AI

Wanting Yang, Zehui Xiong, Tony Q. S. Quek, and Xuemin Shen

Abstract

In the era of 6G, featuring compelling visions of digital twins and metaverses, Extended Reality (XR) has emerged as a vital conduit connecting the digital and physical realms, garnering widespread interest. Ensuring a fully immersive wireless XR experience stands as a paramount technical necessity, demanding the liberation of XR from the confines of wired connections. In this paper, we first introduce the technologies applied in the wireless XR domain, delve into their benefits and limitations, and highlight the ongoing challenges. We then propose a novel deployment framework for a broad XR pipeline, termed "GeSa-XRF", inspired by the core philosophy of Semantic Communication (SemCom) which shifts the concern from "how" to transmit to "what" to transmit. Particularly, the framework comprises three stages: data collection, data analysis, and data delivery. In each stage, we integrate semantic awareness to achieve streamlined transmission and employ Generative Artificial Intelligence (GAI) to achieve collaborative refinements. For the data collection of multi-modal data with differentiated data volumes and heterogeneous latency requirements, we propose a novel SemCom paradigm based on multi-modal fusion and separation and a GAI-based robust superposition scheme. To perform a comprehensive data analysis, we employ multi-task learning to perform the prediction of field of view and personalized attention and discuss the possible preprocessing approaches assisted by GAI. Lastly, for the data delivery stage, we present a semantic-aware multicast-based delivery strategy aimed at reducing pixel level redundant transmissions and introduce the GAI collaborative refinement approach. The performance gain of the proposed GeSa-XRF is preliminarily demonstrated through a case study.

Introduction

With the compelling visions of metaverses and digital twins, extended reality (XR), as a bridge between the real and virtual worlds, has garnered significant attention. To ensure a fully immersive experience, eliminating the tether of XR stands as a paramount technical imperative. Given the inherent scarcity of radio resources, there exists an urgent necessity for a novel communication framework customized for XR. Opportunely, the emergence of semantic communication offers fresh insights to address this requirement.

The pivotal concept in SemCom is shifting from the traditional concern of "how" to transmit information to the consideration of "what" to transmit, thus streamlining the transmission [1]. While numerous review articles highlight the substantial benefits of SemCom to XR scenarios, the comprehensive implementation methods for the entire XR pipeline, encompassing data collection, analysis, and delivery, have been notably deficient [1], [2]. The existing monolithic SemCom frameworks struggle to flexibly meet the varying data volumes, data modalities, and heterogeneous quality of service requirements. Furthermore, for the most mature deep learning (DL)-based SemCom, the unavoidable error floor during the training inevitably results in the degradation of the user's immersive experience in reconstructed virtual scene at the mobile XR devices [3].

Fortunately, the advent of generative artificial intelligence (GAI) has ushered in a promising opportunity to revolutionize the SemCom framework [4], [5]. In the paradigm shift from SemCom to generative SemCom, the extraction of semantic information evolves into the determination of the PROMPT, providing enhanced flexibility in communication system design, while alleviating transmission burdens [6]. Furthermore, the diffusion model, renowned as one of the foremost GAI technologies due to its exceptional ability in generating high-quality images, shows considerable potential for enhancing high-definition XR experiences. Nonetheless, this advancement comes with the drawback of prolonged inference latency, particularly when tasked with generating content of substantial data volumes. Given that XR itself is a computationally sensitive application, the wholesale transfer of the existing generative SemCom into XR undoubtedly brings more challenges to instant virtual scene rendering. As such, the

Wanting Yang, Zehui Xiong (corresponding author), and Tony Q. S. Quek are with Pillar of Information Systems Technology and Design, Singapore University of Technology and Design, Singapore 487372; Xuemin Shen is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada.

Digital Object Identifier:
10.1109/MNET.2024.3414398
Date of Current Version:
18 November 2024
Date of Publication:
14 June 2024

Simultaneously, we introduce a three-dimensional multi-user quality of experience (QoE) evaluation metrics, encompassing weighted resolution based on tile significance, playback smoothness, and playback synchronization to guide strategy optimization.

strategic integration of GAI into the XR pipeline, capitalizing on its strengths to enhance overall XR performance while minimizing computational burdens, becomes a thoughtful imperative.

To fill the research gaps, we systematically review the currently available technologies, delve into their benefits and limitations, and highlight the ongoing challenges in the XR domain. Based on this, we propose a Semantic-aware XR Deployment Framework Enhanced by GAI, termed "GeSa-XRF". In contrast to the existing purely point-to-point SemCom framework, GeSa-XRF is structured around three distinct functional tasks, each addressing critical aspects across the stages of data collection, analysis, and delivery. Leveraging the innovative potential of SemCom and GAI, semantic awareness and GAI techniques are seamlessly integrated into each task. The specific contributions are outlined as follows:

- In the data collection stage, distinct from the existing multi-modal SemCom approach targeted at decision-making [7], we introduce a novel SemCom paradigm centered on multi-modal fusion and separation for reconstructing multi-modal signals. Furthermore, rather than opting between SemCom and traditional communication [2], we harness the robustness of SemCom and propose a GAI-based superposition scheme to manage the collection of multi-modal data with varied data volumes and heterogeneous latency requirements, with a trade-off among performance, inference latency, and training complexity.
- In the data analysis stage, considering the limitations of computing resources on mobile devices, we employ the tile-based rendering method. Subsequently, by leveraging multi-task learning (MTL) techniques, we develop a unified algorithm for attention assessment and field of view (FoV) prediction. Utilizing the attention assessment results, we assign varying degrees of semantic significance to different tiles. Moreover, we classify the tiles into foreground and background segments based on their semantic importance. For the background tiles, we employ GAI-based proactive preprocessing guided by the FoV prediction results to enhance the overall XR performance.
- In the data delivery stage, we incorporate the semantic significance obtained during the data analysis stage into the delivery strategy to reduce pixel-level redundant transmissions. Specifically, the strategy involves semantic-aware multicast cluster decision, semantic-aware transcoding, and semantic-aware scheduling for the foreground tiles. Simultaneously, we introduce a three-dimensional multi-user quality of experience (QoE) evaluation metrics, encompassing weighted resolution based on tile significance, playback smoothness, and playback synchronization to guide strategy optimization. Furthermore, our case study delves into semantic-aware cluster decision-making and transcoding, primarily showcasing the performance improvements facilitated by GeSa-XRF.

XR Deployment Overview and Challenges

To maintain generality, we focus on a broad three-stage XR deployment pipeline. The available techniques and the challenges in each stage are discussed below:

Triggering Stage: Data Collection

The effectiveness of immersive interactions hinges upon the integration of multi-sensory data exhibiting in distinct modalities with varying data volumes. For instance, data streams like audio and video for environment mapping may require data rates exceeding 100 Mbps. Conversely, data volumes for eye tracking and haptic commands are often just a few bytes. Furthermore, human brain reaction times to visual, auditory, and tactile stimuli differ significantly.[1]

Given the sporadic and random nature of the user behavior, allocating a dedicated frequency bandwidth or semi-persistent scheduling for these data may result in low bandwidth utilization and the weak capability of latency guarantee. To this end, the puncturing scheme and superposition scheme [9] posited for scenarios with the coexistence of enhanced mobile broadband (eMBB) and ultra-reliable and low-latency communications (URLLC) traffic appears to be more suitable for the targeted XR case. Nevertheless, the adoption of such a joint scheduling mechanism inevitably impacts the transmission performance of eMBB, leading to the degradation of the quality of virtual scene construction. Thus, *efficiently orchestrating the transmission of multi-modal data with diverse latency requirements still stands as the foremost challenge in enhancing the XR user experience and realizing its success.*

Intermediate Stage: Data Analysis

Despite understanding user interactions, data analysis can also play an essential role in enhancing the XR performance. Nowadays, the FoV prediction has evolved into an integral technology for wireless XR transmission. However, most of the ongoing refinements only contain eye-tracking data. While some advanced research has integrated video saliency detection into FoV predictions [10], it is important to clarify that view direction is influenced not only by the distribution of attractive objects but also by personal preferences. Nonetheless, few studies establish a connection between individual interests and video content, potentially limiting the generalizability of existing models.

Furthermore, while accurate FoV predictions can effectively reduce the latency experienced by users by proactively transmission, this approach fails to address the issues of limited radio resources in fact, especially in multi-user scenarios. As such, relying solely on FoV predictions proves insufficient. Drawing inspiration from the fundamental concept of SemCom, we believe that optimizing XR sessions should involve strategies aimed at

[1] Public research suggests that motion-to-photon latency should be under 15–20 ms for a seamless experience. Motion-to-sound latency can be tolerated up to 30 ms, while touch-to-feel latency imposes the most stringen requirements, often as low as 1 ms.

Stage	Tech.	Semantic Awareness	← Relationship →	Generative AI
Stage I * Data Collection	Objective	*Efficiently orchestrating the transmission of multi-modal data with diverse QoS requirements*		
	Benefit	• Achieving efficient multi-modal semantic compression • Enhancing the robustness of super-position scheme	Utilizing GAI to further enhance the active adversarial capability of the robust semantic decoder in the superposition framework	• Empowering the decoder to proactively capture and eliminate interference caused by superposition on ongoing transmissions
	Approach	DL-based SemCom paradigm	←	GAI-basd denoising
	Challenge	• The requirement to explore the explainability of semantic containers, determine the optimal location of output semantic features, and improve the performance of multimodal fusion and separation is crucial. • The necessity for pre-training an effective discriminator to effectively evaluate the quality of generated data is vital for enhancing the training stability during fine-tuning of GANs.		
Stage IIData Analysis	Objective	*Performing comprehensive data analyses (beyond FoV prediction) to reducing transmitted data*		
	Benefit	• By linking semantic features within the tiles of the FoV to the user's attention for each type of semantic feature, varying importance can be assigned to individual tiles	By analyzing the semantic features of FoV and user's preferences, the tile can be personalized into background and foreground ones for processing separately	• For the background tiles, the GAI-based predictive rendering techniques can reduce the number of tiles to be transmitted in both the requested FoV as and the next FoV
	Approach	Joint training based MTL	→	Image inpainting and outpainting
	Challenge	• Designing the MTL neural network structure delicately to maximize the sharing of underlying features and conserve computational resources, all while ensuring prediction accuracy, presents a significant challenge. • The varied computing capabilities necessitate personalized trade-offs between computing latency for inpainting and outpainting, as well as transmission latency.		
Stage IIIData Delivery	Objective	*Jointly optimizing transcoding, multicast, and scheduling, aiming to decrease pixel-level redundant transmissions and enhance visual experiences, playback smoothness, and the synchronization of multiple XR sessions*		
	Benefit	• Achieve pixel-level redundancy removal in resource-constrained wireless transmission, by integrating semantic awareness into transcoding and multicast	Leveraging GAI techniques to collaboratively enhance semantic-aware content delivery, incorporating denoising, FoV completion, and personalized enhancements	• Denoising received distorted tiles • Recovering the dropped tiles exceeding latency requirements • Enabling personalized enhancements and refinement
	Approach	• Combinatorial optimization • Decision decomposition based on GAI & large language model	←	• Image recovery • Image inpainting • real-time translation
	Challenge	• Decoupling the high-dimensional multi-constraint optimization problem on joint semantic-aware multicast cluster decision, transcoding, and scheduling poses a significant challenge. • The optimization and offloading of computing consumption and latency for GAI-based collaborative refinements is paramount.		

* It should be noted that at each stage we focus only on the most significant issues. The superposition scheme for uplink data collection proposed in Stage I can also be applied to the multicast-based downlink transmission process in Stage III

TABLE 1. Summary of the role of semantic awareness and GAI in GeSa-XRF.

reducing transmitted data volume. In this sense, *there arises a need to explore reliable and comprehensive data analysis (beyond FoV prediction), with the overarching objective of further reducing transmitted data, all while maintaining the user's immersive experience*.

Feedback Stage: Data Delivery

To enhance the immersive XR environment, timely feedback of multi-sensory data to users is crucial. Among the multiple types of sensory data, visual data is particularly data-intensive, which is always perceived as a formidable obstacle for wireless mobile VR deployments. In recent literature, the utilization of online transcoding emerges as a prominent technique for enhancing user experience [11]. However, the user's attention degree to different tiles in the requested FoV is always neglected and all the tiles are treated uniformly. This oversight results in existing efforts optimizing only up to the point of ensuring video playback smoothness and failing to optimize user experience in a personalized manner.

Meanwhile, for multi-user scenarios, despite the distinct regions of interest for each user, there is considerable overlap of the FoVs requested in a session [11]. Given the inherent broadcast nature of wireless networks, multicast presents itself as a promising efficient transmission technology. While this method is straightforward to implement in a single cell, determining clusters for multicast in the typical heterogeneous network (HetNet) becomes challenging, which is influenced by both the available bandwidth of individual base stations (BSs) and the user distribution. Additionally, due to the lack of consideration for the significance of each tile, current multicast strategies solely address the transmission redundancy at the tile level. Hence, *there is a need to explore a strategy aiming to decrease pixel-level redundant transmissions while accommodating personalized visual experiences, playback smoothness, and the synchronization of multiple XR sessions.*

Exploring GeSa-XRF: Illustrative Insights

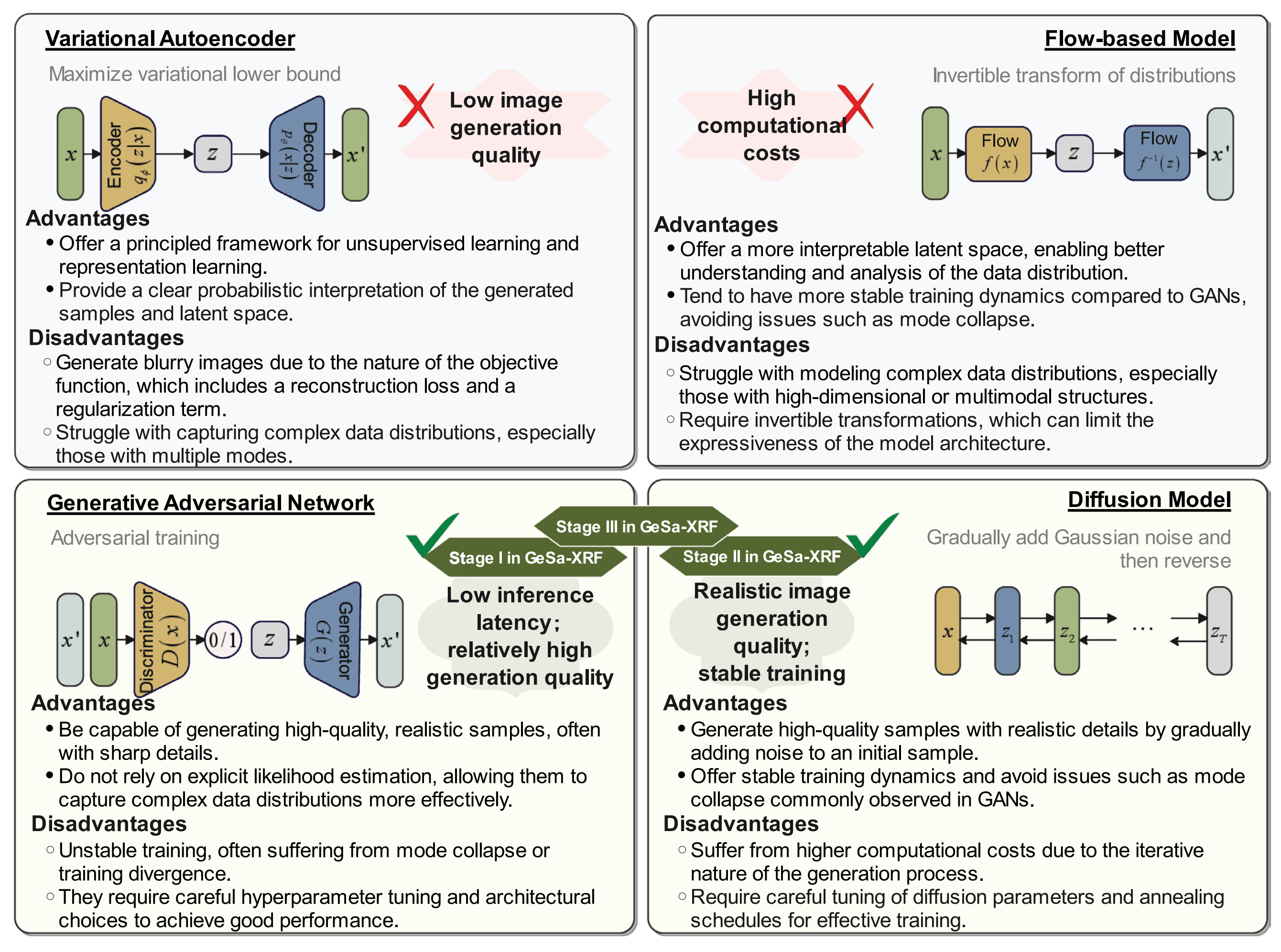
In this section, the specific implementation details of GeSa-XRF in each stage are presented, respectively, where the main roles of semantic awareness and GAI are summarized in Table 1. Meanwhile, the meanstream GAI techniques have been presented in Fig. 1.[2]

Multi-Modal Data Collection

In this subsection, we focus on the orchestration of the multi-modal data collection with diverse latency requirements, which is visualized in Fig. 2.

1) Designing a Multi-modal SemCom Framework for Large Volume Data: First, we target sensory data with large volumes, such as audio and video. Considering the ultra-high

[2] The diagrammatic representation of the model in this figure is from https //lilianweng.github.io/posts/2021-07-11-diffusion-models/.



Advantages

- Offer a principled framework for unsupervised learning and representation learning.
- Provide a clear probabilistic interpretation of the generated samples and latent space.

Disadvantages

○ Generate blurry images due to the nature of the objective function, which includes a reconstruction loss and a regularization term.
○ Struggle with capturing complex data distributions, especially those with multiple modes.

Advantages

- Offer a more interpretable latent space, enabling better understanding and analysis of the data distribution.
- Tend to have more stable training dynamics compared to GANs, avoiding issues such as mode collapse.

Disadvantages

○ Struggle with modeling complex data distributions, especially those with high-dimensional or multimodal structures.
○ Require invertible transformations, which can limit the expressiveness of the model architecture.

Generative Adversarial Network
Adversarial training
x'
x
Discriminator $D(x)$
0/1
z
Generator $G(z)$
x'
Stage III in GeSa-XRF
Stage I in GeSa-XRF
Low inference latency; relatively high generation quality
Stage II in GeSa-XRF
Realistic image generation quality; stable training

Advantages

- Be capable of generating high-quality, realistic samples, often with sharp details.
- Do not rely on explicit likelihood estimation, allowing them to capture complex data distributions more effectively.

Disadvantages

○ Unstable training, often suffering from mode collapse or training divergence.
○ They require careful hyperparameter tuning and architectural choices to achieve good performance.

Advantages

- Generate high-quality samples with realistic details by gradually adding noise to an initial sample.
- Offer stable training dynamics and avoid issues such as mode collapse commonly observed in GANs.

Disadvantages

○ Suffer from higher computational costs due to the iterative nature of the generation process.
○ Require careful tuning of diffusion parameters and annealing schedules for effective training.

FIGURE 1. Summary of mainstream GAI technologies [8].

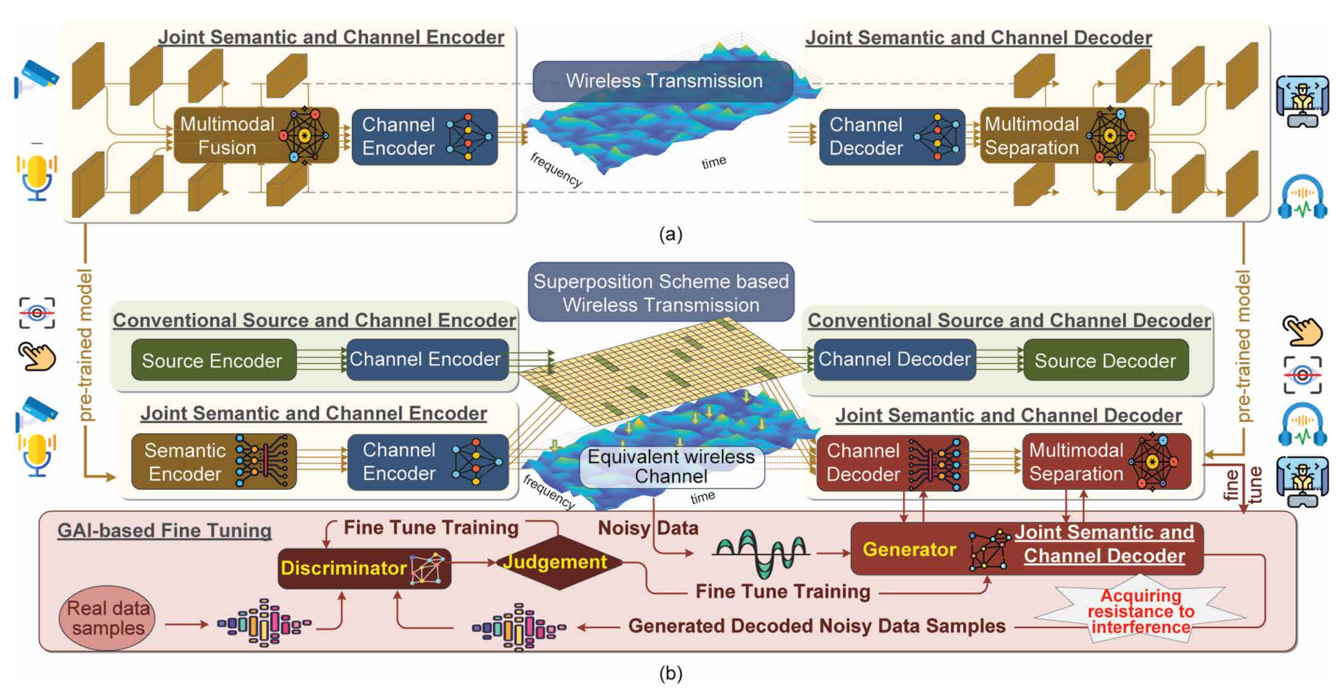


FIGURE 2. GAI-assisted superposition transmission with heterogeneous communication paradigm for multi-modal data collection. a) DL-based multi-modal SemCom framework. b) Robust GAI-assisted superposition scheme for heterogeneous transmission.

latency requirements for uplink transmission, the DL-based multi-modal SemCom stands out the first-choice [7], which allows not only for more efficient semantic compression by removing intermodal redundancies, but also for cross-modal parsing at the XR server to enhance decoding robustness with low inference latency.

Different from the available multi-modal SemCom aimed at decision-making, e.g., visual question answering, the XR server targets at the reconstruction of multi-modal signals. To this end, a multi-modal SemCom framework based on multi-modal fusion and separation is proposed as shown in Fig. 2(a). Inspired by the DL-based SemCom framework designed for uni-modal data reconstruction, the semantic encoders and decoders herein adopt a symmetric structure, both of which consist of two integral components. One corresponds to the neural network for reconstructing individual unimodal data, where the encoder part acts as a layer-structured semantic container, while the decoder part functions as a cascading semantic reasoner. The other relates to the multi-modal fusion or separation module. At the transmitter, the fusion module processes the semantic features output by certain layers of the individual semantic containers, facilitating semantic fusion to eliminate redundant information. Conversely, the separation module at the receiver takes the fused data, parsing multiple semantic features, and conveys them to the corresponding semantic reasoners for incremental enhancement of semantic inference layer by layer. Moreover, the realization of an effective fusion and separation pair necessitates the identification of appropriate semantic features, implying an optimal location for output semantic features within the semantic container.

2) Designing a GAI-assisted Superposition Scheme for Coexistence of Multi-modal Data: According to existing literature [3], DL-based SemCom exhibits greater robustness compared to conventional bit-based communications and can achieve high-quality data reconstruction even at low signal-to-noise ratios. Therefore, the multi-model SemCom proposed above holds the promise to endow the superposition with the resilience to withstand inter-user interference. With this in mind, we conceive a novel superposition scheme for the coexistence of multi-modal sensory data as shown in Fig. 2(b).

Specifically, the data with small data volume, like haptic information, employ conventional communications to mitigate the computational latency, which are transmitted to the XR server by overlaying the semantic transmission described in the section "Designing a Multi-Modal SemCom Framework for Large Volume Data." Such the superposition scheme can be identified as power-domain non-orthogonal multiple access. Due to the end-to-end training fashion, it is important to note that the decoding of the semantic signal cannot involve a direct extraction from the superimposed semantic and bit signals [12]. In this sense, at the destination, this scheme follows the "bits-to-semantics successive interference cancellation ordering", as bit communications require no prior training. Furthermore, to strengthen the ability of SemCom to cope with superposition interference, we can employ GAI to fine-tune the semantic decoder, with relatively low training complexity. Considering its single forward pass characteristics and relative high generation quality, we resort to the generative adversarial network (GAN) to further refine the joint channel and semantic decoder, avoiding introducing much extra latency. Therein, the pre-trained joint channel and semantic decoder in the section "Designing a Multi-modal SemCom Framework for Large Volume Data" is treated as the initial generator in GAN. Simultaneously, to augment the decoder's active denoising capabilities, a discriminator is introduced. The learning object of the discriminator is to distinguish the distorted data and the high-quality data. Against this, the learning objective of the generator is to deceive the discriminator to get a higher score. In other words, the well-trained discriminator can act as a fitting process of implicit functions used to guide semantic decoding optimization, which can reflect the relationship between the compressed semantic information and the quality of the reconstructed data.

Considering its single forward pass characteristics and relative high generation quality, we resort to the generative adversarial network (GAN) to further refine the joint channel and semantic decoder, avoiding introducing much extra latency.

Multi-Task Data Analysis

In this subsection, we delve deeper into the potential of the data analysis stage to streamline the transmission and improve the overall performance of the XR.

1) Joint FoV and Attention Value Prediction based on Multi-task Learning Techniques: As discussed in the section "Intermediate Stage: Data Analysis," the personalized interests of the users are ignored in existing FoV predictions, which causes the FoV prediction model to be retrained for every new XR video. To address this concern, we propose to introduce the factor of user-object-attention (UOA) value into the FoV prediction, which can lay the foundation for personalized video saliency prediction without the need for retraining. Meanwhile, according to our prior work [13], the prediction of UOA values also holds a crucial role in assigning semantic significance to tiles within the FoV, which assists in optimizing transcoding, multicasting, and personalized experience. However, our initial approach in UOA prediction solely relies on the user-object-attention level dataset for Metaverse research. In fact, the shift of the FoV can also serve as an indicator of the user's attention to various objects. In other words, FoV prediction and user-object-attention prediction are intricately connected tasks.

Consequently, we resort to the utilization of MTL techniques [14] to train a model capable of simultaneously performing both prediction tasks, leveraging shared information to enhance overall prediction performance as shown in Fig. 3(a). To establish a suitable dataset, a mapping that delineates the correspondence between objects and tiles within the FoV needs to be created. This enables the fusion of the two

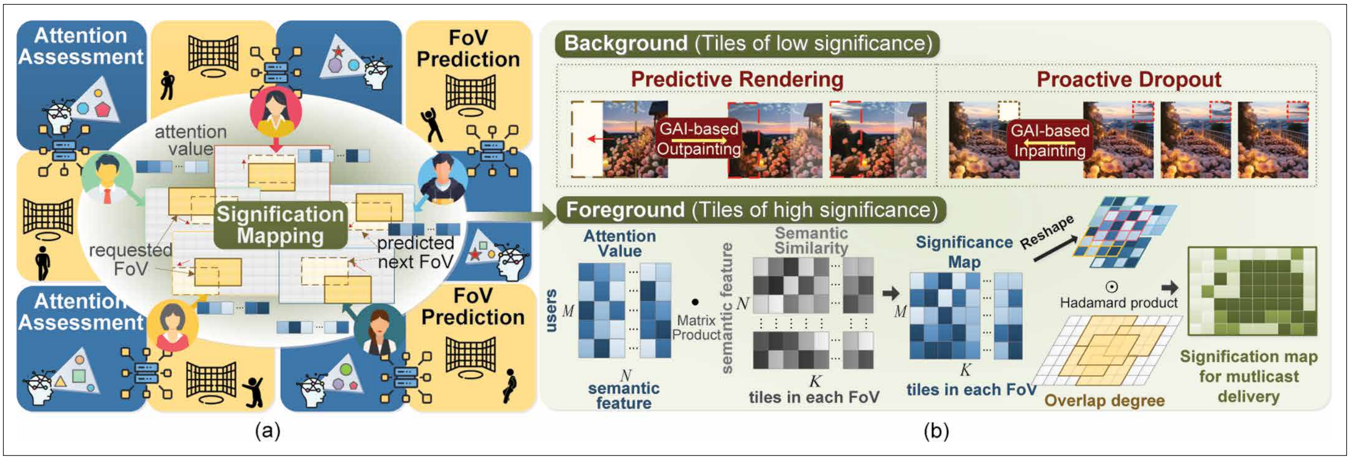


FIGURE 3. Multi-task learning based on beyond FoV prediction for GAI-based preprocessing and semantic significance mapping. a) Joint FoV prediction and attention assessment based on MTL. b) Preprocessing for background and foreground tiles.

datasets dedicated to each prediction task, facilitating multi-task training. Additionally, to recognize potential shared underlying patterns between the two tasks, the neural network architecture should incorporate shared layers to extract low-level common features, which is followed by two task-specific blocks for each prediction task capturing unique patterns related to individual tasks. To maintain an overall optimization objective, we propose the creation of a joint loss function that combines losses from both tasks. The expansion on how the prediction results enhance XR session performance will be explored in the section "Semantic signification based GAI-assisted preprocess and multi-user tile significance mapping."

2) Semantic signification based GAI-assisted preprocess and multi-user tile significance mapping: Based on the inclusion or exclusion of objects of interest to the user, we propose to categorize tiles into two types for separate processes as shown in Fig. 3(b).

- **Background tile:** Background tiles used to create the backdrop or distant scenery. Due to substantial overlap between adjacent FoVs, devices can employ advanced GAI techniques to conduct outpainting for predictive rendering of background information in the unit of tile in advance. Considering the more relaxed downlink latency requirements resulting from proactive pre-processing, we favour the choice of the diffusion model, e.g., the stable diffusion given its excellent image generation details. Herein, the XR server only needs to transmit the FoV prediction results to the users, which can enhance the smoothness of user experiences during FoV transitions without consuming extra computational and transmission resources for the XR server. Similarly, within the requested FoV, background tiles can also be proactively dropped, since they can be rendered at the XR device using diffusion-based inpainting techniques based on received surrounding tiles. However, for specific scenarios, given the differentiated computing capabilities of the devices, a trade-off between the computing latency and the transmission latency should be carefully studied.
- **Foreground tile:** Foreground tiles are used to represent interactive elements or objects that are closer to the player's viewpoint. Concerning the foreground tiles with objects of interest to users, our focus lies in generating a significance map of the tiles based on the user requests. This map serves as a guide for semantic-aware delivery, as discussed in the section "Multi-User Data Delivery." Given that users exhibit varying levels of attention to the same object and different levels of attention to distinct objects, we first propose the derivation of a personalized significance map. This map relies on the matrix product of UOA values across all semantic features and the distribution of semantic features within each tile in the requested FoV [13]. Meanwhile, in a multicasting scenario, tiles requested by a greater number of users should also receive higher significance, given that the loss of them may affect multiple users at the same time. Thus, the ultimate significance map for all transmitted tiles is derived through the Hadamard product of the aggregated multi-user significance map corresponding to their requested FoV and the overlapping degree of the FoV requested by all users. The implications of the multi-user tile significance map on semantic-aware delivery are further explored in the section "Multi-User Data Delivery."

Multi-User Data Delivery

In this subsection, we propose to integrate semantic awareness into the online transcoding strategy as well as multicast delivery, to diminish pixel-level redundancy.

1) Designing a semantic-aware multicast delivery scheme for enhancing experience and playback synchronization: In this work, we consider a typical HetNet architecture with one macro BS (MBS) and several small BSs (SBSs) as shown in Fig. 4(a). All the BSs assume a synchronized discrete-time system with slots. In HetNet, the SBSs generally share the same segment of frequency bandwidth and the MBS operates at a different frequency bandwidth. Meanwhile, all the SBSs can be covered by the MBS, and there is no overlap between SBSs. In the considered multicast system, the resource allocation targets are no longer users but rather tiles requested in individual BSs. In this context, we define a multicast transmission about a tile as a cluster, which is specified by the associated BS, the transcoded resolution, and the transmission slot as shown in Fig. 4(b). During the delivery, the following three types of constraints are considered.

- To avoid redundant transmission, a tile within a given BS can only be transmitted at most once.
- For tiles requested by the user only covered by the MBS, the cluster should be served by MBS. For tiles requested by users covered by SBSs, the cluster is either served only by the MBS or by each of the relevant SBSs separately.
- In each slot, the available resources at each BS should be able to support the transmission.

For guiding the delivery optimization as illustrated in Fig. 4(d), we propose a three-dimensional experience evaluation metric, as depicted in Fig. 4(c). The first is the weighted resolution based on the tile semantic significance acquired in Stage II. The second is the playback smoothness, which is jointly determined by transmission and computing latency. The last is playback synchronization, which can minimize the time-shift between multiple users to ensure an immersive viewing experience for all the users in the virtual world. However, in achieving the optimal experience, the scheduling, transcoding, and multicast decisions are coupled with each other, which makes the optimization problem intractable. Therefore, we propose a possible way to decouple the high-dimensional optimization problem, which can unfold along three steps:

Another feature of MBS is its capability of simultaneous transmission to users covered by different SBSs, which suggests that the clusters relate to the tiles requested in different SBSs

- **Decoupling of Multicast:** The reason for starting with this lies in two distinctive roles of MBS in HetNet. One is the only choice for users uncovered by SBSs, which restricts the clusters associated with these users to be served by MBS in priority. Another feature of MBS is its capability of simultaneous transmission to users covered by different SBSs, which suggests that the clusters relate to the tiles requested in different SBSs. In addition, it is desirable to balance the ratio of total significance scores to available bandwidth between BSs in order to keep the tiles with the same

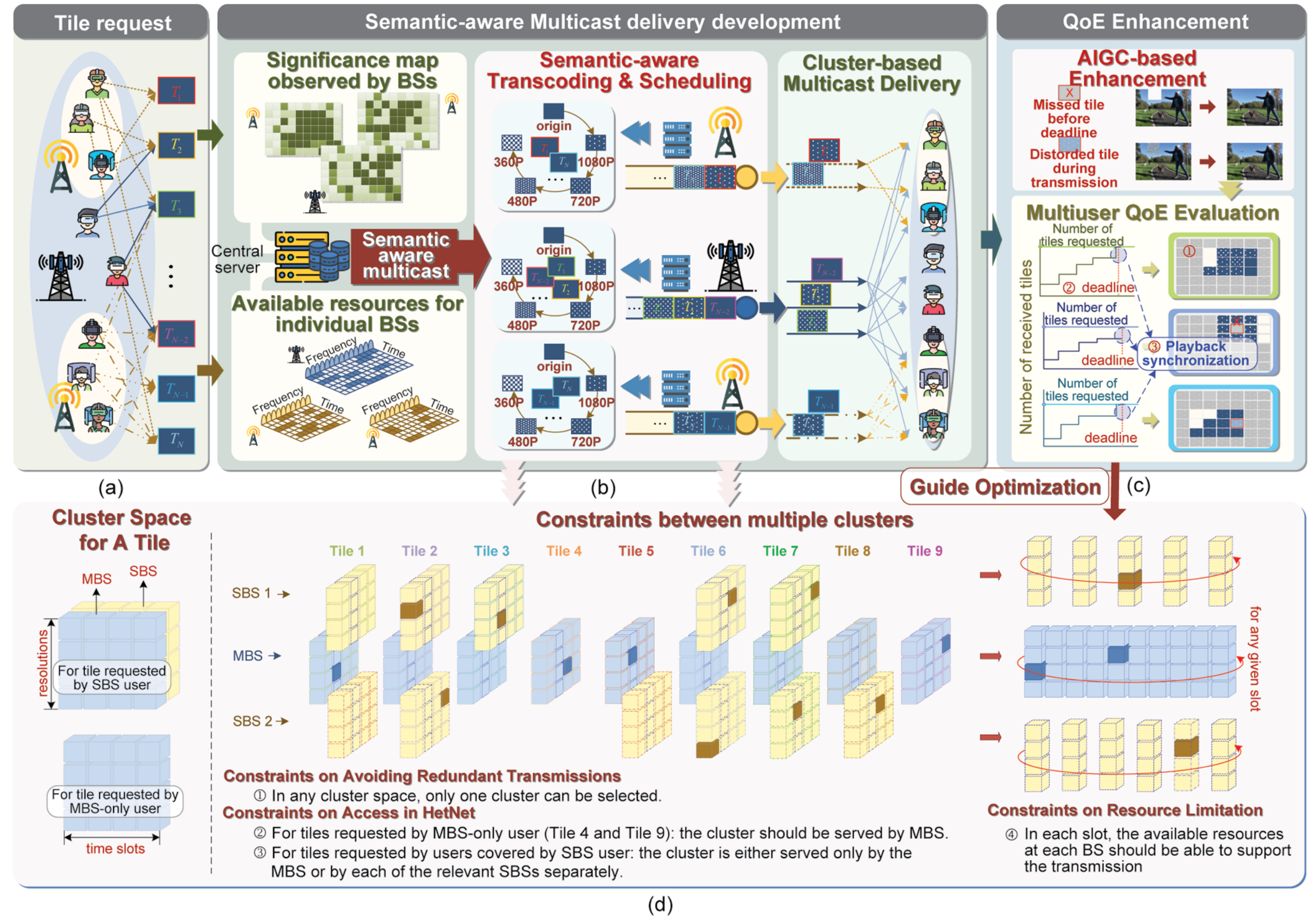

Tile request
Semantic-aware Multicast delivery development
QoE Enhancement
T_1
T_2
T_3
T_{N-2}
T_{N-1}
T_N
Significance map observed by BSs
Central server
Semantic aware multicast
Available resources for individual BSs
Frequency
Time
Semantic-aware Transcoding & Scheduling
origin
360P
1080P
480P
720P
Cluster-based Multicast Delivery
AIGC-based Enhancement
Missed tile before deadline
Distorded tile during transmission
Multiuser QoE Evaluation
Number of received tiles
Number of tiles requested
①
② deadline
deadline
③ Playback synchronization
(a)
(b)
Guide Optimization
(c)
Cluster Space for A Tile
MBS
SBS
resolutions
For tile requested by SBS user
For tile requested by MBS-only user
time slots
Constraints between multiple clusters
Tile 1
Tile 2
Tile 3
Tile 4
Tile 5
Tile 6
Tile 7
Tile 8
Tile 9
SBS 1
MBS
SBS 2
for any given slot
Constraints on Avoiding Redundant Transmissions
① In any cluster space, only one cluster can be selected.
Constraints on Access in HetNet
② For tiles requested by MBS-only user (Tile 4 and Tile 9): the cluster should be served by MBS.
③ For tiles requested by users covered by SBS user: the cluster is either served only by the MBS or by each of the relevant SBSs separately.
Constraints on Resource Limitation
④ In each slot, the available resources at each BS should be able to support the transmission
(d)

FIGURE 4. Tile significance map based semantic-aware XR content delivery with GAI collaborative refinements. a) System model for HetNet. b) Illustration for semantic-aware multicast delivery. c) GAI-based collaborative refinement and performance evaluation. d) Optimization problem formulation for semantic-aware multicast delivery.

significance in different BSs to be transcoded into a similar resolution level.

- **Decoupling of Transcoding:** After Step 1, the coupling between BSs in the resource allocation is released. Then, the online transcoding can performed within each BS separately based on their significance map. With the objective of the total weighted resolution, a natural idea is to give higher resolution level to the more important tile. However, specific decision-making scenarios still need to be determined taking into account the resources available at the BS.
- **Decoupling of Scheduling:** At last, scheduling decisions can be determined by jointly considering the significance map and the transmission progress of the tiles requested by individual users, with the aim of guaranteeing the playback synchronization between multiple users. Moreover, for scenarios with a high degree of dynamic change in wireless resources, the available resources for each time slot also need to be considered in the design of the algorithm.

2) Advancing experience with GAI-assisted denoising, FoV Completion and personalized enhancement: During wireless transmission, transmitted data are susceptible to noise interference, resulting in distorted data that compromises the immersive experience. Furthermore, resource limitation may occasionally prevent the successful transmission of all the requested tiles to the user before playback, leading to playback lag or incomplete construction of virtual scenes. Consequently, exploring corresponding remedies becomes imperative.

Given diffusion powerful generation capabilities and flexible deployment, it emerges as the primary choice for enhancing experience. The forward process of the diffusion model initiates with an input data sample and progressively introduces noise, mirroring the interference experienced by data transmitted in a wireless channel. Conversely, in the reserve process, starting with the noisiest data version, the model iteratively denoises the data, aiming to recover the original data, with a parallel objective to the denoising process in communication systems. While the literature showcases notable denoising capabilities for images [15], the determination of the minimal number of required denoising steps for specific noisy data warrants further investigation to reduce computing latency. Simultaneously, thanks to the success of inpainting techniques, the lost tile can be recovered. Considering the dual factors of computational latency and generation of tile fidelity, GAN and diffusion model can be chosen on a case-by-case basis to make a trade-off between the both factors. In addition to the compensation for transmission losses, GAI can open avenues for personalized user experiences, such as real-time voice translation and personalized ambiance rendering.

Case Study: Streamlined Multicast-Based Data Delivery in GeSa-XRF

In this case study, we focus on the implementation of semantic-aware multicast-based image delivery process in GeSa-XRF, specifically concerning semantic-aware multicast decision and semantic-aware transcoding, which determines the source content to be transmitted as well as the transmission bandwidth. The multi-modal SemCom can be implemented following the results of the strategy.

Scenario Setup

Without loss of generality, we consider a HetNet with one MBS, two SBSs, and five users. Specifically, each SBS covers two users and one user is only covered by the MBS. For a given XR session, we develop the delivery strategy based on the image of real scenes, Jewel Changi Airport, shown in Fig. 5(a). According to the overlap of the FoVs and the common semantic feature that people have across FoVs, we derive the semantic significance maps for each BS, respectively. The tiles highlighted in the three maps refer to the tiles requested by the covered users. The darker the color of the tile, the more important it is. To ensure the playback smoothness, all the requested tiles should be delivered to the users within an allowable latency. In the subsequent development of the strategy, we assume the available bandwidth for the MBS SBS 1, and SBS 2 to be 200 Mbps, 150 Mbps, and 100 Mbps, respectively.

Strategy Development and Performance Analysis

Following the decoupling steps outlined in the section "Multi-User Data Delivery," we begin by identifying the cluster associated tiles requested by user 3 (covered solely by the MBS), designating them to be served by the MBS. To validate our proposed framework, we introduce two cluster grouping methods: semantic-aware multicast and conventional multicast. In the conventional approach, our goal is to achieve load balance, that is, making the ratio of the number of tiles to be sent to each BS to the available bandwidth as equal as possible. Conversely, in the semantic-aware approach, we strive for the similar ratio of the total importance score of the tiles to be sent and the available bandwidth across different BSs, to ensure that the tiles with the same significance to be transcoded into the similar resolution level. The semantic-aware multicast decision results can be found in Fig. 5(b). Both the approaches can be implemented based on conditional linear programming with constraints. After the cluster decision, we perform the transcoding strategy for each BS, respectively. Similarly, we consider the two methods: semantic-aware and conventional transcoding. In semantic-aware transcoding, we aim to optimize the total resolution level weighted by the significance value, while in conventional transcoding, we just optimize the total resolution, regardless of the significance. To solve the above two combinatorial optimization problems, we adopt the genetic algorithm. Fig. 5(c) indicates that average resolution levels by significance under semantic-aware transcoding are generally higher than those under conventional transcoding. The metrics of PSNR and LPIPS achieved by the semantic-aware and semantic-unaware transcoding are 29.175 dB and 28.934 dB, 0.379 and 0.421, respectively. Visual comparisons of

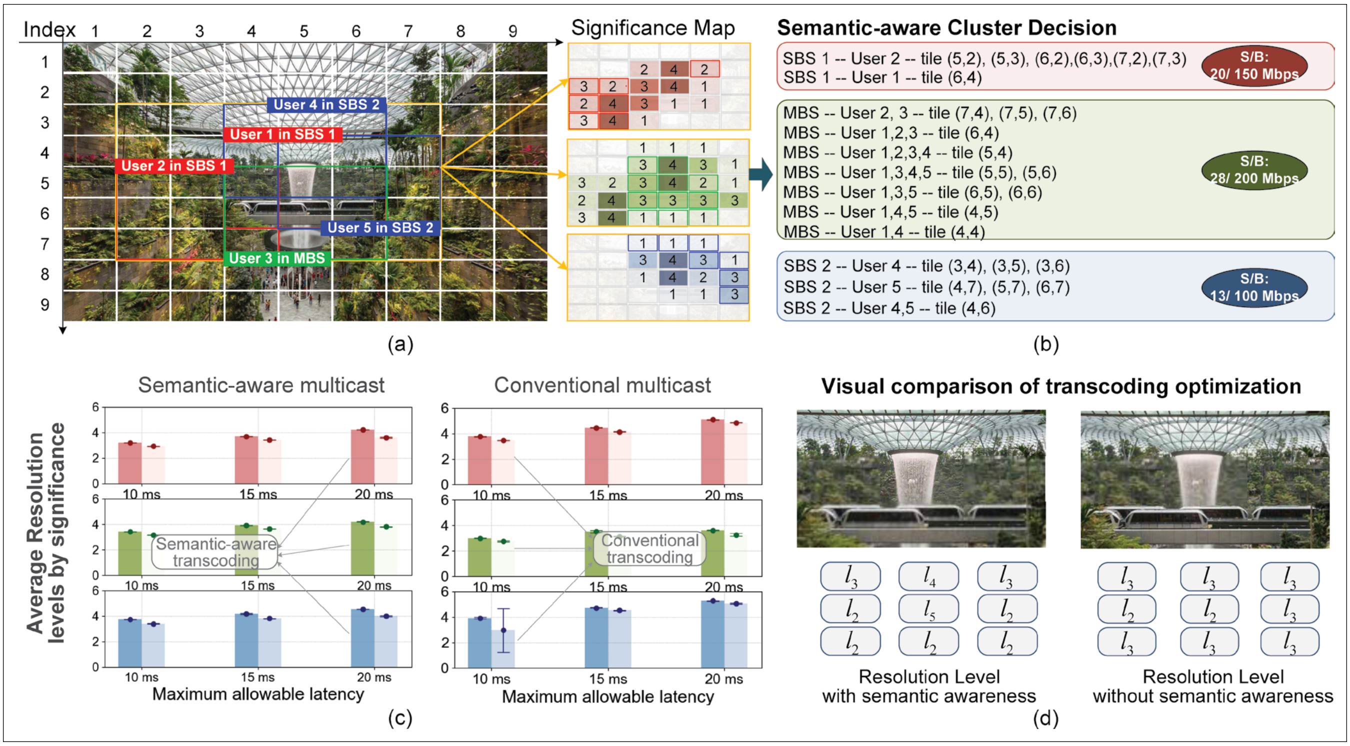


FIGURE 5. Implementation of semantic-aware multicast-based delivery strategy in case study. a) Image of real scenes, Jewel Changi Airport and significance map. b) Results of semantic-aware multicast decision. c) Performance comparison under semantic-aware and conventional multicast with semantic-aware and conventional transcoding. d) Visual comparison of semantic- aware and conventional transcoding optimization, where l_5 to l_1 represent the original resolution, 75%, 50%, 25% and 15% original resolution, respectively.

the two transcoding optimization methods are presented in Fig. 5(d). Furthermore, through a comparison of simulation results between semantic-aware multicast and conventional multicast, we observe a higher similarity in average resolution levels by significance across the three BSs in semantic-aware multicast. This observation establishes a basis for rationalizing multicast distribution of XR content in heterogeneous networks with multi-BS cooperation.

Conclusion and Future Work

In this paper, we have conducted a thorough exploration of efficient wireless XR transmission technologies. Specifically, we have seamlessly integrated semantic awareness to optimize data transmission, leveraging GAI for collaborative refinements and proposed a novel framework called GeSa-XRF. In the future work, in the data collection stage, we intend to involve a more in-depth exploration about the integration of the cutting-edge technologies in 6G, e.g., massive MIMO, in the superposition scheme. During the data analysis stage, our focus will be on optimizing the number of proactively rendered and dropped tiles , taking into account the communication and computational resources at cloud-edge-end architecture in 6G. In data delivery stage, we will explore the optimal transmission strategy, concurrently considering three key experience evaluation metrics. Additionally, we aim to investigate more efficient decoupling strategies for high-dimensional optimization problems leveraging the task reasoning capabilities of large language model techniques.

Acknowledgment

This work was supported in part by the National Research Foundation, Singapore and Infocomm Media Development Authority under its Future Communications Research and Development Programme, in part by SUTD under Grant SRG-ISTD-2021-165, in part by SUTD-ZJU IDEA under Grant SUTD-ZJU (VP) 202102, in part by the Ministry of Education, Singapore, under its SMU-SUTD Joint under Grant 22-SIS-SMU-048, and in part by SUTD Kickstarter Initiative under Grant SKI 20210204.

References

[1] W. Yang et al., "Semantic communications for future internet: Fundamentals, applications, and challenges," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 1, pp. 213–250, 1st Quart., 2023.
[2] C. Wang et al., "Adaptive semantic-bit communication for extended reality interactions," *IEEE J. Sel. Topics Signal Process.*, vol. 17, no. 5, pp. 1080–1092, Sep. 2023.
[3] Z. Qin et al., "Semantic communications: Principles and challenges," 2021, *arXiv:2201.01389*.
[4] E. Grassucci, S. Barbarossa, and D. Comminiello, "Generative semantic communication: Diffusion models beyond bit recovery," 2023, *arXiv:2306.04321*.
[5] A. D. Raha et al., "Generative AI-driven semantic communication framework for NextG wireless network," 2023, *arXiv:2310.09021*.
[6] W. Yang et al., "Semantic change driven generative semantic communication framework," 2023, *arXiv:2309.12775*.
[7] Z. Qin et al., "Survey of research on multimodal semantic communication," *J. Commun.*, vol. 44, no. 5, pp. 28–41, 2023.
[8] R. Zhang et al., "Generative AI-enabled vehicular networks: Fundamentals, framework, and case study," 2023, *arXiv:2304.11098*.
[9] M. Darabi et al., "Hybrid puncturing and superposition scheme for joint scheduling of URLLC and eMBB traffic," *IEEE Commun. Lett.*, vol. 26, no. 5, pp. 1081–1085, May 2022.

[10] J. Li et al., “Spherical convolution empowered viewport prediction in 360 video multicast with limited FoV feedback,” *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 19, no. 1, pp. 1–23, Jan. 2023.

[11] L. Zhong et al., “A multi-user cost-efficient crowd-assisted VR content delivery solution in 5G- and-beyond heterogeneous networks,” *IEEE Trans. Mobile Comput.*, vol. 22, no. 8, pp. 4405–4421, Jun. 2023.

[12] X. Mu and Y. Liu, “Exploiting semantic communication for non-orthogonal multiple access,” *IEEE J. Sel. Areas Commun.*, vol. 41, no. 8, pp. 2563–2576, Aug. 2023.

[13] H. Du et al., “Attention-aware resource allocation and QoE analysis for metaverse xURLLC services,” *IEEE J. Sel. Areas Commun.*, vol. 41, no. 7, pp. 2158–2175, Jul. 2023.

[14] R. M. Samant et al., “Framework for deep learning-based language models using multi-task learning in natural language understanding: A systematic literature review and future directions,” *IEEE Access*, vol. 10, pp. 17078–17097, 2022.

[15] Y. Xie et al., “Diffusion model for generative image denoising,” 2023, *arXiv:2302.02398*.

Biographies

Wanting Yang (wanting_yang@sutd.edu.sg) received the B.S. and Ph.D. degrees from the Department of Communications Engineering, Jilin University, Changchun, China, in 2018 and 2023, respectively. She is currently a Research Fellow with the Singapore University of Technology and Design. Her research interests include wireless semantic communication, generative artificial intelligence, learning, martingale, and predictive resource allocation. She served as a Technical Program Committee Member for flagship conferences, such as WCNC, GLOBECOM, and VTC.

Zehui Xiong (Senior Member, IEEE) (zehui_xiong@sutd.edu.sg) received the Ph.D. degree from Nanyang Technological University (NTU), Singapore. He is currently an Assistant Professor with the Singapore University of Technology and Design, and also an Honorary Adjunct Senior Research Scientist with the Alibaba-NTU Singapore Joint Research Institute, Singapore. He is currently serving as the Editor and the Guest Editor for many leading journals. He is also serving as the Associate Director of Future Communications R&D Programme.

Tony Q. S. Quek (Fellow, IEEE) (tonyquek@sutd.edu.sg) received the B.E. and M.E. degrees in electrical and electronics engineering from the Tokyo Institute of Technology, Tokyo, Japan, in 1998 and 2000, respectively, and the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2008. He is currently the Cheng Tsang Man Chair Professor and a Full Professor with the Singapore University of Technology and Design (SUTD). He also serves as the Head of ISTD Pillar, the Sector Lead of the SUTD AI Program, and the Deputy Director of the SUTD-ZJU IDEA. He is currently serving as an Editor for IEEE Transactions on Wireless Communications and an Elected Member of the IEEE Signal Processing Society SPCOM Technical Committee. He was an Executive Editorial Committee Member for the IEEE Transactions on Wireless Communications, an Editor of IEEE Transactions on Communications, and an Editor for the IEEE Wireless Communications Letters.

Xuemin (Sherman) Shen (Fellow, IEEE) (sshen@uwaterloo.ca) received the B.Sc. degree from Dalian Maritime University, China, in 1982, and the M.Sc. and Ph.D. degrees in electrical engineering from Rutgers University, NJ, USA, in 1987 and 1990, respectively. He is currently a Professor and the University Research Chair with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research interests inclue mobility and resource management, UWB wireless communications networks, wireless network security, and vehicular Ad Hoc and sensor networks. He received the Excellent Graduate Supervision Award in 2006 and the Outstanding Performance Award in 2004 and 2008 from the University of Waterloo, and the Premier’s Research Excellence Award (PREA) in 2003 from the Province of Ontario, Canada. He serves as the Founding Area Editor for IEEE Transactions on Wireless Communications and the Editor-in-Chief for *Peer-to-Peer Networking and Application*. He is a Registered Professional Engineer of Ontario, Canada, and a Distinguished Lecturer of IEEE Communications Society.