

Review

Open Access



A review on multimodal communications for human-robot collaboration in 5G: from visual to tactile

Zhuorui Wang¹, Mingkai Chen¹ , Qian Liu²

¹The Key Laboratory of Broadband Wireless Communication and Sensor Network Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, Jiangsu, China.

²Department of Computer Science and Technology, Dalian University of Technology, Dalian 116024, Liaoning, China.

Correspondence to: Dr. Mingkai Chen, the Key Laboratory of Broadband Wireless Communication and Sensor Network Technology, Nanjing University of Posts and Telecommunications, No.66 Xinmofan Road, Nanjing 210003, Jiangsu, China. E-mail: mkchen@njupt.edu.cn

How to cite this article: Wang, Z.; Chen, M.; Liu, Q. A review on multimodal communications for human-robot collaboration in 5G: from visual to tactile. *Intell. Robot.* **2025**, *5*(3), 579-606. <https://dx.doi.org/10.20517/ir.2025.30>

Received: 23 Feb 2025 **First Decision:** 9 May 2025 **Revised:** 22 Jun 2025 **Accepted:** 23 Jun 2025 **Published:** 8 Jul 2025

Academic Editor: Simon Yang **Copy Editor:** Pei-Yun Wang **Production Editor:** Pei-Yun Wang

Abstract

With collaborative advances in wireless communication, artificial intelligence, and sensor technologies, robotic systems are undergoing a revolutionary evolution from single-function actuators to intelligent task processing platforms. In complex dynamic environments, the limitations of conventional unimodal perception have become increasingly apparent, struggling to meet the precision requirements for object attribute recognition and environmental interaction. In the future, deep-integrated multimodal perception technologies will emerge as a predominant trend, where cross-modal communication between vision and tactile sensing represents a critical breakthrough direction for enhancing robotic environmental cognition. Currently, research on multimodal visual-tactile communication remains scarce. Therefore, this paper conducts a comprehensive survey of this emerging field. First, this paper systematically summarizes mature video and tactile communication frameworks. Subsequently, this paper analyzes current implementations of single-modal streaming transmission for visual and tactile data, thereby investigating the state-of-the-art in multimodal visual-tactile communication. Finally, this paper briefly explores the promising prospects of visual-tactile communication technology, highlighting its transformative potential to enable context-aware robotic manipulation and adaptive human-robot collaboration.

Keywords: AI, video communication, multi-model communication, Tactile Internet, human-robot collaboration



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.



1. INTRODUCTION

With the advent of the 5G era, robotic applications have become increasingly prevalent in diverse domains. From industrial manufacturing, healthcare services, and logistics transportation to the domestic assistance and educational sectors, robotic technologies demonstrate substantial potential and practical value. Driven by rapid advances in artificial intelligence (AI), sensor technologies, and computational capabilities, robots have evolved from rudimentary single-function mechanical arms to intelligent and autonomous service systems. Future robotic systems are anticipated to exhibit not only enhanced decision-making capacities for complex tasks but also greater adaptability in human-robot interactions (HRI). Crucially, the continuous development of sensing and communication technologies will enable robots to perceive and interpret intricate environments with greater precision, enabling more efficient and accurate task execution. Within this development trajectory, environmental perception and interactive capabilities have emerged as pivotal challenges in improving robotic performance. As two primary modalities for external perception, vision and tactile sensing provide complementary dimensions of environmental cognition through the capture and interpretation of optical and mechanical signals, respectively. The effective integration and transmission of this multimodal information represents a critical frontier to achieving advanced robotic functionalities.

Visual systems endow robotic platforms with the capacity to “perceive” environmental contexts. Through integrated cameras, depth sensors, and computer vision algorithms, robotic systems achieve real-time object recognition, motion tracking, 3D mapping, and vision-guided navigation/obstruction avoidance. In autonomous vehicles, for example, visual perception forms the foundation for lane detection and pedestrian identification, while in industrial quality inspection, high-precision vision systems enable micrometer-level defect detection in manufactured components. The integration of deep learning technologies has significantly expanded the frontiers of visual perception, allowing robotic agents to extract semantic information from complex scenes and predict dynamic variations. Nevertheless, visual signals remain susceptible to environmental interference (e.g., illumination fluctuations and occlusions) and inherently lack direct access to object physical properties (e.g., hardness, texture details). These limitations underscore the necessity of tactile sensing as a complementary perceptual modality.

Tactile systems enable robotic devices to “sense” mechanical properties and interaction states through integrated force, pressure, and slip detection sensors. In minimally invasive surgical robots, for example, haptic feedback enables surgeons to discern variations in tissue stiffness, thus preventing excessive force application. Similarly, in industrial assembly lines, tactile information ensures precise grasping of fragile components by robotic manipulators. In contrast to visual modalities, tactile signals exhibit high spatio-temporal resolution and physical immediacy, providing real-time monitoring of contact forces, surface deformation, and other dynamic interaction parameters. However, the inherent sparsity and noise-prone nature of tactile data, coupled with high-bandwidth transmission requirements for distributed sensor arrays, present unique challenges in signal communication and computational processing.

However, the increasing complexity of robotic application scenarios has made the sole reliance on unimodal perception (visual or tactile) inadequate for contemporary operational demands. To achieve enhanced task execution efficiency and intelligent responsiveness, robotic systems require the integration of cross-modal sensory information. Within this paradigm shift, communication technologies play a pivotal role, particularly in multirobot collaboration, teleoperation systems, and large-scale data transmission applications, where real-time data exchange and coordinated operations mandate robust, low-latency communication architectures. As fundamental robotic communication paradigms, visual and tactile data transmission systems have been extensively investigated and implemented on various platforms. These technologies not only provide real-time transmission capabilities for spatio-temporal visual cues and haptic

feedback but also serve as critical enablers for interrobot coordination and telemanipulation reliability.

Emerging research is increasingly focused on integrating multimodal perception with communication architectures, particularly within the domain of visuo-tactile interaction. By synergizing visual and tactile information flows, robotic systems achieve comprehensive environmental awareness and improve response precision. Such multimodal visuo-tactile communication frameworks not only augment perceptual capabilities but also significantly improve operational dexterity in human-robot collaboration and unstructured environments. Although preliminary frameworks have been established in existing studies, two predominant bottlenecks persist. Firstly, optimization of unimodal transmission techniques (e.g., video encoding, haptic signal compression) remains predominantly modality-specific, failing to address cross-modal coordination requirements that lead to suboptimal resource allocation in shared communication channels. Secondly, the theoretical foundations for spatiotemporal alignment of heterogeneous data streams and cross-modal semantic fusion mechanisms lack systematic formulation, particularly in dynamic scenarios involving time-varying sensory uncertainty.

Although current research on video and haptic communication has been relatively mature, there are few reviews on multimodal visual touch communication technology and current development. Therefore, this paper investigates this gap to meet the urgent need for the rapid development of multimodal services and robots in the future.

The structure of this paper is shown in [Figure 1](#): This paper first reviews the rich research achievements in the field of video communication and tactile communication, then analyzes the disadvantages of single-mode communication in multimode services, and investigates the current research status of multimode visual touch communication in detail. These include a vision-based tactile sensor, visual-tactile data processing, visual-tactile data transmission, and visual-tactile data reconstruction. In addition, this paper will focus on the current hot large model and investigate the application of large models in the field of visual touch communication. Finally, the future development prospect and application of multimodal visual touch communication in robot field are briefly analyzed.

2. VIDEO COMMUNICATION

2.1. Video coding

The development of video coding technology reflects the dual drive of international cooperation and market demand. International Telecommunication Union - Telecommunication Standardization Sector (ITU-T) and International Organization for Standardization/International Electrotechnical Commission (ISO/IEC) have gradually built a modern video coding standard system through independent research and development and joint research. Among them, H.261, as the first practical standard, laid the foundation for the industry^[1]. Even today, H.261 remains useful in extremely low-bandwidth environments, such as certain military or remote sensing applications, where hardware constraints limit codec complexity. Moving Picture Experts Group - Part 1 (MPEG-1) promoted the popularity of video compact disc (VCD) and opened the era of civil use of digital video^[2]. MPEG-1 still finds limited use in legacy systems, archival formats, and simple embedded devices where computational efficiency outweighs quality demands. H.262/MPEG-2 became the core technology of the digital television revolution and set the technical standard in broadcasting^[3]. Due to its wide deployment, H.262 continues to be used in standard-definition digital television broadcasting and some satellite and cable TV systems, especially in regions with legacy infrastructure. H.264/MPEG-4 AVC, as a landmark standard, with its excellent compression efficiency, not only fully replaced H.262 in traditional fields, such as Blu-ray and high-definition broadcasting, but also expanded to mobile video, video conferencing, and other emerging application scenarios, becoming the

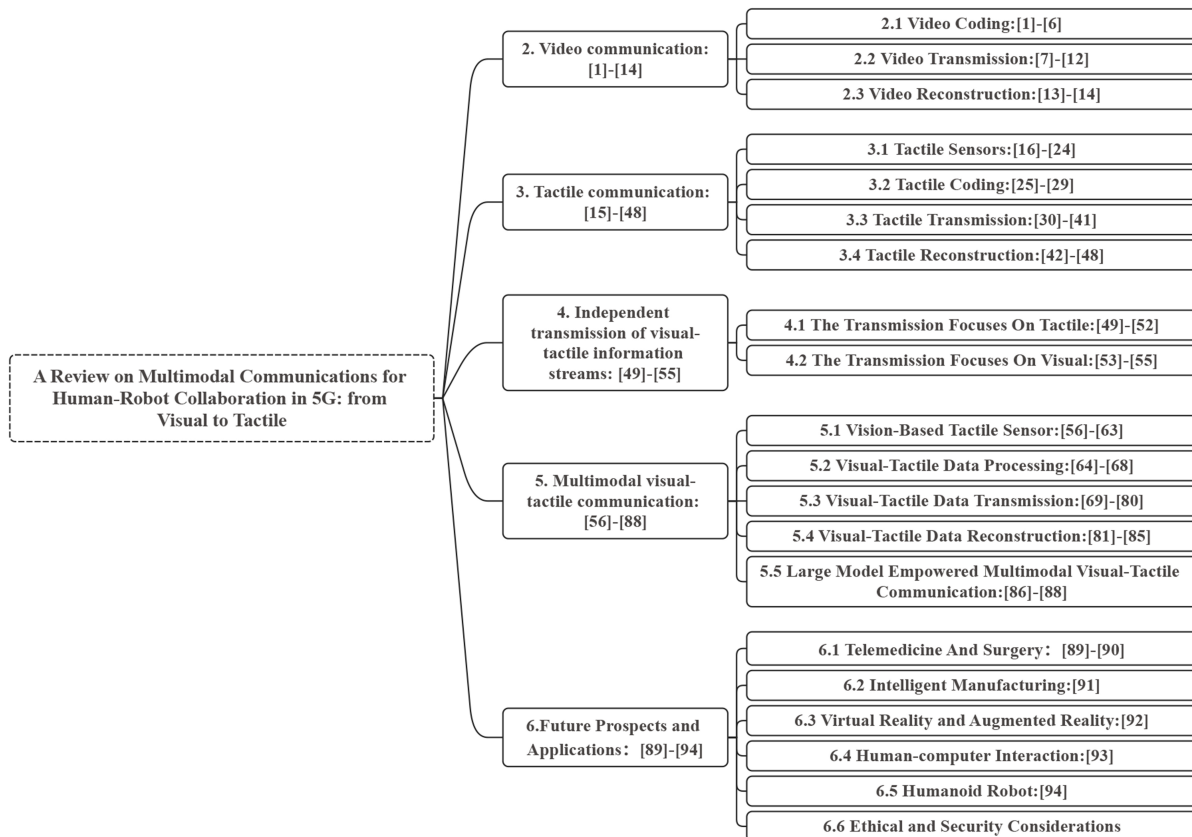


Figure 1. The structure of the paper.

mainstream technology in the field of digital video^[4].

However, with the popularity of ultra-high definition (4K/8K), multiview/stereoscopic video, and the rise of new media forms such as virtual reality (VR) and augmented reality (AR), the pressure on network transmission bandwidth is increasing exponentially. It puts more stringent requirements on the efficiency of the coding: the new generation of standards needs to achieve a reduction in bit rate of more than 50% while maintaining real-time. In addition, the requirements of robust transmission and multi-screen adaptive coding in heterogeneous network environments are driving the evolution of standard formulation from simply pursuing compression performance to multidimensional optimization. This evolution of demand led directly to the development of H.265/HEVC in 2013^[5]. Compared with H.264, H.265 significantly improves coding efficiency by introducing a more efficient parallel processing architecture and a more refined block coding technology, which promotes the popularity of ultra-high-definition 4K video.

In 2020, VCEG and MPEG further developed the versatile video coding (VVC) standard^[6], which achieves a bit rate reduction of approximately 50% while maintaining the same video quality. VVC has significantly expanded the application support range, adding support for computer-generated content (CGC), screen content (SCC), high dynamic range (HDR), and multilayer/multiview coding, while optimizing the encoding scheme of immersive media such as 360° panoramic video, providing technical support for the evolution of future media forms.

2.2. Video transmission

With the rapid development of video streaming technology, multiple protocols have emerged. The real-time messaging protocol (RTMP), primarily used for live streaming, relies on Flash technology and offers low-latency advantages suitable for real-time streaming. However, due to privacy concerns associated with Flash and its gradual obsolescence, RTMP is no longer widely used for audience-facing streaming transmission. Nevertheless, it remains valuable in closed-loop systems such as professional broadcasting workflows and surveillance networks, where its low-latency performance and compatibility with legacy encoding equipment are still advantageous. The advanced transport satellite protocol (ATSP), designed for low-latency broadcasting, does not support real-time streaming data transmission and is currently limited to specific fields such as security monitoring^[7]. In particular, ATSP's deterministic transmission features continue to make it suitable for mission-critical scenarios such as military communications or satellite-based telemetry, where stability and fixed bandwidth allocation are prioritized over adaptability.

Currently, dynamic adaptive streaming over HTTP (DASH), developed by the Moving Picture Experts Group (MPEG), has become the mainstream streaming protocol. It dynamically adjusts video quality according to network conditions, ensuring smooth playback across varying bandwidths^[8]. Using chunked transmission and standard HTTP protocols, DASH is compatible with existing network infrastructure, allowing simple deployment and low costs. Recent research has further improved DASH. For example, Wang *et al.* proposed an edge computing-based adaptive wireless video transcoding framework^[9], which deploys transcoding servers near base stations (BS) to significantly reduce the traffic of the core network and enhance the user experience under time-varying network conditions. Souane *et al.* introduced a deep reinforcement learning-based DASH streaming method that models the streaming process as a Markov decision process (MDP) and optimizes decisions via long short-term memory (LSTM) neural networks, allowing stable and high-quality video streaming^[10].

The surge in mobile data traffic and the dense deployment of BS have increased energy consumption for networks and user devices, while increasing operational costs. To address this, Abou-zeid *et al.* proposed an energy-efficient predictive green streaming (PGS) optimization framework^[11]. Its phased heuristic algorithm jointly optimizes resource allocation, video quality, and BS on/off strategies, reducing BS energy consumption by up to 85% while maintaining user experience. Zhou *et al.* proposed an efficient content distribution system (ECDS) based on device-to-device (D2D) communication for smart cities^[12]. Combining theoretical analysis with technical implementation, ECDS utilizes localized information for distributed decision making, addressing the high energy consumption of traditional cellular networks in content delivery.

2.3. Video reconstruction

In various stages of video signal acquisition, compression, transmission, and display, quality degradation inevitably occurs, which makes perceptual assessment of video quality critical. Common metrics for video quality assessment include the peak signal-to-noise ratio (PSNR) and the structural similarity index measure (SSIM). PSNR reflects the relative intensity of noise or distortion in video frames, while SSIM quantifies the visual structural similarity between reconstructed video and original content. Despite their limitations in fully capturing human perceptual experience, PSNR and SSIM remain widely used in industrial benchmarking, codec development, and hardware-level testing, due to their simplicity, reproducibility, and low computational overhead. In particular, PSNR is still a standard metric in many embedded systems and real-time evaluation scenarios where computational resources are constrained. Min *et al.* conducted extensive subjective and objective video quality assessment research covering both generic videos and specific categories such as streaming media, user-generated content, 3D, virtual and AR, HDR, high frame rate and visual applications^[13]. This work includes reviews of subjective evaluation tools and databases,

classification studies of general objective methods, and analyses of specialized assessment models for specific scenarios and emerging domains. In video transmission system evaluation, quality of experience (QoE) has become a crucial metric. Zhao *et al.* provide an overview of QoE modeling considering influential factors in end-to-end video transmission chains, QoE evaluation (including subjective testing and objective QoE monitoring) and QoE management strategies for video transmission over different types of networks^[14].

3. TACTILE COMMUNICATION

The human tactile system acquires tactile information related to contact through numerous receptors, with the information obtained generally categorized into two main types. Tactile information based on kinesthetic perception, also known as kinesthetic information, and tactile information based on cutaneous perception, referred to as tactile information. Kinesthetic information typically originates from the joints and ligaments within the human body, which encompass parameters such as force, velocity, and angular velocity. Tactile information includes various modalities such as static pressure, vibration, friction, surface texture, skin stretching, three-dimensional shape perception, thermal sensation, and pain perception^[15].

3.1. Tactile sensors

The human tactile system perceives mechanical stimuli via specialized receptors, which inspire the design goals of artificial tactile sensors: sensing forces from 0.01 to 10 N, achieving sub-millisecond response times, and maintaining spatial resolutions as fine as 1 mm in sensitive regions such as the fingertips^[16].

To fulfill these requirements, common tactile sensing mechanisms include capacitive, piezoresistive, piezoelectric, optical, and magnetic/inductive methods. Capacitive sensors, as demonstrated by Chen *et al.*^[17] and Rana *et al.*^[18], leverage microstructured dielectric layers and high-permittivity nanomaterials to enhance sensitivity and dynamic range. However, they remain vulnerable to parasitic capacitance, temperature variation, and require complex circuitry. Piezoresistive sensors offer simple, cost-effective designs, with the ability to detect both normal and shear forces^[19], but often suffer from noise, hysteresis, and higher power consumption. Piezoelectric devices respond rapidly to pressure changes and support high-frequency dynamic sensing, as shown by Liu *et al.*, though they are limited to detecting time-varying signals due to charge leakage^[20]. Optical sensors provide high spatial resolution and robustness against electromagnetic interference, employing methods such as light intensity variation^[21] and speckle pattern analysis^[22], though they can be bulky and alignment-sensitive. Magnetic and inductive sensing approaches use flexible magnetic nanocomposites or structured films to detect force-induced field changes^[23,24], enabling multi-axis sensing and potential integration with learning-based reconstruction.

As summarized in Table 1, these sensor types differ in terms of precision, stability, power consumption, and dynamic response. These characteristics have direct implications for tactile communication, influencing data rate requirements, signal-to-noise ratios, encoding schemes, and energy efficiency in transmission systems.

3.2. Tactile coding

3.2.1. Kinesthetic coding

Since kinesthetic signals are typically applied in closed-loop communication scenarios, they demand stringent network latency support, requiring encoders to strike an optimal balance between signal sampling and transmission. Consequently, kinesthetic encoder design has consistently followed two methodologies: real-time teleoperation (TPTO) and non-real-time encoding.

Table 1. Tactile sensor types and their communication implications

| Sensor type | Key features | Communication implications |
|----------------|-------------------------------------|---|
| Capacitive | High resolution, sensitive to noise | Requires high bandwidth; sensitive to EMI; needs robust encoding |
| Piezoresistive | Simple, low-cost, noisy output | Low data rate; limited accuracy; tolerant to low-complexity protocols |
| Piezoelectric | Dynamic response, no static sensing | Suits event-driven schemes; poor for continuous streaming |
| Optical | High accuracy, EMI-immune | High data throughput; stable in noisy environments |
| Inductive | High output, power-hungry | Supports longer range; energy cost impacts protocol design |

EMI: Electromagnetic interference.

Real-time teleoperation aims to make human operators imperceptible to intermediary technologies in practical applications. This necessitates codecs to adopt sample-by-sample processing schemes to ensure signal transmission transparency. A prominent approach in this domain is perceptual dead zone-based encoding. The perceptual dead zone represents a region below a defined perceptual threshold, where tactile samples falling within this zone are discarded due to insignificant signal variations^[25]. Predictive coding serves as a complementary real-time encoding strategy, where estimation algorithms predict future tactile sample values based on prior data. This allows transmission of only tactile samples that deviate significantly from predicted values, conserving communication bandwidth without substantial information loss^[26]. Combining these methods further reduces tactile data latency. For instance, Xu *et al.* proposed an enhanced kinesthetic encoding algorithm integrating dead zone encoding with piecewise linear prediction based on local linear features of position and velocity signals^[27]. This method leverages the sparse distribution characteristics of force signals, significantly improving encoding efficiency.

For non-real-time encoding, while increased latency enhances kinesthetic information encoding efficiency, it introduces stability risks to communication systems. To mitigate this, control mechanisms are typically incorporated. Differential pulse code modulation (DPCM) and adaptive differential pulse code modulation (ADPCM) are widely adopted in such codecs, effectively controlling distortion while achieving high compression ratios. Additionally, algorithms such as discrete cosine transform (DCT), fast discrete cosine transform (FDCT), and wavelet packet transform (WPT) have been applied to single-degree-of-freedom kinesthetic systems.

3.2.2. Tactile coding

Unlike kinesthetic communication, tactile information transmission operates in an open loop manner and does not impose stringent latency requirements. Hassen *et al.* proposed PVC-SLP, a perceptual coding method that uses the tactile sensitivity model of affective somatosensory feedback (ASF)^[28]. By introducing sparsity constraints and residual coefficients into linear prediction, along with optimizing residual and predictor coefficients, this approach effectively preserves the critical features of tactile signals. The diagram of PVC-SLP encoder and decoder is shown in Figure 2.

In addition, tactile coding schemes include waveform coding and parameter coding. Steinbach *et al.* suggested that waveform coding could utilize orthogonal linear transforms such as the DCT or discrete wavelet transform (DWT) to convert vibrotactile signals into alternative domains^[29]. Quantization eliminates signals with amplitudes below predefined thresholds, reducing redundancy in raw data before entropy encoding. They also argued that tactile features, such as friction and roughness of object surfaces, can be extracted to form parametric representations. These feature vectors, captured from materials, are transmitted to remote tactile rendering frameworks to replicate tactile perceptions in virtual environments.

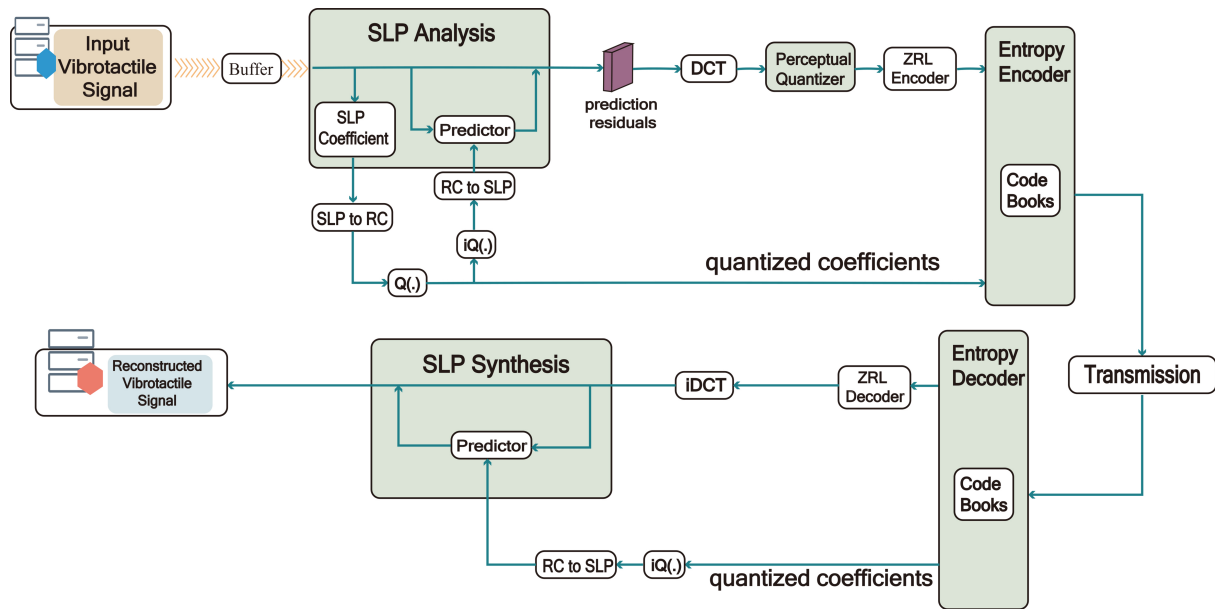


Figure 2. The proposed PVC-SLP coding method^[28].

While significant progress has been made in both kinesthetic and tactile coding techniques, several unresolved research problems remain. Current kinesthetic coding methods often adopt uniform encoding strategies across position, velocity, and force signals, despite their differing statistical and perceptual characteristics. This limits the adaptability and efficiency of real-time teleoperation systems, especially under varying task conditions or high degrees of freedom. Non-real-time approaches, although achieving higher compression rates, may compromise system stability and responsiveness. On the other hand, tactile coding schemes, including perceptual models such as PVC-SLP and transform-based methods, still face challenges in generalization across users and contexts. Most tactile codecs operate in isolation, without considering multimodal integration (MMI) or real-time constraints. Moreover, the trade-offs between compression efficiency, perceptual fidelity, and computational latency are not yet well-balanced for deployment in real-world, latency-sensitive applications. Future research should focus on developing adaptive, perceptually aware, and multimodal coding frameworks that can dynamically respond to user context, device constraints, and network conditions.

3.3. Tactile transmission

3.3.1. Tactile communication architecture

Huang *et al.* proposed a consensus-based peer-to-peer distributed control architecture to enable multiuser tactile interactions over partially connected and unreliable internet communication networks^[30]. This architecture addresses challenges such as heterogeneous latency and packet loss by locally simulating shared deformable virtual objects and synchronizing these local replicas across the Internet using consensus control algorithms, thereby ensuring consistent tactile experiences among users. Schuwerk *et al.* implemented a region-of-interest (RoI)-based communication framework for the shared haptic virtual environment (SHVE)^[31]. By dynamically forming communication groups based on users' current RoIs and leveraging the Age of Information (AoI) metric to reduce communication load, combined with deadband techniques to further lower update rates, this decentralized hybrid architecture excels in large-scale SHVE. It effectively minimizes network traffic while maintaining high-fidelity tactile interaction consistency among users.

The IEEE P1918.1 Standardization Group proposed a universal standardized architecture for the Tactile Internet^[25]. Designed with modularity and generality, this architecture interoperates with diverse network connectivity options, including wired and wireless networks, as well as dedicated and shared network technologies, making it suitable for most tactile communication and Tactile Internet research. Its core components include the tactile edge and network domains, integrating functional entities such as sensors, actuators, gateways, network controllers, and support engines (SEs). The group also defined multiple interfaces (e.g., access interfaces, tactile interfaces) to facilitate interactions between critical entities. Although compatible with existing technologies such as 5G and ultra-reliable low-latency communication (URLLC), the architecture still faces limitations in physical propagation distance.

To reduce geographical latency between user terminals and tactile servers, mainstream approaches leverage mobile edge cloud and cloud edge computing. Edge computing decentralizes computational resources to the network edge, such as BS or near end-user devices, enabling local processing of tactile, kinesthetic, and sensor data (e.g., tactile rendering, trajectory prediction) directly at the source. This significantly reduces cloud transmission delays, meeting the 1–10 millisecond ultra-low latency requirements of Tactile Internet (TI) scenarios. Local caching further minimizes redundant data transfers, improving real-time performance. A prime example is the MEC-enhanced cellular architecture^[32], which focuses on short-cycle real-time data analysis to drastically reduce latency. Proximity to end users allows edge nodes to efficiently filter and analyze data, improving processing efficiency. Devices operate at lower energy levels by offloading computations, while users benefit from high-quality personalized services due to cached content at the edge^[33–35].

Hou *et al.* introduced an AI-integrated edge computing communication framework, combining prediction and decision-making modules to maximize resource utilization while satisfying latency and reliability constraints^[36]. The incorporation of federated learning algorithms ensures that raw tactile data remain localized, preserving privacy and security. Wei *et al.* proposed a novel QoE-driven Tactile Internet architecture tailored for smart cities^[37]. Using a fast and reliable QoE management framework based on the broad learning system (BLS), this architecture achieves high QoE performance with low computational complexity and high availability.

3.3.2. Tactile communication protocol

The low-latency characteristics of tactile data conflict with the requirements of traditional transmission protocols: TCP offers high reliability, but introduces undesirable jitter through its congestion control mechanisms, making it unsuitable for real-time streaming. UDP, while ideal for real-time transmission, lacks congestion control and data integrity guarantees. The adaptive medical emergency service transport protocol (AMESETP), designed for real-time medical services, addresses these limitations by monitoring packet loss and frame delay variations at the receiver. Through feedback via ACK frames, the sender adjusts frame rates and data volume per frame. As demonstrated by its creators, AMESETP outperforms UDP and approaches real-time transport protocol (RTP) in performance. However, AMESETP lacks rapid adaptability and requires more granular service rate tiers and precise network condition detection mechanisms^[38].

Gokhale *et al.* proposed Haptics over IP (HoIP), an application layer protocol that digitizes tactile signals for adaptive sampling and employs a multithreaded architecture on transmitters and receivers to minimize processing delays. This approach achieved low-latency tactile data transmission while preserving perceptual quality. Two years later, they refined HoIP as a peer-to-peer tactile communication protocol, successfully reducing round-trip tactile data latency to below 30 milliseconds^[39,40].

The fuzzy real-time transport protocol for sensory data (FRTPS)^[41] is built on RTP by integrating fuzzy logic control. It dynamically adjusts transmission rates based on receiver-side metrics such as packet loss rate and frame delay variations, offering flexible rate adaptation strategies tailored for real-time multisensory data transmission.

Despite considerable advancements in tactile communication architectures and protocols, several critical challenges remain unresolved. Existing distributed and peer-to-peer architectures effectively address network heterogeneity and scalability, yet struggle with unpredictable latency and packet loss over unreliable networks, limiting their applicability in highly dynamic environments. Although edge computing and AI-integrated frameworks promise ultra-low latency and enhanced resource management, their deployment faces challenges related to interoperability, privacy preservation, and real-time decision accuracy. On the protocol side, conventional transmission protocols (TCP, UDP) fail to fully meet tactile data demands, while specialized protocols such as AMESETP and HoIP demonstrate improvements but require further refinement for adaptability, fine-grained network awareness, and robustness. Moreover, the integration of fuzzy logic and other adaptive rate control mechanisms shows potential but needs more extensive validation across diverse tactile use cases. Future research should focus on developing holistic tactile transmission solutions that combine adaptive, privacy-aware architectures with flexible, context-sensitive protocols capable of guaranteeing ultra-low latency, high reliability, and QoE in real-world tactile Internet scenarios.

3.4. Tactile reconstruction

3.4.1. Tactile information display

The reproduction of tactile information relies primarily on force feedback devices and sensors that employ vibration, ultrasonic, and electrostatic actuation mechanisms. Force-feedback devices integrate sensors and actuators controlled by DC motors. Sensors provide information about the position and orientation of the device in virtual/real environments. When the device interacts with objects, actuators show the corresponding forces or torques to the user.

Phung *et al.* proposed a novel touch-responsive bidirectional haptic display featuring integrated tactile sensors and actuators. This system dynamically delivers tactile feedback based on touch signals, surpassing human perceptual thresholds to convey rich tactile sensations^[42]. In wearable haptic displays, Uramune *et al.* introduced HaPouch, a wearable tactile device that addresses the limitations of traditional pneumatic haptic displays that require bulky air tubes and heavy compressors. Using liquid-gas phase-transition actuators, HaPouch provides tactile feedback with a response time of a few seconds, as validated through sensory evaluations. This innovation enables applications such as human-machine interfaces for force interaction^[43]. Zhu *et al.* developed TapeTouch, a handheld shape-shifting device capable of dynamically rendering varied shapes and softness in real time. Designed for VR applications requiring rich tactile feedback, such as virtual object manipulation, surgical training simulations, and remote operations, TapeTouch enhances immersive interaction through its adaptive physical properties^[44].

3.4.2. Tactile signal evaluation

The quality of tactile signals has historically been evaluated through subjective experiments. To date, only limited attention has been devoted to developing objective quality assessment methods for haptic communication. The most commonly used objective metrics remain simple, mathematically defined measures, such as the PSNR or mean squared error (MSE). Sakr *et al.* introduced the Haptic Perceptually Weighted PSNR (HPW-PSNR), which incorporates psychophysical models of tactile perception into the calculation process, offering a more accurate evaluation method^[45]. This HPW-PSNR employs mathematical

formulations and fuzzy logic principles to account for the perceptual significance of tactile signal degradation.

Perceptual mean squared error (PMSE) is an improved error metric that transforms the reference and compressed force signals into perceptual space using Fechner's Law. This approach better reflects human perceptual differences in tactile signals than simple numerical discrepancies, accounting for the non-linear characteristics of human sensory systems to align evaluation results with actual human experiences^[46].

A highly successful alternative for image quality assessment, the SSIM, has also been adapted for tactile quality and other multimodal signals such as video and audio. Hassen *et al.* proposed Haptic SSIM (HSSIM)^[47], which integrates human tactile perception models with the SSIM framework to more accurately predict perceived quality in force feedback signals. However, above studies focus solely on the feedback of force, position, and velocity, neglecting the feedback of the vibrotactile. To address this, Liu *et al.* designed a hybrid objective metric that combines SNR and SSIM^[48]. They first demonstrated the applicability of SNR and SSIM to vibrotactile quality assessment and developed the MULTi-Stimulus Hidden Reference Test (MUSHRT) protocol for subjective evaluations. The experimental results showed that the hybrid metric outperforms the standalone SNR and SSIM in time-varying scenarios, which proves to be effective in assessing vibrotactile signals.

Although significant progress has been made in tactile reconstruction technologies, both in hardware devices and signal evaluation methods, several challenges remain. Current force-feedback and vibrotactile devices still face limitations in size, response speed, and the richness of tactile sensations they can deliver, which affects the realism and immersion of user experiences. Striking a balance between portability and high performance continues to be a key design challenge. On the evaluation side, while perception-based objective metrics have been increasingly proposed to better capture human tactile experiences, most focus primarily on force and position signals, with limited comprehensive coverage of complex vibrotactile feedback. Subjective testing methods remain labor-intensive and difficult to standardize, hindering the generalizability and comparability of results. Future work should emphasize the development of efficient, multimodal tactile devices alongside the refinement of objective assessment frameworks that encompass multiple tactile dimensions, thereby driving improvements in tactile communication system performance and real-world applications.

4. SEPARATE TRANSMISSION OF VISUAL AND TACTILE STREAMS

Video and tactile signals exhibit significant differences in transmission latency, jitter, and reliability requirements, making effective transmission and processing of these signals a highly challenging problem. Current visual-tactile transmission schemes often focus on a single modality, leveraging lossless transmission of one modality to assist the other. For example, some approaches prioritize high-fidelity visual data to enhance tactile prediction accuracy, while others rely on precise tactile feedback to optimize visual compression or rendering efficiency.

4.1. The transmission focuses on tactile information

She *et al.* proposed a short-frame structure where each frame contains only the control signal phase, significantly reducing transmission latency^[49]. By optimizing resource allocation strategies, including transmission power, bandwidth, and duration, this design minimizes resource consumption while satisfying latency constraints. Nielsen *et al.* integrated multiple communication interfaces [such as Wi-Fi, long-term evolution (LTE), and high-speed packet access (HSPA)] and reduced overall transmission delays by optimizing data distribution weights across these interfaces^[50]. Since retransmissions increase latency,

traditional Hybrid Automatic Repeat reQuest (HARQ) protocols are suboptimal in 5G networks. Kotaba *et al.* combined non-orthogonal multiple access (NOMA) with HARQ, allowing concurrent transmission of retransmitted and new data packets from multiple users within the same time-frequency resource block^[51]. This approach reduces queueing delays and resource waste, achieving notable performance improvements. Tanveer *et al.* used reinforcement learning to optimize handover management in 5G ultradense small cell networks^[52], intelligently reducing handover latency and signaling overhead.

These methods effectively lower transmission latency, achieve higher spectral efficiency and system capacity under limited resources, and enhance tactile information reliability. However, they do not address the high-throughput demands of video data.

4.2. The transmission focuses on visual information

Given the substantial throughput demands of the video data, Yuan *et al.* proposed a distributed channel power allocation scheme based on an iterative stackelberg matching game^[53]. This approach enables efficient resource allocation without relying on global channel state information (CSI), maximizing the throughput of D2D pairs while suppressing interference to cellular links in D2D-enabled cellular networks. Zhang *et al.* integrated AI with D2D technologies, proposing advances such as D2D-enhanced mobile edge computing (MEC), D2D-enabled intelligent network slicing and NOMA-D2D-based cognitive networks^[54]. These innovations address the high-throughput requirements critical for 6G networks. Bennis *et al.* introduced a comprehensive framework optimized for low-latency and high-reliability communication^[55]. By leveraging techniques such as short transmission time intervals (TTIs), HARQ, and edge computing, the framework reduces latency. Simultaneously, multi-connectivity, redundant transmission, and network slicing enhance reliability, ensuring robust performance in dynamic network environments.

5. MULTIMODAL VISUAL-TACTILE COMMUNICATION

5.1. Vision-based tactile sensor

Previous sections have summarized advances in tactile sensors, but their limited measurement units and low-resolution tactile imaging capabilities restrict comprehensive information acquisition. Consequently, research focus has shifted to vision-based tactile sensors, which capture microscopic deformation data during object contact through imaging systems. Computer vision algorithms then extract tactile features from these images, enabling higher spatial resolution and richer tactile information compared to traditional tactile sensors.

Yamaguchi *et al.* developed a vision-based tactile sensor for robotic fingers, featuring a transparent multilayer structure that allows simultaneous capture of tactile data and external visual information during contact^[56]. This capability is critical for tasks such as cutting and grasping. In 2013, Yuan *et al.* introduced GelSight, a standard vision-based tactile sensor that employs red, green and blue (RGB) light sources arranged symmetrically around a central axis to generate uniformly illuminated tactile images^[57]. Later in 2018 and 2021, Donlon *et al.* iterated on this design with GelSlim^[58] and GelSight Wedge^[59], progressively achieving lighter designs and improved reconstruction accuracy. In 2020, Gomes *et al.* from the University of Liverpool extended the GelSight concept to create GelTip, a fingertip-shaped sensor that provides full-coverage tactile perception on the outer surface of the finger, including the tip and sides, mimicking human tactile capabilities^[60]. Fan *et al.* proposed ViTacTip, which integrates a biomimetic tip to amplify tactile signals during interactions. This design naturally fuses visual and tactile data within a single device, producing hybrid tactile-visual images^[61].

Another category involves binocular or multi-camera imaging systems. Kuppuswamy *et al.* from Toyota Research Institute developed the Soft Bubble sensor, which uses an internal RGBD camera to track 3D deformations of a soft membrane structure^[62]. Zhang *et al.* introduced Tac3D, a vision-based tactile sensor for measuring 3D contact surface geometry and force distribution. Using virtual binocular vision technology, Tac3D offers advantages such as simple structure, low computational cost, and affordability^[63]. Table 2 lists the structure design, tactile sensing functions and tactile modes of mainstream visual and tactile sensors.

5.2. Visual-tactile data processing

Recently, object recognition has emerged as a critical focus in visual-tactile communication, where the fusion of visual and tactile learning demonstrates significant potential for performance improvement. However, cross-modal fusion of visual and tactile features faces multiple heterogeneity challenges. In terms of spatiotemporal characteristics, visual systems rely on low-frequency discrete sampling (such as 30 Hz RGB images), while tactile systems capture dynamic contact forces and vibrations through high-frequency continuous acquisition (up to kHz-level rates), resulting in difficulties in spatiotemporal feature alignment. Representationally, visual data predominantly resides in high-resolution 2D pixel space, whereas tactile data combines 1D temporal signals (such as force trajectories) with low-resolution tactile images, creating structural disparities. At the perceptual semantic level, vision emphasizes global geometry and texture, while touch focuses on local material properties (such as stiffness and friction coefficients) and microscopic topography. Despite their complementary functions, a semantic gap hinders effective collaboration between the two modalities. These challenges, temporal asynchrony, representational heterogeneity, and semantic asymmetry, render traditional feature-level fusion methods insufficient to establish intrinsic cross-modal correlations, necessitating joint representation frameworks to bridge deep modal differences.

To align interaction information between visual and tactile modalities for object recognition, Liu *et al.* proposed a kernel sparse coding method^[64]. This method uses kernel sparse coding to map data to high-dimensional feature space, and introduces group sparsity constraint (L2,1 norm) to jointly optimize sparse coding results of visual and tactile modes, thus fusing multimodal information at the group level. Lee *et al.* introduced a generative adversarial network (GAN) that utilizes correlations between vision and touch, capable of generating realistic visual images from tactile inputs or synthesizing tactile images from visual inputs^[65]. However, these methods merely concatenate visual and tactile information without accounting for their unique interaction dynamics. Addressing this, Wei *et al.* developed an alignment-based multiscale fusion model. As can be seen in Figure 3, the method learns to align single modal features by visual-tactile contrast, and integrates visual and tactile features with different resolutions using a multi-scale Transformer fusion module. The fused features are then processed by element-by-element summation and MLP header to achieve object classification^[66]. Babadian *et al.* designed dedicated temporal feature extraction architectures, TactileNet for tactile data and VisionNet for visual data, and built a feature fusion framework on this basis^[67].

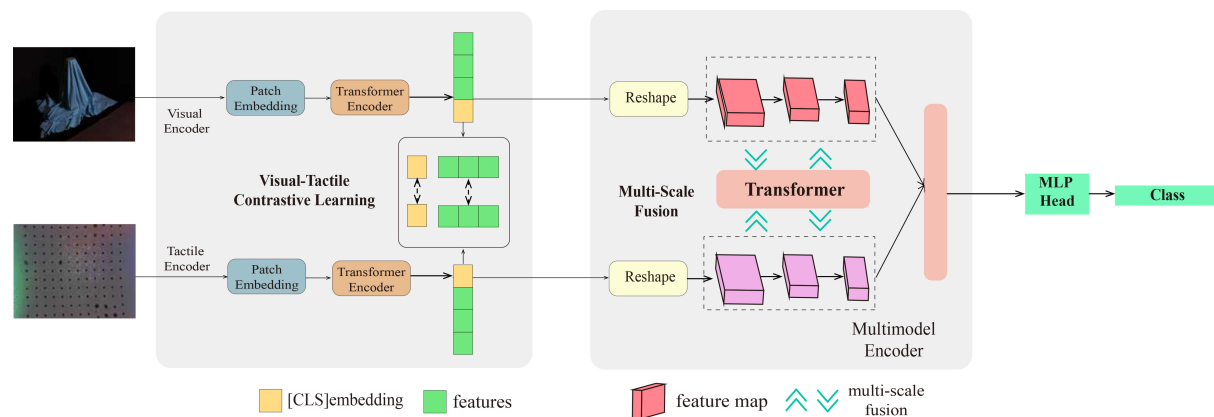
Cross-modal transfer learning has also emerged as a promising research direction due to its efficient data utilization, strong cross-domain generalization, and model stability. Falco *et al.* proposed a transfer learning approach that identifies shared representations between visual and tactile modalities and defines domain-adaptive features^[68], achieving cross-domain feature transfer and object classification. Experimental results validate its outstanding performance in real-time processing and accuracy.

Visual-tactile data processing encounters significant challenges that impede effective multimodal fusion. A key issue is the spatiotemporal misalignment between modalities - visual data is typically captured at low

Table 2. Mainstream visual and tactile sensor structure design, tactile sensing function and tactile mode

| Sensor | Structural form | Contact colloid layer | Illuminant | Camera system | Force estimation |
|-----------------------------|-----------------|-----------------------|------------|-------------------|------------------|
| GelSight ^[57] | Fingertip | Array marker | RGB | Monocular | Neural network |
| GelSlim ^[58] | Fingertip | Array marker | RGB | Monocular | Finite element |
| GelTip ^[60] | Finger | Unmarked point | RGB | Monocular | None |
| Soft Bubble ^[62] | Palm | Dense marker | Monochrome | Monocular + depth | None |
| Tac3D ^[63] | Palm | Array marker | Monochrome | Virtual binocular | Finite element |

RGB: Red, green and blue.

**Figure 3.** The proposed alignment-based multiscale fusion model^[66].

frequencies, while tactile data involves high-frequency continuous signals, complicating temporal synchronization. Additionally, representational heterogeneity poses difficulties: visual data consists mainly of high-resolution 2D images, whereas tactile data combines 1D temporal signals with low-resolution tactile images, creating structural disparities that hinder unified feature extraction. Beyond this, a semantic gap exists - vision focuses on global geometry and texture, while touch emphasizes local material properties and microscopic details. Many current fusion methods simply concatenate features without fully modeling the dynamic interaction between modalities, limiting their ability to capture intrinsic cross-modal correlations. Thus, designing joint representation frameworks that reconcile modality-specific traits while leveraging complementarities remains a key open problem. Cross-modal transfer learning shows promise for improving generalization and data efficiency but still faces challenges adapting robustly across varied scenarios. Real-time processing demands add further complexity, requiring a balance between computational efficiency and accuracy. Moreover, the lack of large-scale, well-annotated synchronized datasets constrains the development of advanced models. Tackling these challenges is critical to unlocking the full potential of visual-tactile integration in object recognition and related fields.

5.3. Visual-tactile data transmission

As discussed in Chapter 4, most visual-tactile communication methods focus on a single modality and do not address both visual and tactile data simultaneously. To address significant differences in transmission latency, jitter, and reliability between visual and tactile signals, Zhou *et al.* proposed a scheduling and transmission framework that takes advantage of the distinct characteristics of visual and tactile streams^[69]. For tactile dominant services, tactile data is embedded in time-frequency resource blocks originally allocated to audio and video streams, reducing tactile transmission latency while improving the utilization of network resources. In another approach, a heterogeneous stream cotransmission framework was introduced to pretransmit future tactile signals via prediction mechanisms, lowering latency while

compensating for reliability losses through D2D communication. This design satisfies the URLLC requirements of tactile signals and the enhanced mobile broadband (eMBB) demands of video streams, with experimental results showing reduced power consumption along with the meeting of heterogeneous cross-modal requirements^[70].

Wu *et al.* further developed a cross-modal stream transmission architecture that leverages intra-modal redundancy and inter-modal correlations. By designing adaptive stream scheduling and joint uplink/downlink resource allocation schemes, this method optimizes the symmetry of bidirectional tactile transmission to meet the low latency, high reliability and high performance requirements for multimodal services^[71]. Tong *et al.* proposed a transmission strategy incorporating proactive packet dropping and cross-modal recovery, which selectively discards packets from specific modalities based on user QoE metrics and reconstructs lost signals at the receiver using cross-modal recovery models. This approach effectively alleviates network congestion while improving interaction transparency and user experience^[72].

To generalize across diverse multimodal service scenarios, Wei *et al.* advanced a universal cross-modal transmission strategy. At the transmitter, a mutual transmission mechanism between visual and tactile signals eliminates redundancy through intermodal correlations, for example, by reducing the sampling rates of the visual frame during transmission of tactile signals (H2VC) or using visual redundancy to minimize transmission of tactile data (V2HC)^[73]. Suo *et al.* tailored three transmission strategies for different service types: Strategy A prioritizes audio-video streams via D2D or cellular modes while multiplexing tactile streams to conserve spectrum; Strategy B focuses on tactile-dominant services using D2D or relay modes for latency reduction; Strategy C balances audio-video bandwidth and tactile latency demands with flexible mode selection. These strategies are integrated into a joint mode selection and resource allocation framework, optimized via an improved NSGA-III multi-objective algorithm. The framework dynamically allocates bandwidth, power, and relay nodes to maximize audio-video rates, minimize tactile latency, and improve throughput and energy efficiency, demonstrating effectiveness in addressing diverse multimodal requirements through simulations^[74].

The achievements in the field of video streaming transmission indicate that edge computing is a potentially effective way to achieve the goal of large-scale multimedia streaming transmission. Typically, these studies focus on the three key characteristics of edge computing, namely data caching, computing processing, and communication interaction. Li *et al.* introduced adaptive streaming technology and utilized the active caching function of edge computing for video streaming transmission, enabling the caching of multiple representations for the same video^[75]. This cache optimization framework not only solves the heterogeneity problem of users, but also achieves a higher cache performance improvement than the ordinary video file caching scheme. It can optimally allocate the cache resources of edge servers, not only among different videos but also among multiple representations of the same video, in order to alleviate network congestion. In order to manage the network resources at the edge of multimedia services reasonably, Li *et al.* proposed a novel mobile edge cache placement optimization framework, which is applicable to the video-on-demand system based on adaptive streams^[76]. It considers the bitrate-distortion characteristics of different video representations and maximizes the reduction of the average video distortion for all users through integer linear programming (ILP). It simultaneously meets the constraints of the backhaul link, the storage capacity of the edge server, and the user transmission and initial startup delays. Chen *et al.* studied the caching of computing tasks in the edge cloud for the first time, analyzed the influence of task popularity, task content size and required computing power on the caching strategy, and provided the optimal caching strategy for computing tasks^[77]. Gao *et al.* proposed an edge intelligence architecture that integrates Communication, Caching, Computation and Control^[78]. It aims to address the diverse service requirements and strict

technical constraints existing in the transmission process of heterogeneous modal data. This work introduces AI methods to enhance system performance: The privacy and security of D2D communication and device-to-edge are guaranteed respectively by using blockchain and federated learning. The content popularity prediction and efficient cache scheduling for dynamic environments are realized by combining transfer learning and collaborative filtering. The UE-Edge collaborative computing mechanism is proposed to reduce energy consumption and latency, and a unified optimization control model is constructed to jointly schedule communication, cache and computing resources. To enhance the system's adaptability to network dynamics and service heterogeneity, the author designed A deep reinforcement learning method based on the attention mechanism (A-DRL) to achieve online optimization decision-making for resource allocation. The implementation of this 4C-based autonomous strategy facilitates the attainment of a desirable cross-modal transmission performance.

In traditional cross-modal stream transmission mechanisms, a unified scheduling strategy is often adopted, which neglects the semantic and spatiotemporal correlations between modalities. As a result, it becomes challenging to guarantee the extreme latency and reliability requirements of haptic streams while maintaining the quality of other modalities such as video. To address this issue, Yuan *et al.* proposed a content-aware cross-modal stream transmission (CCST) architecture^[79]. By mining the content relevance between video and haptic data, CCST enables haptic streams to preempt video resources that are highly correlated with them. At the receiver side, a content reconstruction model is employed to compensate for video degradation caused by preemption. Specifically, the authors formulated a joint scheduling model with the objective of maximizing video utility under the constraint of haptic service rate, and designed an online resource allocation algorithm adaptive to dynamic channel conditions. Experimental results demonstrated that this strategy ensures high reliability of haptic service while effectively mitigating the degradation of video quality due to resource competition, thereby significantly enhancing overall system performance. This work validates the potential of leveraging inter-modal semantic correlations for transmission scheduling and expands the design paradigm of cross-modal communication mechanisms.

Building upon this insight into inter-modal coordination, Gao *et al.* further explored the application of cross-modal communication in the context of holographic video streaming, which refers to the transmission of volumetric video content that supports realistic, three-dimensional immersive visual experiences. Due to the extremely high data rates, resource contention, and network dynamics involved, holographic video presents considerable challenges for wireless transmission. To this end, Gao *et al.* developed a comprehensive optimization framework encompassing visual saliency prediction, extreme haptic compression, and rate-adaptive scheduling^[80]. At the encoding layer, a visual saliency modeling method was proposed that integrates audio and haptic cues to improve the prediction accuracy of user attention regions, thereby enabling more efficient compression of holographic video. Additionally, a modality-correlated haptic encoding architecture was designed to achieve perceptually lossless compression, alleviating resource conflicts between haptic and holographic video streams. At the transmission layer, the authors introduced a reinforcement learning-based priority-aware rate adaptation mechanism, which dynamically adjusts transmission strategies based on both the importance of video tiles and real-time network conditions. This approach significantly enhances user immersion and system robustness. As the first study to systematically incorporate multimodal collaboration as a central optimization principle in the holographic video context, this work offers critical technical support for next-generation immersive media services.

Visual-tactile data transmission faces significant challenges due to the inherent heterogeneity of the two modalities, particularly in latency, jitter, and reliability requirements. The asynchronous and distinct

characteristics of visual and tactile streams complicate unified scheduling and resource allocation, making it difficult to simultaneously guarantee ultra-low latency and high reliability for tactile data while maintaining high-quality visual transmission. Existing transmission frameworks often overlook the semantic and spatiotemporal correlations between modalities, resulting in suboptimal bandwidth utilization and quality degradation, especially under network congestion. Furthermore, the dynamic and unpredictable nature of wireless environments demands adaptive and intelligent transmission strategies that can respond in real-time to varying channel conditions and user demands - yet such mechanisms remain underdeveloped. The efficient integration of edge computing for caching, processing, and communication introduces additional complexity, as optimal resource management must balance computational overhead, energy consumption, and latency constraints across modalities. Moreover, with emerging applications such as holographic video streaming that require extremely high data rates and tight coordination between visual and haptic information, current transmission methods face scalability and robustness limitations. Privacy and security considerations also add another layer of challenge, as cross-modal transmission protocols must safeguard sensitive tactile and visual data without compromising performance. These open questions highlight the need for novel cross-modal scheduling algorithms, semantic-aware resource allocation, robust predictive encoding schemes, and adaptive learning-based frameworks to meet the stringent demands of next-generation immersive multimedia communication systems.

5.4. Visual-tactile data reconstruction

At the receiver end, ensuring the integrity of tactile and visual signals and enhancing the quality of reconstructed information is critical, making visual-tactile signal reconstruction a key focus. This process leverages the received signal data to provide high-quality restored images to end users.

Wei *et al.* highlighted that tactile signals are often bursty and unpredictable, highly susceptible to interference and noise, leading to significant quality degradation at the receiver^[81]. In addition, tactile signals may be entirely absent in scenarios where tactile sensors or acquisition devices are unavailable. To address these challenges, Wei *et al.* proposed audio-visual aided haptic reconstruction (AVHR), a cloud-edge collaborative method. By leveraging large-scale audio-visual databases in the cloud for knowledge learning and transferring this knowledge to edge nodes, AVHR employs multimodal fusion to extract shared semantics from audio and visual signals and then uses an autoencoder architecture to synthesize tactile signals. Experimental results demonstrate improved reconstruction efficiency and quality while meeting low-latency, high-reliability tactile transmission requirements, making it suitable for applications such as remote industrial control and haptic-enhanced reality.

Chen *et al.* introduced a vision-aided cross-modal tactile reconstruction method that maps spatio-temporal visual features (e.g., deformation, contact regions) to tactile signals, which is shown in Figure 4. Combining low-rank decomposition, attention mechanisms, and lightweight convolutional neural networks (CNNs), this approach achieves efficient tactile force reconstruction, addressing traditional limitations such as reliance on complex networks and sensitivity to lighting conditions^[82].

A Clustering-guided Triplet-constrained deep clustering network (CT-DCN) was proposed to explore fine-grained subcategory information in tactile signals^[83]. Through triplet-constrained learning, shared semantics among audio, visual, and tactile signals are extracted, enabling the generation of high-quality tactile signals under weak supervision (lacking fine-grained labels) and weak pairing (ambiguous cross-modal correspondences). This method mitigates interference during transmission and synthesizes “virtual” tactile signals, making it ideal for immersive multimodal services.

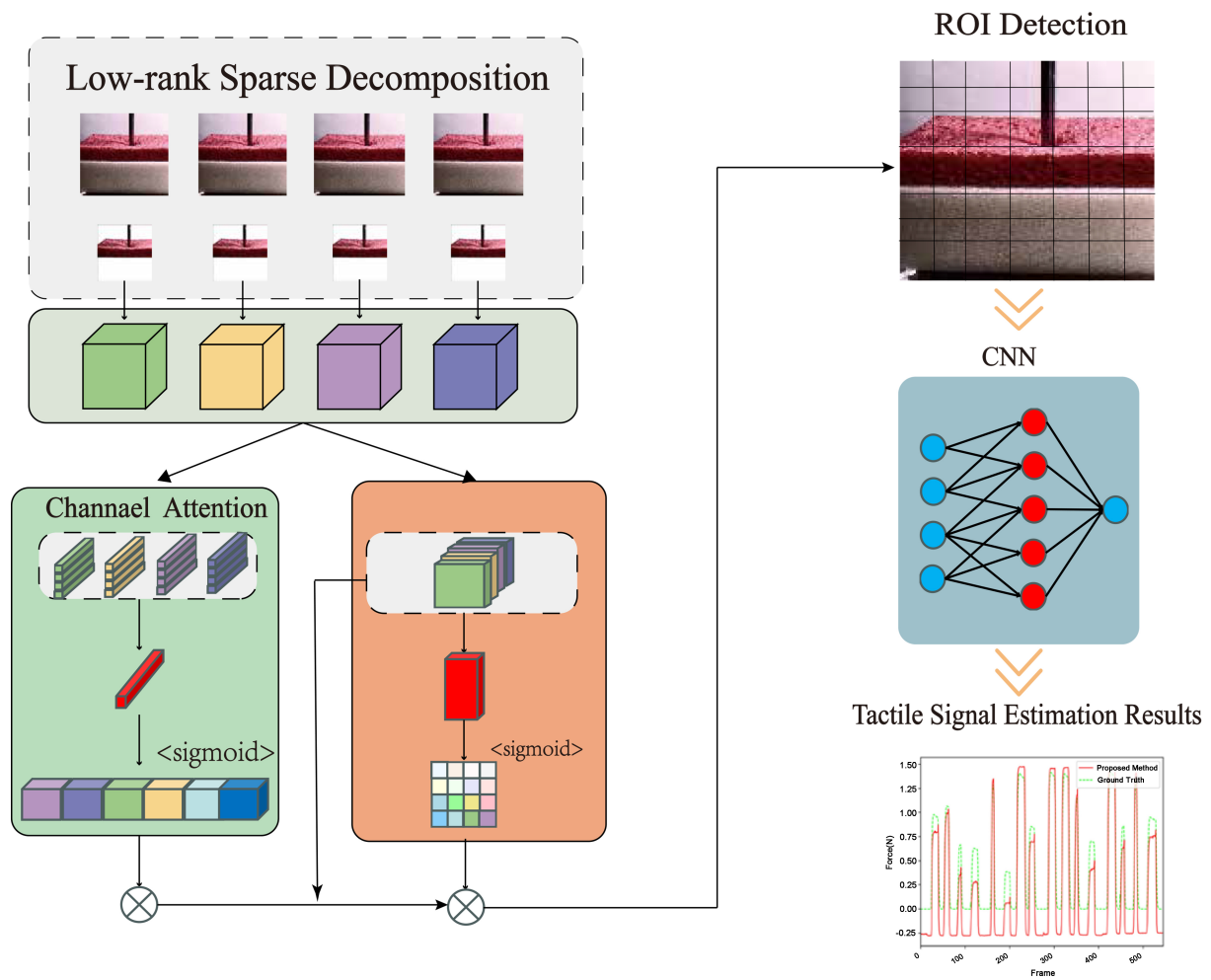


Figure 4. The proposed cross-modal reconstruction method^[82].

Chen *et al.* developed a masked autoencoder with denoising adversarial networks (MAE-D) for signal reconstruction. By integrating autoencoders and GANs, MAE-D achieves robust multimodal signal recovery under poor network conditions, compensating for packet loss, a critical capability for teleoperated robotics^[84].

Human subjective perception remains pivotal in visual-tactile communication, yet few studies prioritize perceptual-driven reconstruction. Wei *et al.* exploited the intrinsic links between auditory, tactile, and visual signals in human perception to reconstruct delayed or degraded visual signals^[85]. Specifically, they designed a perception-driven cross-modal reconstruction framework, including a time-frequency masking-based auditory-tactile redundancy reduction mechanism at the transmitter to minimize non-visual data transmission, and an auditory-haptic fused visual reconstruction (AHFVR) method at the receiver to restore impaired visual signals. Evaluations on standard multimodal datasets and real-world platforms confirmed superior visual reconstruction quality and enhanced user immersion. Future research should further explore human perceptual mechanisms to refine reconstruction methodologies.

Ensuring the integrity and high quality of reconstructed visual and tactile signals at the receiver remains a significant challenge due to the inherently bursty and unpredictable nature of tactile data, which is highly

vulnerable to noise and interference during transmission. In many practical scenarios, tactile signals may be partially or completely lost, especially when tactile sensing devices are unavailable or malfunctioning, further complicating reliable signal recovery. Although recent approaches leverage visual information to aid tactile reconstruction and employ advanced techniques such as attention mechanisms and lightweight neural networks, these methods still face limitations in handling varying environmental conditions, real-time processing constraints, and robustness to incomplete or degraded input data. The use of weakly supervised learning to extract fine-grained tactile semantics with limited labeled data shows potential but requires further development to effectively manage ambiguous cross-modal correspondences and transmission disturbances. Furthermore, integrating denoising and generative adversarial frameworks has improved reconstruction performance under poor network conditions, yet balancing reconstruction accuracy with computational efficiency and latency remains unresolved. Crucially, the role of human perceptual characteristics in guiding reconstruction algorithms has been underexplored. Aligning reconstruction processes with human sensory integration could significantly enhance subjective experience and immersion, suggesting a vital direction for future research. Overall, advancing adaptive, perception-driven visual-tactile reconstruction frameworks that ensure high-fidelity, low-latency recovery in realistic communication scenarios while optimizing resource usage represents a key open challenge for the field.

5.5. Visual-tactile communication with large models

The core of multimodal visual-tactile communication lies in integrating visual information with haptic feedback to enable more natural and immersive interactive experiences. With the rapid advancement of AI, particularly the emergence of large-scale models, new opportunities have arisen for this field. In multimodal visual-tactile communication, large-scale models are primarily applied to cross-modal signal processing and fusion, dynamic signal modulation and optimization, and personalized interaction enhancement. By learning intrinsic correlations between visual and tactile modalities, these models leverage deep learning techniques such as GANs or autoencoders to convert visual information into tactile signals or assist in the recovery and enhancement of haptic signals. In addition, large-scale models can analyze environmental parameters and user behavior in real time, dynamically adjusting the intensity, frequency, and patterns of the tactile signal to mitigate interference factors such as environmental noise and device limitations, thus improving the signal reliability and user experience. Furthermore, they generate personalized visual-tactile feedback based on user preferences and behavioral patterns, optimizing signal presentation to enhance interaction naturalness and immersion.

Currently, Wei *et al.* proposed an AI-driven multimodal communication framework that addresses data privacy through federated learning, optimizes visual-tactile signal cotransmission through reinforcement learning, and achieves cross-modal signal fusion and recovery using transfer learning, significantly improving communication performance and user experience^[86]. Farooq *et al.* introduced magnetorheological fluids as a modulation medium for tactile signals and combined them with deep learning algorithms to dynamically optimize tactile transmission, enhancing the reliability and integrity of haptic feedback in noisy environments^[87]. Cheng *et al.* presented a framework integrating multimodal large language models (MLLMs) with integrated sensing and communication (ISAC) systems^[88]. Using multimodal encoders, word embedding layers, and multihead attention mechanisms, this approach processes and fuses multimodal sensing data and communication control signals. Through a beam prediction case study, Lipkova *et al.* demonstrated MLLM's superior generalization capabilities and performance over traditional methods^[89]. However, challenges such as theoretical boundary gaps, the design of cloud-edge-device collaboration mechanisms, and data privacy and security concerns remain unresolved.

In the field of large-scale model-driven multimodal visual-tactile communication, current research remains nascent. Critical areas of improvement include optimizing model efficiency for low-latency requirements, enhancing multimodal data fusion and alignment, strengthening data privacy and security, refining cloud-edge-device coordination, and tailoring model performance for specific applications. Further exploration is essential to address these challenges and advance technology.

6. FUTURE PROSPECTS AND APPLICATIONS

6.1. Telemedicine and surgery

By integrating multimodal perception and feedback mechanisms, visual-tactile communication technology establishes a bidirectional bionic interaction framework for telemedicine. Using the synergy between tactile sensing and stereo vision, this technology enables low-latency transmission of high-precision operational commands and biomechanical data, effectively addressing the critical limitation of missing tactile feedback in traditional telesurgery. In clinical applications, dynamic force field modeling and adaptive control strategies significantly enhance the safety of minimally invasive procedures and tissue preservation. Coupled with novel network architectures, it facilitates cross-regional medical collaboration and demonstrates rapid-response capabilities in emergency scenarios.

Current advancements focus on multiphysical quantity perception fusion and VR mapping, progressively overcoming limitations in tactile information dimensionality and biomechanical tissue simulation accuracy through optimized haptic encoding and deep integration with digital twin systems. An emerging trend is the convergence of visual-tactile sensing with high-dimensional clinical data to create intelligent and adaptive surgical systems. For instance, AI-driven multimodal data integration has shown remarkable promise in enhancing diagnosis and intraoperative decision-making. A representative study by Lipkova *et al.* demonstrates that fusing heterogeneous medical data substantially improves predictive modeling in oncology, uncovering novel biomarkers and enabling more personalized treatment strategies^[89]. This illustrates how combining sensory data with high-level clinical knowledge can create a more context-aware and patient-specific surgical experience.

Extending this idea, Shao *et al.* further emphasize that MMI of medical imaging, genetic profiles, and clinical records using AI can significantly improve mutation status prediction, which is vital for precision oncology^[90]. Such predictive capabilities, when embedded into the surgical workflow, can inform real-time treatment decisions - such as targeted resections or intraoperative adjustments - based not only on visual and tactile cues but also on predicted molecular characteristics of tissue.

These developments point to a broader trajectory in which telemedicine systems evolve from purely sensor-driven platforms to context-rich, data-centric environments that support personalized interventions. However, achieving this integration poses challenges in aligning heterogeneous data modalities across spatial and temporal scales, ensuring data interoperability, and building trustworthy AI decision frameworks. Addressing these technical bottlenecks is crucial to realizing the full potential of intelligent remote surgery.

Future developments aim to build cross-modal perception loops and intelligent edge computing platforms, expanding tactile feedback from static force perception to dynamic physical fields such as temperature and texture. In addition, future telesurgical systems must incorporate standardized safety verification protocols, AI-driven anomaly detection, and hierarchical trust mechanisms to ensure stable operations under uncertain network or physiological conditions.

6.2. Intelligent manufacturing

Visual-tactile communication technology has emerged as a pivotal force in the intelligent transformation of industrial robotics by enabling real-time interaction and closed-loop control of multidimensional perceptual data. This technology establishes a cross-modal perception system that integrates vision, force sensing, and pose coordination, allowing operators to remotely control robots with high precision while receiving dynamic contact force field distributions and high-resolution visual feedback. Such capabilities significantly improve operational safety and environmental adaptability in complex scenarios. In extreme environments, the system achieves sensitive responses to micrometer-level contact force variations through real-time force field modeling and adaptive control strategies.

Technological advancements are now expanding from single-mode force feedback to multi-physical quantity fusion perception. By integrating novel sensing architectures with multimodal information processing frameworks, the system enables a comprehensive analysis of object surface properties (e.g., texture characteristics) and thermodynamic parameters. In parallel, the integration of visual-tactile data with large-scale industrial datasets - including sensor readings, production logs, and textual records - is becoming essential for enabling intelligent decision-making at the system level. Recent work by Wang *et al.* demonstrates that multimodal large language models (MLLMs) can effectively unify these diverse data sources through semantic tokenization and cross-modal alignment, providing real-time decision support in complex manufacturing scenarios^[91]. Their Transformer-based framework supports dynamic anomaly detection, predictive maintenance, and visual question answering, revealing the potential of combining physical interaction data with knowledge-based reasoning for enhanced manufacturing intelligence.

This suggests that future visual-tactile systems should not only focus on sensor-level feedback control but also interface with high-level decision modules powered by MLLMs. By linking real-time physical interactions to contextual understanding of production workflows, such systems can support closed-loop decision-making under uncertainty. However, these advancements face challenges such as domain-specific pretraining data scarcity, real-time multimodal reasoning latency, and cross-device standardization for data exchange.

These breakthroughs not only provide innovative solutions for precision assembly and hazardous environment operations but also lay a technical foundation for upgrading smart manufacturing toward autonomous decision making and highly interactive modes. Ultimately, the convergence of tactile sensing, machine vision, and large-scale multimodal cognition marks a significant step toward building highly resilient and context-aware industrial systems capable of intelligent collaboration with human operators.

6.3. VR and AR

Visual-tactile communication technology significantly enhances the interactive realism and environmental immersion of VR systems through multimodal sensory fusion mechanisms. Using collaborative modeling of haptic rendering algorithms and physics engines, this technology reconstructs surface textures, deformation characteristics, and mechanical responses of virtual objects, creating embodied interaction experiences. In professional training applications, high-fidelity reproduction of tissue elasticity is achieved in surgical simulations using biomechanical haptic feedback. In educational contexts, it deepens operational cognitive understanding in virtual experiments through tactile-visual synergy.

A notable implementation of such MMI is seen in the haptic-enabled VR learning framework proposed by Sanfilippo *et al.*, which leverages auditory, visual, and tactile feedback to improve engagement and presence in virtual educational tasks^[92]. By augmenting the commercial Valve Knuckles EV3 controllers with vibrotactile actuators, and integrating them with the Unity game engine for real-time 3D interaction

rendering, the system provides low-cost, scalable access to immersive learning environments. Human-subject experiments using this setup - focused on tasks such as waste-sorting training - demonstrated clear improvements in spatial presence, realism, and cognitive involvement as measured by the Igroup Presence Questionnaire (IPQ).

Current technological breakthroughs focus on flexible electronic skins and neuroinspired encoding theories, with the aim of addressing spatio-temporal synchronization challenges in cross-modal perception. Nonetheless, limitations persist in aligning vibrotactile feedback intensity and spatial targeting with visual deformation cues, especially when interaction occurs over deformable or dynamic objects. Additionally, maintaining low-latency response loops while integrating high-resolution haptic rendering in Unity remains a technical hurdle, particularly on standalone or mobile VR platforms. Looking ahead, the integration of distributed haptic network architectures will propel the evolution of metaverse interaction paradigms toward multiphysical coupling dimensions, enabling seamless integration of tactile, thermal, and kinesthetic feedback. Future research is likely to explore more deeply the standardization of multimodal VR frameworks using cross-platform engines such as Unity, in combination with open APIs and modular actuator kits, to further democratize the deployment of haptics-enhanced immersive learning and simulation environments.

6.4. Human-robot interaction

Visual-tactile communication technology, by integrating multimodal sensing and feedback, is gradually reshaping the interface between humans and intelligent machines - not only in prosthetics and wearables, but also in broader HRI contexts. As robots are increasingly deployed in shared environments, the ability to understand and respond to human motion with fluidity and precision becomes essential. Studies have shown that while vision provides anticipatory awareness - detecting human trajectories and predicting potential collisions - tactile feedback remains indispensable for regulating contact forces in real time. However, most conventional systems treat these modalities in isolation, often reacting only before or after a collision occurs. This fragmented approach misses the continuous nature of physical interaction.

Rather than relying on reactive control alone, emerging work has begun to integrate visual and tactile cues within unified, hierarchical frameworks. Xu *et al.*, for instance, developed a system in which vision modulates robot velocity based on proximity, while tactile sensors adjust compliance in response to contact intensity^[93]. Their robot, ViTaR, embeds this coordination within an event-driven mechanism that selectively activates tactile responses when visual cues suggest high-risk proximity, thus minimizing both network overhead and response latency. In repeated HRI tests, this system not only reduced impact forces significantly but also demonstrated that only a continuously adaptive, multimodal control loop can match the fluidity and safety demands of real-world collaboration.

This shift - from segmented to continuous perception - raises new technical challenges. Synchronizing distributed tactile arrays in real time, ensuring low-latency actuation across high-DOF systems, and interpreting whether a contact is accidental or intentional all require systems that not only sense but understand context. These are not just engineering hurdles - they are perceptual and cognitive ones as well.

At the same time, research into neural-compatible interfaces and multimodal feedback encoding is driving the development of prosthetics that can both interpret fine-grained motion and provide tactile return. As flexible electronics become capable of capturing not just pressure but also temperature, humidity, and shear force, the boundary between machine response and human intuition continues to blur - bringing HRI ever closer to natural, adaptive, and safe coexistence.

6.5. Humanoid robot

Visual–tactile fusion is redefining the capabilities of humanoid robots, enabling them to interact with their environments in ways that go beyond mechanical precision to incorporate perception, adaptation, and shared understanding. The anthropomorphic form of humanoid platforms offers a unique advantage: it allows for intuitive alignment between human demonstrations and robotic motion, making them ideal candidates for teaching-through-interaction paradigms.

Recent work by Xu *et al.* introduces a human-in-the-loop natural teaching approach that leverages visual–tactile communication to bridge perception and action^[94]. In this paradigm, a demonstrator uses wearable sensors and VR interfaces to guide a humanoid robot in real time, sharing both visual context and motion data. The robot, built on the InMoov platform, replicates multi-DOF movements based on this shared input. Here, visual perception conveys environmental understanding, while tactile mapping supports precise motion imitation - together forming a cross-modal channel that enables the robot to learn tasks in a way that resembles human observation and mimicry.

This approach demonstrates how visual–tactile systems are not just tools for feedback, but active components in knowledge transfer - encoding semantics, refining alignment, and reducing ambiguity in interaction. Rather than relying solely on large datasets or offline training, such frameworks open a path toward more natural, efficient robot learning grounded in shared perception.

Looking forward, visual–tactile communication will play an increasingly central role in enabling humanoid robots to operate autonomously in real-world, unstructured environments. Unlike rigid sensor–action loops, visual–tactile fusion supports bidirectional, adaptive interaction: robots can interpret complex visual scenes while continuously adjusting their physical behavior based on real-time tactile cues. This is particularly crucial in social settings - such as caregiving, service robotics, or collaborative manufacturing - where physical contact carries rich contextual meaning that must be correctly interpreted and safely managed.

Moreover, as neuromorphic tactile sensors, embedded vision systems, and edge AI technologies continue to evolve, future humanoids will be capable of maintaining multimodal situational awareness with minimal latency and power consumption. This will allow them to engage in fine motor tasks, interpret social touch, and even learn from subtle corrections during physical teaching. Ultimately, visual–tactile communication will serve not only as a sensing interface, but as a cognitive bridge - linking perception, memory, intention, and embodiment into a unified behavioral loop. In this sense, it forms a foundational layer for the emergence of genuinely adaptive, intuitive, and socially integrated humanoid agents.

6.6. Ethical and security considerations

As visual-tactile communication technologies advance toward real-world deployment, particularly in high-stakes domains such as healthcare, surveillance, and HRI, ethical, privacy, and security concerns become increasingly critical. In telemedicine and remote surgery, the transmission of high-fidelity tactile and visual data may involve sensitive physiological and behavioral information. Ensuring the confidentiality and integrity of such data necessitates robust encryption, secure edge-cloud collaboration, and privacy-preserving techniques such as federated learning. Failure to address these issues could lead to breaches of patient confidentiality and erode trust in automated healthcare systems.

Beyond data protection, ethical concerns also arise regarding user consent, autonomy, and transparency - especially in scenarios involving continuous monitoring or physical feedback. In surveillance or caregiving

applications, for instance, visual-tactile systems may capture intimate or involuntary interactions, raising questions about informed consent and potential misuse. Furthermore, long-term exposure to artificial haptic feedback - particularly in immersive environments - may introduce unforeseen physiological or psychological effects, such as sensory desensitization or cognitive overload.

Deployment risks also include reliability and safety under uncertain or degraded network conditions. In tactile-critical systems such as robotic surgery or industrial teleoperation, latency, jitter, or signal degradation can lead to unintended behavior or safety hazards. In addition, algorithmic biases embedded in perception or feedback models - such as inconsistent tactile interpretation across demographic groups or underrepresented tissue types - pose risks to fairness and inclusiveness.

To mitigate these challenges, future research should integrate ethical principles into system design from the outset, emphasizing human-centered transparency, safety assurance, and adaptive fail-safe mechanisms. The establishment of domain-specific regulatory frameworks and multidisciplinary oversight - including input from ethicists, clinicians, and policy-makers - is essential to ensure that visual-tactile communication technologies are deployed responsibly, equitably, and with societal trust.

7. CONCLUSIONS

This paper provides a comprehensive overview of current research in multimodal visuo-tactile communication, aiming to support future advancements in robotic applications and multimodal communication technologies. To achieve this, this paper first reviews the state-of-the-art in unimodal communication systems, including video and tactile communication. Subsequently, this paper analyzes the performance limitations of independent transmission strategies for multimodal data streams in existing implementations. The study further investigates critical components of visuo-tactile communication systems, encompassing sensor design, data processing architectures, transmission protocols, and signal reconstruction methodologies. Finally, this paper explores potential development trajectories for multimodal communication, emphasizing its transformative role in enabling high-fidelity HRI and intelligent environmental adaptation.

DECLARATIONS

Authors' contributions

Made substantial contributions to conception and design of the study and performed data analysis and interpretation: Wang, Z.; Chen, M.

Performed data acquisition and provided administrative, technical, and material support: Liu, Q.

Availability of data and materials

Not applicable.

Financial support and sponsorship

None.

Conflicts of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

© The Author(s) 2025.

REFERENCES

1. Lei, S. M.; Chen, T. C.; Sun, M. T. Video bridging based on H.261 standard. *IEEE. Trans. Circuits. Syst. Video. Technol.* **1994**, *4*, 425–37. [DOI](#)
2. Sikora, T. MPEG digital video-coding standards. *IEEE. Signal. Process. Mag.* **1997**, *14*, 82–100. [DOI](#)
3. Hartung, F.; Girod, B. Digital watermarking of MPEG-2 coded video in the bitstream domain. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany. Apr 21–24, 1997. pp. 2621–4. [DOI](#)
4. Marpe, D.; Wiegand, T.; Sullivan, G. The H.264/MPEG4 advanced video coding standard and its applications. *IEEE. Commun. Mag.* **2006**, *44*, 134–43. [DOI](#)
5. Sullivan, G. J.; Ohm, J.; Han, W.; Wiegand, T. Overview of the high efficiency video coding (HEVC) standard. *IEEE. Trans. Circuits. Syst. Video. Technol.* **2012**, *22*, 1649–68. [DOI](#)
6. Bross, B.; Wang, Y.; Ye, Y.; et al. Overview of the versatile video coding (VVC) standard and its applications. *IEEE. Trans. Circuits. Syst. Video. Technol.* **2021**, *31*, 3736–64. [DOI](#)
7. Muhammad, M.; Kasmis, F.; De Cola, T. Advanced transport satellite protocol. In *2012 IEEE Global Communications Conference (GLOBECOM)*, Anaheim, USA. Dec 03–07, 2012. IEEE; 2012. pp. 3299–304. [DOI](#)
8. Stockhammer, T. Dynamic adaptive streaming over HTTP -: standards and design principles. In *Proceedings of the Second Annual ACM Conference on Multimedia Systems. Association for Computing Machinery*; 2011. pp. 133–44. [DOI](#)
9. Wang, D.; Peng, Y.; Ma, X.; et al. Adaptive wireless video streaming based on edge computing: opportunities and approaches. *IEEE. Trans. Serv. Comput.* **2019**, *12*, 685–97. [DOI](#)
10. Souane, N.; Bourenane, M.; Douga, Y. Deep reinforcement learning-based approach for video streaming: dynamic adaptive video streaming over HTTP. *Appl. Sci.* **2023**, *13*, 11697. [DOI](#)
11. Abou-Zeid, H.; Hassanein, H. S.; Valentin, S. Energy-efficient adaptive video transmission: exploiting rate predictions in wireless networks. *IEEE. Trans. Veh. Technol.* **2014**, *63*, 2013–26. [DOI](#)
12. Zhou, L.; Wu, D.; Chen, J.; Dong, Z. Greening the smart cities: energy-efficient massive content delivery via D2D communications. *IEEE. Trans. Ind. Inf.* **2018**, *14*, 1626–34. [DOI](#)
13. Min, X.; Duan, H.; Sun, W.; Zhu, Y.; Zhai, G. Perceptual video quality assessment: a survey. *Sci. China. Inf. Sci.* **2024**, *67*, 4133. [DOI](#)
14. Zhao, T.; Liu, Q.; Chen, C. W. QoE in video transmission: a user experience-driven strategy. *IEEE. Commun. Surv. Tutorials.* **2017**, *19*, 285–302. [DOI](#)
15. Antonakoglou, K.; Xu, X.; Steinbach, E.; Mahmoodi, T.; Dohler, M. Toward haptic communications over the 5G Tactile Internet. *IEEE. Commun. Surv. Tutorials.* **2018**, *20*, 3034–59. [DOI](#)
16. Simpkins, A. Robotic tactile sensing: technologies and system (Dahiya, R.S. and Valle, M.; 2013) [On the Shelf]. *IEEE. Robot. Automat. Mag.* **2013**, *20*, 107. [DOI](#)
17. Chen, C. C.; Chang, P. Z.; Shih, W. P. Flexible tactile sensor with high sensitivity utilizing botanical epidermal cell natural micro-structures. In *SENSORS, 2012 IEEE*, Taipei, Taiwan. Oct 28–31, 2012. IEEE; 2012. p. 1–4. [DOI](#)
18. Rana, A.; Roberge, J.; Duchaine, V. An improved soft dielectric for a highly sensitive capacitive tactile sensor. *IEEE. Sensors. J.* **2016**, *16*, 7853–63. [DOI](#)
19. Pyo, S.; Lee, J.; Kim, M.; et al. Development of a flexible three-axis tactile sensor based on screen-printed carbon nanotube-polymer composite. *J. Micromech. Microeng.* **2014**, *24*, 075012. [DOI](#)
20. Liu, W.; Yu, P.; Gu, C.; Cheng, X.; Fu, X. Fingertip piezoelectric tactile sensor array for roughness encoding under varying scanning velocity. *IEEE. Sensors. J.* **2017**, *17*, 6867–79. [DOI](#)
21. Massaro, A.; Spano, F.; Cazzato, P.; La Tegola, C.; Cingolani, R.; Athanassiou, A. Robot tactile sensing: gold nanocomposites as highly sensitive real-time optical pressure sensors. *IEEE. Robot. Automat. Mag.* **2013**, *20*, 82–90. [DOI](#)
22. Fujiwara, E.; Paula, F. D.; Wu, Y. T.; Santos, M. F. M.; Schenkel, E. A.; Suzuki, C. K. Optical fiber tactile sensor based on fiber specklegram analysis. In *2017 25th Optical Fiber Sensors Conference (OFS)*, Jeju, Korea. Apr 24–28, 2017. IEEE; 2017. p. 1–4. [DOI](#)
23. Alfadhel, A.; Kosel, J. Magnetic nanocomposite cilia tactile sensor. *Adv. Mater.* **2015**, *27*, 7888–92. [DOI](#) [PubMed](#)
24. Yan, Y.; Hu, Z.; Yang, Z.; et al. Soft magnetic skin for super-resolution tactile sensing with force self-decoupling. *Sci. Robot.* **2021**, *6*, eabc8801. [DOI](#)
25. Holland, O.; Steinbach, E.; Prasad, R. V.; et al. The IEEE 1918.1 “Tactile Internet” Standards Working Group and its Standards. *Proc. IEEE.* **2019**, *107*, 256–79. [DOI](#)
26. Sakr, N.; Georganas, N. D.; Zhao, J. Human perception-based data reduction for haptic communication in Six-DoF telepresence systems. *IEEE. Trans. Instrum. Meas.* **2011**, *60*, 3534–46. [DOI](#)
27. Xu, Y.; Huang, Q.; Zheng, Q.; Fang, Y.; Zhao, T. Perception-based prediction for efficient kinesthetic coding. *IEEE. Signal. Process.*

- Lett.* **2024**, *31*, 2530-4. DOI
28. Hassen, R.; Gulecyuz, B.; Steinbach, E. PVC-SLP: perceptual vibrotactile-signal compression based-on sparse linear prediction. *IEEE. Trans. Multimedia.* **2021**, *23*, 4455-68. DOI
 29. Steinbach, E.; Strese, M.; Eid, M.; et al. Haptic codecs for the Tactile Internet. *Proc. IEEE.* **2019**, *107*, 447-70. DOI
 30. Huang, K.; Lee, D. Consensus-based peer-to-peer control architecture for multiuser haptic interaction over the internet. *IEEE. Trans. Robot.* **2013**, *29*, 417-31. DOI
 31. Schuwerk, C.; Chaudhari, N.; Steinbach, E. An area-of-interest based communication architecture for shared haptic virtual environments. In *2013 IEEE International Symposium on Haptic Audio Visual Environments and Games (HAVE)*, Istanbul, Turkey. Oct 26-27, 2013. IEEE; 2013. pp. 57-62. DOI
 32. Ateya, A. A.; Vybornova, A.; Kirichek, R.; Koucheryavy, A. Multilevel cloud based Tactile Internet system. In *2017 19th International Conference on Advanced Communication Technology (ICACT)*, PyeongChang, Korea. Feb 19-22, 2017. IEEE; 2017. pp. 105-10. DOI
 33. Hu, Z.; Zheng, Z.; Wang, T.; Song, L.; Li, X. Caching as a service: small-cell caching mechanism design for service providers. *IEEE. Trans. Wirel. Commun.* **2016**, *15*, 6992-7004. DOI
 34. Ansari, N.; Sun, X. Mobile edge computing empowers Internet of Things. *IEICE. Trans. Commun.* **2018**, *E101.B*, 604-19. DOI
 35. Kiani, A.; Ansari, N. Toward hierarchical mobile edge computing: an auction-based profit maximization approach. *IEEE. Internet. Things. J.* **2017**, *4*, 2082-91. DOI
 36. Hou, Z.; She, C.; Li, Y.; Niyato, D.; Dohler, M.; Vucetic, B. Intelligent communications for Tactile Internet in 6G: requirements, technologies, and challenges. *IEEE. Commun. Mag.* **2021**, *59*, 82-8. DOI
 37. Wei, X.; Duan, Q.; Zhou, L. A QoE-driven Tactile Internet architecture for smart city. *IEEE. Network.* **2020**, *34*, 130-6. DOI
 38. Kokkonis, G.; The Society of Digital Information and Wireless Communication. An open source architecture of a wireless body area network in a medical environment. *Int. J. Digit. Inf. Wirel. Commun.* **2016**, *6*, 63-77. DOI
 39. Gokhale, V.; Dabeer, O.; Chaudhuri, S. HoIP: haptics over Internet protocol. In *2013 IEEE International Symposium on Haptic Audio Visual Environments and Games (HAVE)*, Istanbul, Turkey. Oct 26-27, 2013. IEEE; 2013. pp. 45-50. DOI
 40. Gokhale, V.; Chaudhuri, S.; Dabeer, O. HoIP: a point-to-point haptic data communication protocol and its evaluation. In *2015 Twenty First National Conference on Communications (NCC)*, Mumbai, India. Feb 27 - Mar 01, 2015. IEEE; 2015. p. 1-6. DOI
 41. Kontogiannis, S.; Kokkonis, G. Proposed fuzzy real-time haptics protocol carrying haptic data and multisensory streams. *Int. J. Comput. Commun. Control.* **2020**, *15*, DOI
 42. Phung, H.; Hoang, P. T.; Jung, H.; Nguyen, T. D.; Nguyen, C. T.; Choi, H. R. Haptic display responsive to touch driven by soft actuator and soft sensor. *IEEE/ASME. Trans. Mechatron.* **2021**, *26*, 2495-505. DOI
 43. Uramune, R.; Ishizuka, H.; Hiraki, T.; Kawahara, Y.; Ikeda, S.; Oshiro, O. HaPouch: a miniaturized, soft, and wearable haptic display device using a liquid-to-gas phase change actuator. *IEEE. Access.* **2022**, *10*, 16830-42. DOI
 44. Zhu, L.; Jiang, X.; Shen, J.; Zhang, H.; Mo, Y.; Song, A. TapeTouch: a handheld shape-changing device for haptic display of soft objects. *IEEE. Trans. Vis. Comput. Graph.* **2022**, *28*, 3928-38. DOI PubMed
 45. Sakr, N.; Georganas, N. D.; Zhao, J. A perceptual quality metric for haptic signals. In *2007 IEEE International Workshop on Haptic, Audio and Visual Environments and Games*, Ottawa, Canada. Oct 12-14, 2007. IEEE; 2007. pp. 27-32. DOI
 46. Chaudhari, R.; Steinbach, E.; Hirche, S. Towards an objective quality evaluation framework for haptic data reduction. In *2011 IEEE World Haptics Conference*, Istanbul, Turkey. Jun 21-24, 2011. IEEE; 2011. pp. 539-44. DOI
 47. Hassen, R.; Steinbach, E. HSSIM: an objective haptic quality assessment measure for force-feedback signals. In *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, Cagliari, Italy. May 29 - Jun 01, 2018. IEEE; 2018. p. 1-6. DOI
 48. Liu, X.; Dohler, M.; Deng, Y. Vibrotactile quality assessment: hybrid metric design based on SNR and SSIM. *IEEE. Trans. Multimedia.* **2020**, *22*, 921-33. DOI
 49. She, C.; Yang, C.; Quek, T. Q. S. Radio resource management for ultra-reliable and low-latency communications. *IEEE. Commun. Mag.* **2017**, *55*, 72-8. DOI
 50. Nielsen, J. J.; Liu, R.; Popovski, P. Ultra-reliable low latency communication using interface diversity. *IEEE. Trans. Commun.* **2018**, *66*, 1322-34. DOI
 51. Kotaba, R.; Manchon, C. N.; Balercia, T.; Popovski, P. How URLLC can benefit from NOMA-based retransmissions. *IEEE. Trans. Wirel. Commun.* **2021**, *20*, 1684-99. DOI
 52. Tanveer, J.; Haider, A.; Ali, R.; Kim, A. An overview of reinforcement learning algorithms for handover management in 5G ultra-dense small cell networks. *Appl. Sci.* **2022**, *12*, 426. DOI
 53. Yuan, Y.; Yang, T.; Feng, H.; Hu, B. An iterative matching-stackelberg game model for channel-power allocation in D2D underlaid cellular networks. *IEEE. Trans. Wirel. Commun.* **2018**, *17*, 7456-71. DOI
 54. Zhang, S.; Liu, J.; Guo, H.; Qi, M.; Kato, N. Envisioning device-to-device communications in 6G. *IEEE. Network.* **2020**, *34*, 86-91. DOI
 55. Bennis, M.; Debbah, M.; Poor, H. V. Ultrareliable and low-latency wireless communication: tail, risk, and scale. *Proc. IEEE.* **2018**, *106*, 1834-53. DOI
 56. Yamaguchi, A.; Atkeson, C. G. Combining finger vision and optical tactile sensing: reducing and handling errors while cutting vegetables. In *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*, Cancun, Mexico. Nov 15-17, 2016.

- IEEE; 2016; pp. 1045–51. [DOI](#)
57. Yuan, W.; Dong, S.; Adelson, E. H. GelSight: high-resolution robot tactile sensors for estimating geometry and force. *Sensors* **2017**, *17*, 2762. [DOI](#) [PubMed](#) [PMC](#)
58. Donlon, E.; Dong, S.; Liu, M.; Li, J.; Adelson, E.; Rodriguez, A. GelSlim: a high-resolution, compact, robust, and calibrated tactile-sensing finger. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Madrid, Spain. Oct 01–05, 2018. IEEE; 2018. pp. 1927–34. [DOI](#)
59. Wang, S.; She, Y.; Romero, B.; Adelson, E. GelSight wedge: measuring high-resolution 3D contact geometry with a compact robot finger. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, Xi'an, China. May 30 – Jun 05, 2021. IEEE; 2021. pp. 6468–75. [DOI](#)
60. Gomes, D. F.; Lin, Z.; Luo, S. GelTip: a finger-shaped optical tactile sensor for robotic manipulation. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Las Vegas, USA. Oct 24 2020 – Jan 24 2021, 2021. IEEE; 2021. pp. 9903–9. [DOI](#)
61. Fan, W.; Li, H.; Si, W.; Luo, S.; Lepora, N.; Zhang, D. ViTacTip: design and verification of a novel biomimetic physical vision-tactile fusion sensor. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, Yokohama, Japan. May 13–17, 2024. IEEE; 2024. pp. 1056–62. [DOI](#)
62. Kuppuswamy, N.; Alspach, A.; Uttamchandani, A.; Creasey, S.; Ikeda, T.; Tedrake, R. Soft-bubble grippers for robust and perceptive manipulation. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Las Vegas, USA. Oct 24 2020 – Jan 24 2021. IEEE; 2021. pp. 9917–24. [DOI](#)
63. Zhang, L.; Wang, Y.; Jiang, Y. Tac3D: a novel vision-based tactile sensor for measuring forces distribution and estimating friction coefficient distribution. *arXiv* **2022**, arXiv:2202.06211. <https://arxiv.org/abs/2202.06211>. (accessed 26 Jun 2025)
64. Liu, H.; Yu, Y.; Sun, F.; Gu, J. Visual–tactile fusion for object recognition. *IEEE. Trans. Automat. Sci. Eng.* **2017**, *14*, 996–1008. [DOI](#)
65. Lee, J. T.; Bollegala, D.; Luo, S. “Touching to See” and “Seeing to Feel”: robotic cross-modal sensory data generation for visual-tactile perception. In *2019 International Conference on Robotics and Automation (ICRA)*, Montreal, Canada. May 20–24, 2019. IEEE; 2019. pp. 4276–82. [DOI](#)
66. Wei, F.; Zhao, J.; Shan, C.; Yuan, Z. Alignment and multi-scale fusion for visual-tactile object recognition. In *2022 International Joint Conference on Neural Networks (IJCNN)*, Padua, Italy. Jul 18–23, 2022. IEEE; 2022. p. 1–8. [DOI](#)
67. Babadian, R. P.; Faez, K.; Amiri, M.; Falotico, E. Fusion of tactile and visual information in deep learning models for object recognition. *Inf. Fusion.* **2023**, *92*, 313–25. [DOI](#)
68. Falco, P.; Lu, S.; Natale, C.; Pirozzi, S.; Lee, D. A transfer learning approach to cross-modal object recognition: from visual observation to robotic haptic exploration. *IEEE. Trans. Robot.* **2019**, *35*, 987–98. [DOI](#)
69. Zhou, L.; Wu, D.; Chen, J.; Wei, X. Cross-modal collaborative communications. *IEEE. Wirelless. Commun.* **2020**, *27*, 112–7. [DOI](#)
70. Yang, L.; Wu, D.; Zhou, L. Heterogeneous stream scheduling for cross-modal transmission. *IEEE. Trans. Commun.* **2021**, *69*, 6037–49. [DOI](#)
71. Wu, D.; Zhou, L. Cross-modal stream transmission: architecture, strategy, and technology. *IEEE. Wirelless. Commun.* **2024**, *31*, 134–40. [DOI](#)
72. Tong, Q.; Wei, W.; Liu, C.; Zhou, X.; Zhang, Y.; Wang, D. Cross-modal transmission with active packet loss and restoration for Tactile Internet. *IEEE. Commun. Mag.* **2024**, *62*, 70–6. [DOI](#)
73. Wei, X.; Liao, J.; Zhou, L.; Sari, H.; Zhuang, W. Toward generic cross-modal transmission strategy. *IEEE. Trans. Commun.* **2024**, *72*, 6059–72. [DOI](#)
74. Suo, Y.; Chen, Y.; Gao, Y.; Wei, X. Dynamic transmission mode selection for multi-modal services. *IEEE. Commun. Lett.* **2023**, *27*, 911–5. [DOI](#)
75. Li, L.; Shi, D.; Hou, R.; Chen, R.; Lin, B.; Pan, M. Energy-efficient proactive caching for adaptive video streaming via data-driven optimization. *IEEE. Int. Things. J.* **2020**, *7*, 5549–61. [DOI](#)
76. Li, C.; Toni, L.; Zou, J.; Xiong, H.; Frossard, P. QoE-driven mobile edge caching placement for adaptive video streaming. *IEEE. Trans. Multimedia.* **2018**, *20*, 965–84. [DOI](#)
77. Chen, M.; Hao, Y.; Hu, L.; Hossain, M. S.; Ghoneim, A. Edge-CoCaCo: toward joint optimization of computation, caching, and communication on edge cloud. *IEEE. Wirelless. Commun.* **2018**, *25*, 21–7. [DOI](#)
78. Gao, Y.; Wei, X.; Kang, B.; Chen, J. Edge intelligence empowered cross-modal streaming transmission. *IEEE. Network.* **2021**, *35*, 236–43. [DOI](#)
79. Yuan, Z.; Wei, X.; Zhou, L.; Zhuang, W. Content-aware cross-modal stream transmission. *IEEE. Wirelless. Commun. Lett.* **2024**, *13*, 2507–11. [DOI](#)
80. Gao, Y.; Wang, T.; Zhou, L.; Zhuang, W. CRoss-MODAL communications for holographic video streaming. *IEEE. Wirelless. Commun.* **2025**, *32*, 96–102. [DOI](#)
81. Wei, X.; Shi, Y.; Zhou, L. Haptic signal reconstruction for cross-modal communications. *IEEE. Trans. Multimedia.* **2022**, *24*, 4514–25. [DOI](#)
82. Chen, M.; Xie, Y. Cross-modal reconstruction for tactile signal in human-robot interaction. *Sensors* **2022**, *22*, 6517. [DOI](#) [PubMed](#) [PMC](#)
83. Yang, Z.; Wang, H.; Shi, Y.; Ye, L.; Wei, X. Fine-grained audio-visual aided haptic signal reconstruction. *IEEE. Signal. Process. Lett.* **2024**, *31*, 1349–53. [DOI](#)

84. Chen, Y.; Li, A.; Wu, D.; Zhou, L. Toward general cross-modal signal reconstruction for robotic teleoperation. *IEEE. Trans. Multimedia.* **2024**, *26*, 3541-53. [DOI](#)
85. Wei, X.; Yao, Y.; Wang, H.; Zhou, L. Perception-aware cross-modal signal reconstruction: from audio-haptic to visual. *IEEE. Trans. Multimedia.* **2023**, *25*, 5527-38. [DOI](#)
86. Wei, X.; Zhou, L. AI-enabled cross-modal communications. *IEEE. Wireless. Commun.* **2021**, *28*, 182-9. [DOI](#)
87. Farooq, A.; Rantala, J.; Raisamo, R.; Hippula, A. Haptic mediation through artificial intelligence: magnetorheological fluid as vibrotactile signal mediator. In *2022 Symposium on Design, Test, Integration and Packaging of MEMS/MOEMS (DTIP)*, Pont-a-Mousson, France. Jul 11-13, 2022. IEEE; 2022. p. 1-4. [DOI](#)
88. Cheng, L.; Zhang, H.; Di, B.; Niyato, D.; Song, L. Large language models empower multimodal integrated sensing and communication. *IEEE. Commun. Mag.* **2025**, *63*, 190-7. [DOI](#)
89. Lipkova, J.; Chen, R. J.; Chen, B.; et al. Artificial intelligence for multimodal data integration in oncology. *Cancer. Cell.* **2022**, *40*, 1095-110. [DOI](#) [PubMed](#) [PMC](#)
90. Shao, J.; Ma, J.; Zhang, Q.; Li, W.; Wang, C. Predicting gene mutation status via artificial intelligence technologies based on multimodal integration (MMI) to advance precision oncology. *Semin. Cancer. Biol.* **2023**, *91*, 1-15. [DOI](#)
91. Wang, T.; Zhang, B.; Jiang, D.; Li, D. A multimodal large language model framework for intelligent perception and decision-making in smart manufacturing. *Sensors* **2025**, *25*, 3072. [DOI](#) [PubMed](#) [PMC](#)
92. Sanfilippo, F.; Blažauskas, T.; Girdžiūna, M.; Janonis, A.; Kiudys, E.; Salvietti, G. A multi-modal auditory-visual-tactile e-learning framework. In: Sanfilippo F, Granmo O, Yayilgan SY, Bajwa IS, editors. *Intelligent technologies and applications*. Cham: Springer International Publishing; 2022. pp. 119-31. [DOI](#)
93. Xu, C.; Zhou, Y.; He, B.; et al. An active strategy for safe human-robot interaction based on visual-tactile perception. *IEEE. Syst. J.* **2023**, *17*, 5555-66. [DOI](#)
94. Xu, W.; Li, X.; Gong, L.; et al. Natural teaching for humanoid robot via human-in-the-loop scene-motion cross-modal perception. *IR.* **2019**, *46*, 404-14. [DOI](#)