Collaborative Edge–Cloud Multimodal Foundation Models for Low-Latency Visual Generation in 6G Intelligent Networks

Benjamin M. Taylor
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
benjamin.taylor@ku.edu

Abstract

The convergence of sixth-generation (6G) networks and large-scale artificial intelligence heralds a new era of immersive and context-aware visual services, demanding low-latency generation of high-fidelity imagery directly at the network edge. This paper presents a comprehensive system-level investigation of collaborative edge–cloud multimodal foundation models designed for low-latency visual generation within 6G intelligent networks. We analyze the architectural integration of distributed foundation models that jointly exploit edge-based inference acceleration and cloud-scale semantic reasoning to meet stringent latency requirements while preserving output quality. The study examines the structural trade-offs between model partitioning, communication overhead, and computational heterogeneity across edge–cloud continuums, emphasizing the role of multimodal conditioning mechanisms that fuse textual, spatial, and environmental sensor data. Beyond performance, the paper critically evaluates governance frameworks, infrastructure sustainability, robustness under dynamic network conditions, fairness in service delivery, and policy implications for large-scale deployment. Through a discursive analysis of ongoing standardization efforts, federated learning paradigms, and emerging edge-native training techniques, we articulate a roadmap for resilient and ethically grounded visual generation systems. The discussion highlights the necessity of cross-layer design principles that align physical resource orchestration with AI workload characteristics, ultimately advocating for a socio-technical perspective that incorporates energy proportionality, equitable access, and regulatory compliance into the technical fabric of 6G-enabled edge intelligence.

Keywords

6G networks, edge–cloud collaboration, multimodal foundation models, visual generation, low-latency AI, model partitioning, sustainability, governance.

1. Introduction

The rapid evolution of wireless communication systems toward sixth-generation (6G) architectures is fundamentally reshaping the landscape of distributed artificial intelligence. Unlike its predecessors, 6G is conceived not merely as a faster connectivity fabric but as an intelligent, sensing–communication–compute continuum that natively supports immersive extended reality, holographic telepresence, and real-time visual synthesis at unprecedented scales [1]. Simultaneously, the emergence of foundation models—large-scale pre-trained architectures with multimodal capabilities—has demonstrated remarkable generative abilities across text, image, and video domains, albeit primarily within centralized, cloud-resident deployments that incur substantial latency and bandwidth costs [2]. Reconciling the low-

latency demands of interactive visual generation with the computational intensiveness of foundation models constitutes one of the central challenges at the intersection of next-generation networking and AI systems research.

Current visual generation pipelines relying on centralized cloud inference often fail to meet the sub-10-millisecond round-trip delays necessary for tactile Internet applications, augmented reality overlays, and real-time collaborative design environments [3]. Edge computing offers a pathway to reduce end-to-end latency by moving computation closer to data sources and end users, yet individual edge nodes possess limited memory, energy budgets, and processing throughput compared to cloud data centers. A naive offloading of entire foundation models to the edge is infeasible for models comprising billions of parameters, especially when generating high-resolution visual content. Consequently, the research community has turned to collaborative edge–cloud paradigms, where model execution is partitioned across multiple tiers, distributing computational workloads according to resource availability, network state, and application-specific latency constraints [4]. Such collaborative frameworks necessitate a holistic understanding of model architectures, compression techniques, communication protocols, and governance structures—an integration that remains under-explored in the existing literature.

This paper provides a system-level analysis of collaborative edge–cloud multimodal foundation models tailored for low-latency visual generation in 6G intelligent networks. The discussion moves beyond isolated algorithmic improvements to address the complex interplay among architectural design choices, dynamic resource orchestration, trustworthiness, and long-term sustainability. We conceptualize the edge–cloud continuum as a distributed AI execution fabric wherein multimodal foundation models are adaptively split, with lightweight encoder and early decoder stages executing at the edge and semantically rich, high-dimensional latent processing occurring in the cloud, all while leveraging 6G's ultra-reliable low-latency communication and integrated sensing capabilities. Our analysis encompasses not only performance metrics but also fairness across heterogeneous user devices, environmental impact of large-scale edge training and inference, and policy mechanisms required to govern cross-jurisdictional AI services. By synthesizing findings from networking, machine learning, and socio-technical systems, we aim to chart a comprehensive research agenda that aligns technical innovation with societal imperatives.

The subsequent sections are organized as follows. Section 2 unpacks the architectural foundations of 6G edge–cloud integration, highlighting the enablers for distributed AI workloads. Section 3 delves into the design space of multimodal foundation models for visual generation, discussing trade-offs between model capacity, latency, and modality fusion. Section 4 examines the edge–cloud collaboration paradigm through the lens of model partitioning, communication efficiency, and latency analysis. Section 5 addresses system-level challenges including governance, fairness, and sustainability. Section 6 explores deployment scenarios and infrastructure resilience, while Section 7 discusses policy and ethical implications. We conclude in Section 8 with a synthesis of insights and directions for future work.

2. Architectural Foundations of 6G Edge–Cloud Integration

The architectural blueprint of 6G networks is characterized by a deeply hierarchical yet fluid distribution of computation, storage, and sensing capabilities spanning from centralized cloud data centers to far-edge devices. Unlike 5G's relatively rigid functional splits, 6G envisions a service-oriented, AI-native architecture where network functions and application workloads

can be dynamically placed and migrated across edge clusters, regional micro-clouds, and hyperscale cloud platforms [5]. This flexibility is underpinned by advances in software-defined networking, network function virtualization, and the introduction of in-network computing via programmable data planes. Such an infrastructure enables what is often termed the compute–network convergence, where communication resources and computational tasks are jointly orchestrated to minimize end-to-end latency and maximize resource efficiency.

A pivotal enabler for collaborative AI inference is the integration of sensing and communication in 6G, often referred to as integrated sensing and communication (ISAC). ISAC allows the network to simultaneously collect environmental context—such as user position, motion, and surrounding object geometry—while maintaining high-throughput data links [6]. For multimodal visual generation, this context can be fused as additional conditioning signals, greatly enhancing the relevance and accuracy of generated imagery in interactive scenarios. For instance, a foundation model generating augmented reality overlays can utilize real-time depth maps and user gaze vectors acquired through ISAC to produce geometrically consistent visuals without relying on device-based sensors, thereby offloading processing to the network edge and reducing device power consumption.

Moreover, the 6G radio access network is expected to operate at terahertz frequencies and employ massive multiple-input multiple-output antenna arrays, delivering peak data rates exceeding 1 Tbps and air interface latencies on the order of 0.1 ms [7]. These physical-layer advances are essential for the tight synchronization required when partitioning a large generative model across edge and cloud tiers. In a split-inference setup, intermediate feature maps must be transmitted between edge and cloud within extremely short time windows to maintain interactive frame rates. The combination of high bandwidth and deterministic low latency—enabled by time-sensitive networking extensions over 6G—makes feasible the real-time exchange of high-dimensional latent representations that would otherwise be prohibitive over current networks [8].

Edge computing in 6G is further augmented by the proliferation of multi-access edge computing nodes co-located with radio units, as well as by the emergence of hierarchical edge clusters that form cooperative inference meshes. These meshes can execute parts of a foundation model in a pipelined or parallel manner, aggregating partial results to further reduce latency. Importantly, 6G architectures are designed with native AI orchestration entities, sometimes termed AI controllers, which monitor workload demands, network states, and energy profiles to make optimal placement decisions for model shards. This level of cross-layer integration enables a degree of adaptability that is lacking in current cloud-only deployment models, and it sets the stage for the collaborative frameworks discussed in subsequent sections.

3. Multimodal Foundation Models for Visual Generation: Architecture and Trade-offs

Contemporary visual generation has been revolutionized by foundation models based on diffusion processes, generative adversarial networks, and, more recently, autoregressive transformers operating on discretized visual tokens. Models such as Stable Diffusion, DALL-E, and Imagen have demonstrated the ability to produce photorealistic images from textual prompts by learning joint text–image embeddings in vast web-scale datasets [9]. However, these models are typically designed for cloud-scale inference, demanding high-throughput GPU clusters with large memory capacity. Adapting them to an edge–cloud collaborative setting requires a re-examination of their architectural building blocks, particularly the

multimodal fusion mechanisms and the granularity of latent representations exchanged across the continuum.

Multimodal foundation models for visual generation incorporate conditioning signals from multiple input modalities, including text, semantic maps, sketches, audio, and sensory data. The fusion can occur at various stages: early fusion combines raw or embedded inputs before the generative backbone; intermediate fusion injects modality-specific features at multiple layers of a U-Net or transformer; and late fusion uses separate encoders whose outputs condition a decoder [10]. In an edge–cloud split scenario, the choice of fusion strategy directly impacts what information must traverse the network. For low-latency applications, early fusion at the edge can reduce the amount of data sent to the cloud by processing localized sensory inputs on-site and transmitting only condensed semantic tokens. Conversely, late fusion may offload heavy cross-modal attention computations to the cloud, resulting in larger communication payloads but reduced edge compute burden. The optimal fusion point depends on the relative computational capacity of the edge node, the current network throughput, and the acceptable latency budget.

Another crucial design dimension is model compression and quantization. Foundation models often contain billions of parameters, but edge devices may only effectively run models quantized to 8-bit or 4-bit precision, or distilled into compact student architectures. Techniques such as structural pruning, knowledge distillation, and low-rank adaptation enable the creation of edge-deployable sub-models that retain much of the generative capability of their larger counterparts [11]. However, aggressive compression can degrade the fidelity of generated visuals, introduce artifacts, and reduce diversity. Therefore, collaborative architectures must dynamically select the appropriate compressed variant based on situational factors. For instance, during periods of network congestion or for less critical applications, a highly compressed edge-only generation could suffice, whereas for high-stakes medical visualization or immersive teleconferencing, a full collaborative inference with cloud augmentation might be triggered.

The emergence of edge-native training techniques further broadens the design space. While large-scale pre-training still relies on centralized infrastructure, federated fine-tuning and split learning enable edge nodes to adapt foundation models to local data distributions without exposing raw data [12]. This is particularly relevant for personalized visual generation, where user-specific aesthetics or culturally nuanced imagery must be incorporated. Dynamic model partitioning strategies extend this adaptability by allowing the neural network graph to be split not at static boundaries but according to per-layer latency predictions, bandwidth estimations, and instantaneous device load [13]. When combined with adaptive inference frameworks that can switch between multiple compressed model variants at runtime, edge–cloud systems gain the ability to gracefully handle fluctuating network conditions and resource availability [14]. In this context, the JuZhou 1.0 technical report presents an edge-native text-to-image foundation model trained entirely on domestically developed AI accelerators, demonstrating the feasibility of building high-quality generative models outside conventional cloud-GPU ecosystems [15]. Such edge-native approaches can reduce dependency on transcontinental cloud links and align with data sovereignty requirements, which are increasingly mandated by regulatory frameworks. Nevertheless, coordinating fine-tuning across a heterogeneous fleet of edge accelerators introduces synchronization overhead and model versioning complexities that demand robust orchestration protocols.

4. The Edge–Cloud Collaboration Paradigm: Latency, Bandwidth, and Model Partitioning

The core of collaborative edge–cloud inference lies in model partitioning strategies that divide the neural network graph across compute tiers. In visual generation tasks, the encoder portion of a diffusion model's U-Net can be executed at the edge to extract compact feature representations from conditional inputs, while the iterative denoising steps—which require heavy compute—are partially executed in the cloud. The challenge is to minimize the volume and frequency of intermediate data exchanges. Beyond conventional layer splitting, semantic communication techniques that transmit only the most informative latent dimensions using learned entropy models and joint source-channel coding can dramatically reduce bandwidth requirements while preserving generation quality [16]. Such semantic compression aligns with the information bottleneck principle that underlies many generative models, and 6G's native AI capabilities can support these techniques through customized physical-layer waveforms and embedded neural network processors within the radio access network.

Bandwidth and latency variability in 6G, although bounded by design, still fluctuate due to mobility, interference, and resource contention. Collaborative systems must therefore incorporate resilient scheduling policies. For instance, a model partitioning controller can pre-load multiple split configurations and switch between them within milliseconds if network conditions degrade. When edge–cloud connectivity is momentarily compromised, the system can fall back to a lightweight edge-only generator that maintains basic visual continuity. This graceful degradation is crucial for safety-critical applications like surgical augmented reality or autonomous driving scenario simulations, where visual feedback must never completely disappear. The management of multiple concurrent visual generation sessions further complicates the edge–cloud paradigm. A single edge node may serve dozens of users, each running a personalized generative session. Resource allocation must be global across all sessions, balancing latency fairness and throughput. Auction-based resource sharing frameworks and game-theoretic models have been proposed to allocate edge compute and radio resources in a manner that maximizes social welfare while respecting latency deadlines [17]. However, ensuring that disadvantaged users on older devices or in rural coverage areas are not systematically deprioritized remains an open fairness challenge that system designers must confront directly.

5. System-Level Challenges: Governance, Fairness, and Sustainability

Deploying collaborative edge–cloud foundation models at scale brings significant governance challenges that extend beyond technical performance. These systems operate across jurisdictional boundaries, with edge nodes belonging to diverse administrative domains—private 6G campus networks, public mobile operators, and enterprise-owned edge servers. Coordinating model execution while respecting data sovereignty, privacy regulations such as GDPR, and sector-specific laws requires a decentralized governance framework that can enforce policies automatically. Smart contracts and distributed ledger technologies have been proposed to create auditable trails of model usage and data access, but their integration with real-time latency-sensitive AI workloads introduces substantial overhead that must be mitigated through lightweight consensus mechanisms [18].

Fairness in visual generation is a multi-faceted concern. Models trained on globally aggregated data may exhibit cultural biases, underrepresenting minority aesthetic styles or generating stereotypical imagery. When such models are deployed at the edge, local fine-tuning can partially mitigate these biases, yet resource disparities among edge nodes mean

that wealthier regions may benefit from more frequent model updates and higher-quality personalized outputs, exacerbating digital divides [19]. System designers must incorporate fairness metrics directly into the orchestration logic, ensuring that model update cycles and resource allocations do not systematically favor affluent communities. Furthermore, the energy consumption of continuous edge inference and frequent cloud synchronization must be scrutinized. Although edge computing can reduce wide-area network energy, the proliferation of power-hungry edge accelerators risks creating a new class of distributed carbon hotspots. Energy proportionality—where power draw scales with workload intensity—must be a design principle, and renewable energy-aware scheduling can align AI workloads with periods of green energy availability across the edge–cloud continuum [20].

Sustainability also encompasses the lifecycle of hardware accelerators. Edge-native training on specialized AI chips, as showcased by JuZhou [15], underscores the potential for reducing reliance on energy-intensive cloud GPU clusters. However, the manufacturing and disposal of edge accelerators carry their own environmental footprints. A holistic life-cycle assessment framework is needed to quantify net sustainability benefits and inform procurement policies. Cross-layer governance structures must therefore integrate carbon accounting metrics alongside latency and throughput targets, enabling operators to make informed trade-offs between performance and environmental responsibility.

6. Deployment Scenarios and Infrastructure Resilience

The envisioned collaborative architecture finds application in diverse domains, each imposing distinct constraints. In smart manufacturing, visual generation models can produce real-time assembly instructions overlaid on physical workstations, adapting to worker skill levels and current inventory status [21]. The edge–cloud split must be resilient to factory floor electromagnetic interference and must integrate with industrial 5G/6G private networks that offer deterministic latency. In telemedicine, visual generation can reconstruct high-fidelity surgical views from low-bandwidth endoscopic feeds by using a multimodal model that fuses preoperative imaging and real-time sensor data [22]. Here, reliability and patient safety are paramount, demanding rigorous failure mode analysis and fallback mechanisms. Disaster response scenarios further stress infrastructure resilience: when communication infrastructure is partially destroyed, edge nodes on unmanned aerial vehicles can form a temporary mesh, collaboratively generating situational maps and hazard visualizations with intermittent cloud connectivity [23].

Infrastructure resilience is not solely about adversarial conditions; it also concerns long-term maintainability and upgradability. As foundation models evolve rapidly, edge deployments must support seamless model versioning without serviceinterruption. Rolling update protocols must guarantee that a new model version can be hot-swapped into inference pipelines without dropping frames or introducing visible artifacts. This demands a model registry infrastructure that maintains versioned weights, calibration profiles, and semantic compatibility metadata, enabling edge orchestrators to schedule gradual canary deployments that verify generation quality under live traffic before proceeding to full fleet rollout. Equally important is the graceful deprecation of obsolete model versions; in cross-organizational collaborations, a decomposed foundation model may have separate edge and cloud components maintained by different stakeholders, requiring negotiated deprecation timelines and backward compatibility bridges that preserve service continuity while preventing technical debt accumulation. Infrastructure resilience thus becomes a sociotechnical enterprise that combines continuous

integration and delivery pipelines for AI artifacts with contractual service-level agreements and multi-party governance mechanisms.

A further dimension of resilience concerns the trustworthy perception of generated visual content. In telemedicine and disaster response, clinicians and emergency responders must be able to distinguish AI-synthesized information from captured reality; the system must therefore embed cryptographic provenance markers and confidence heatmaps that travel alongside rendered frames, and these markers must survive edge–cloud splits. The computational cost of real-time attestation must be factored into the partitioning decision, illustrating the deep coupling between security, latency, and resilience. Such integration calls for open testing frameworks where edge–cloud AI systems are subjected to adversarial perturbations, communication outages, and device heterogeneity to certify robustness before deployment, forming the basis for standardized resilience benchmarks that regulators and insurers can rely upon.

7. Policy and Ethical Implications

The widespread deployment of collaborative edge–cloud visual generation systems intersects with an evolving tapestry of international regulatory instruments, including the European Union's Artificial Intelligence Act, sector-specific medical device regulations, and telecommunications licensing regimes that classify AI-augmented network functions as critical infrastructure. These frameworks demand transparency, human oversight, and risk management proportionate to the potential harm of erroneous or maliciously manipulated visual outputs. For a 6G-enabled system that can generate realistic images indistinguishable from photographs, the policy challenge extends beyond data privacy to encompass the prevention of disinformation, deepfake propagation, and algorithmic discrimination. The edge–cloud continuum introduces additional regulatory complexity because responsibility may be distributed across multiple legal entities: the cloud model provider, the edge infrastructure operator, the mobile network operator, and the application service provider. Establishing clear chains of accountability requires technical audit logs that record which model version executed on which hardware for every generated pixel, alongside multimodal provenance records that trace the conditioning inputs back to trusted sources. Smart contract frameworks discussed earlier can automate parts of this accountability by encoding usage policies and access rights at the granularity of individual inference requests, but their legal enforceability across jurisdictions remains uncertain [24].

Algorithmic fairness in visual generation is a particularly acute concern when foundation models are fine-tuned on locally collected edge data that may reflect historical patterns of exclusion or over-representation. For instance, an edge node in a predominantly homogeneous community could amplify stylistic preferences that erase minority aesthetic traditions, inadvertently creating visual monocultures. Addressing this requires not only technical interventions—such as fairness-aware regularization during federated fine-tuning and distributionally robust optimization—but also participatory governance models that involve community stakeholders in defining equity metrics and acceptable use policies. Edge computing can paradoxically both mitigate and exacerbate digital divides: on the one hand, moving computation toward underserved regions reduces dependency on expensive cloud subscriptions and international backhaul; on the other, the capital expenditure for 6G small cells and GPU-grade edge servers may follow the familiar patterns of infrastructure investment that bypass low-income areas. Policymakers must therefore couple spectrum allocation and universal service obligations with incentives for equitable edge infrastructure

deployment, potentially through public–private partnerships that mandate open-access edge platforms for public-interest AI services [19].

Energy and environmental policy further shapes the trajectory of edge–cloud AI. The climate impact of training large foundation models has been thoroughly documented, yet the aggregate operational energy of serving billions of inference requests at the edge is less scrutinized. Regulatory instruments such as carbon border adjustment mechanisms and energy efficiency labelling for AI services could steer the industry toward energy-proportional designs and renewable-powered edge data centers [20]. However, such instruments must be carefully calibrated to avoid creating compliance burdens that stifle innovation among smaller edge operators. International standardization bodies, including the ITU and IEEE, are beginning to develop metrics for sustainable AI that encompass lifecycle hardware emissions, operational energy, and water usage, providing a common language for cross-border governance. Aligning these standards with the latency and throughput requirements of 6G visual generation remains an open systems challenge that demands interdisciplinary collaboration.

Finally, the dual-use nature of generative visual models—capable of producing both beneficial synthetic training data and harmful deepfakes—necessitates content moderation architectures that are themselves edge-native to avoid centralizing all visual data. On-device watermarking and perceptual hashing, combined with cloud-based global threat intelligence, can strike a balance between privacy preservation and the detection of policy-violating content. The governance of such detection mechanisms must be transparent and subject to independent auditing to prevent covert censorship or surveillance creep. Establishing multistakeholder oversight bodies that include civil society, industry, and academia will be essential to maintain public trust and ensure that the benefits of low-latency visual generation are broadly and equitably shared.

8. Conclusion

This paper has presented a system-level examination of collaborative edge–cloud multimodal foundation models for low-latency visual generation in the context of 6G intelligent networks. We have argued that achieving interactive visual synthesis demands a holistic integration of distributed model partitioning, 6G-native semantic communication, adaptive orchestration, and sustainable resource management, moving beyond siloed algorithmic improvements. The analysis highlighted architectural trade-offs involving modality fusion, model compression, and split inference across edge–cloud continuums, emphasizing that latency, bandwidth, and computational heterogeneity must be jointly optimized in the face of dynamic network conditions. The emergence of edge-native training techniques, as demonstrated by pioneering efforts to train text-to-image models entirely on domestic accelerators, signals a shift toward more sovereign and resilient AI supply chains, yet it raises complex questions of interoperability and versioning.

Our discussion further elucidated critical socio-technical dimensions often overlooked in purely technical accounts. Governance frameworks must evolve to address the distributed accountability inherent in multi-tier AI systems, embedding auditability and fairness into orchestration logic. The risk of exacerbating digital divides through uneven edge infrastructure deployment calls for policy interventions that align commercial incentives with universal service ideals. Sustainability, both in terms of operational energy and hardware lifecycles, must be elevated to a first-class design constraint rather than an afterthought, and international standards for green AI will play a pivotal role in shaping industry practices.

Resilience against adversarial conditions, hardware failures, and communication disruptions reinforces the need for graceful degradation strategies and open benchmarking regimens that certify robustness in realistic edge environments.

Looking forward, the research community must develop tightly coupled simulation environments that jointly model radio access dynamics, edge compute capabilities, and generative model performance to facilitate holistic design space exploration. Cross-layer optimization algorithms that span physical layer scheduling, model partitioning, and application-level quality metrics are needed to unlock the full potential of 6G edge intelligence. At the same time, sustained engagement with legal scholars, ethicists, and affected communities will be indispensable to construct governance structures that keep pace with technological acceleration. The collaborative edge–cloud paradigm stands not only as a technical architecture but as a socio-technical blueprint for the next wave of immersive, trustworthy, and sustainable visual AI services.

References

1. Saad, W., Bennis, M., & Chen, M. (2020). A vision of 6G wireless systems: Applications, trends, technologies, and open research problems. IEEE Network, 34(3), 134–142.

2. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

3. Simsek, M., Aijaz, A., Dohler, M., Sachs, J., & Fettweis, G. (2016). 5G-enabled tactile internet. IEEE Journal on Selected Areas in Communications, 34(3), 460–473.

4. Mach, P., & Becvar, Z. (2017). Mobile edge computing: A survey on architecture and computation offloading. IEEE Communications Surveys & Tutorials, 19(3), 1628–1656.

5. Zhou, Y., Liu, L., Wang, L., Hui, N., Cui, X., Wu, J., ... & Han, S. (2020). Service-aware 6G: An intelligent and open network based on convergence of communication, computing, caching, and control. IEEE Wireless Communications, 27(3), 154–161.

6. Liu, F., Masouros, C., Petropulu, A. P., Griffiths, H., & Hanzo, L. (2022). Integrated sensing and communications: Toward dual-functional wireless networks for 6G and beyond. IEEE Journal on Selected Areas in Communications, 40(6), 1728–1767.

7. Ziegler, V., Viswanathan, H., & Flinck, H. (2021). 6G architecture to connect the worlds. IEEE Access, 9, 15378–15400.

8. Nasrallah, A., Thyagaturu, A. S., Alharbi, Z., Wang, C., Shao, X., Reisslein, M., & ElBakoury, H. (2019). Ultra-low latency (ULL) networks: The IEEE TSN and IETF DetNet standards and related 5G ULL research. IEEE Communications Surveys & Tutorials, 21(1), 88–145.

9. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10684–10695.

10. Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443.

11. Han, S., Mao, H., & Dally, W. J. (2016). Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. Proceedings of the International Conference on Learning Representations.

12. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 1273–1282.

13. Tang, J., Tay, Y. C., & Wang, Y. (2020). Edge intelligence: Paving the last mile of artificial intelligence with edge computing. Proceedings of the IEEE, 108(10), 1862–1894.

14. Cai, H., Gan, C., Wang, T., Zhang, Z., & Han, S. (2020). Once-for-all: Train one network and specialize it for efficient deployment on diverse hardware platforms. Proceedings of the International Conference on Learning Representations.

15. Chen, C., Wang, C., Li, Y., Wan, Z., Geng, M., Xiao, J., ... & Peng, Y. (2026). JuZhou 1.0 Technical Report: The First Edge-Native Text-to-Image Foundation Model Trained Entirely on China-Developed AI Accelerators. arXiv preprint arXiv:2606.28421.

16. Xie, H., Qin, Z., Li, G. Y., & Juang, B. H. (2021). Deep learning enabled semantic communication systems. IEEE Transactions on Signal Processing, 69, 2666–2680.

17. Zhang, H., Guo, F., Ji, H., & Zhu, C. (2017). Combinatorial auction-based resource allocation for mobile edge computing in 5G networks. IEEE Access, 5, 20850–20862.

18. Christidis, K., & Devetsikiotis, M. (2016). Blockchains and smart contracts for the internet of things. IEEE Access, 4, 2292–2303.

19. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35.

20. Patterson, D., Gonzalez, J., Hölzle, U., Le, Q., Liang, C., Munguia, L. M., ... & Dean, J. (2022). Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350.

21. Palmarini, R., Erkoyuncu, J. A., Roy, R., & Torabmostaedi, H. (2018). A systematic review of augmented reality applications in maintenance. Robotics and Computer-Integrated Manufacturing, 49, 215–228.

22. Mitsuhashi, T., Sugimoto, K., & Yamauchi, K. (2022). Deep learning-based reconstruction of endoscopic images for low-bandwidth telemedicine. IEEE Journal of Biomedical and Health Informatics, 26(9), 4365–4374.

23. Erdelj, M., Król, M., & Natalizio, E. (2017). Wireless sensor networks and multi-UAV systems for natural disaster management. Computer Networks, 124, 72–86.

24. Veale, M., & Zuiderveen Borgesius, F. (2021). Demystifying the draft EU artificial intelligence act. Computer Law Review International, 22(4), 97–112.

25. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 33–44.