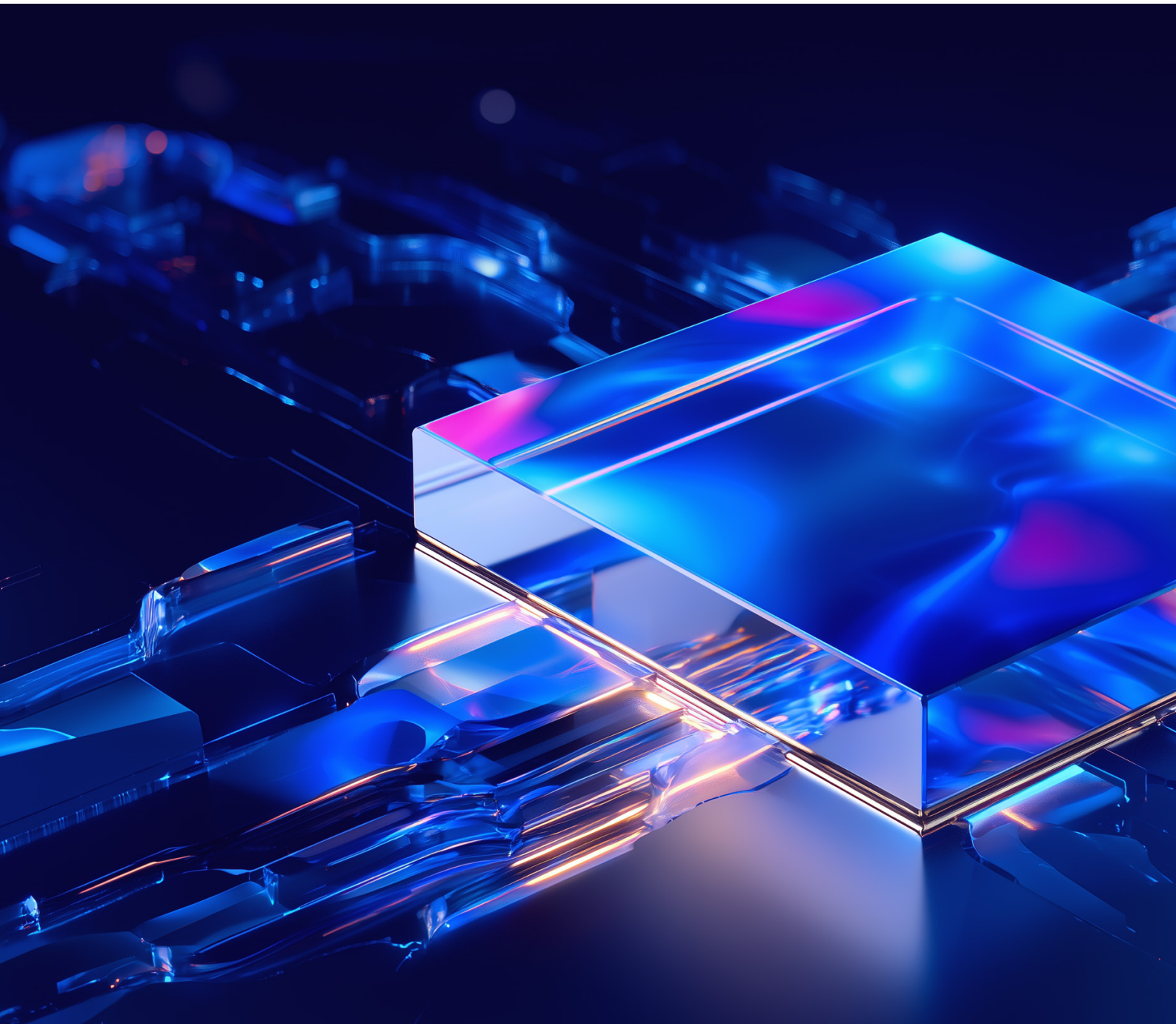
softserve

EDGE AI FOR REAL-TIME INTELLIGENCE AT SCALE

SoftServe Edge AI Kit powers instant, personalized experiences at the edge



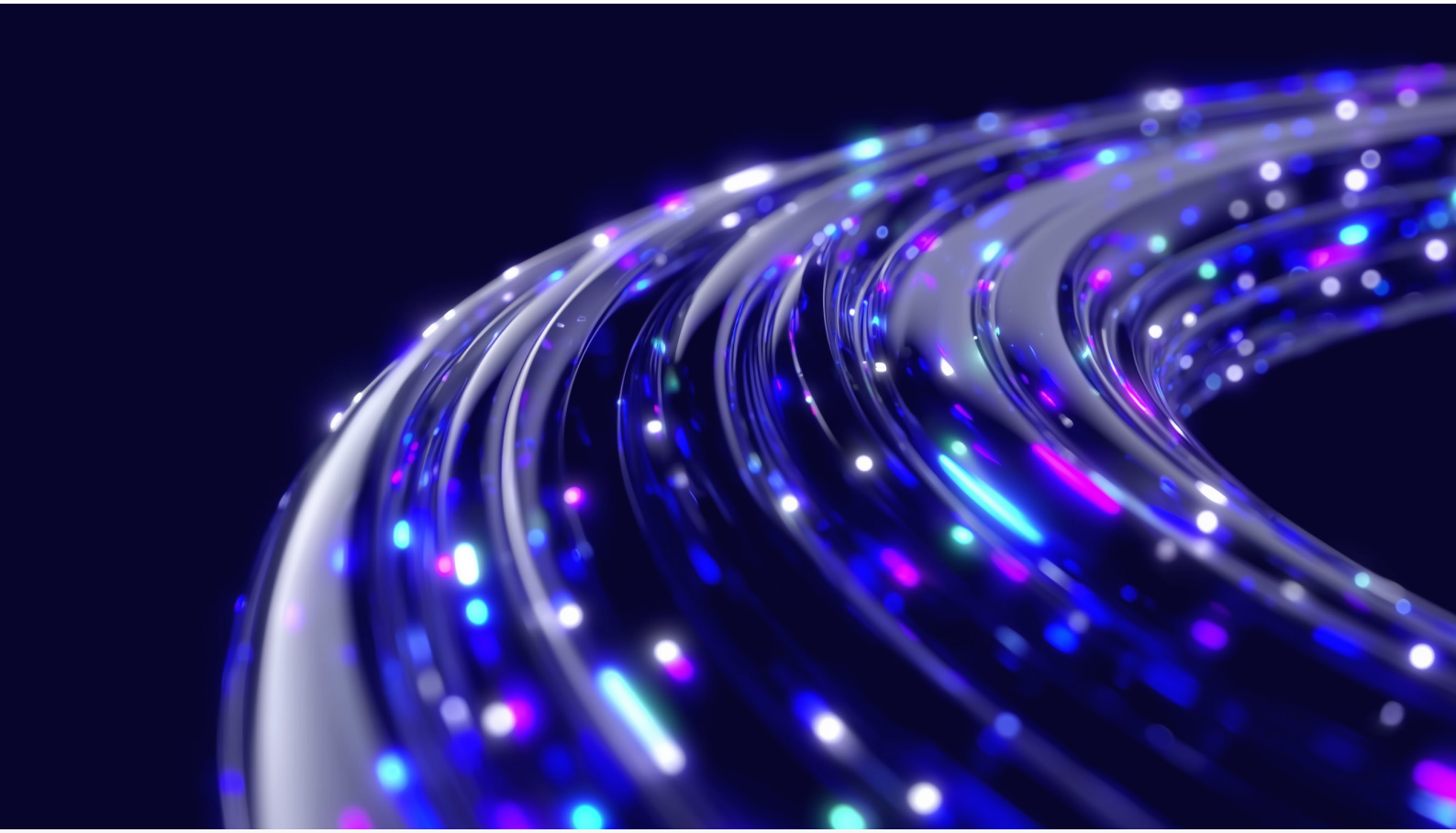
AI is moving closer to where data is created — at the edge. From AR wearables to industrial IoT devices, enterprises need real-time insights without relying on constant cloud connectivity. An Edge AI Kit delivers precisely that: hybrid local–cloud inference that enables lightning-fast performance, reduced data transfer, and enhanced privacy.

Traditional AI systems depend heavily on the cloud. While centralized models are powerful, they introduce latency, bandwidth costs, and privacy concerns — critical obstacles for experiences that demand instant, context-aware responses.

Edge AI addresses that by executing AI models locally on devices, minimizing the round-trip time to the cloud. With hybrid inference, businesses can combine the speed of local compute with the power of the cloud, optimizing performance, efficiency, and security.

The key benefits of edge AI are:

Speed	Cost efficiency	Resilience	Privacy
Process data in < 200 ms for real-time responsiveness.	Cut cloud bandwidth needs by up to 10×.	Maintain performance even in low-connectivity environments.	Keep sensitive data on-device while leveraging cloud insights securely.



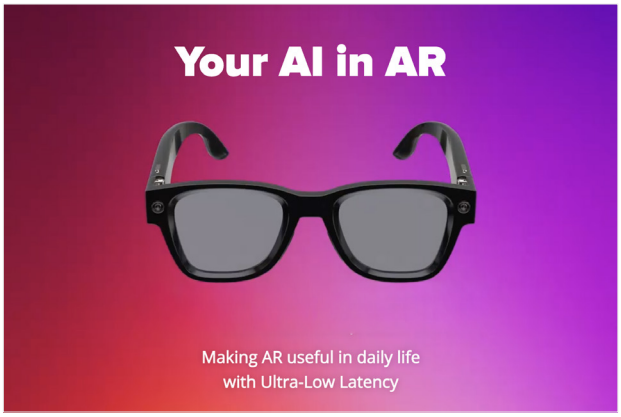
Edge AI's real-world impact

Edge AI is already delivering measurable outcomes in a variety of industries. The examples below show how moving inference closer to the data source improves speed, reduces costs, and strengthens privacy without sacrificing functionality.

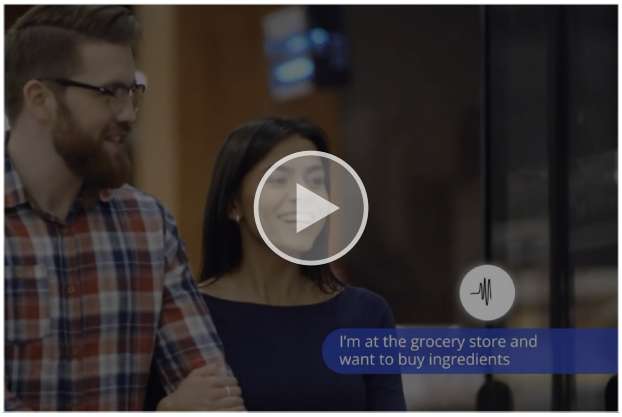
Case study spotlight: Store Assistant

A leading global AR/VR technology company sought to take in-store experiences to a new level using **AI-powered smart glasses**. Traditional cloud-based solutions couldn't meet the required sub-second response times, resulting in lag, high data costs, and degraded user engagement.

An Edge AI Kit enabled a **Store Assistant** — an AI-powered AR shopping experience that runs directly on smart glasses. By combining on-device SLMs for speech and vision with AWS Local Zones for contextual reasoning, the system delivered hybrid inference in under 200 milliseconds.



AI-powered AR Shopping with Store Assistant



Meet Your Voice-Activated Shopping Companion

The result is real-time, natural interactions, guiding customers through aisles, recognizing products, and enabling seamless checkouts, all without dependency on constant cloud access. It delivers:

Voice activation for hands-free, conversational interactions.	**Real-time navigation** and personalized recommenda-tions.	**Instant product recognition,** allergen alerts, and promotional triggers.	**Up to 10× reduction** in cloud data transfer, reducing latency and cost.	**Scalable architecture** for multi-industry adaptation.

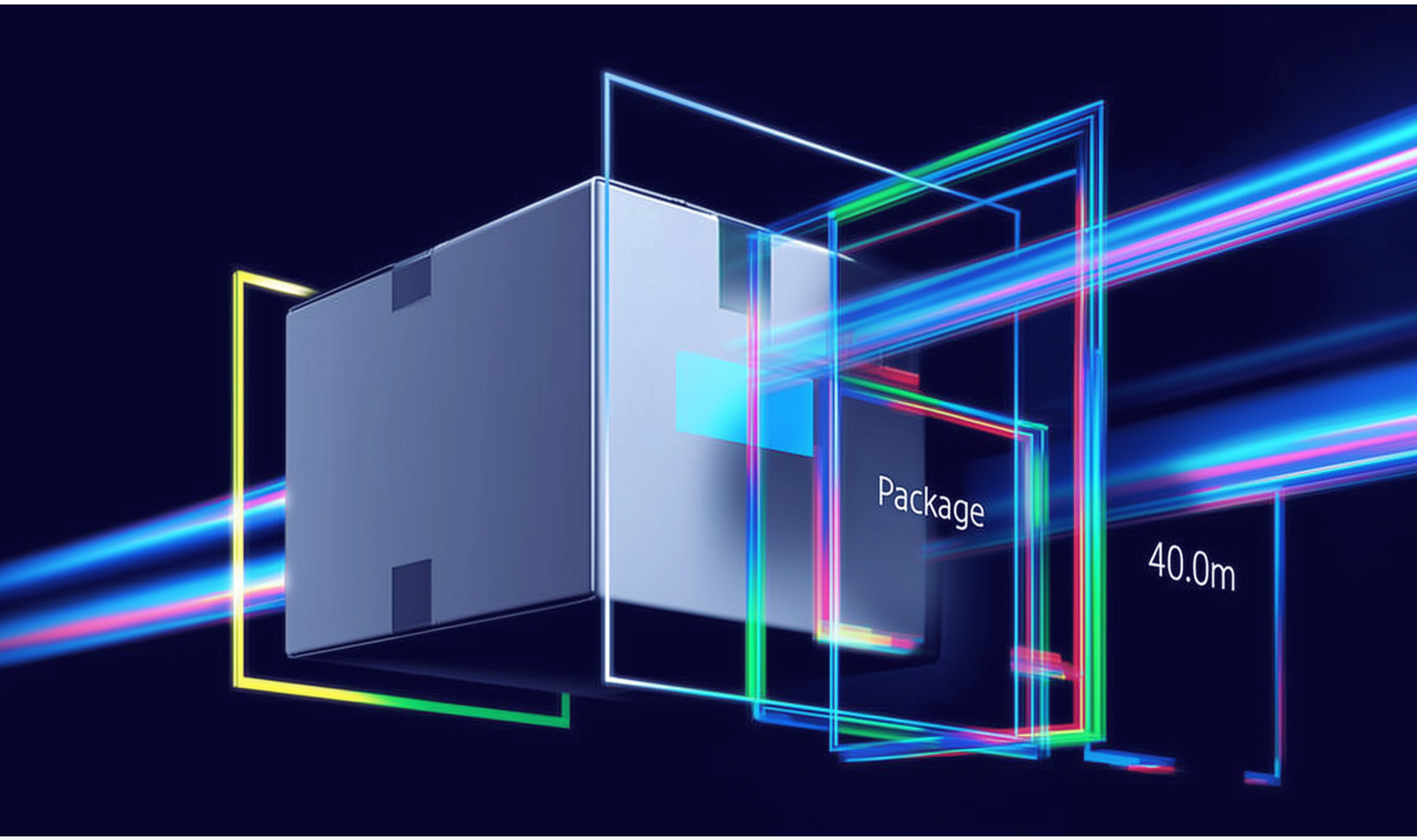
Case study spotlight: On-device image compliance

A leading player in logistics and delivery hardware, focusing on privacy-first logistics operations, needed to automate visual compliance directly on handheld devices. In last-mile delivery, couriers must take proof-of-delivery photos for every package they drop off. But these photos must comply with strict privacy and data-protection rules — they can't show people, addresses, or any sensitive information.

Traditionally, such validation happens in the cloud, which means delay, data transfer costs, and dependency on connectivity. The client wanted to move this capability directly onto the device to make it instant, private, and independent from the network.

A real-time image compliance pipeline running fully on a Qualcomm chip using the SNPE framework combines several models for feature extraction and segmentation, object detection, and text recognition. Once a courier takes a photo, the device automatically detects and blurs any sensitive areas — all locally, within about 400 milliseconds. The result is:

Instant compliance: Real-time validation and masking on-device.	**Offline operation:** Works without network connectivity.	**Reduced bandwidth:** No cloud upload required.	**Privacy-by-design:** Sensitive data never leaves the device.



Bring Gen AI to edge devices

SoftServe's Edge AI Kit is a modular platform for deploying real-time Gen AI and machine-learning models directly on edge devices such as phones, wearables, kiosks, or medical equipment.

Built on Qualcomm hardware and AWS Local Zones, the Edge AI Kit combines on-device processing with cloud scalability to deliver real-time Gen AI applications across industries. It offers:

- **Hybrid local–cloud inference:** Dynamic model routing between local hardware and AWS Local Zones for optimal latency.
- **Optimized SLM stack:** Lightweight, quantized small-language models designed for constrained compute environments.
- **Cloud-agnostic architecture:** Seamless integration with AWS and Qualcomm ecosystems; extendable to other clouds.
- **Cross-industry flexibility:** Adaptable from consumer electronics and retail to healthcare, logistics, and manufacturing.

The architecture behind the Edge AI Kit connects edge devices and cloud resources through a coordinated workflow. It balances local processing for speed with cloud-based reasoning for context, creating an adaptive system that supports real-time interaction and efficient resource use.

Edge layer (Qualcomm chips):

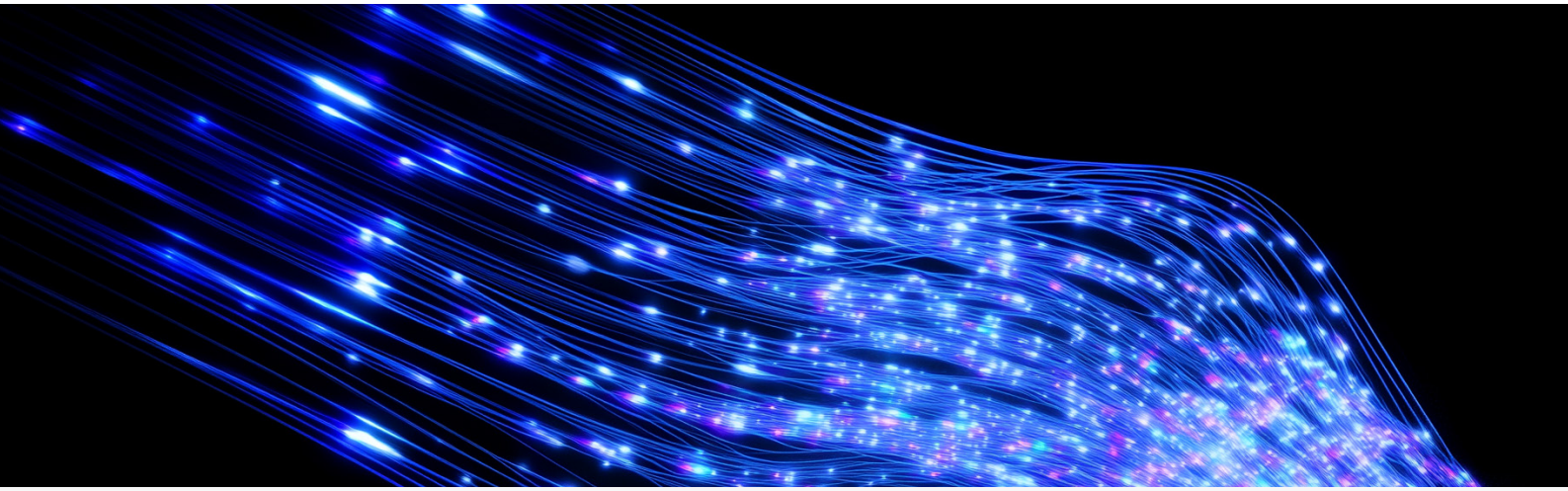
- Qualcomm chipsets for local compute.
- On-device SLMs for object detection, voice activation, and navigation logic.
- Optimized for power efficiency and low latency.

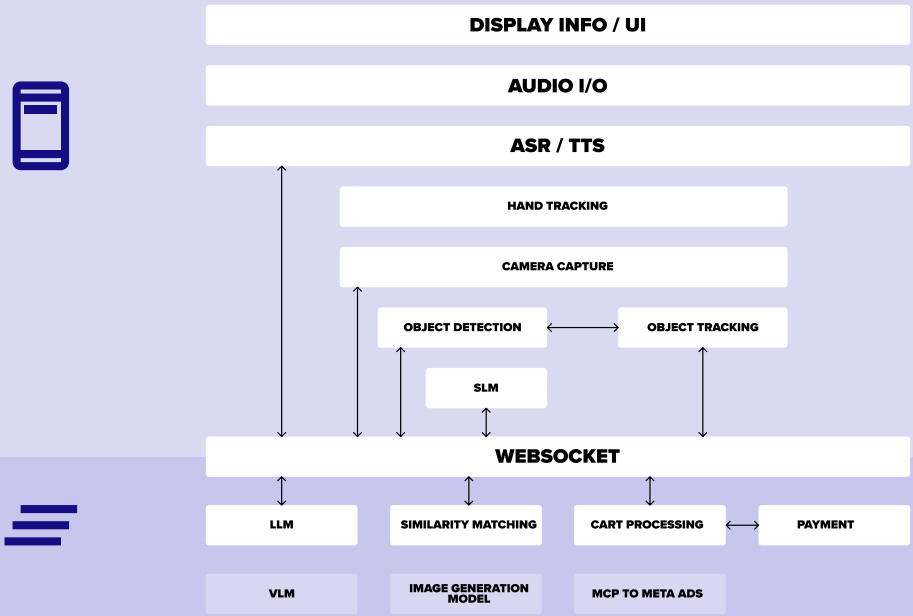
Cloud layer (AWS Local Zones):

- LLMs and Gen AI models for semantic understanding and recommendations.
- Real-time analytics for user interaction tracking and adaptive flows.
- Secure, encrypted data exchange and regional scalability.

Data flow & security highlights

- End-to-end encryption across edge-to-cloud communication.
- WebSocket-based low-latency streaming for continuous vision and speech data.
- Modular design for integration with third-party systems.





Edge AI Kit's workflow pipeline

The Edge AI Kit enables a unified hybrid inference workflow, connecting edge devices and cloud intelligence through an adaptive orchestration layer. Each step is designed to process data efficiently, maintain context, and return actionable results without compromising speed or privacy.

Step 1. Input acquisition (Edge layer)

Sensors, cameras, and microphones on the device capture multimodal data — vision, audio, or telemetry — and preprocess it locally using optimized pipelines for noise reduction and feature extraction.

Step 2. Local inference (Edge layer)

Quantized small language and vision models execute on-device, handling real-time tasks such as object recognition, speech intent detection, or anomaly identification. Latency is typically under 200 ms.

Step 3. Contextual enrichment (Cloud layer)

Only relevant metadata or embeddings are sent to AWS Local Zones (or other regional clouds), where larger models perform contextual reasoning, semantic matching, or generative responses.

Step 4. Feedback & orchestration (Edge ↔ Cloud)

Results are streamed back to the device via WebSocket for continuous, low-latency interaction. The orchestration engine dynamically decides whether inference runs locally or in the cloud, based on workload, connectivity, and latency thresholds.

Step 5. Action execution (Edge layer)

The device triggers the appropriate response — visual overlay, voice feedback, automation command, or user notification — completing a closed intelligence loop at the edge.

Returns show a bright future for SoftServe Edge AI Kit

The results demonstrate how edge-based inference improves operational efficiency and user experience. These metrics confirm that moving AI closer to the data source reduces latency, lowers bandwidth requirements, and supports scalable deployment, regardless of industry:

- **Latency**: < 200 ms real-time AI inference.
- **Efficiency**: Up to 10× less data sent to the cloud.
- **User Experience**: Seamless, low-latency interaction loop.
- **Scalability**: Architecture ready for multi-region deployment.
- **Adaptability**: Applicable to industries beyond retail — healthcare, manufacturing, smart cities, and more.

The Edge AI Kit is a horizontal platform for intelligent, low-latency AI deployment across diverse domains. Potential extensions include:

Smart wearables	Industrial IoT	Automotive & mobility	Public spaces
Real-time health monitoring and diagnostics.	Predictive maintenance and defect detection at the edge.	Driver assistance and on-board AI assistants.	Personalized digital signage and contextual advertising.

A competitive edge

SoftServe's Edge AI Kit proves that the future of AI lies in hybrid intelligence, where the immediacy of edge processing combines with the power of the cloud. The Store Assistant Demo shows how this fusion delivers real-world value today through smarter, faster, and more personalized experiences for businesses and end users alike.

Ready to accelerate your edge AI strategy?

Speak with SoftServe's edge AI experts to explore how this technology can be adapted to your use case.

CONTACT US

About Us

SoftServe is a premier IT consulting and digital services provider.

We expand the horizon of new technologies to solve today's complex business challenges and achieve meaningful outcomes for our clients. Our boundless curiosity drives us to explore and reimagine the art of the possible. Clients confidently rely on SoftServe to architect and execute mature and innovative capabilities, such as digital engineering, data and analytics, cloud, and AI/ML.

Our global reputation is gained from more than 30 years of experience delivering superior digital solutions at exceptional speed by top-tier engineering talent to enterprise industries, including high tech, financial services, healthcare, life sciences, retail, energy, and manufacturing.

Visit our website, blog, LinkedIn, Facebook, and X (Twitter) pages for more information.

AUSTIN HQ

201 W. 5th Street, Suite 1550
Austin, TX 78701
+1 866 687 3588 (USA)
Toll Free: +1 866 687 3588

LONDON

30 Cannon Street
London EC4 6XH
United Kingdom
+44 203 807 01 41

info@softserveinc.com
www.softserveinc.com

softserve