

# Alibaba says Qwen 4 is in training and unveils its Zhenwu V900 AI chip

Source copy kept by wheresthe.ai for the story "Alibaba says Qwen 4 is in training and unveils its Zhenwu V900 AI chip" (<https://wheresthe.ai/d/cn-alibaba-apsara-qwen-4-zhenwu-2026-09/>).

Source: <https://www.alizila.com/alibaba-clouds-2026-apsara-conference-full-stack-ai-roadmap-along-with-global-market-expansion-plan/>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: 1d73cd00556f01ea47428d8bad2df91125aa5e082ba93b995e25f81d4fc8706c

---

Support Turkey Earthquake Emergency Relief

Support Turkey Earthquake Emergency Relief

-->

By

Alizila Staff

|

Published on Sept. 24, 2026

At the 2026 Apsara Conference, Alibaba Cloud's annual flagship technology conference, Alibaba unveiled a roadmap for its full-stack AI strategy spanning chips, cloud infrastructure, models and agents, alongside plans for its first cloud regions in Türkiye, Finland and the Netherlands and a target of more than 20 GW of global data center capacity by 2032.

"The theme of this year's Apsara Conference is 'Intelligence Goes Beyond,'" said Joe Tsai, Chairman of Alibaba Group. "This is the true meaning of 'Intelligence Goes Beyond': guiding AI from technological breakthroughs toward value creation.

"Today, the total volume of Machine Thinking is less than 3% of all Human Thinking. If that volume eventually scales to 1,000x human capacity, the simple math tells us: Machine Thinking still has an enormous growth runway. With this in mind, our target is that by 2032, the global data center capacity operated by Alibaba Cloud will surpass 20GW, fueling the industry's exponentially rising demand for AI," said Eddie Wu, CEO of Alibaba Group.

Here's why these announcements matter – and how they fit into a broader push to make AI more scalable, practical and accessible for businesses.

Qwen 4 in training

At the conference, Alibaba revealed that its next-generation model, Qwen 4, is currently in training. The company further announced its roadmap for the upcoming Qwen 4.5 and Qwen 5 model series, projected to scale up to 5 to 10 trillion parameters.

Alibaba has revealed progress in RSI (Recursive Self-Improvement) driven by empirical feedback. Over a month of fully automated runs, Qwen3.8-Max completed 33 iterative cycles, boosting its Artificial Analysis score from 40 to 45. In a chip design experiment, the model underwent over 60 hours of self-improvement, making more than 10,000 EDA tool calls to produce production-grade chip bus modules, reducing chip area by 42% with zero compromise in performance.

Multimodal models from Speech, Audio to Vision

Alibaba introduced multimodal model updates spanning speech, audio and vision. The company debuted Qwen3.8-LiveTranslate, a simultaneous interpretation model that reduces latency (LAAL) nearly 20% from 2.8 to 2.3 seconds, and unveiled Qwen-Audio-3.1-TTS-Next, an audio generation model capable of creating complete cinematic soundscapes from a text script for audiobooks, film and television, podcasts and games.

Qwen-Image 3.1, an image-generation model optimized for creative design and e-commerce marketing, is set to

launch later this year, cutting visual design turnaround from hours to seconds.

Bringing the models closer to general users, Alibaba has launched Qwen Intelligence , a business-facing full-stack agentic solution optimized for smartphones. It provides phone makers with access to a Qwen-powered agent platform to support next-generation AI phones that can handle complex, cross-app tasks reliably.

#### New silicon from T-Head

T-Head, Alibaba's chip design unit, has unveiled the Zhenwu V900, its latest AI training and inference processor. The Zhenwu V900 delivers three times the performance of its predecessor, the Zhenwu M890, featuring 216 GB of GPU memory and 1,200 GB/s of inter-chip bandwidth with native support for FP8 and FP4. It is scheduled for mass production and commercial release in the first quarter of 2027. T-Head's Zhenwu AI chips have been serving over 650 customers across industries including automobile, finance, energy and manufacturing.

Alibaba also unveiled its upgraded supernode server, integrating the Zhenwu V900 processor, ICN Switch, Panmai SmartNIC and Zhenyue SSD controller chip, supporting a supernode cluster comprising up to 500,000 cards.

Its roadmap for next-generation proprietary CPUs tailored for agentic AI tasks includes the Yitian 720 and the Yitian 730, scheduled for launch in 2027; the Yitian 730 is the first CPU built on T-Head's proprietary microarchitecture, with up to a 40% increase in SPECint2017/GHz performance over the Yitian 710.

#### An agentic cloud

Alibaba Cloud unveiled a suite of upgrades centered on its agentic cloud strategy, built around three core scenarios – model, harness and context.

The AI Native Cloud powers large-scale model training and inference. In real-world post-training of Qwen models, its Platform for AI (PAI) completed state-of-the-art model training in just five days; the new generation system Cloud Parallel File Storage (CPFS) cuts enterprise AI storage costs by 69%; and HPN 8.0 Pro supports over 130,000 network ports at 800G speeds in a single cluster.

The Agent Native Cloud enables enterprise-grade agent deployment, operation and security, led by AgentCore to build, run and manage AI agents throughout their lifecycle and the Agent Security Center for full-lifecycle security and compliance.

The Context Engine provides real-time data and long-term memory. Agent Context connects a company's documents, business systems, chat records and multimodal data into a single foundation, cutting token usage by up to 67%. OpenLake has been upgraded into a unified, multi-modality data lakehouse, reducing total costs by 38% and cutting query response times by 40%.

#### Taking the stack global

The conference's second day turned to international business. Over the next 12 months, Alibaba Cloud will establish its first cloud regions in Türkiye, Finland and the Netherlands, while expanding its data center footprint in Malaysia, Germany, the United Arab Emirates, France and Hong Kong. It currently operates 107 availability zones in 31 regions.

"We are expanding our global cloud footprint to bring computing resources closer to customers and partners, while continuing to evolve our AI technologies to help enterprises deploy models, optimize multi-model workloads and create new AI-powered applications," " " said Dr. Feifei Li, Chief Technology Officer and President of International Business of Alibaba Cloud Intelligence .

Alibaba Cloud also launched three new AI solutions to help international businesses build and scale production-grade applications. Smart Studio, an out-of-the-box product, helps businesses build dedicated Model-as-a-Service (MaaS) platforms in minutes, with its inference architecture delivering up to 505% higher throughput than standard open-source frameworks on the same hardware. Smart Fusion abstracts multi-model deployment into a single endpoint, helping reduce token spending by around 50%. Smart Video is an AI video agent for short-drama studios and advertisers that automatically creates videos up to one hour long, with support for more than 20 languages.

#### Customers turn AI into real-world impact

Alibaba's cloud and AI technologies are helping customers and partners across industries address diverse

business needs.

Panasonic Digital (Shanghai) has partnered with Alibaba Cloud to accelerate its comprehensive digital and AI transformation. Leveraging Alibaba Cloud's Bailian (Model Studio) to securely access the Qwen models, Panasonic has built a custom capability hub that powers specialized AI agents across manufacturing, supply chain and corporate functions. The collaboration has achieved an 80% reduction in documentation time through automated meeting minutes, a 10-fold increase in contract review efficiency using the "Smart Compare" agent.

Loomi Entertainment Group (LEG), a Malaysia-based entertainment company, has partnered with Alibaba Cloud to accelerate the adoption of AI-generated content (AIGC) among creators and businesses in Malaysia and the wider region. LEG has integrated Alibaba's visual generation models, including Wan and HappyHorse through Model Studio, into Imaginary, its proprietary cinematic AI engine. Additionally, the two companies have trained more than 100 local content creators.

With next-generation models in training, purpose-built silicon heading to production and an agentic cloud built for the next wave of AI, Alibaba Cloud is positioning its full stack as the foundation for enterprises moving AI from experimentation into everyday operations.

Alongside its enterprise offerings, Alibaba also unveiled a series of AI hardware products at the conference, including an agentic computer, its latest AI glasses, earbuds and an upgraded note-taking device, designed to bring AI capabilities into everyday life and boost productivity at the workplace and beyond.

Explore comprehensive details about the 2026 Apsara Conference at [2026 Apsara Conference: Press Kit](#)

[AI Alibaba Alibaba Cloud Apsara Conference](#)

[Reuse this content](#)

[Never Miss a Story Subscribe to Our Newsletter](#)

[Popular Stories](#)

[AI QwenWork Unveiled E-Commerce Tool To Help Merchants in China Boost Efficiency](#)

[AI Agenda Joe Tsai: Open Source Is Europe's AI Opportunity](#)

[AI Alibaba DAMO Academy Unveils New AI Models to Advance Cancer Diagnoses](#)

[AI Alibaba Unveils Agentic Computer, AI Wearables and More at 2026 Apsara Conference](#)