CIC灼识

Global On-Device AI Inference Chip Industry

Executive Summary

The rapid expansion of smart devices deployment has led to an exponential increase in perception data. This surge has sharply raised the demand for front-end processing and real-time local computation. Traditional cloud-based architectures, limited by latency and computing capacity, are increasingly falling short of current application needs. As a result, AI-powered devices must advance in both high-precision perception and efficient computation.

In this context, on-device AI inference chips have grown more important than ever. By integrating AI models with smart device perception technologies, these chips form a real-time closed-loop system of sensing, computing, and execution. They enable AI-driven analysis and decision-making directly on the device using physical data such as texts, images, videos, and audios, significantly reducing reliance on cloud resources.

Table of Contents

1. Market Overview

1.1 Market Definition

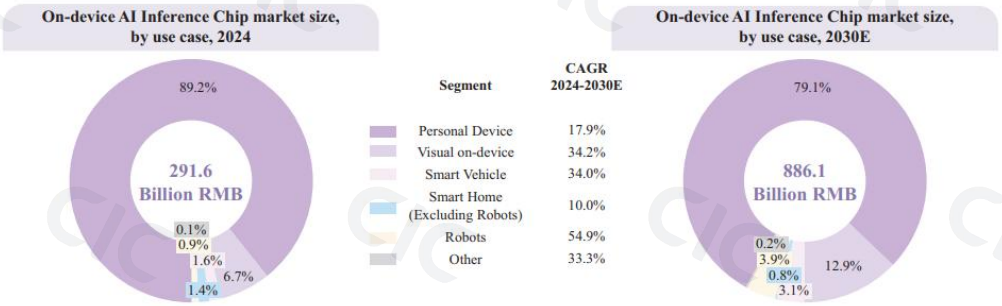
On-device AI inference chips are used directly in end-user devices, such as consumer electronics like smartphones, smart vehicles, and smart home appliances.

Surging perception data volumes have driven more demands for on-device AI inference chips. By integrating AI models with smart device perception technologies, these chips form a real-time closed-loop system of sensing, computing, and execution. They enable AI-driven analysis and decision-making directly on the device using physical data such as texts, images, videos, and audios, significantly reducing reliance on cloud resources.

1.2 Market Size and Shipment

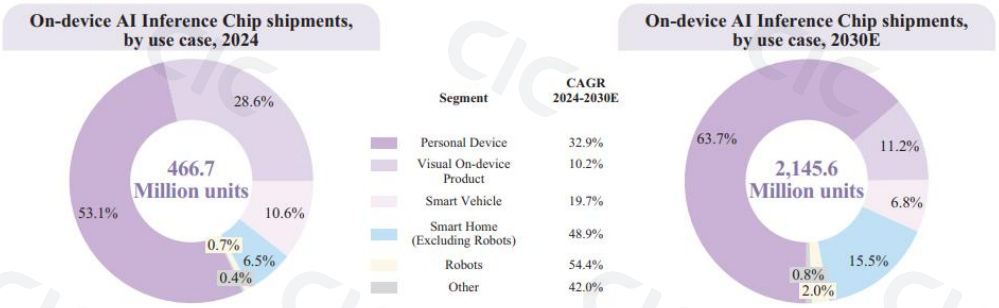
The on-device AI inference chip market is broad. Global on-device AI inference chip market size on various segments is expected to grow fast, with CAGRs of 54.9%, 34.2%, 34.0%, 17.9% and 10.0% for robots, visual on-device product, smart vehicle, personal devices and smart home,respectively.

Global On-device AI Inference Chip Market Size by Use Case, 2024-2030E

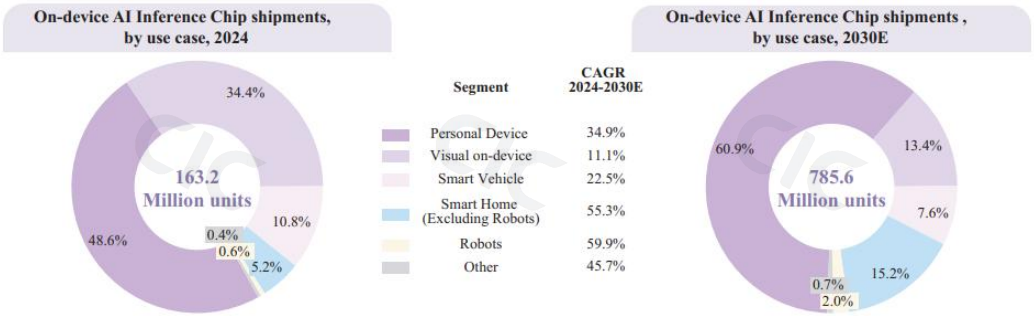


From 2024 to 2030, Global on-device AI inference chip shipments on robots, smart home, personal devices, smart vehicle, and visual on-device product are projected to achieve shipment CAGRs of 54.4%, 48.9%, 32.9%, 19.7% and 10.2%, while their counterparts in China are expected to grow at the rates of 59.9%, 55.3%, 34.9%, 22.5% and 11.1%, respectively.

Global On-device AI Inference Chip Shipments by Use Case, 2024-2030E



China's On-device AI Inference Chip Shipments by Use Case, 2024-2030E (Million)



Notes:

Personal device includes consumer electronics such as smartphone, wearable device and XR which refers to head-mounted and wearable devices designed for immersive interactive experiences, encompassing applications in virtual reality (VR) and augmented reality (AR).

Visual on-device product includes visual perception devices in different scenarios, such as industrial, urban, household and other.

Smart vehicle refers to vehicles equipped with intelligent driving functions.

Smart home refers to household applications such as TV, speaker, and air conditioner.

Robots include industrial robot, vacuum cleaner, food delivery robots and etc.

Other includes industrial testing equipment and intelligent industrial control system.

Source: CIC Reports, interviews with market participants, industry publications, government statistics, listed companies' public filings, news

2. Fast-growing product segment: Visual On-device Inference Chip

2.1 Market Definition of Visual On-device AI Inference Chip

Visual on-device products are designed to process only a single type of visual input. They commonly include public safety cameras, dash cams, smart locks, machine vision inspection equipments, etc. Visual on-device AI inference chips are typically categorized by performance into low-end and mid-to-high-end categories. While low-end chips typically feature less than 1 TOPS (Tera Operations Per Second) in computing power, mid-to-high-end chips exceed this threshold.

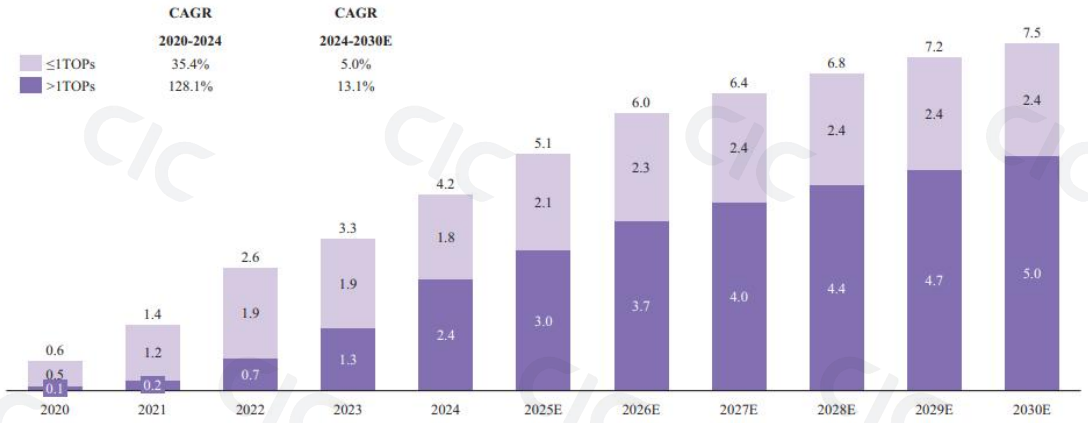
2.2 Market Size and Shipment of Visual AI Inference Chip

As demand continues to rise for high resolution, intelligent, and low-latency processing in visual applications, the market is rapidly shifting toward mid-to-high-end chips. These chips are emerging as the fastest-growing product segment.

In 2024, the global market size for mid-to-high-end visual on-device AI inference chips is expected to increase from RMB2.4 billion in 2024 to RMB5.0 billion in 2030, representing a CAGR of 13.1%.

Global Market Size of Visual On-device AI Inference Chips by TOPS, 2020-2030E

(RMB Billion)



Correspondingly, global shipments of mid-to-high-end chips reached 34.8 million units, accounting for approximately 26.0% of the total. By 2030, this number is expected to rise to 99.9 million units, with the share expanding to 41.3%.

Global Shipments of Visual On-device AI Inference Chips by TOPS, 2020-2030E

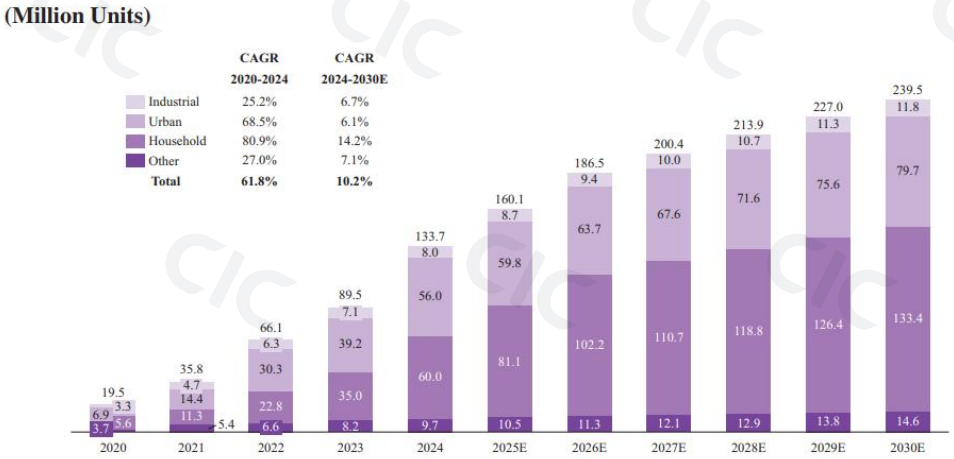
(Million Units)

Market segmentation by application scenario provides another critical dimension for calculation.

The global shipment of visual on-device AI inference chips is

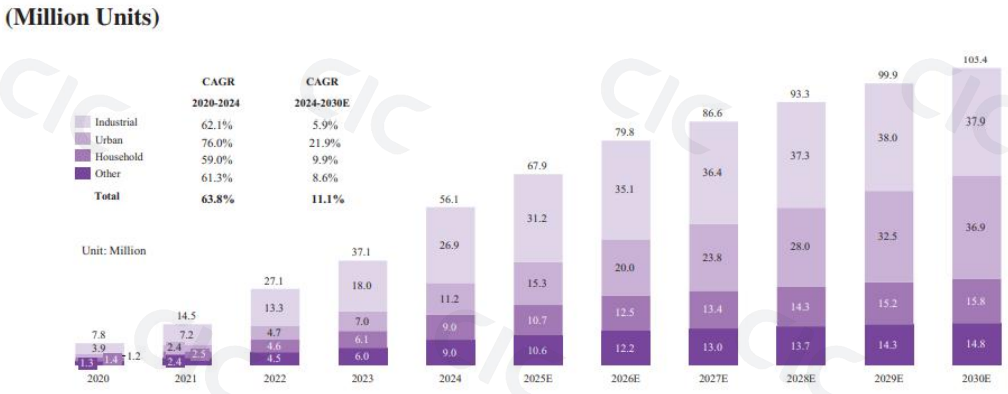
expected to grow significantly from 133.7 million units in 2024 to 239.5 million units in 2030, representing a CAGR of 10.2%.

Global Shipments of Visual On-device AI Inference Chips by Application Scenario, 2020-2030E



China's shipment of visual on-device AI inference chip is expected to reach 105.4 million units by 2030. These chips have found widespread adoption across diverse domains, including household, urban, industrial, and other application areas.

China's Shipments of Visual On-device AI Inference Chips by Application Scenario, 2020-2030E



Source: CIC Reports

2.3 Key Drivers and Trends in Visual On-device AI Inference Chip

Intelligent transformation: Conventional image and video systems no longer meet the real-time, precise, and intelligent requirements of modern household, urban and industrial environments. To enable capabilities such as facial recognition, behavior analysis, and anomaly detection in real time, powerful and energy-efficient AI chips are essential, driving increasing adoption of visual on-device computing chips.

Rising data processing demand: Devices increasingly require high-resolution, multimodal perception, which is lifting the demand for heterogeneous chip architectures that integrate NPUs, ISPs, and DSPs, providing the computing backbone for complex AI tasks like image enhancement, semantic understanding, and voice interaction.

Policy support: China's Notice on Promoting the Development of the "Internet of Everything" in Mobile Internet of Things encourages innovation and industrialization in chip and module technologies. Other initiatives, including the New Infrastructure Construction Plan and the 14th Five-Year Plan for Intelligent Manufacturing, further

strengthen the policy framework that supports large-scale deployment of on-device AI inference chips.

About CIC

CIC is a professional consulting firm offering tailored end-to-end support across the full investment and financing lifecycle. The firm boasts a world-leading track record in guiding landmark first-in-sector IPOs across global markets, alongside unrivaled reach and in-depth coverage capabilities across specialized niche market segments.

CIC helps enterprises refine scalable business models and craft compelling capital narratives to enable seamless access to global capital markets, while serving as a trusted due diligence partner to investment institutions. It delivers granular industry insights and direct access to subject matter experts, empowering clients to identify high-value opportunities and mitigate critical risks effectively.

CIC team maintains deep, real-time market intelligence across a diverse set of sectors—including financial services, artificial intelligence, big data, internet, high technology, healthcare, education, entertainment, consumer goods, transportation and logistics, energy and power, environmental and building technology,

chemicals, industrial manufacturing, and agriculture—delivering unparalleled access to sector-specific, actionable insights.

CIC Reports & Industry Overview

At CIC, we employ a rigorous, multi-method research framework, combining primary and secondary sources to underpin our analysis. Primary research involves in-depth engagements with industry thought leaders and practitioners, particularly in supply chain finance. Secondary research synthesizes publicly available datasets from authoritative bodies, including the National Bureau of Statistics of the People's Republic of China, the State Administration of Financial Regulation (SAFR, formerly the China Banking and Insurance Regulatory Commission), the China Securities Regulatory Commission (CSRC), and public company filings. We apply proprietary data analytics frameworks to process collected information, validating findings through cross-referencing data from multiple research streams to ensure analytical rigor and reliability.

All statistical data presented is verifiable and grounded in information available as of the date of this report.

Extracts are refined summaries of in-depth CIC industry research reports, highlighting supply and demand trends, key growth drivers, R&D trends and future outlook, etc. of various segmented fields, integrating multi-dimensional insights such as expert interviews, market surveys and industry data analysis.

Disclaimer

This report (the "Report") is prepared by CIC based on information available as of the date hereof. The Report is furnished strictly for informational and reference purposes only and is not intended to, nor shall it be construed as, being definitive or conclusive. Nothing contained herein shall constitute or be deemed to constitute investment advice, a recommendation, or an offer, solicitation or inducement to engage in any investment activity. CIC hereby expressly disclaims any and all liabilities for any loss, damage or claims of any nature howsoever arising, whether directly or indirectly, from the use of or reliance upon any information contained in the Report.

Contact CIC

For more information about this report or to learn more about CIC services, please visit CIC official website, or email us at marketing@cninsights.com.