How Qualcomm is Powering AI-driven Multimedia at the Edge

Ning Bi

VP, Engineering

Qualcomm Technologies, Inc.

Qualcomm

Agenda

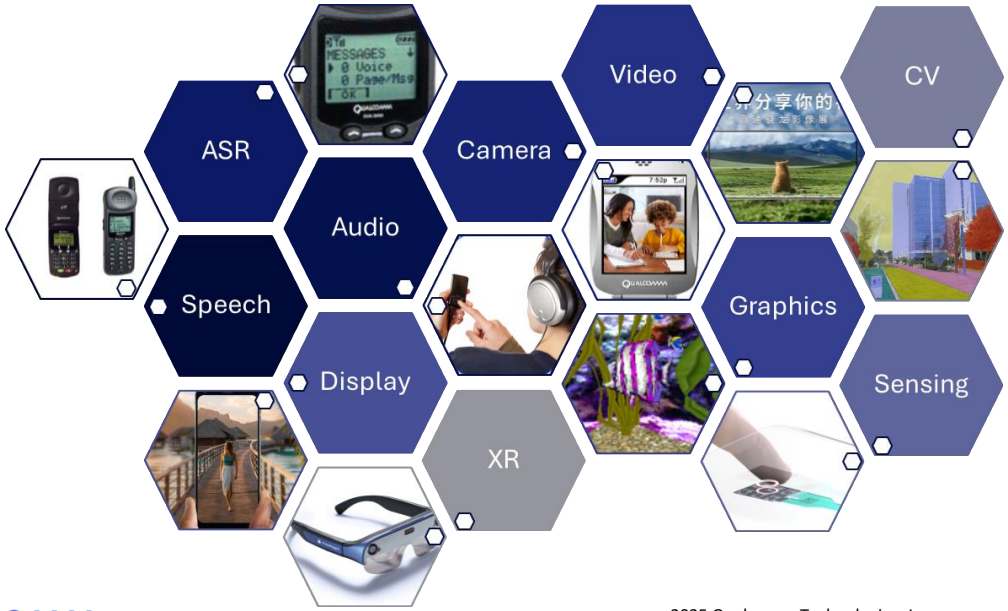
- Embedded multimedia driven by AI
- Evolution of multimedia at the edge
- Enabling AI solutions efficiently at the edge
- Ecosystem from Qualcomm Technologies

Embedded multimedia driven by AI

From mobile station modem to Snapdragon® SoCs

Growth of multimedia to support the business development of Qualcomm Technologies

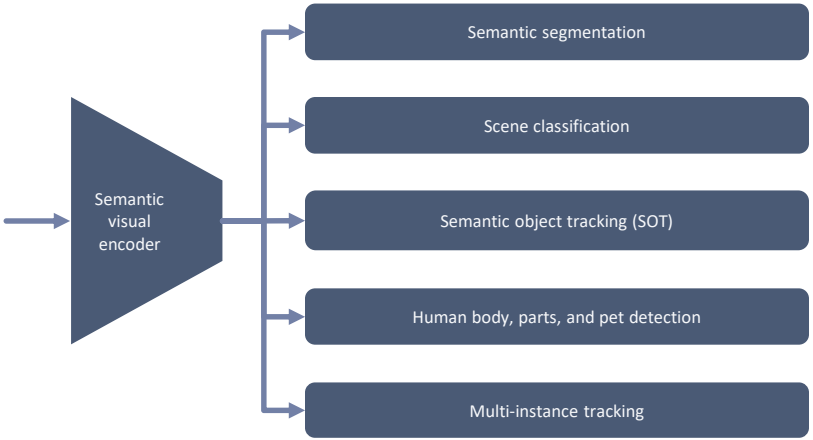
- Multimedia content creation, distribution and consumption
- Consumption in dual ways: device ⇆ user
- Revolutionizing the processes by AI



From algorithm-engineering to data-driven approach

AI enabled a new generation of multimedia solutions on edge

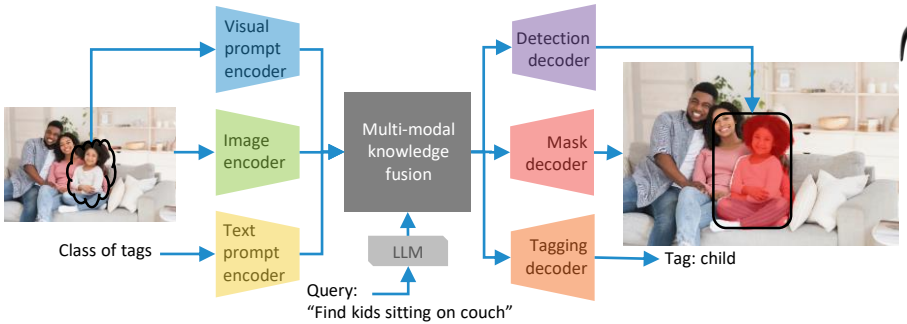
- Example on visual object tracking
  - CV algorithm: Touch-to-track (T2T) by using KCF, KL flow, and updated templates with BRIEF+NCC
  - Data-driven: Semantic object tracking (SOT) by a NN semantic visual encoder with multiple attentions
  - Doubled accuracy in benchmark
- Edge implementation on Snapdragon®
  - NN running on Qualcomm® Hexagon™ NPU, Qualcomm® Adreno™ GPU, or Qualcomm® Oryon™ CPU
  - CV algorithms with dedicated hardware on the Engine for Visual Analytics (EVA)



From artificial perception to artificial intelligence

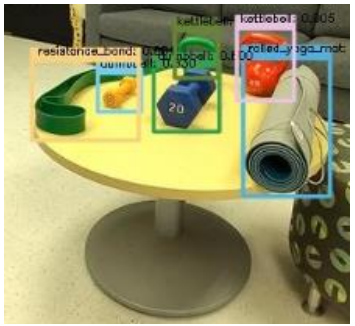
Enable machines to sense or to make analysis for prediction

- Artificial perception (AP)
  - Auditory, visual, and other sensing technologies enabling machines to acquire surrounding information
  - Examples in computer vision technologies: object detection, tracking, classification, etc.
- Artificial intelligence (AI)
  - Goal of AI: mimic core cognitive activities of human thinking
  - The NN foundation models trained on large amount of data applied in multi-tasks show possibilities
  - Vision language model (VLM) with cross-modal knowledge fusion
- Examples on edge
  - Open vocabulary segmentation, detection and tagging
  - AI glasses in recall, assistant, and forensic search



Open vocabulary segmentation, detection and tagging

AI recall

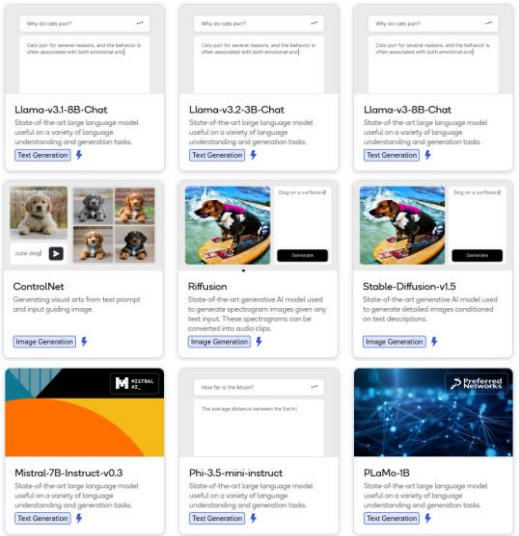


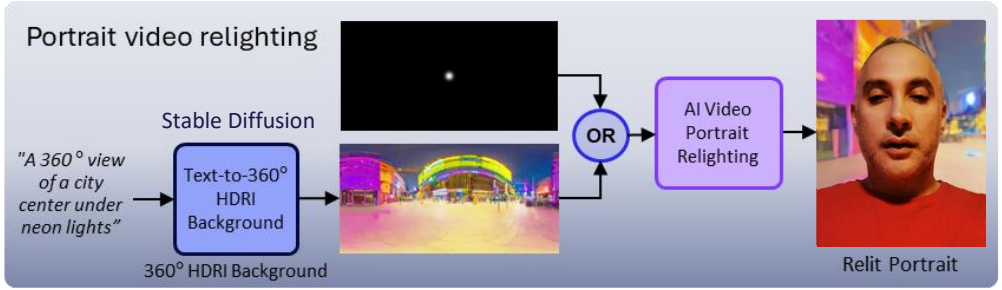
AI assistant

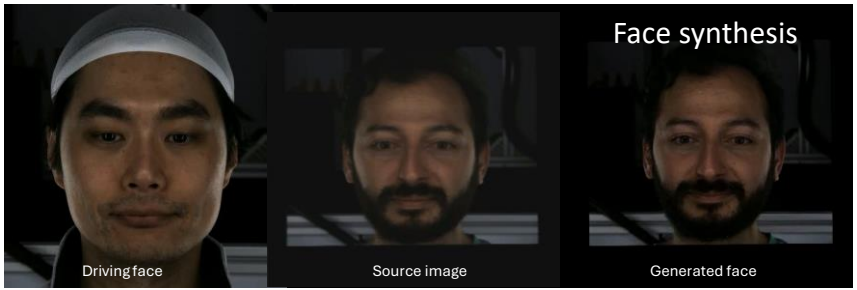
From AI to generative AI

Create new content, like text, audio, image, video, or code, from “emergent” abilities

- Embedded gen AI in text generation
  - Llama-v3.2-3B-Chat, Llama-v3-8B-Chat, Phi-3.5-mini-instruct, etc.
  - 10+ LLM models available at Qualcomm AI Hub
- Embedded gen AI in image generation
  - ControlNet, Riffusion, Stable-Diffusion-v2.1, etc.
  - Example app: Background augmentation in Portrait Video Relighting
- Gen AI in human biometric data generation
  - AI model training data synthesis, e.g., same ID with different appearance
  - Applied in human biometric authentication to avoid issue of data privacy





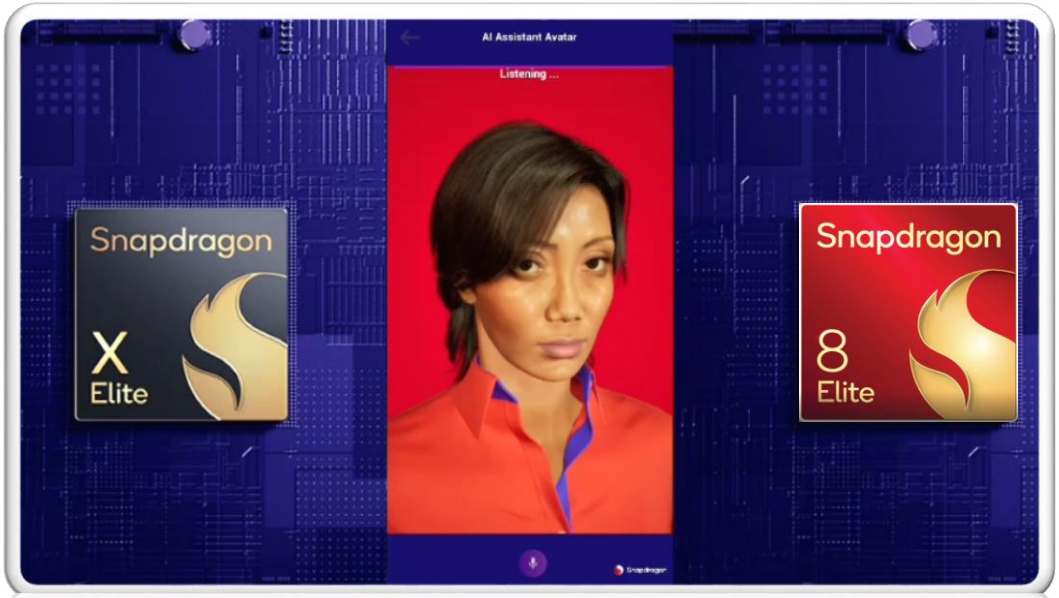


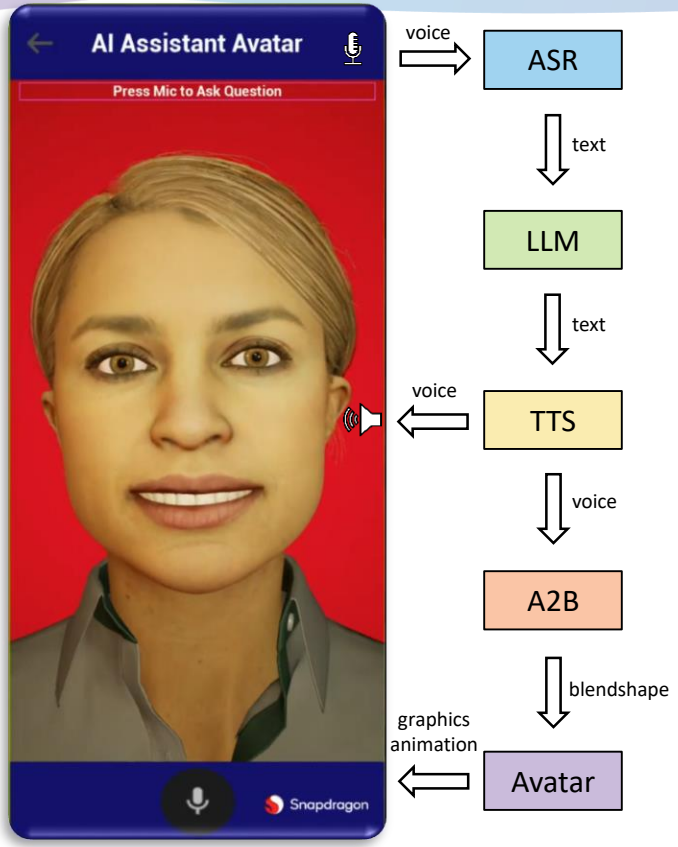
Evolution of multimedia at the edge

From smartphone to personal AI assistant

AI assistant avatar driven by gen AI with real-time human face animation

- Talking avatar to embody an LLM as an Agentic AI
  - In Snapdragon® 8 Elite mobile platform for smartphone
  - In Snapdragon® X Elite platform for AI PC
  - Applications as AI assistant avatar, translator, info kiosk, etc.

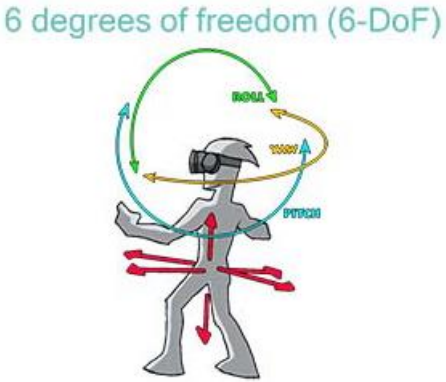




From stereo sound to spatial audio rendering

AI-driven spatial audio with head tracking

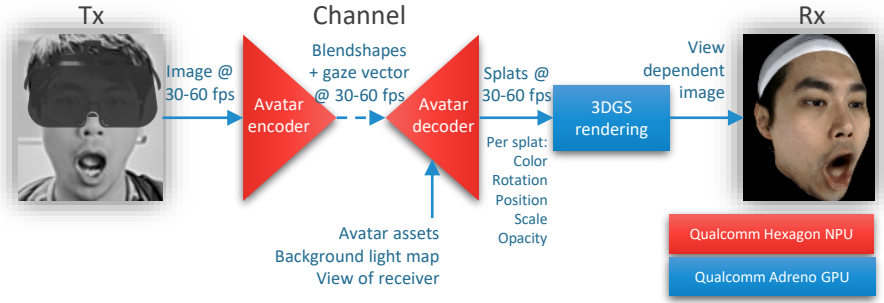
- Reconstruction of sound field
  - Human hearing allows us to precisely identify the location of sounds.
  - Human brain interprets location cues according to how sound interacts with the shape of ears and heads.
  - Spatial audio technology emulates acoustic interaction to stimulate the mind into perceiving sound in 3D.
  - Over earbuds or headphones, the emulation relies on sound filters of Head Related Transfer Function (HRTF).
- Head tracking to align the sound field with the world coordinates
  - By using 6-DoF motion tracking in VR HMD or AR glasses
  - By using facial landmarks and head pose tracking to link a display with earbuds or headphones

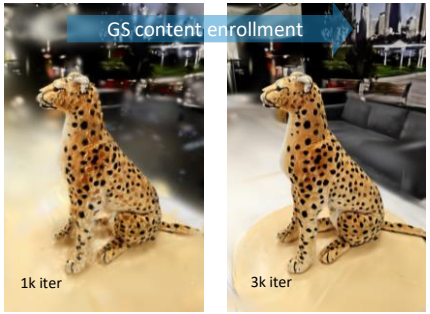


From 2D video to 3D spatial media by Gaussian splatting

AI-driven spatial visual media

- Reconstruction of the real 3D world
  - Picture and video is 2D projection of the 3D real world
  - Rebuild a photorealistic virtual 3D world
- Gaussian splatting (GS) representation
  - Content capturing and creation
  - Content transition and rendering
- Relightable GS avatar
  - Facial expression encoder
  - Blandshapes over the communication channel
  - Avatar decoder, 3D GS rendering and animation



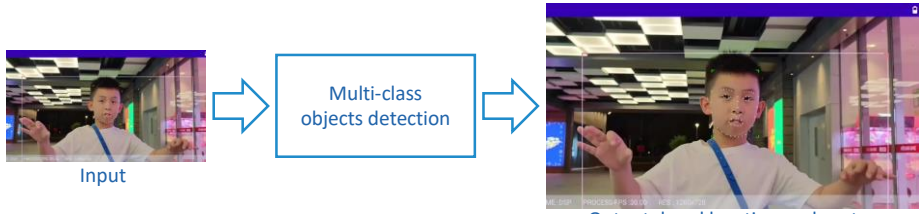


Enabling AI solutions efficiently at the edge

From multi-stage to one-stage model

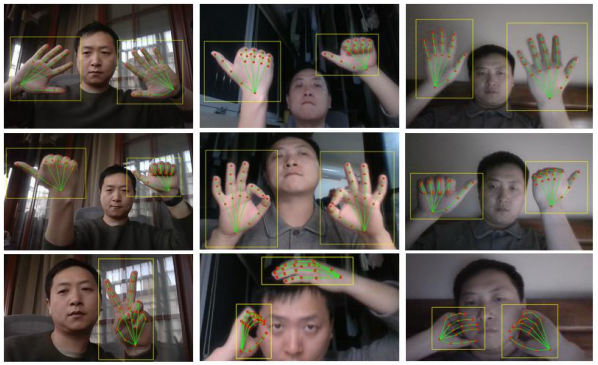
End-to-end approach for faster and low-power solution

- Many CV tasks may involve
  - Multiple steps, e.g., detection, tracking, classification, etc. and/or
  - Multiple pipelines from different cameras and/or sensors, such as ADAS
- Data-driven approach for an end-to-end solution
  - Hand gesture recognition
    - Multi-stage: hand detection, skeleton, & gesture classification
    - One-stage: detection of multiple hand gestures

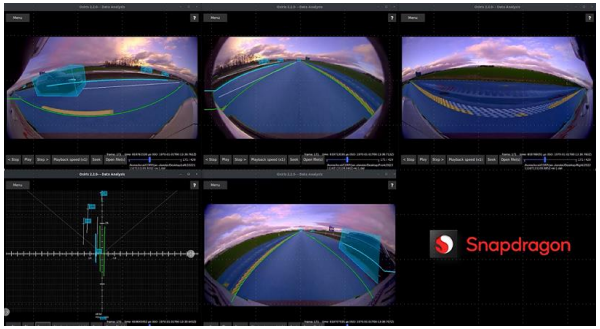


  - BEV in ADAS
    - BEVFormer (birds eye view) networks with multi-camera inputs to reach high accuracy at the system level
    - Representing surroundings using BEV perception in downstream modules for planning and control

Hand gestures



BEVFormer

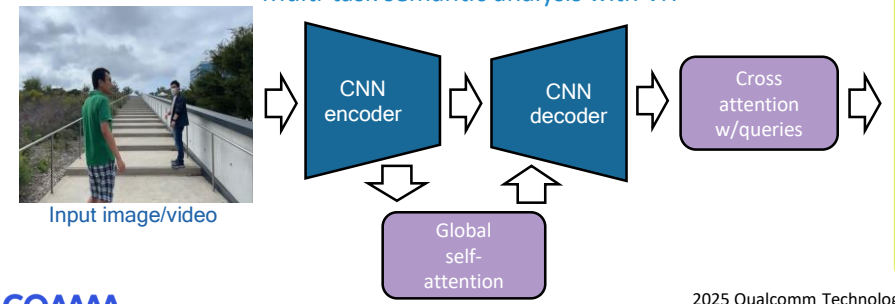


From single-task to multi-task neural networks

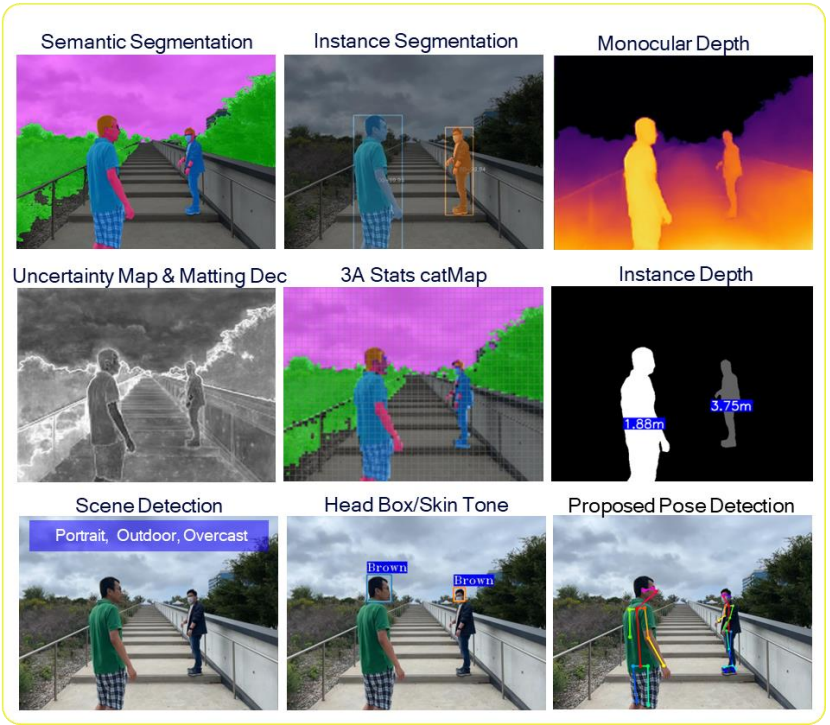
Shared backbone for CV feature concurrency

- Visual encoder-decoder architecture with transformer
  - Self-attention for accurate semantic segmentation
  - Cross attention with queries for multiple other tasks
- Multi-task NN as a foundation model
  - Semantic and instance segmentation
  - Monocular and instance depth
  - Uncertainty map for smoother bokeh
  - Head detection and skin tone classification
  - Scene classification
  - Body pose detection

Multi-task semantic analysis with ViT



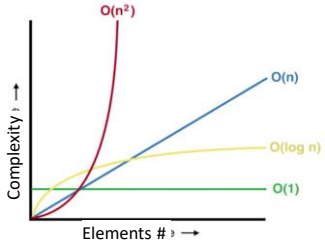
Output: multi-tasks for AI camera



From transformer to Mamba-like linear attention

Fitting better on the edge with compute, memory and power limitations

- Advantage of transformer with multi-head attention
  - Popular NN architecture for LLM and vision models
  - “Attention is all you need.” $O=\text{Softmax}(QK^T)V$
  - Challenge when not having enough memory – giving up multi-scale attention
- Mamba-like linear attention
  - Linear attention by selective state space model
  - Attention complexity reduced from $O(N^2)$ to $O(N)$
  - Able to use multi-scale attention
- Improvement in video segmentation
  - Approach 1: video semantic segmentation based on transformer
  - Approach 2: video semantic segmentation based on Mamba-like linear attention with 6% improvement

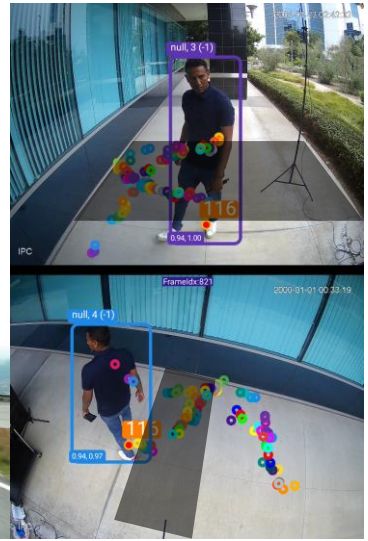


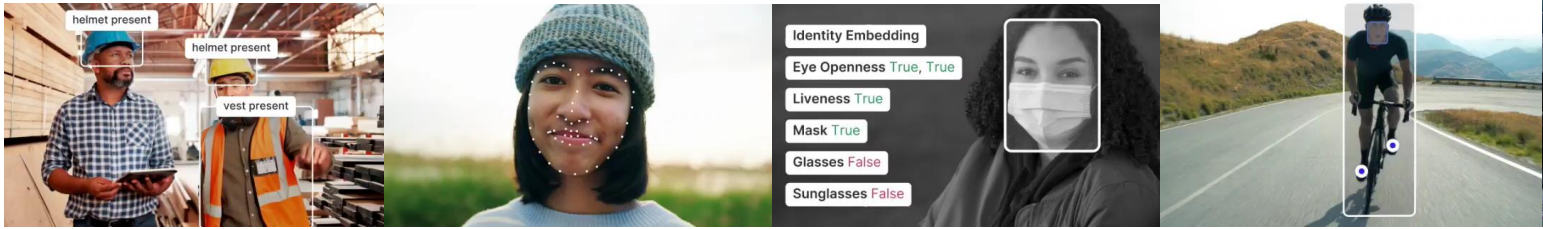
Ecosystem from Qualcomm Technologies

From OEM reference design to ecosystem expansion

New business development demands more on AI-driven multimedia technologies

- Goals for multimedia application development on the edge
  - Support ASIC design, prove SoC functionalities, enable reference design and SaaS, and grow ecosystem
- Ecosystem expansion
  - Qualcomm AI Hub: 150+ pre-optimized AI models (from both open source and internal) on devices
  - Qualcomm® Intelligent Multimedia SDK: Gstreamer framework with micro-services for typical CV applications
  - Qualcomm® Computer Vision SDK: 400+ CV functions accelerated on mobile devices
  - Snapdragon® Vision API: CV functions accelerated by HW on mobile devices
- Qualcomm AI Hub models: custom built vision solutions
  - Sample: restricted zone monitoring:
    - Detection of feet turns a seemingly complicated 3D reasoning problem into a 2D point check to tell if the zone is stepped into or simply occluded by body.





From closed system to open source

Create your solution with our ecosystem of models, cloud, runtime, and SDK partners

EMBEDDED VISION SUMMIT

Open source at Qualcomm

AI Model Efficiency Toolkit (AIMET)

A library that provides advanced model quantization techniques that enable ML models to run efficiently on neural processing units.

Qualcomm® AI Hub Apps

Sample apps and tutorials to help deploy machine learning models on Qualcomm Technologies' SoCs.

Qualcomm AI Hub Models

State-of-the-art machine learning models optimized for deployment on Qualcomm Technologies' SoCs.

Qualcomm® Cloud AI SDK

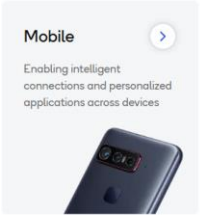
Models, tutorials, and samples for AI processing in the cloud using the Qualcomm Cloud AI 100.

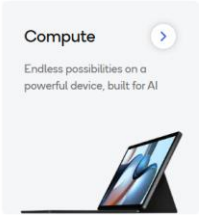
Qualcomm® Linux® Software Stack

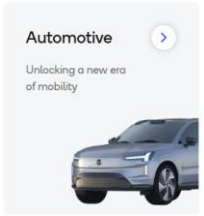
A suite of software, tools, and resources to develop IoT applications and products, with a focus on advanced, accelerated AI.

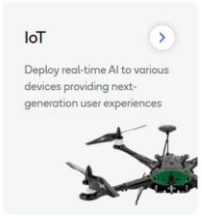
Qualcomm AI Hub simplifies deploying AI models for vision, audio, and speech applications to edge

- The key features of Qualcomm AI Hub for Mobile, Compute, Automotive and IoT platforms
  - 150+ AI models deployed on edge devices https://www.qualcomm.com/videos/qualcomm-ai-hub
  - Simple steps to deploy your own AI models on devices









Qualcomm Edge Impulse simplifies customization and training of AI models to edge

- Powering IoT developers with edge AI

From simulation to edge deployment

Sample case: Low Power Computer Vision Challenges (LPCVC) at CVPR 2025 workshop

- The LPCVC 2025 with 3 tracks
  - Image classification under various lighting conditions and formats
  - Open-vocabulary segmentation with text prompt
  - Monocular depth estimation
- Sponsored by Qualcomm® Technologies and benchmarked on Qualcomm AI Hub
  - Hundreds submissions being uploaded on Snapdragon® 8 Elite or Snapdragon® X Elite for benchmarking
  - Winning models will be published at Qualcomm AI Hub in open source
- Organize the 8th Workshop on Efficient Deep Learning for Computer Vision at CVPR 2025
  - LPCVC results will be announced in the workshop

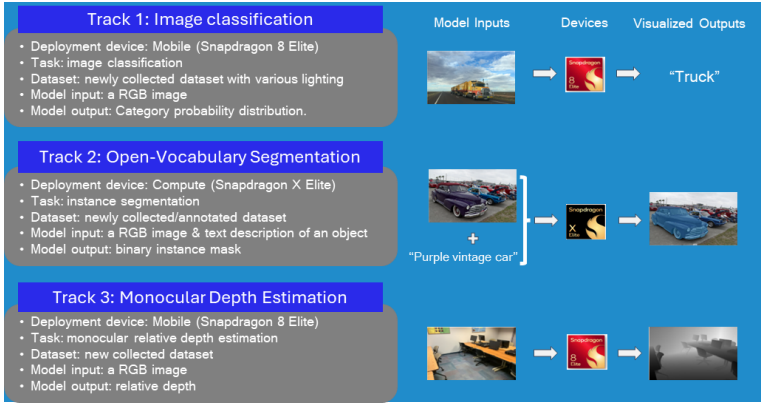
Collaborators

Purdue University, Loyola University Chicago, Lehigh University
IEEE Computer Society

Sponsor

Qualcomm

Champion	Second	Third
$1,500 + AI PC	$1,000	$500



Resources

- Qualcomm AI Hub
  - https://aihub.qualcomm.com
- Qualcomm Intelligent Multimedia Software Development Kit (IM SDK) Reference
  - https://docs.qualcomm.com/bundle/publicresource/topics/80-70018-50/overview.html
- Qualcomm Computer Vision SDK
  - https://www.qualcomm.com/developer/software/qualcomm-fastcv-sdk
- Qualcomm Engine for Visual Analytics (EVA) Simulator for Mobile
  - https://www.qualcomm.com/search#q=eva&tab=all
- Snapdragon Ride SDK
  - https://www.qualcomm.com/developer/software/snapdragon-ride-sdk?redirect=qdn
- Snapdragon Spaces SDK
  - https://www.qualcomm.com/snapdragon-spaces-sdk-download
- Qualcomm® Cloud AI SDK
  - https://www.qualcomm.com/developer/software/cloud-ai-sdk
- Qualcomm Hexagon NPU SDK
  - https://www.qualcomm.com/developer/software/hexagon-npu-sdk
- Qualcomm Neural Processing SDK | Qualcomm Developer
  - https://www.qualcomm.com/developer/software/neural-processing-sdk-for-ai
- AI Model Efficiency Toolkit (AIMET) | Qualcomm Developer
  - https://www.qualcomm.com/developer/software/ai-model-efficiency-toolkit
- Qualcomm Linux | Qualcomm
  - https://www.qualcomm.com/developer/software/qualcomm-linux

Welcome to visit

Qualcomm

booth # 409

Thank you

Nothing in these materials is an offer to sell any of the components or devices referenced herein.

References in this presentation to "Qualcomm" may mean Qualcomm Incorporated, Qualcomm Technologies, Inc., and/or other subsidiaries or business units within the Qualcomm corporate structure, as applicable. Qualcomm Incorporated includes our licensing business, QTL, and the vast majority of our patent portfolio. Qualcomm Technologies, Inc., a subsidiary of Qualcomm Incorporated, operates, along with its subsidiaries, substantially all of our engineering, research and development functions, and substantially all of our products and services businesses, including our QCT semiconductor business.

Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries.
Qualcomm patented are licensed by Qualcomm Incorporated.

Follow us on:

For more information, visit us at qualcomm.com & qualcomm.com/blog