

The logo for the 2024 Embedded VISION Summit is centered on the left side of the slide. It features a white octagonal background with a colorful, multi-layered border in shades of purple, blue, green, yellow, and orange. The text "2024" is at the top, "embedded" is below it, "VISION" is in large, bold, dark blue letters with a gradient, and "SUMMIT" is at the bottom in a smaller, dark blue font.

2024
embedded
VISION
SUMMIT®

What's Next in On-Device Generative AI

Jilei Hou

Vice President, Engineering
Qualcomm Technologies, Inc.

Today's agenda

Trends in generative AI and
why on device is key

Efficiency techniques to bring
generative AI on device

Toward agents and
embodied AI at the edge

Q&A





Transformers
are key and
extending to
more modalities

Multi-camera and LIDAR aligned for bird's-eye-view

Enable enhanced perception of the world for autonomous vehicles, robots, and more using cross-view attention

Wireless multimodal fusion in deepSense 6G

Understand environments better by combining GPS, camera, and mmWave RF using transformers to improve mmWave beam management

Robotics with GATr

Enable robots to efficiently learn complex dexterous skills in 3D spaces from cameras through use of geometric algebra transformers (GATr)

MODALITY AND USE CASE

Voice UI

Voice is a natural and intuitive interface for conversation

Large multimodal models

Utilizing more sensing input modalities to better understand the world

Video & 3D

Generating content for a richer and more realistic experience



Generative AI capabilities continue to increase

Agents

Execute multi-step tasks with reasoning autonomously to achieve a goal

CAPABILITY AND KPI

Longer context window

Allows in-depth conversations

Personalization

Fine-tuned models customized to consumers, enterprises, or industries (e.g., LoRA)

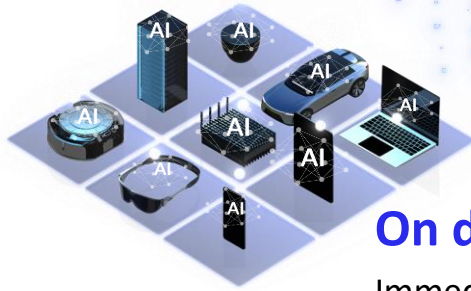
Higher resolution

Process higher fidelity images for better accuracy

To scale, the center of gravity of AI processing is moving to the edge

Hybrid AI

Distribute workloads among cloud and edge/devices to deliver more powerful, efficient, and highly optimized experiences



Central cloud

Ease of development & deployment
Training | Very large models
Aggregation | Absolute performance



Edge cloud (on-prem or nearby)

Immediacy | Reliability | Personalization | Privacy | Security
Fine-tuning | Aggregation

On device

Immediacy | Reliability | Personalization | Privacy | Security
Cost | Energy



Advancements in edge platforms for generative AI and transformers

Multiple axes to optimize AI models
and efficiently run them on hardware



Distillation

Learning weights for a smaller student model,
which mimic a larger teacher model



Quantization & compression

Learning to reduce bit-precision while keeping
desired accuracy



Speculative decoding

Utilizing a large model in concert with
a draft model for a faster token rate



Efficient image & video architecture

Designing smaller neural networks that are
on par or outperform original architecture



Heterogeneous computing

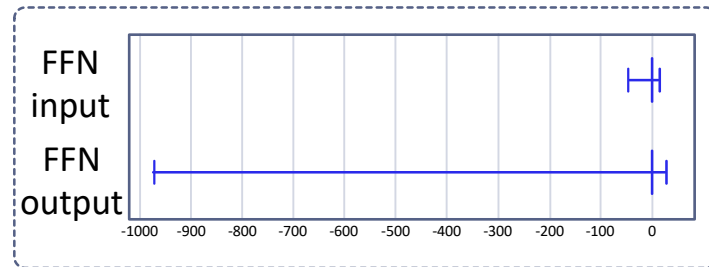
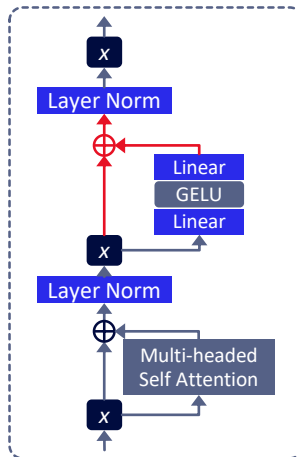
Utilizing the best processor for diverse
AI workloads to improve efficiency

Improving transformer quantization accuracy by reducing outliers

Many modern transformers learn big activation outliers, making them difficult to quantize

This holds for many tasks, training objectives, and models (language encoders/ decoders and vision transformers)

Goal: Address the root cause of the issue and propose a new pre-training protocol to dampen the outliers



How to set quantization grid for residual sum?



Helping attention heads do nothing¹

Strong outliers are related to behavior of attention heads trying to learn “no-op” or a partial update of the residual

To achieve exact zeros in the attention matrix for a no-op, the input to softmax is pushed to be larger and larger during training, causing outliers

Our pretraining methods significantly reduce outliers and improve post-training quantization (PTQ) accuracy

Our proposed methods (training from scratch) applied to **BERT-base**, **OPT-125m** and **ViT-S/16**

Two independent modifications to the attention mechanism allow representing exact zeros (and ones)

1. Clipped softmax

2. Gated attention

Easy to integrate into any transformer model with softmax attention

Model	Method	FP16/32	Max inf. norm	Avg. kurtosis	W8A8
BERT (ppl.↓)	Vanilla	4.49 ±0.01	735 ±55	3076 ±262	1249 ±1046
	Clipped softmax	4.39 ±0.00	21.5 ±1.5	80 ±6	4.52 ±0.01
	Gated attention	4.45 ±0.03	39.2 ±26.0	201 ±181	4.65 ±0.04
OPT (ppl.↓)	Vanilla	15.84 ±0.05	340 ±47	1778 ±444	21.18 ±1.89
	Clipped softmax	16.29 ±0.07	63.2 ±8.8	19728 ±7480	37.20 ±2.40
	Gated attention	15.55 ±0.05	8.7 ±0.6	18.9 ±0.9	16.02 ±0.07
ViT (acc.↑)	Vanilla	80.75 ±0.10	359 ±81	1018 ±471	69.24 ±6.93
	Clipped softmax	80.89 ±0.13	73.7 ±14.9	22.9 ±1.6	79.77 ±0.25
	Gated attention	81.01 ±0.06	79.8 ±0.5	19.9 ±0.3	79.82 ±0.11

Clipped softmax and gated attention are our techniques.
ppl. = perplexity;
acc. = accuracy

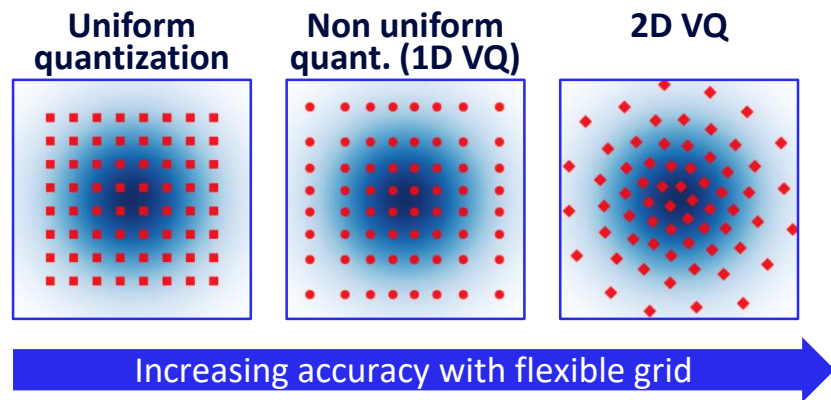
On par or slightly better floating-point performance

Significantly reduced both outlier magnitude and kurtosis

Significantly better PTQ INT8 performance

Vector quantization (VQ) shrinks models while maintaining desired accuracy

Employing non-linear quantization and expanding the dimensionality of the representational grid through VQ



1D quantization requires that each dimension is quantized separately, resulting in a grid.

VQ allows for an arbitrary region of quantization points in a 2D space.

VQ can improve footprint and latency for memory-bound generative AI like LLMs

Setting	BPV ↓	Relative ↓ footprint	Relative ↓ latency
INT4	4	1.00x	1.00x
INT8	8	2.00x	1.93x
2D 2.5B @b512	3	0.75x	0.98x
2D 2.5B @b2048	2.25	0.56x	0.96x
2D 2B @b1024	2.25	0.56x	0.87x
Llamav2-7B 1D 3B @b128	3.5	0.88x	0.96x

"GPTVQ: The Blessing of Dimensionality for LLM Quantization", van Baalen et al., ICML 2024, <https://arxiv.org/abs/2402.15319v1>. VQ feature is coming to AI Model Efficiency Toolkit (AIMET). AIMET is a product of Qualcomm Innovation Center, Inc.

Speculative decoding

Speeds up token rate by trading off compute for bandwidth

- Token generated from draft
- Token checked & accepted by target

Recite the first law of robotics A robot may

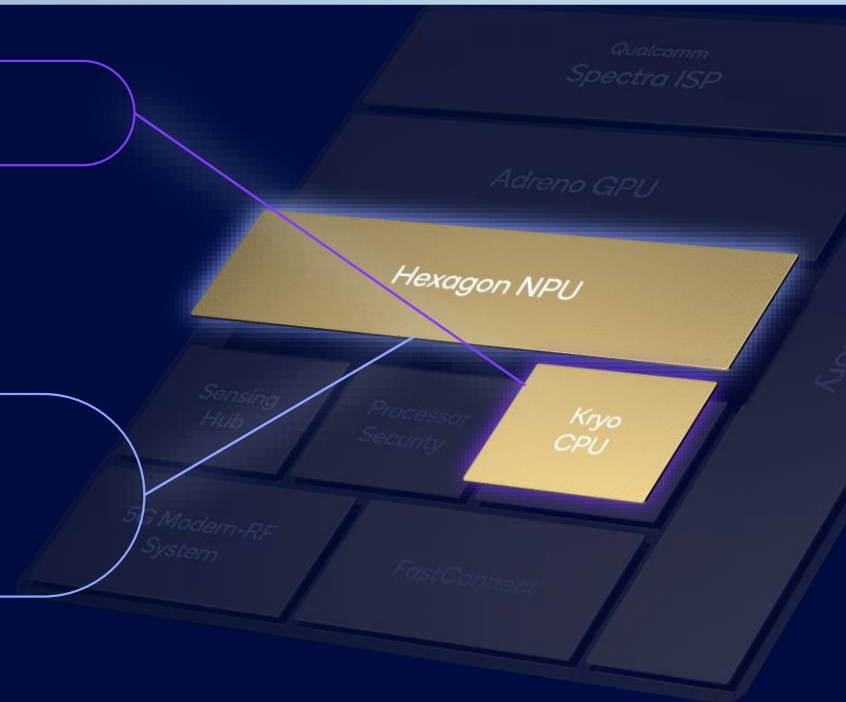
Llama 2 draft

A robot should not

Recite the first law of robotics A robot may not

Llama 2

A robot may not harm



Draft model generates a few speculative tokens at a time

Target model decides which to accept in one pass

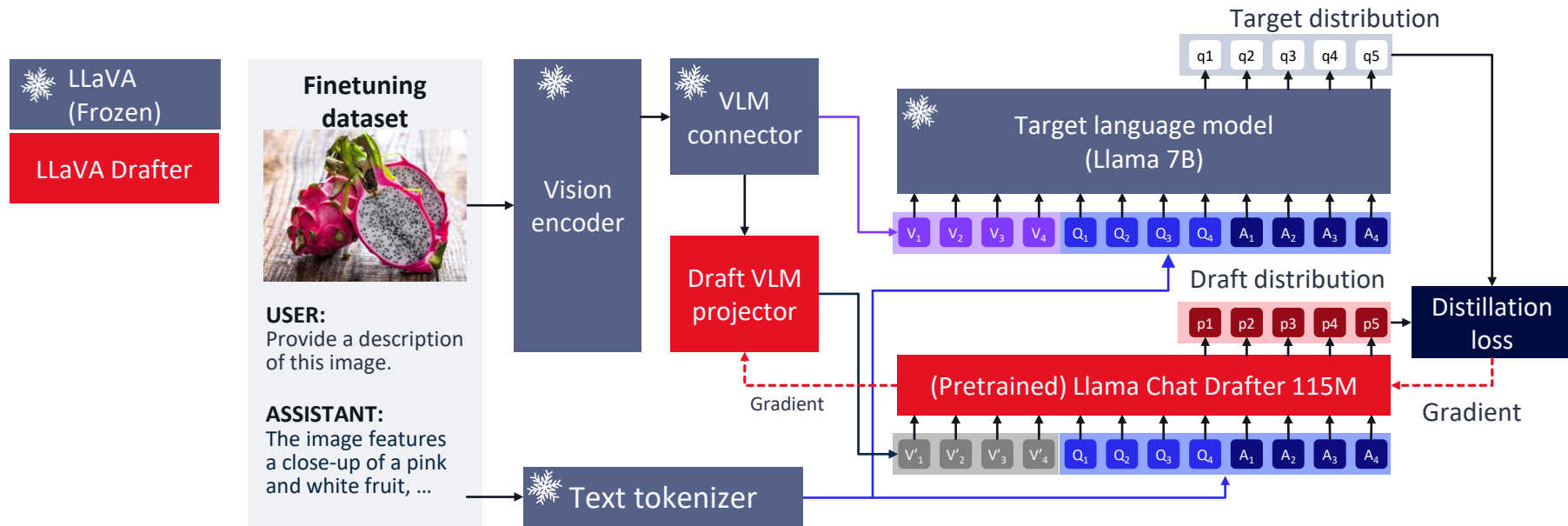
A good draft model predicts with a high acceptance rate

Training the draft model for multimodal LLM speculative decoding

LLaVA as an example of an LMM with vision

Llama Chat Drafter 115M is fine-tuned on LLaVA finetuning dataset using TVD++ distillation loss¹

Additional Draft VLM projector layer is trained for image-language interface in draft model

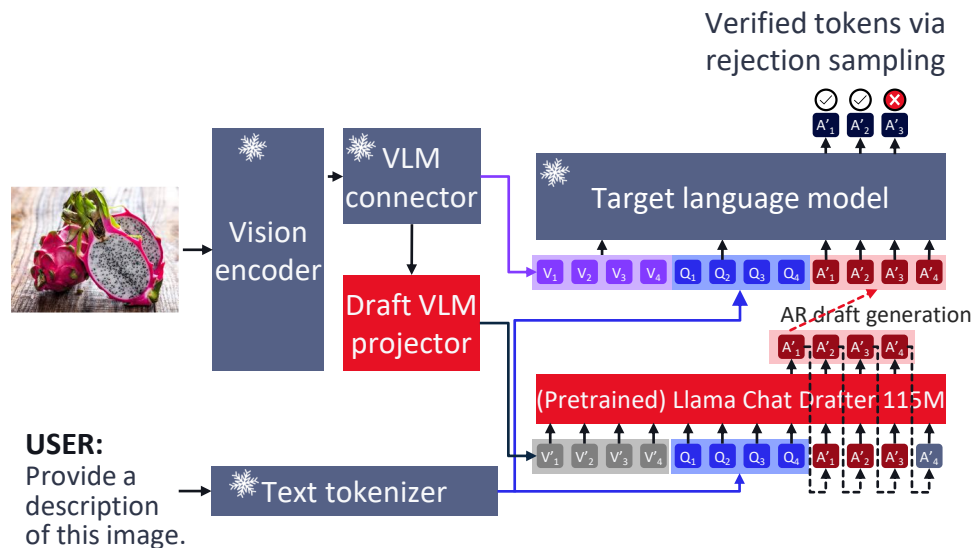


LLaVA draft model finetuning

1: Goel, Raghav, et al. "Direct Alignment of Draft Model for Speculative Decoding with Chat-Fine-Tuned LLMs." (<https://arxiv.org/abs/2403.00858v3>, 2024); LLaVA: large language and vision assistant; VLM: vision language model

Applying speculative decoding to LMMs with vision

Achieving higher token rates for LMMs without sacrificing accuracy



LLaVA: large language and vision assistant; MBSU: memory-bound speed-up; SD: speculative decoding

Generation example (accepted drafts underlined)

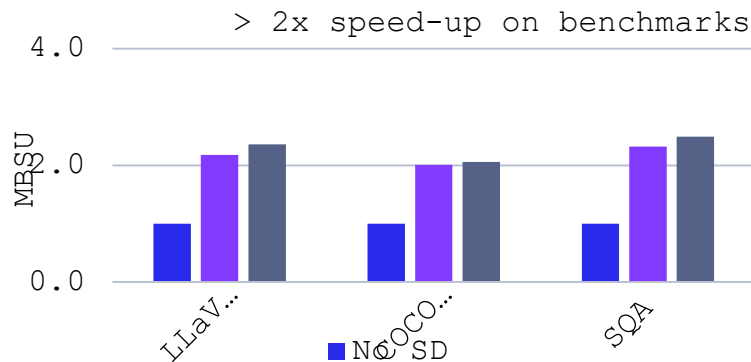


Input prompt:

A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions.

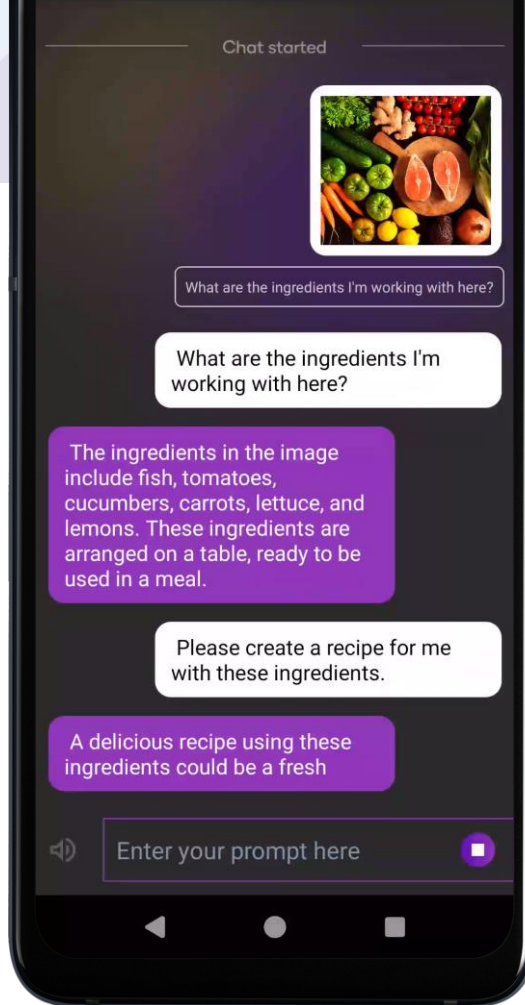
USER: <image> Provide a detailed description of the given image.

Output generation: The image features a close-up of a pink and white fruit, possibly a dragon fruit, with its skin peeled back to reveal the white and black seeds inside. The fruit is placed on a wooden table, and the close-up view highlights the intricate details of the seeds and the fruit's skin. The dragon fruit is the main focus of the image, and it appears to be the only fruit in the scene.



At
MWC
2024

WORLD'S FIRST large multimodal model (LMM) on an Android phone



LLMs can now see

7+ billion parameter LMM, LLaVA,
with text, speech, and image inputs

Multi-turn intuitive conversations
about
an image at a responsive token rate

Full-stack AI optimization to achieve
high performance at low power

Enhanced privacy, reliability,
personalization, and cost with
on-device processing

The potential of generative video editing

Given an input video and a text prompt describing the edit, generate a new video

Input video



Edited video



Prompt: “pink flamingo walking”

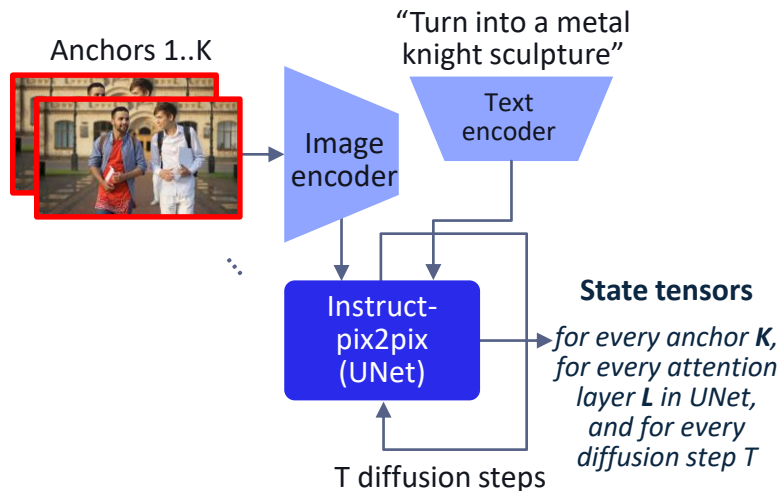
Key challenges:

1. Temporal consistency
2. High computational cost

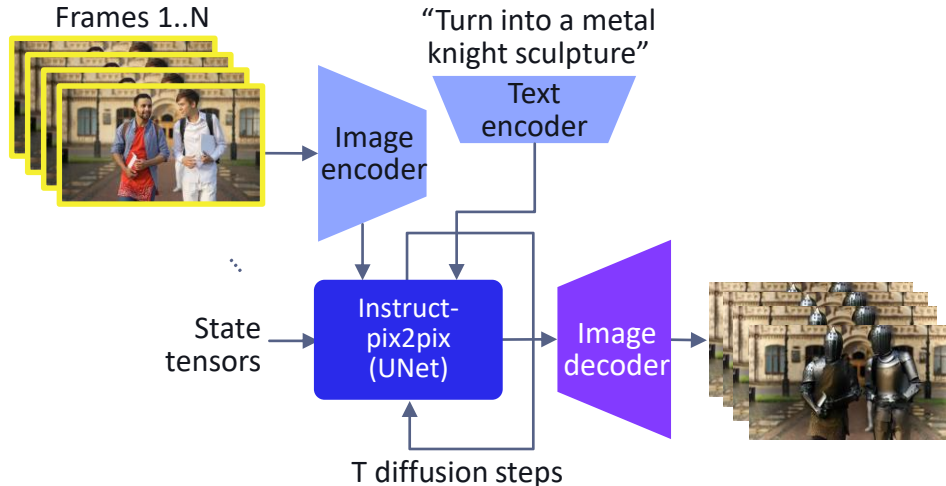
Making generative video methods efficient for on-device AI

Optimizations to FAIRY¹, a video-to-video generative AI model

Stage 1: Extract states from anchor frames



Stage 2: Edit video across remaining frames



Steps to enable
on device

Cross-frame
optimization

Efficient
instructPix2Pix

Image/text
guidance conditioning

1: “FAIRY: Fast Parallelized Instruction-Guided Video-to-Video Synthesis” (<https://arxiv.org/abs/2312.13834>)

Original video



Turn into a marble roman sculpture



Turn into low poly art



Turn into a metal knight sculpture



Change the style to cartoon



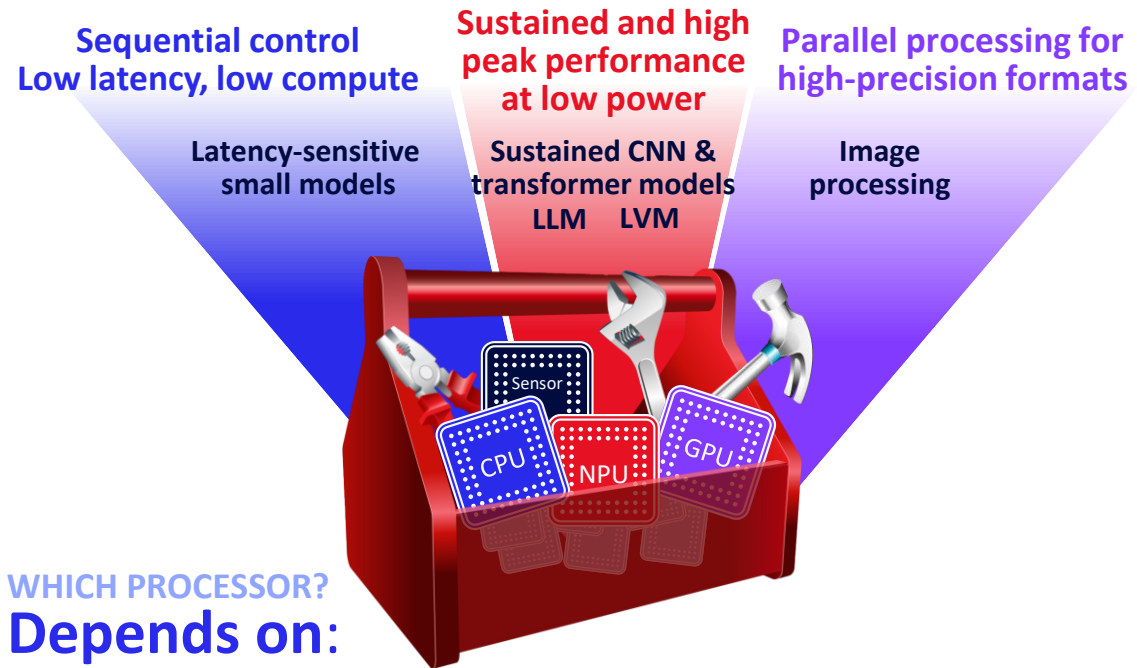
In cubism style



Diverse processors are essential for maximizing performance and power efficiency in generative AI applications

Generative AI use cases across verticals have diverse requirements and computational demands

- On-demand
- Sustained
- Pervasive



LLM: large language model
LVM: language vision model

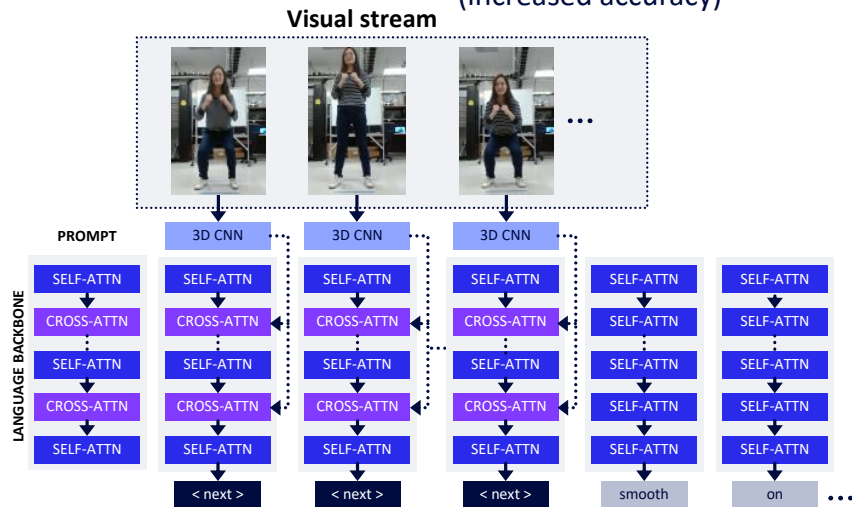
Our situated vision-language model for fitness coaching

Key innovations

Leading results



- A 3D CNN-based vision backbone for processing the vision stream
- A pretrained Llama2-7B language model backbone to generate interactive feedbacks
- A cross-attention-based adapter deeply fusing the two
- End-to-end training for situated visual understanding
- Processing the vision stream (dynamic vs static)
- Introducing action tokens (when/what to say)
- Pre-training the vision backbone (increased accuracy)



Question: Provide an appropriate feedback for the user

Video-LLaMA: We see a young man standing in a kitchen, wearing a red shirt and white shorts. ❌

Video-ChatGPT: The user has successfully demonstrated the ability to perform a balancing act on a pair of blue pools. ❌

Coach-LLaMA: This is awesome. Let's keep the intensity high! ✅

Method	T-F-Score ↑	T-BERT ↑	T-Rouge-L ↑	Mixtral-Score ↑
Video-LLaMA	0.57	0.436	0.029	2.39
Video-ChatGPT	0.57	0.439	0.033	2.72
Coach-Llama (ours)	0.64	0.512	0.115	3.10

What to Say and When to Say it: Video-Language Model and Benchmark for Situated Interactions (2024)

Aimed at the development of interactive multi-modal vision-language models based in the controlled but challenging fitness coaching domain

FIT-Coach benchmark and dataset

A novel interactive visual coaching benchmark and dataset as a test-bed for real-time, real-world situated interaction



Fitness questions dataset

148 exercises	300k short-clip videos	470+ hours
1900 unique participants	1.1M+ high-level question-answer pairs	400k+ fine-grained question- answer pairs

Fitness feedback dataset

9+ hours of fitness coaching session	148 exercises	~3.5 minutes long sessions with 5 to 6 exercises	21 unique participants
---------------------------------------------------------	-------------------------	---------------------------------------------------------------------	-------------------------------------

The path to humanoid robots

We need to take advantage of end-to-end learning

Dexterous
manipulation and
domain transfer

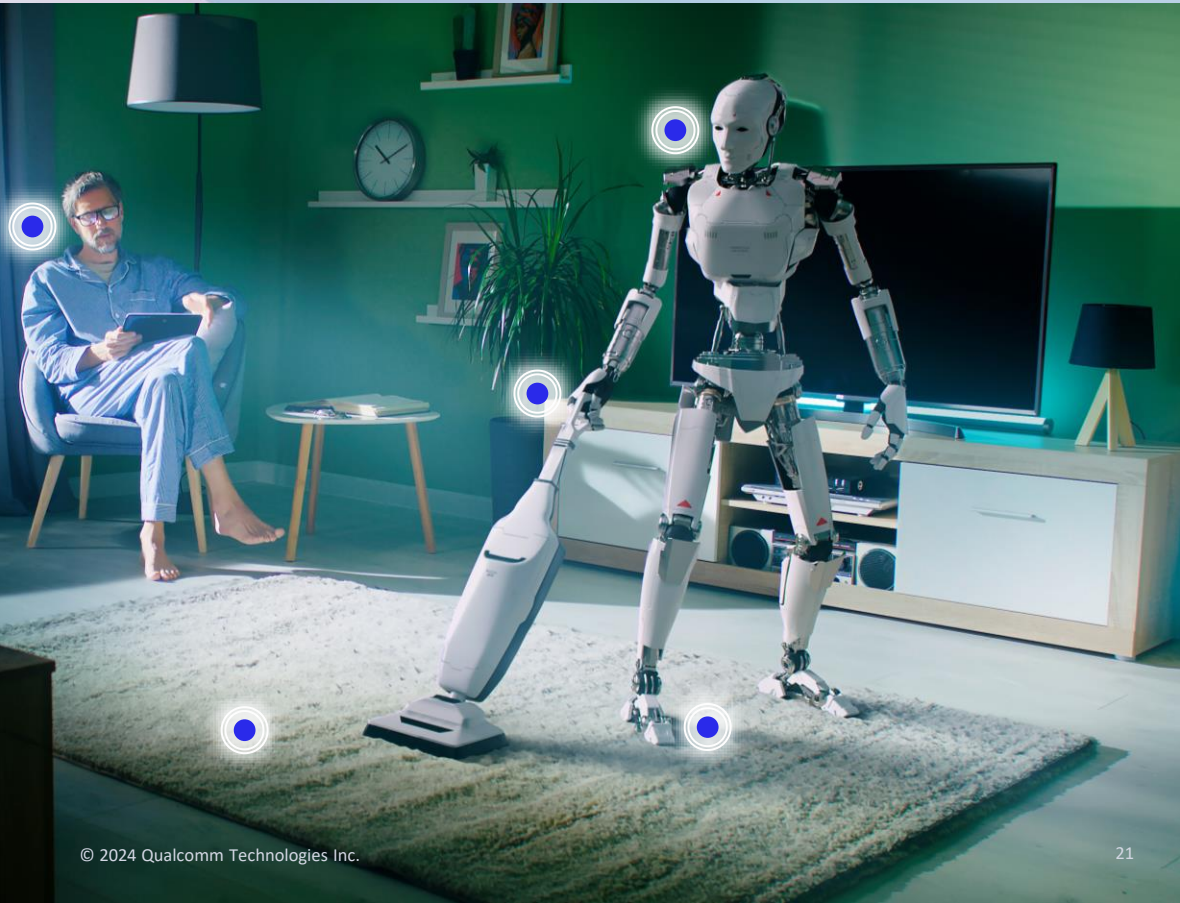
Challenging for
current end-to-end
learning

Situated
understanding
of scenes in live
video streams

Required but
previously ignored
Now, significant
progress

Recognition of
objects in images,
low-level control

Mature solutions
in place



Generative AI capabilities are evolving
and more beneficial on the edge

Advancements in architectures, algorithms,
and heterogeneous computing are enabling
generative AI on the edge

Generative AI agents and systems allow
developers to significantly enhance
applications and enable embodied AI

Qualcomm AI Hub

<https://aihub.qualcomm.com/>



2024 Embedded Vision Summit

May 21st (1:00-4:00pm)

“Accelerating Model Deployment with Qualcomm® AI Hub” – Bhushan Sonawane

May 22nd (1:30-2:00pm)

“OpenCV for High-Performance, Low-Power Vision Applications on Snapdragon” – Xin Zhong

May 23rd (9:50-10:20am)

“What’s Next in On-Device Generative AI” – Jilei Hou

May 23rd (10:20-11:10am)

“Multimodal LLMs at the Edge: Are We There Yet?” – Jilei Hou (Panel session)

May 23rd (1:30-2:00pm)

“Deploying large models on the edge : Success Stories & Challenges” – Vinesh Sukumar

Thank you

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

© Qualcomm Technologies, Inc. and/or its affiliated companies. All Rights Reserved.

Qualcomm is a trademark or registered trademark of Qualcomm Incorporated. Other products and brand names may be trademarks or registered trademarks of their respective owners.

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 technologies are licensed by Qualcomm Incorporated.

Follow us on: [in](#) [X](#) [@](#) [▶](#) [f](#)

For more information, visit us at [qualcomm.com](https://www.qualcomm.com) & [qualcomm.com/blog](https://www.qualcomm.com/blog)

Questions



www.qualcomm.com/research/artificial-intelligence



www.qualcomm.com/news/eng

Connect with us



www.youtube.com/c/QualcommResearch



[@QCOMResearch](https://twitter.com/QCOMResearch)



[https://assets.qualcomm.com/
mobile-computing-newsletter-sign-up.html](https://assets.qualcomm.com/mobile-computing-newsletter-sign-up.html)



[www.slideshare.net/
qualcommwirelessevolution](http://www.slideshare.net/qualcommwirelessevolution)