Qualcomm

What's next in on-device generative AI

Joseph Soriaga

Sr. Director, Technology
Qualcomm Technologies, Inc.

May 30, 2024

Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries.

Today's agenda

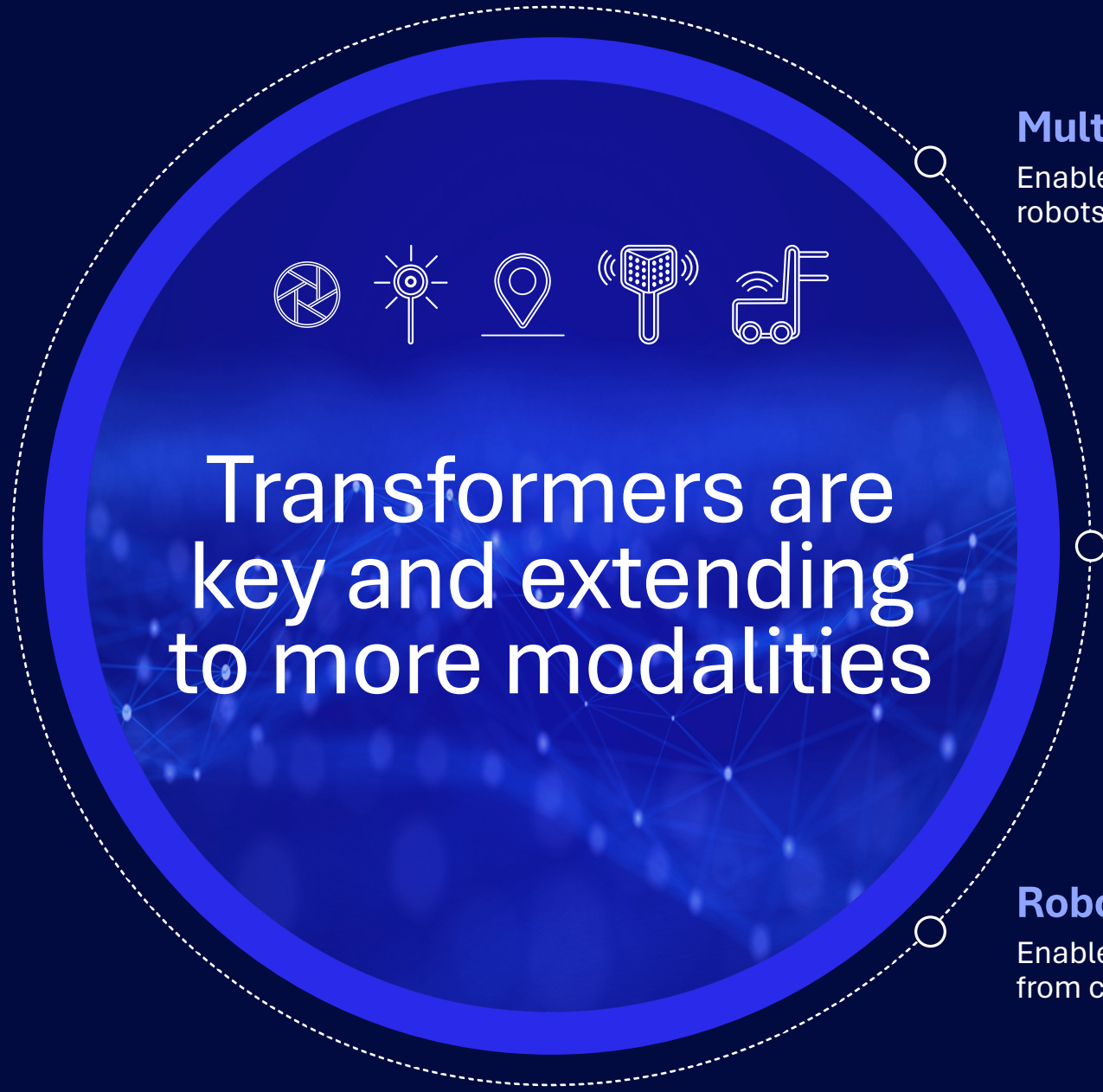
Trends in generative AI and why on device is key

Efficiency techniques to bring generative AI on device

Adaptation and personalization techniques

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

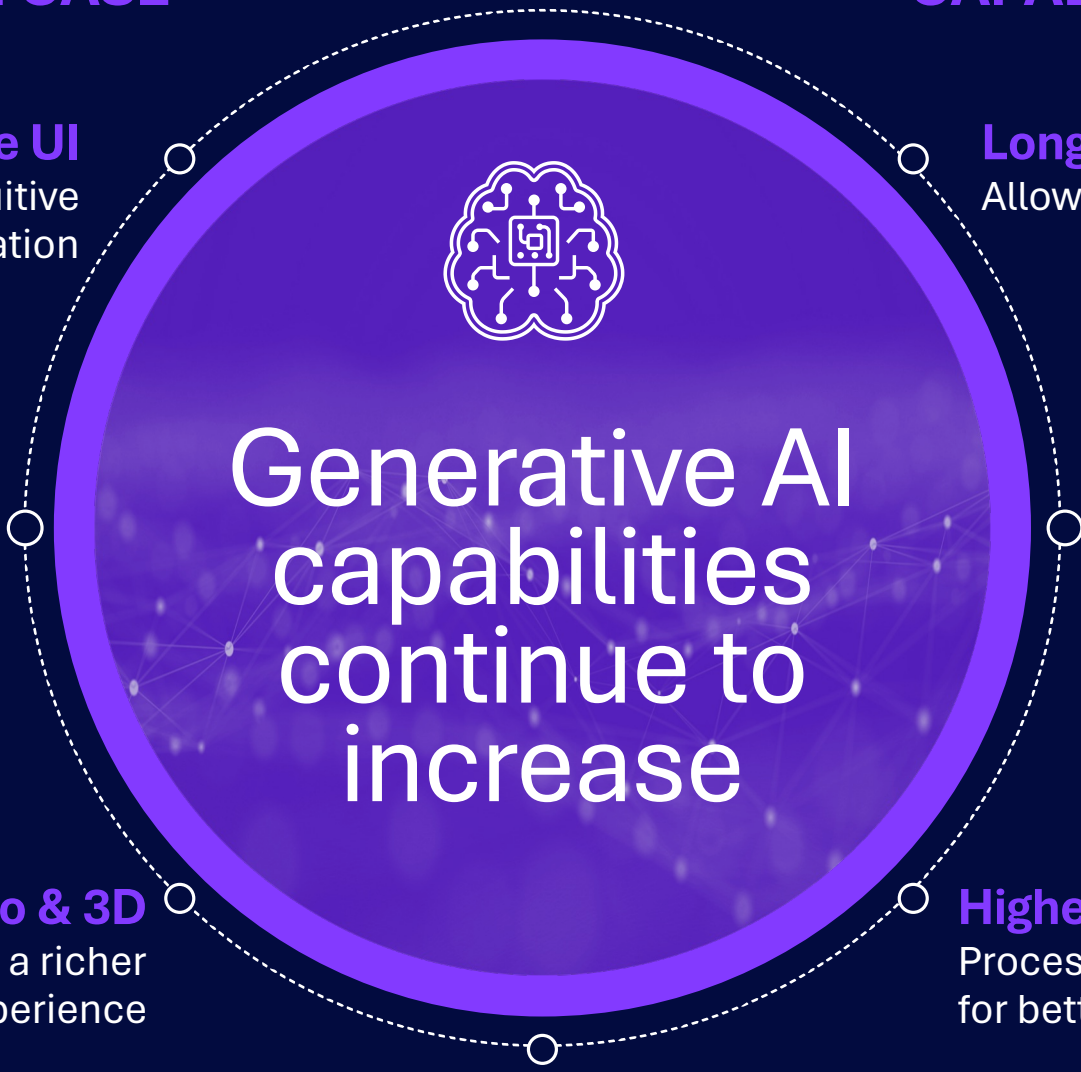
Understand environments better by combining GPS, camera, and mmWave RF using transformers to improve mmWave beam management

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

CAPABILITY AND KPI



Voice UI
Voice is a natural and intuitive interface for conversation

Longer context window
Allows in-depth conversations

Large multimodal models
Utilizing more sensing input modalities to better understand the world

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

Video & 3D
Generating content for a richer and more realistic experience

Higher resolution
Process higher fidelity images for better accuracy

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

LoRA: low-rank adaptation

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

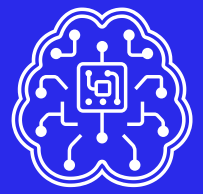
Fine-tuning | Aggregation

On device

Immediacy | Reliability | Personalization | Privacy | Security

Cost | Energy

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

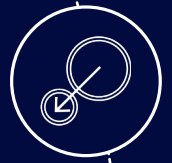


Advancements in edge platforms for generative AI and transformers

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

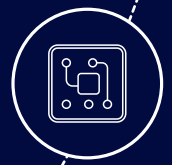
Learning to reduce bit-precision while keeping desired accuracy

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

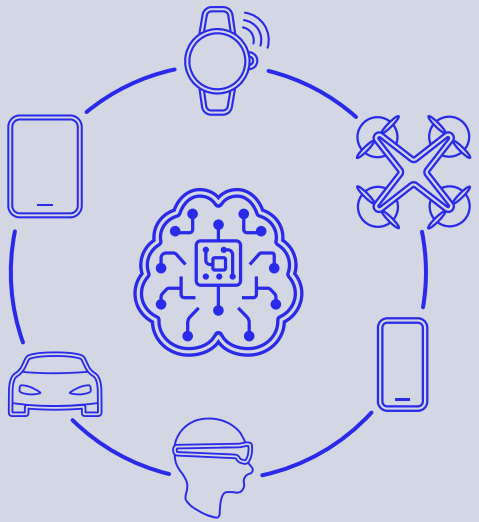


Heterogeneous computing

Utilizing the best processor for diverse AI workloads to improve efficiency

LLM quantization motivations

- A 4x smaller model (i.e., FP16 -> INT4)
- Reduce memory bandwidth and storage
- Reduce latency
- Reduce power consumption



Shrinking an LLM for increased performance while maintaining accuracy is challenging

LLM quantization challenges

- Maintain accuracy of FP published models
- Post-training quantization (PTQ) may not be accurate enough for 4-bit
- The training pipeline (e.g., data or rewards) is not available for quantization aware training (QAT)

LLM: large language model

Knowledge distillation

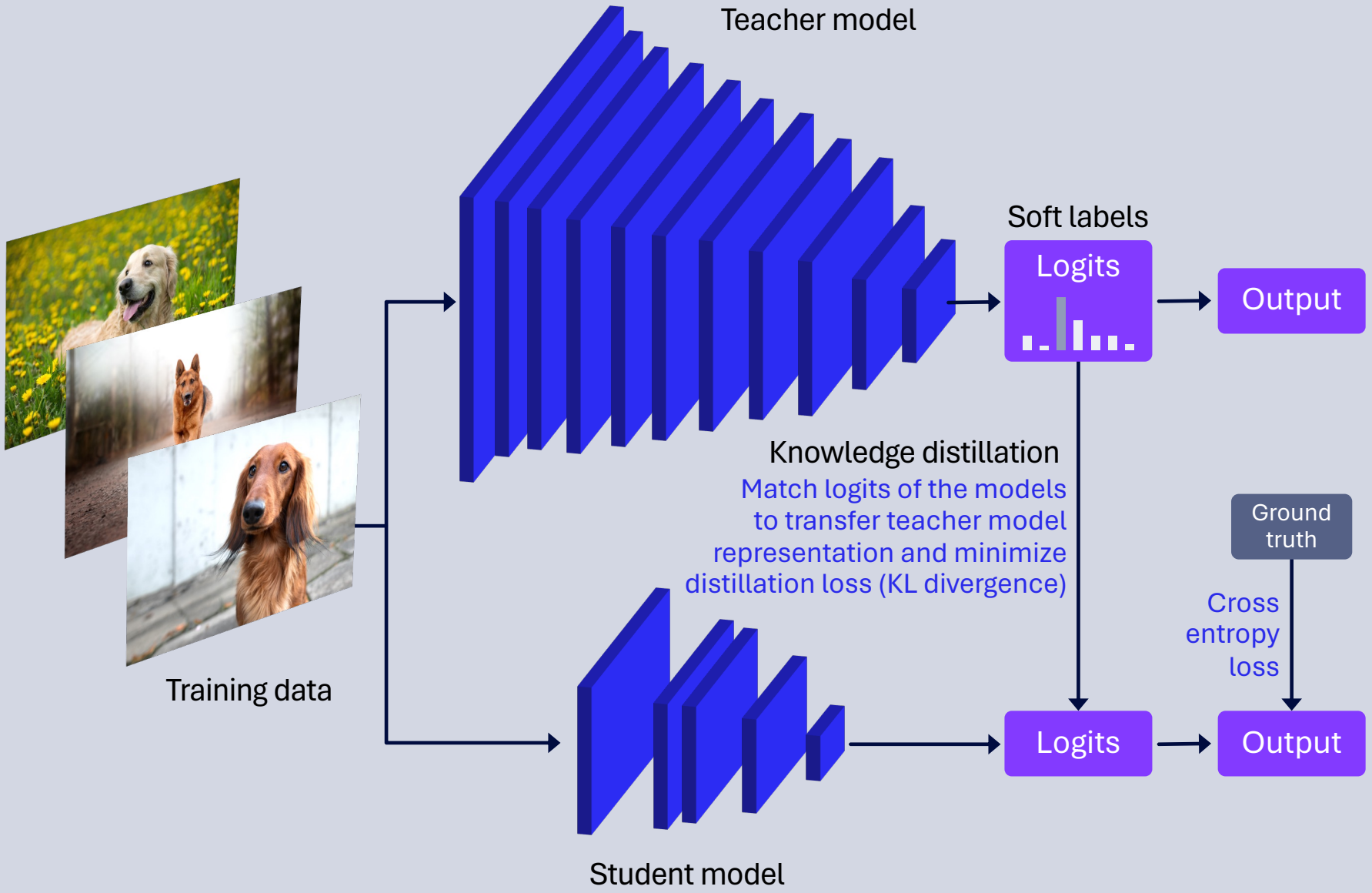
Training a smaller “student” model to mimic a larger “teacher” model

Create a smaller model with fewer parameters

Run faster inference on target deployment

Maintain prediction quality close to the teacher

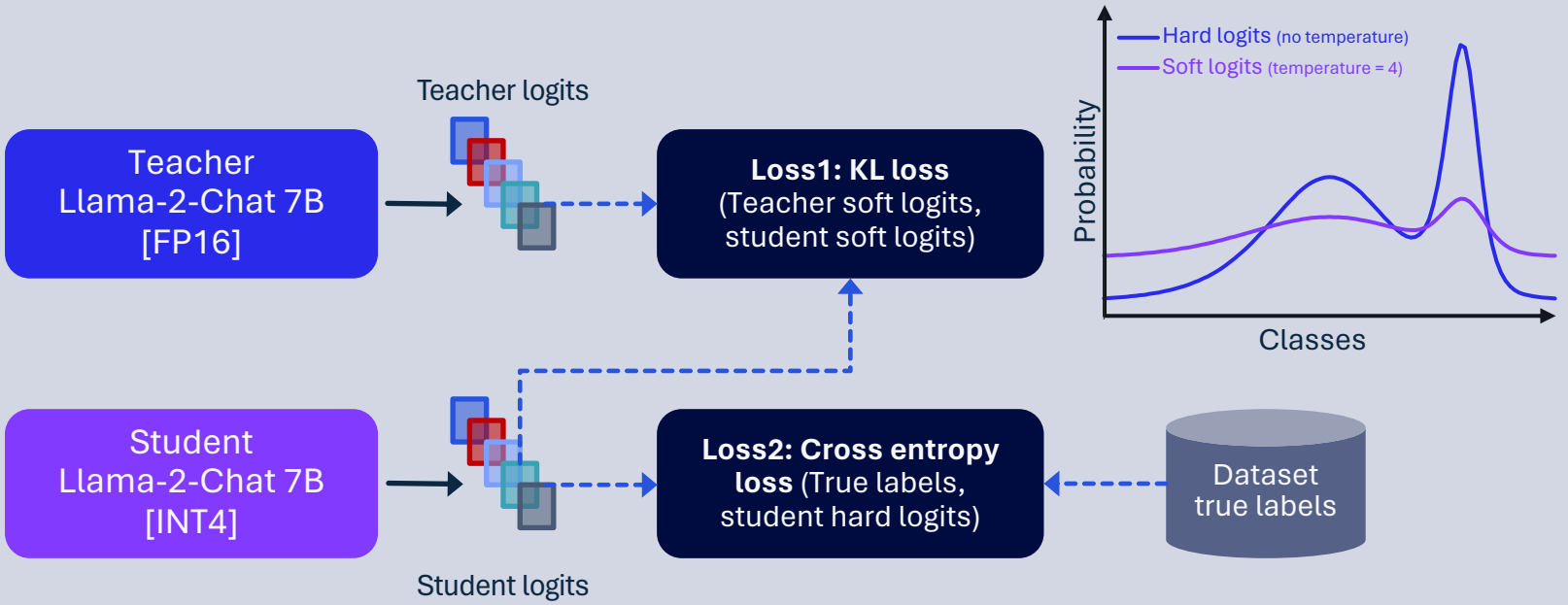
Less training time



Quantization-aware training with knowledge distillation (KD)

Reduces memory footprint while solving quantization challenges of maintaining model accuracy and the lack of original training pipeline

Construct a training loop that can run two models on the same input data



KD loss function combines the KL divergence loss and hard-label based CE loss

<1
Point increase in perplexity[1]

<1%
Decrease in accuracy

4X
Decrease in model size

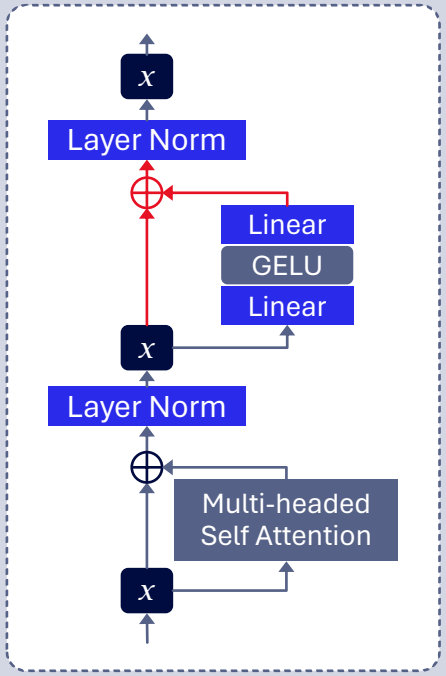
1: Perplexity is average over several test sets, including wikitext and c4 (subset); KL: Kullback–Leibler

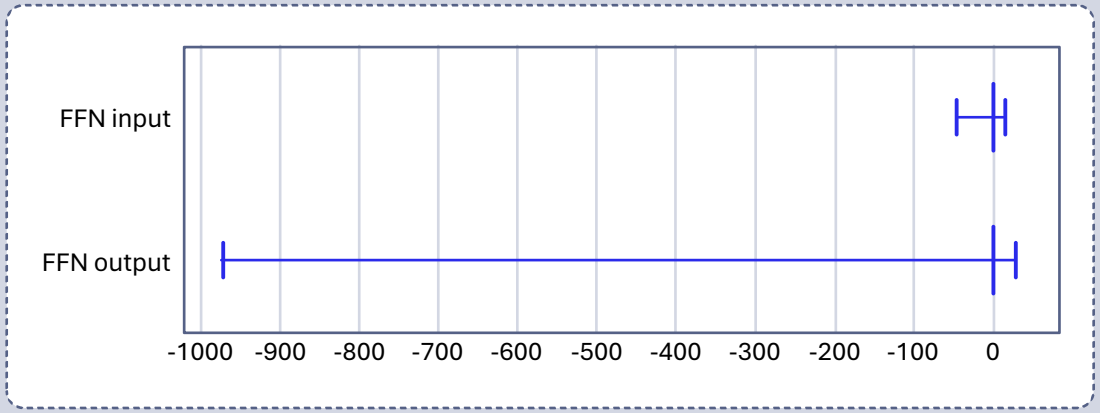
Improving transformer quantization accuracy by reducing outliers

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?

High clipping error →

High rounding error ↑

Helping attention heads do nothing[1]

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

FFN: feed forward network; 1: Quantizable Transformers: Removing Outliers by Helping Attention Heads Do Nothing, NeurIPS 2023, https://export.arxiv.org/abs/2306.12929

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

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

Quantizable Transformers: Removing Outliers by Helping Attention Heads Do Nothing, NeurIPS 2023, https://export.arxiv.org/abs/2306.12929

Our proposed methods (training from scratch)

applied to **BERT-base**, **OPT-125m** and **ViT-S/16**

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

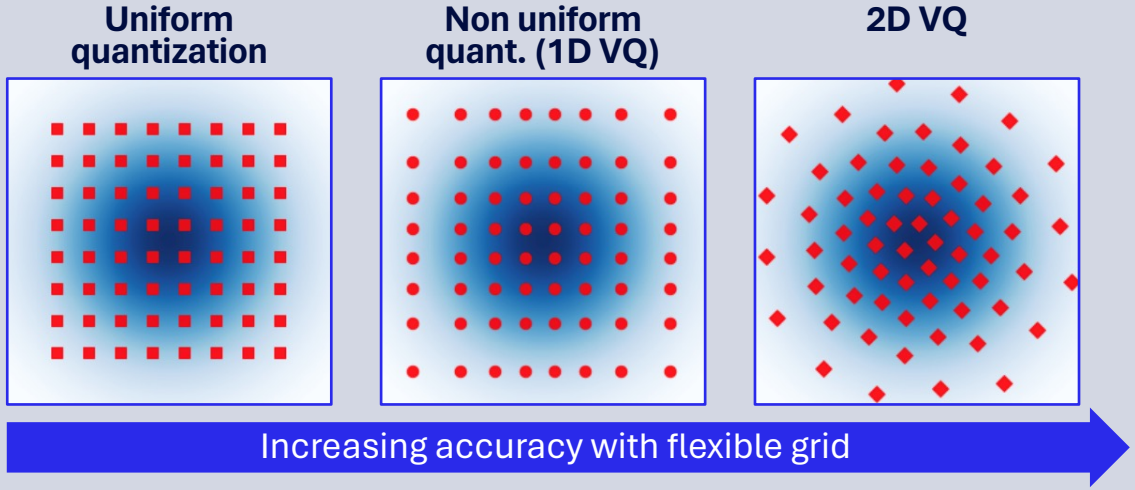
On par or slightly better floating-point performance

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.

Our GPTVQ[1] method achieves state-of-the-art quantization results

Extending GPTQ with vector quantization leads to improved size versus accuracy trade-off

		WikiText2 perplexity ↓			Zeroshot avg accuracy ↑		
		Llama 2-7B	Mistral-7B	Mixtral-MoE 8x7B	Llama 2-7B	Mistral-7B	Mixtral-MoE 8x7B
FP16		5.47	5.25	3.84	70.47	75.69	75.93
2.25 bpv (W2@b64)	SOTA methods	9.62	14.24	10.07	47.51	51.76	48.78
	GPTVQ 1D (ours)	10.08	**9.56**	**8.06**	**51.95**	**55.82**	**57.12**
	GPTVQ 2D (ours)	**7.97**	**10.11**	**6.23**	**59.08**	**56.14**	**63.92**
	GPTVQ 4D (ours)	**7.22**	**7.16**	**5.55**	**61.49**	**64.44**	**66.43**
3.125 bpv (W3@b128)	SOTA methods	6.03	5.83	4.71	66.16	72.24	72.73
	GPTVQ 1D (ours)	**5.98**	**5.76**	**4.59**	**67.61**	71.56	**72.75**
	GPTVQ 2D (ours)	**5.82**	**5.51**	**4.30**	**67.88**	**73.56**	**74.36**

Our GPTVQ method achieves 3.125 bits per value at similar accuracy as INT4 uniform quantization

1. "GPTVQ: The Blessing of Dimensionality for LLM Quantization", van Baalen et al., ICML 2024, https://arxiv.org/abs/2402.15319v1.
SOTA methods: State-of-the-art methods include the best performing of RTN, GPTQ, AWQ, and OmniQuant.
Feature 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

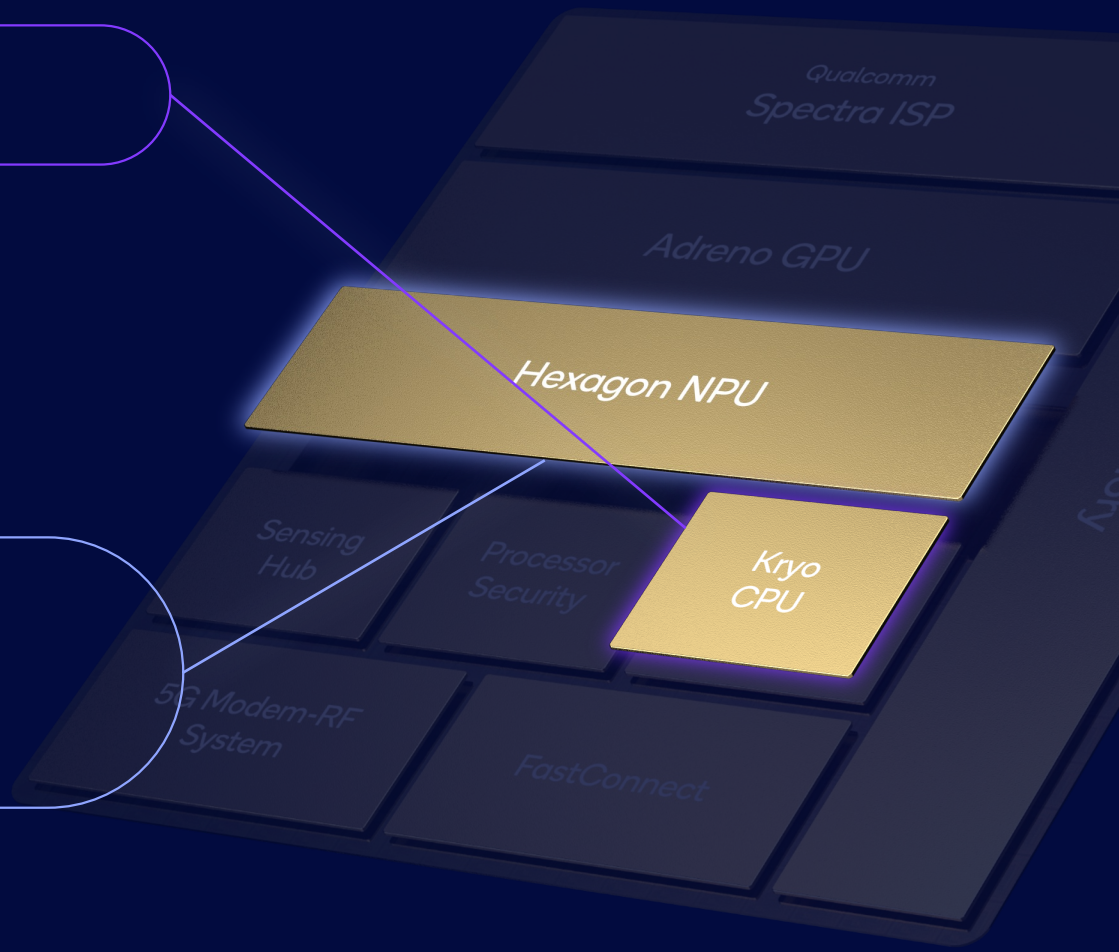
Llama 2 draft

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

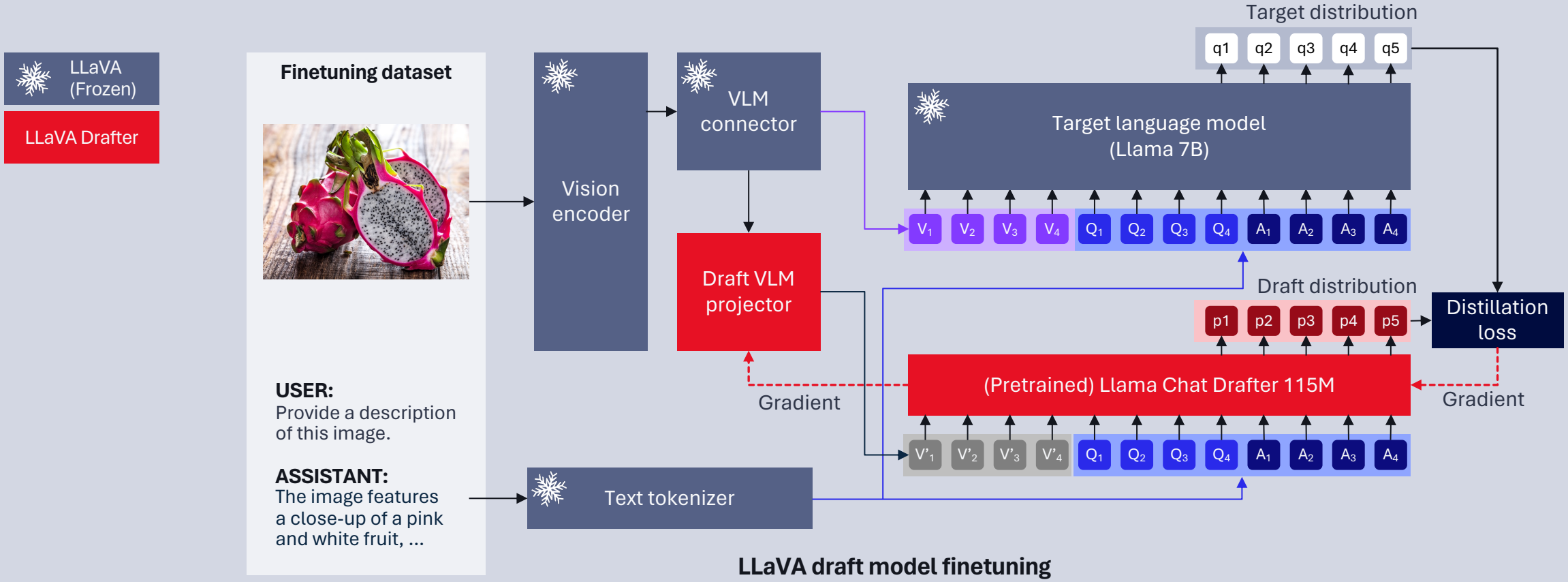
A good draft model predicts with a high acceptance rate

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

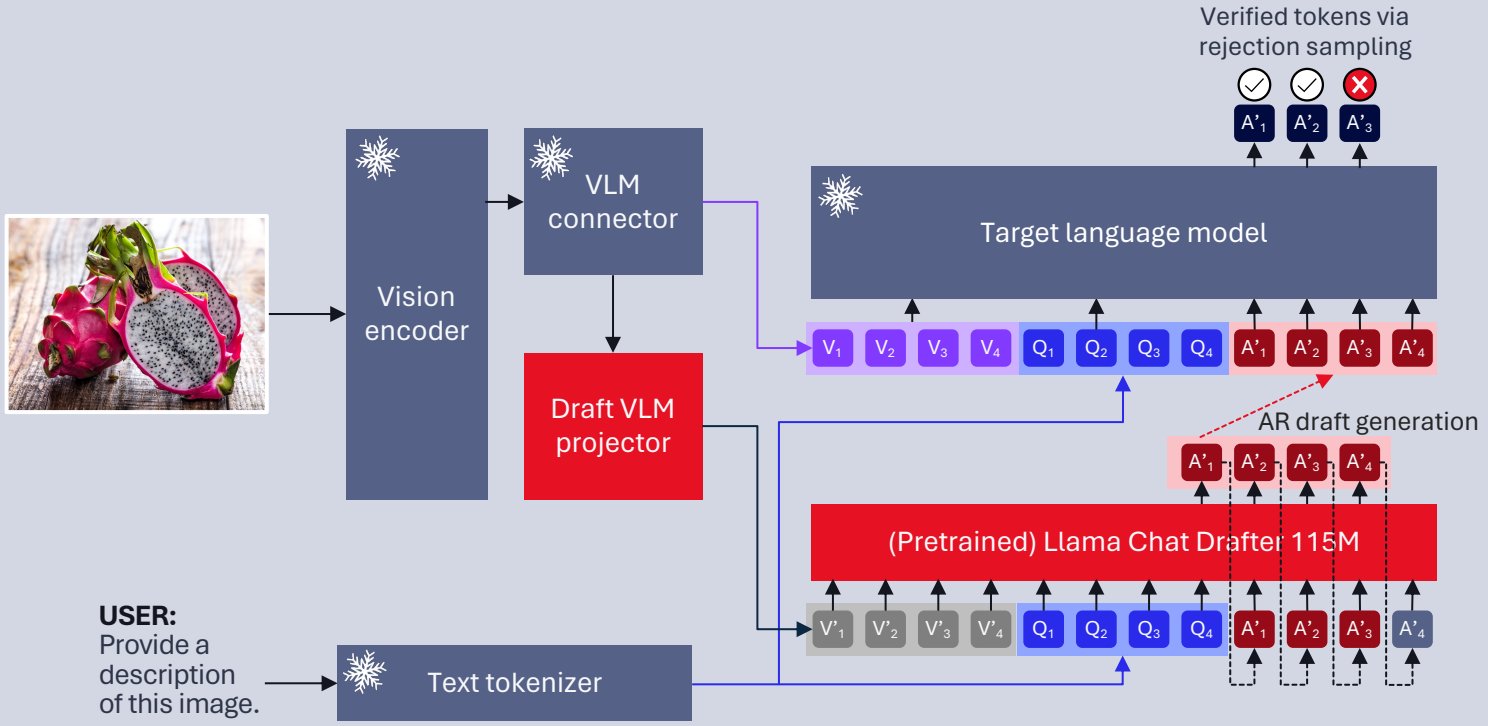


LLaVA draft model finetuning

1: Goel, Raghavv, 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

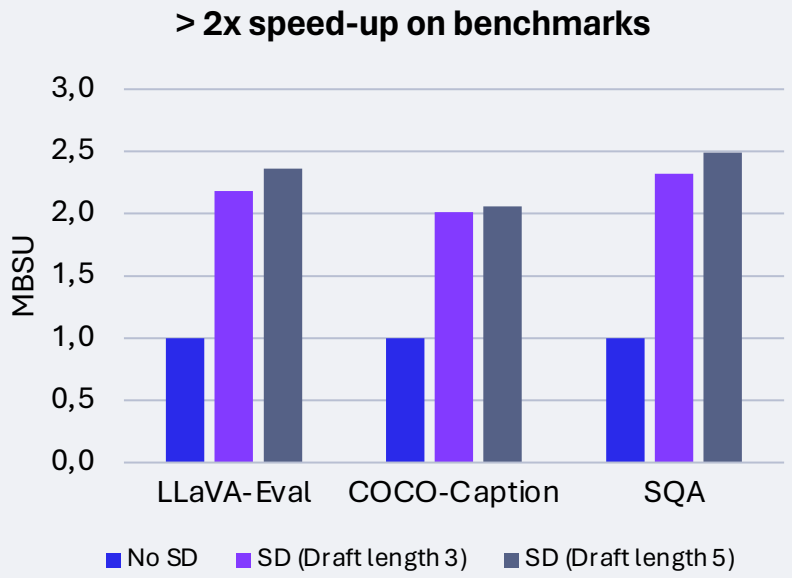
Achieving higher token rates for LMMs without sacrificing accuracy



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.



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

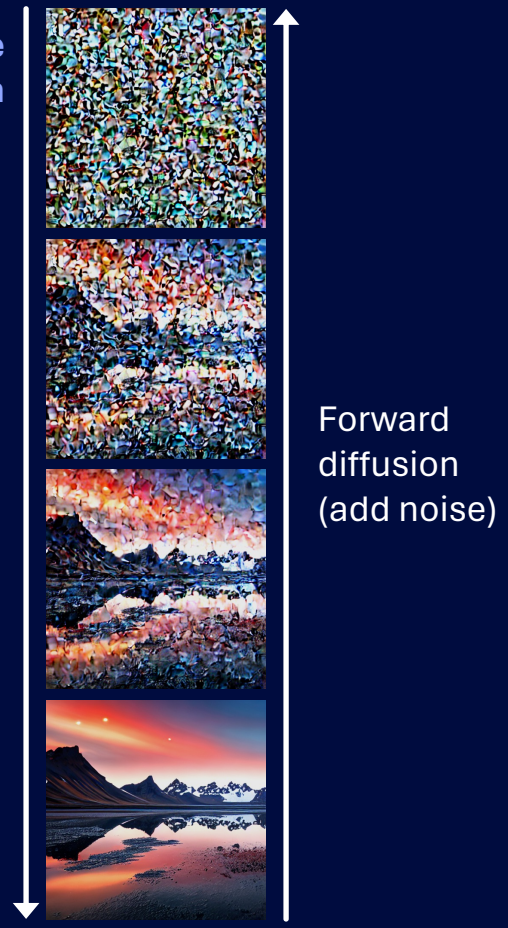
At MWC 2024

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

LLM: large language model; LLaVA: large language and vision assistant

What is diffusion?



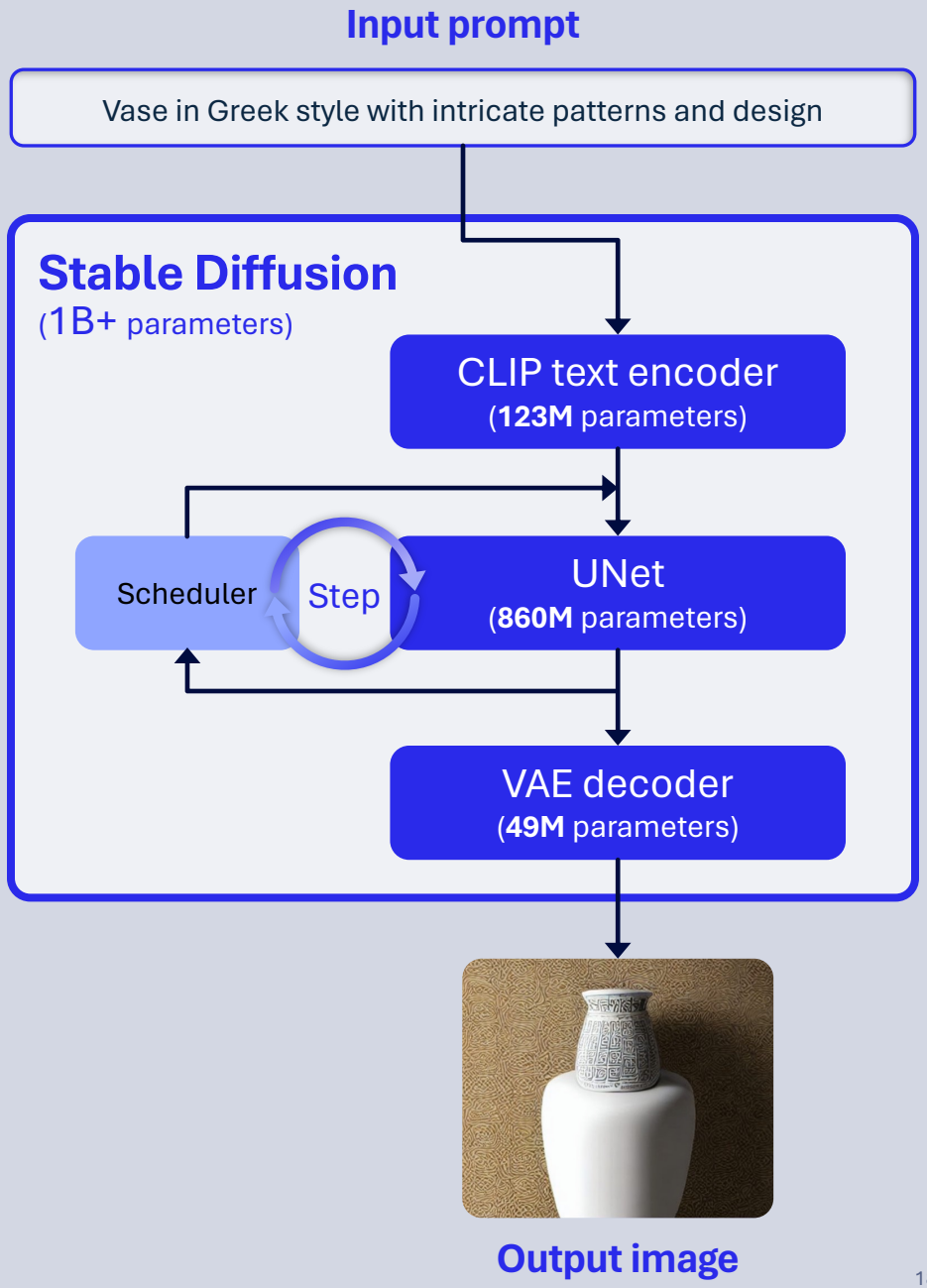
VAE: variational auto encoder;
CLIP: contrastive language-image pre-training

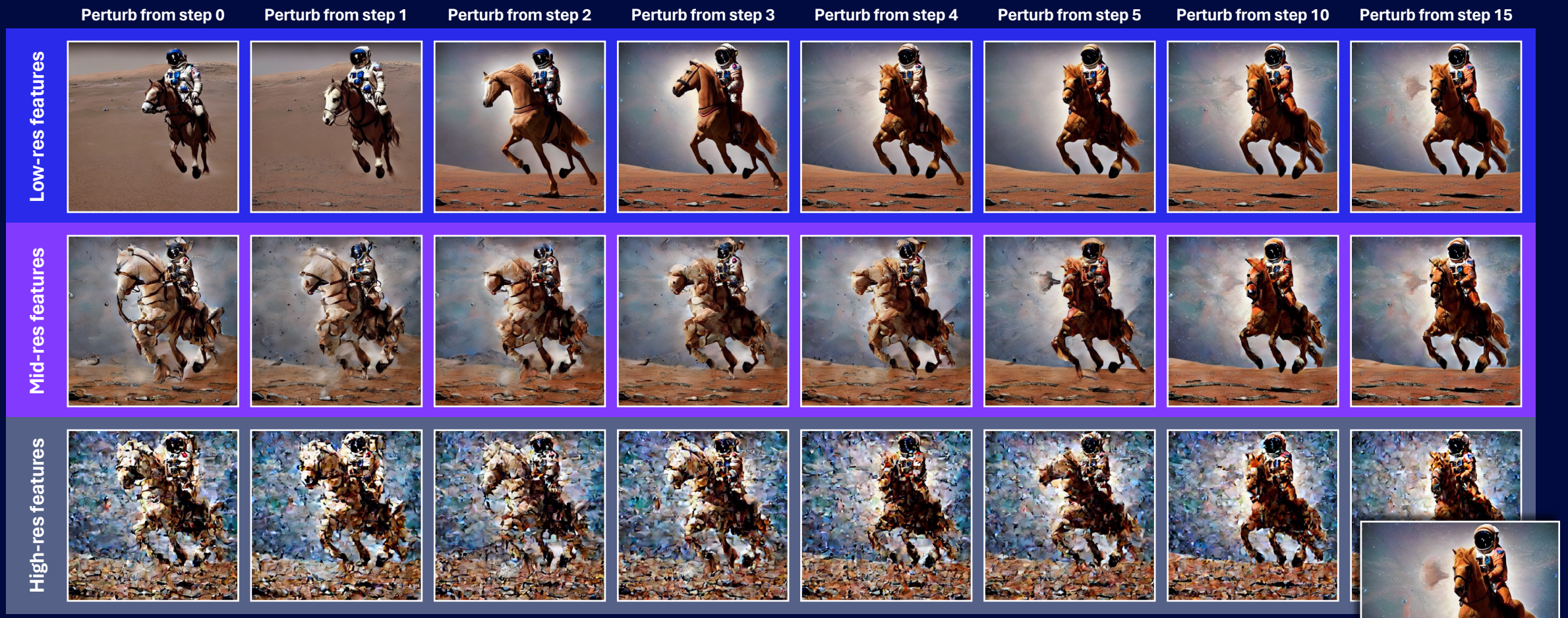
Stable Diffusion architecture

UNet is the biggest component model of Stable Diffusion

Many steps, often 20 or more, are used for generating high-quality images

Significant compute is required



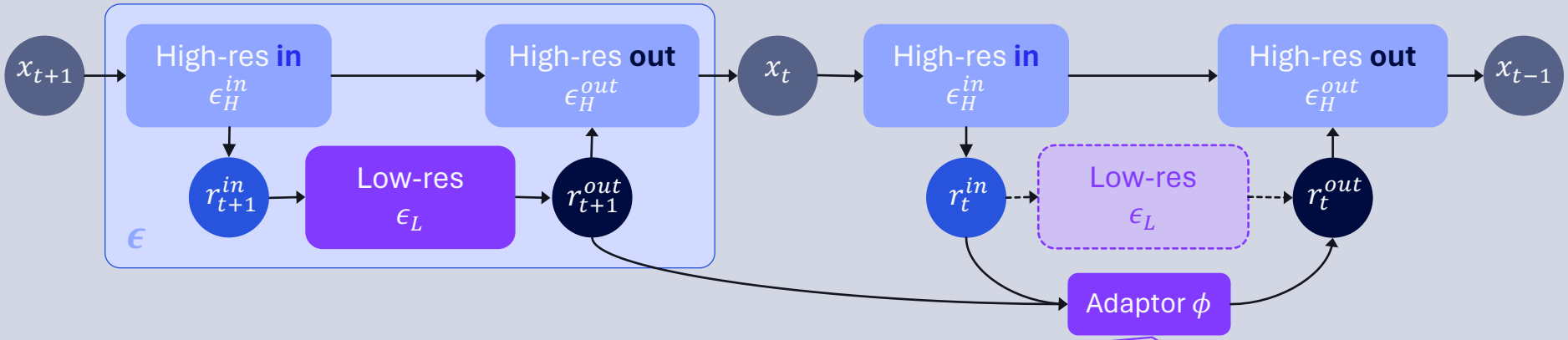


UNet low-resolution features can be perturbed without a noticeable change, whereas small perturbations on UNet high-resolution features degrade the image generation

Prompt: “Image of an astronaut riding a horse on mars"

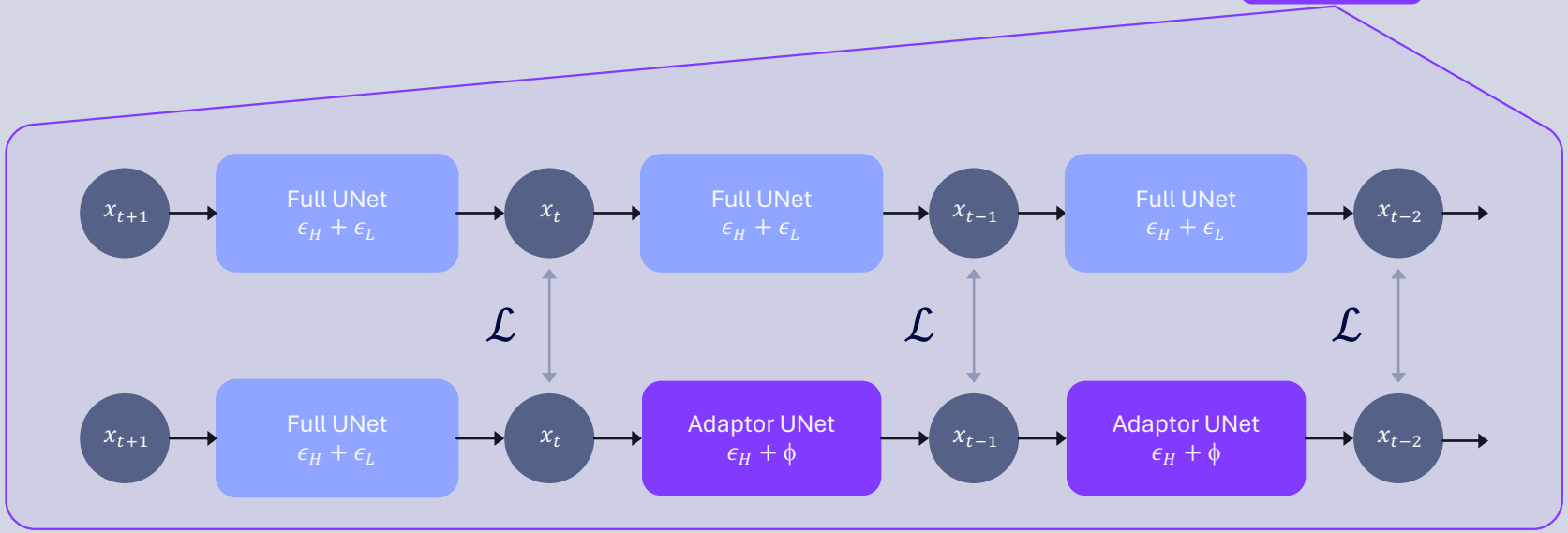
Clockwork architecture

An efficient approximation of low-res features by adapting from previous steps



Training the adaptor

Distillation from a full UNet over all denoising steps



Clockwork leverages the perturbation robustness to save computations and can improve any diffusion model (> 1.4X reduction in FLOPS)

Clockwork Diffusion: Efficient Generation With Model-Step Distillation, https://arxiv.org/abs/2312.08128

The potential of generative video editing

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

Key challenges:

1. Temporal consistency

2. High computational cost

Input video

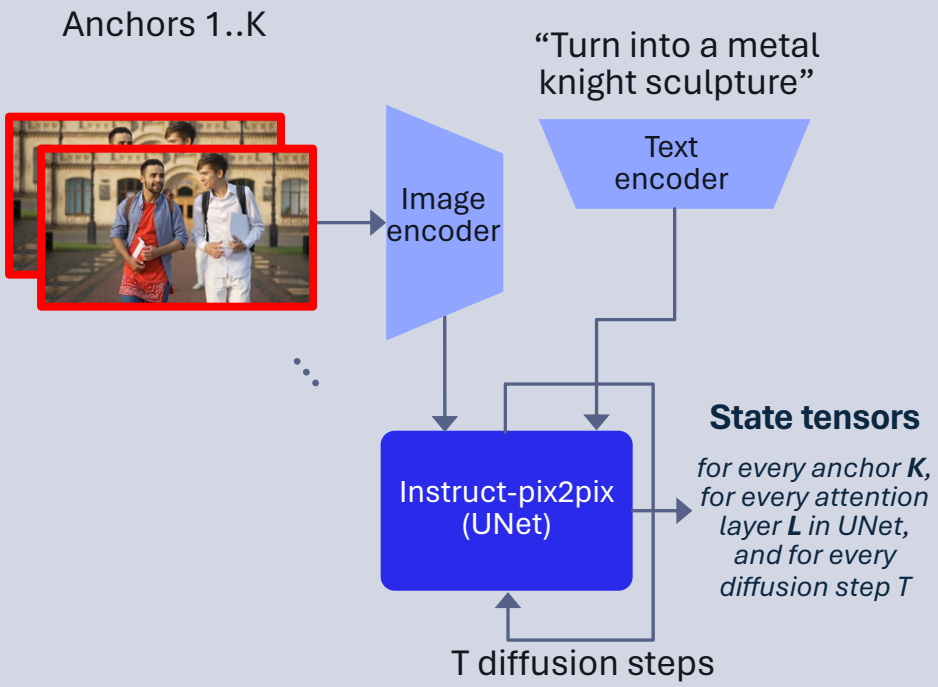
Edited video

Prompt: “pink flamingo walking”

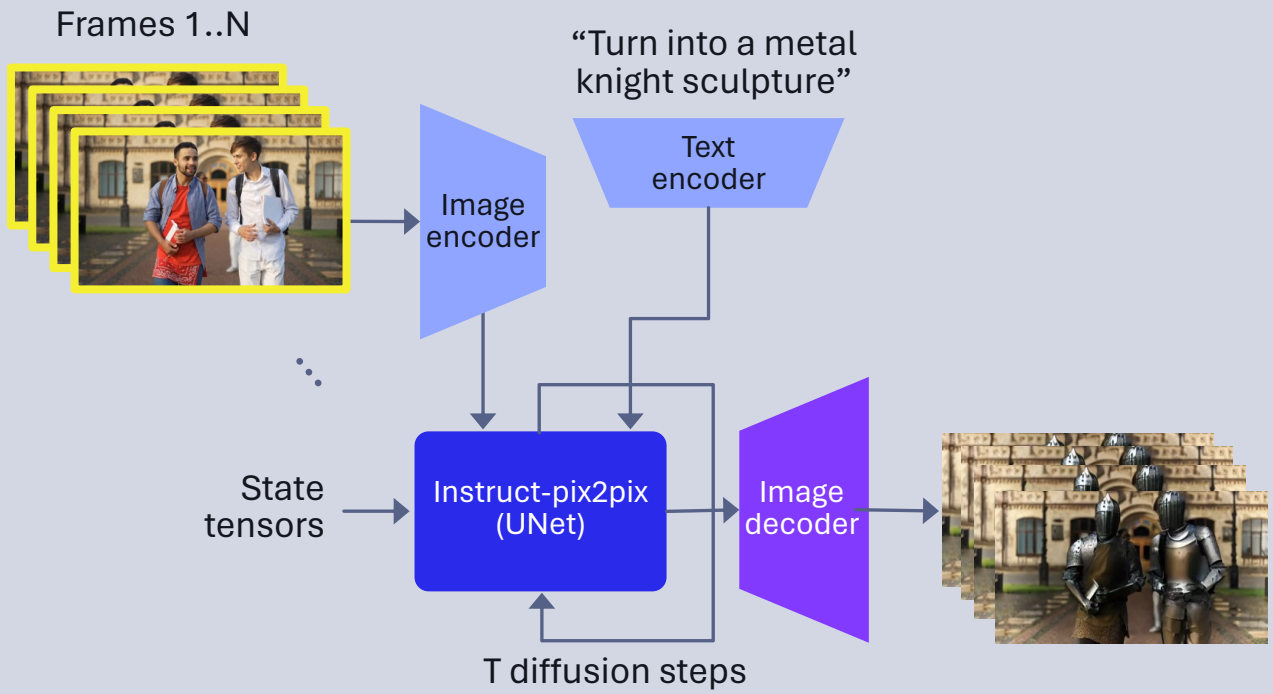
Making generative video methods efficient for on-device AI

Optimizations to FAIRY[1], 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

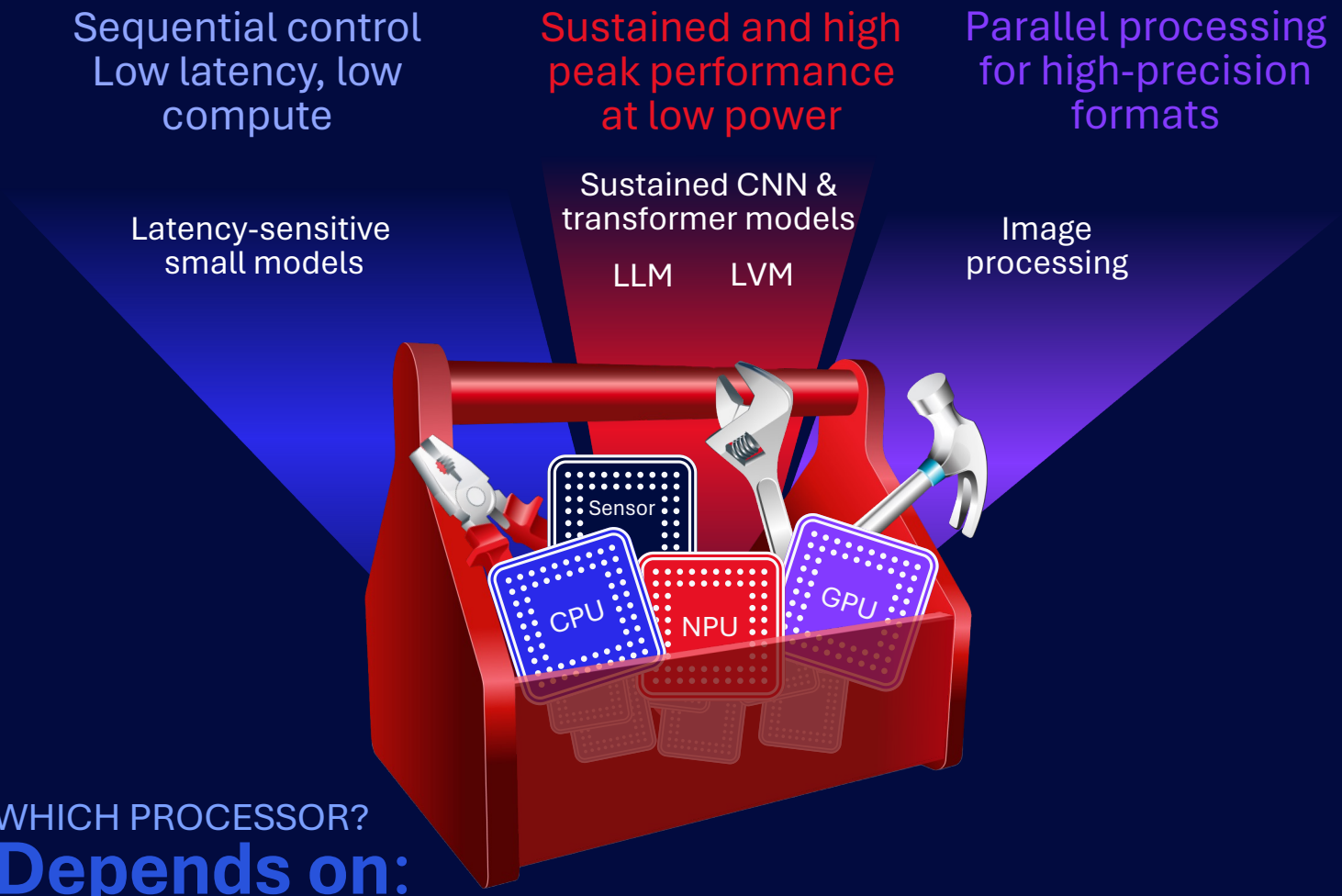
Fast FAIRY results

Making generative video feasible on device through significant reduction in computation and memory

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



WHICH PROCESSOR?

Depends on:

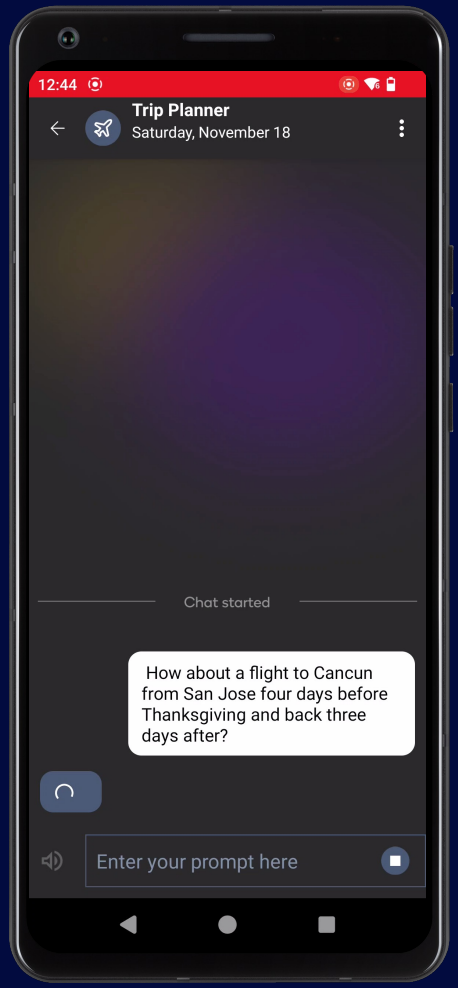
Use case	Device tier	Key performance indicators
Device type	Development time	Developer expertise

LLM: large language model; LVM: language vision model

CASE STUDY

The AI travel assistant

Many complex AI workloads

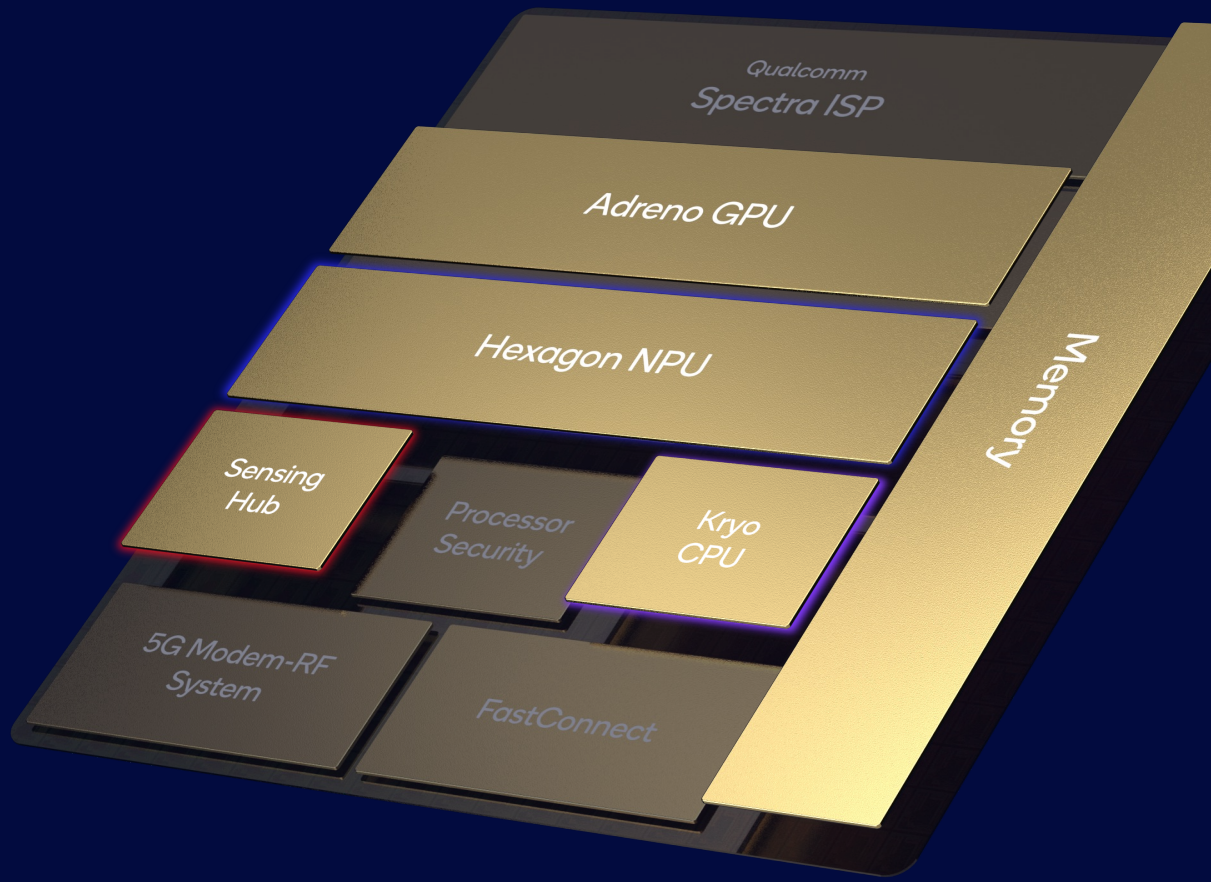


ASR Whisper

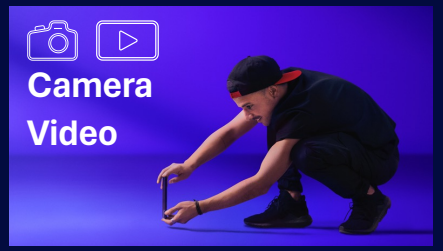
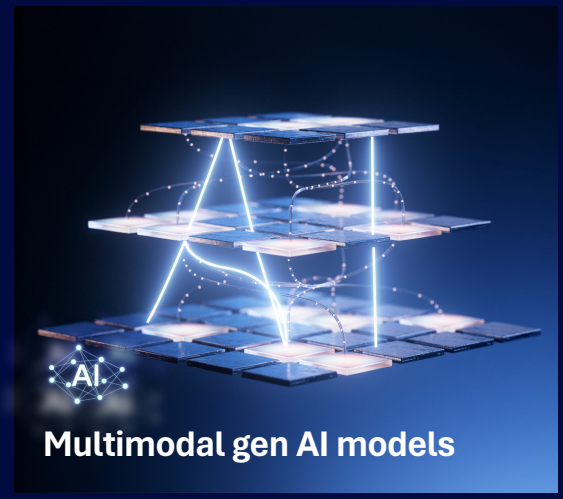
LLM Llama 2

Speech Open-source TTS

Cloud Skyscanner



The need for an NPU

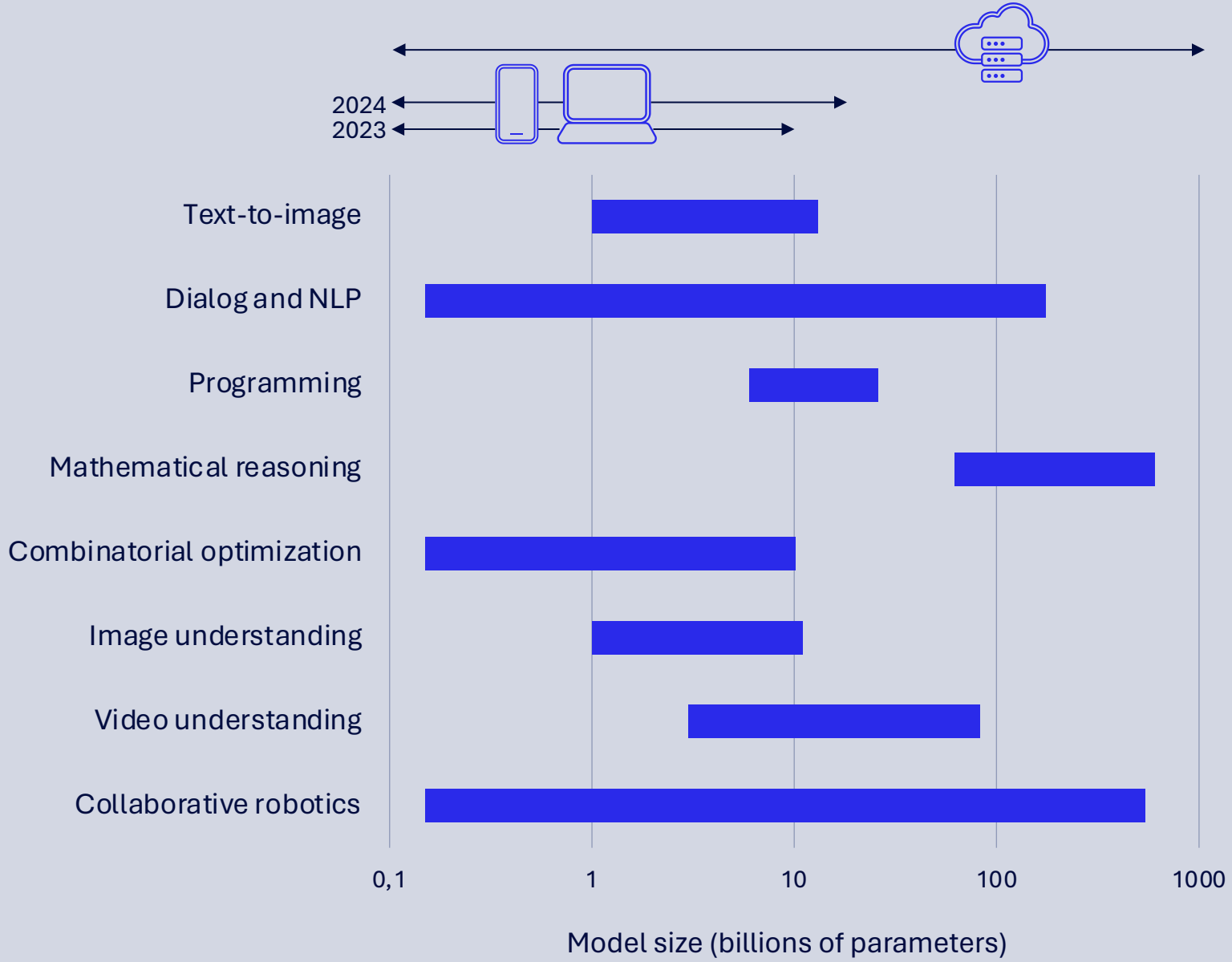
	2015	2016-2022	2023	2023+
Use case		 		
Hardware	Scalar, Vector	Tensor, Scalar, Vector	Micro-tile inferencing, Tensor, Scalar, Vector; Transformer support	Multimodal AI; Transformer support
Models	Simple CNN	Transformer / LSTM/ RNN/CNN	10B LLMs / LVMs	10B++LLMs/LVMs

CNN: convolutional neural network; LSTM: long short-term memory; RNN: recurrent neural network; LLM: large language model; LVM: language vision model

On-device AI can support a variety of Gen AI models

A broad number of Gen AI capabilities can run on device using models that range from **1 to 10 billion** parameters

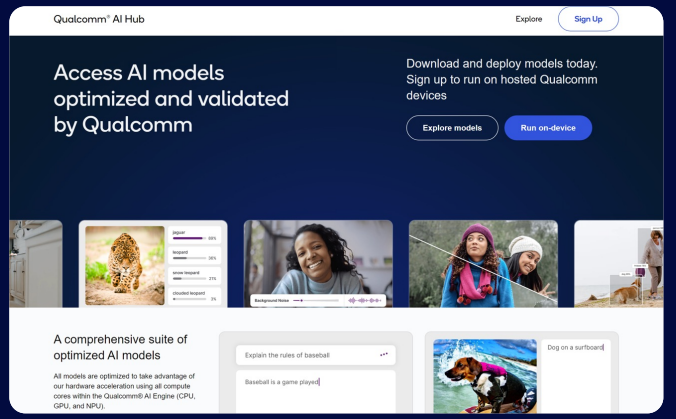
We can run models with over **10 billion parameters on device today** and anticipate this growing substantially **in the coming years**



Assuming INT4 parameters

Qualcomm® AI Hub

Library of fully optimized AI models for deployment across Snapdragon® and Qualcomm® platforms



AIHUB.QUALCOMM.COM

Qualcomm

GitHub

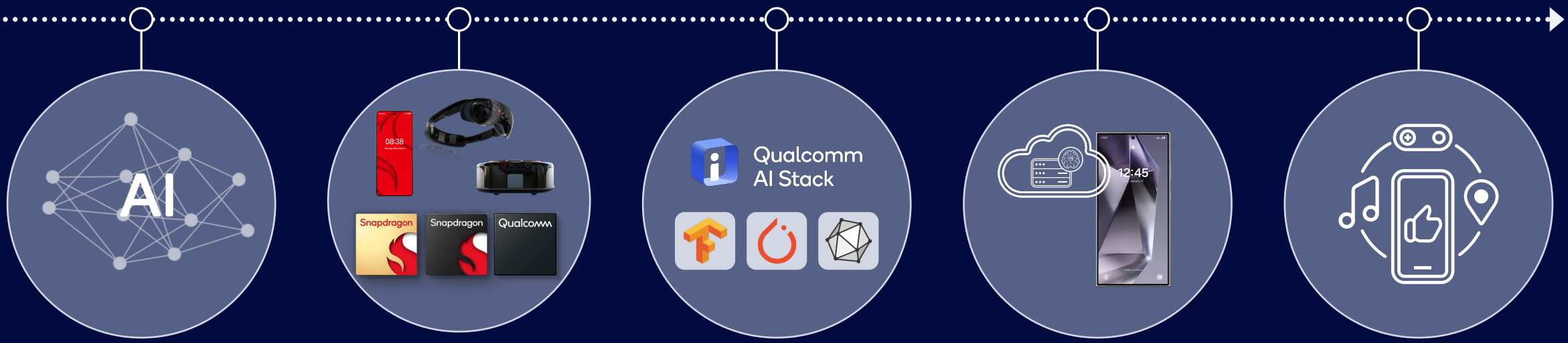
Pick a model
from library or BYOM

Pick a target
device or chipset

Pick a runtime

Test & validate
cloud-hosted device or locally

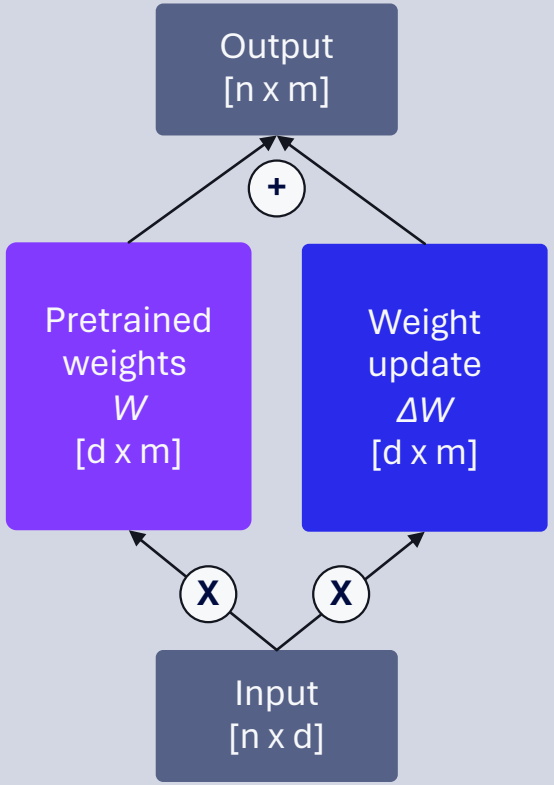
Deploy



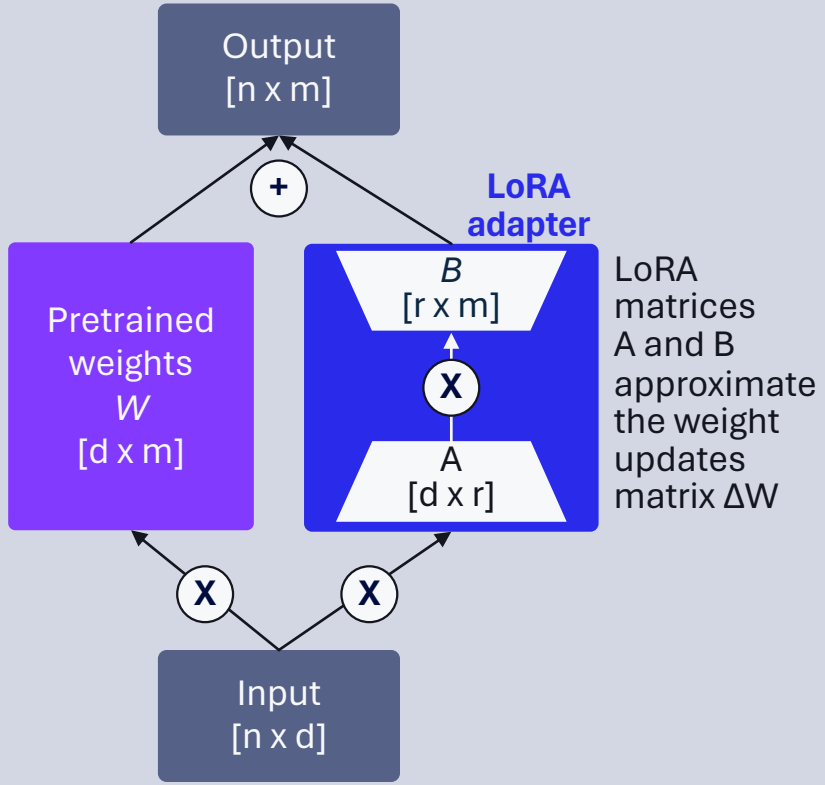
LoRA enables customization of generative AI across use cases

Beyond reducing the computations and memory usage during model fine-tuning, it enables personalized on-device inference at scale

Traditional fine-tuning



LoRA fine-tuning



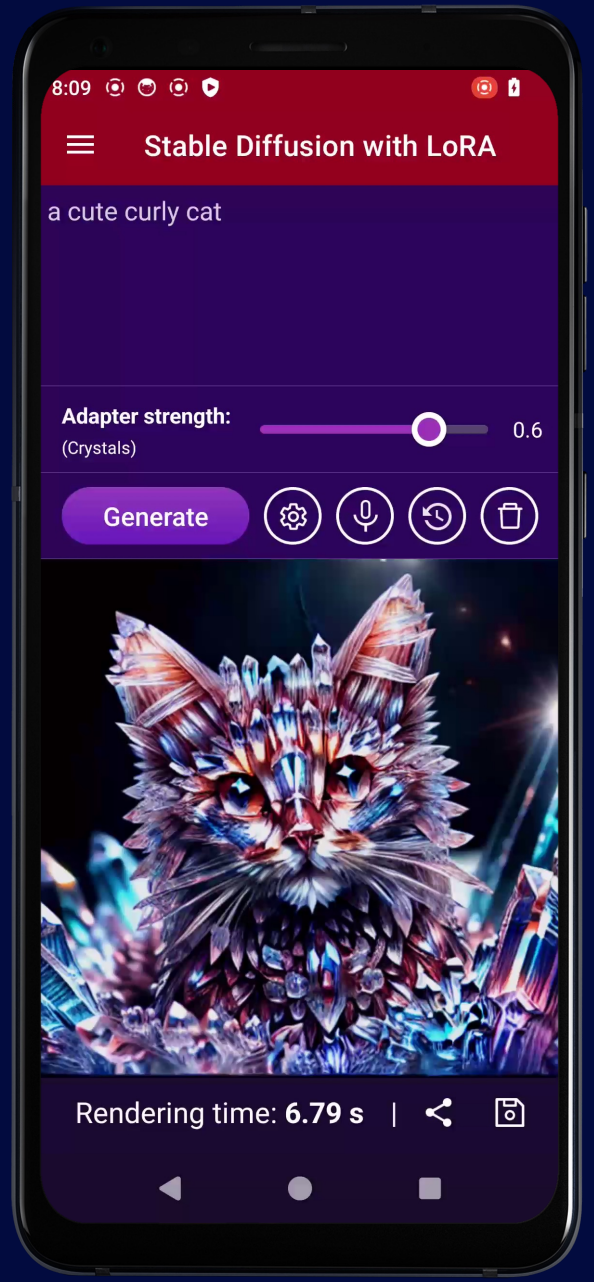
Reduces the number of trainable parameters of fine-tuned AI models (e.g., 2% of model)

Enables greater efficiency, scalability, and customization of on-device generative AI use cases

Multiple LoRA adapters can be used in parallel or swapped out

LoRA: low-rank adaptation

OUR FIRST
low-rank adaptation (LoRA) on an Android phone



1+ billion parameter Stable Diffusion with **LoRA adapter** for customized experiences

Create high-quality custom images based on **personal or artistic preferences**

LoRA enables **scalability** and **customization** of on-device generative AI across use cases

Full-stack AI optimization to achieve high performance at low power and fast switching between adapters

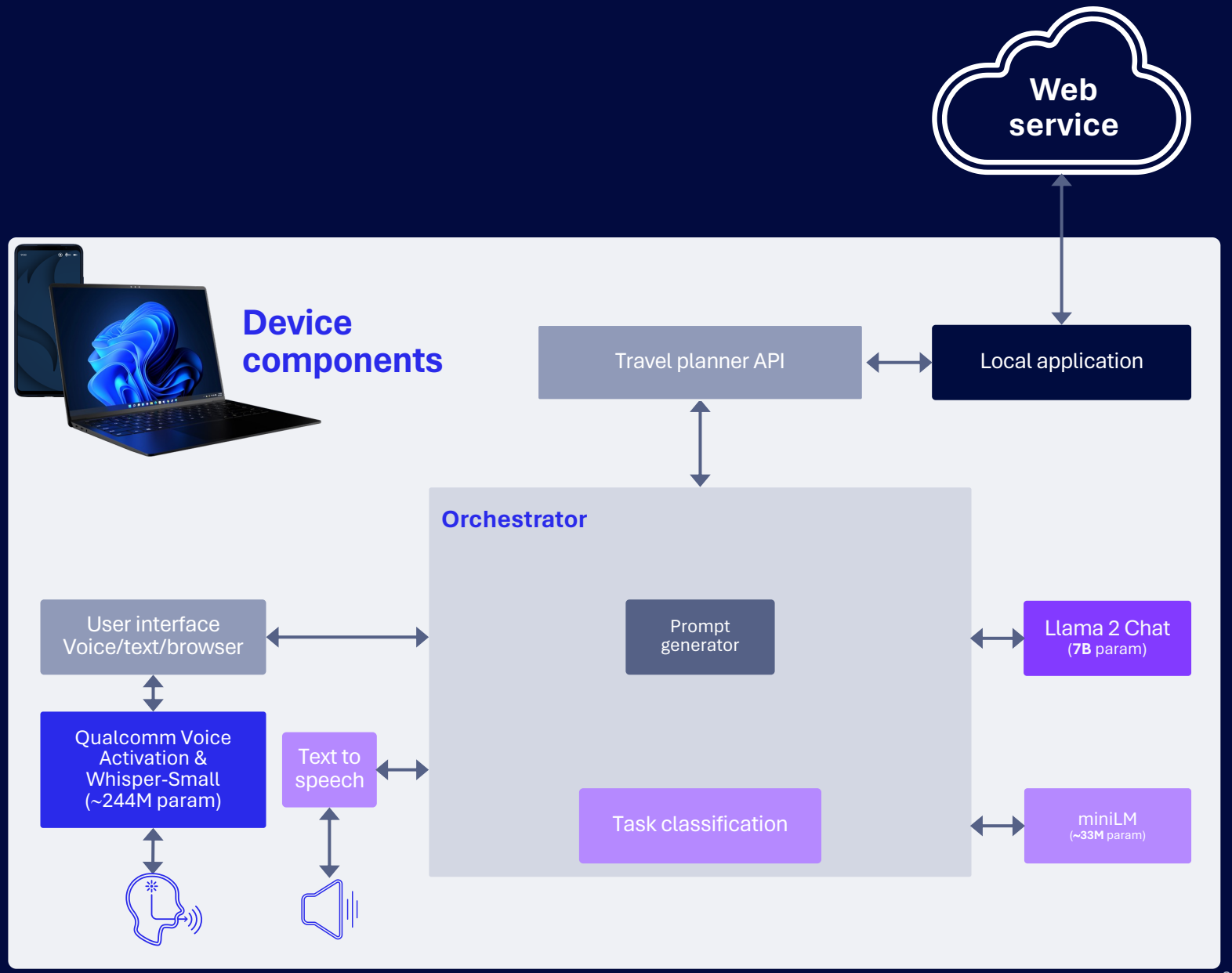
Enhanced privacy, **reliability**, **personalization**, and **cost** with on-device processing

AI assistant enables basic chat and chat-assisted apps on device

Orchestration across different tasks based on user query

Powered by Llama 2 Chat (7B)

Voice UI with Qualcomm® Voice Activation and Whisper-Small (244M)



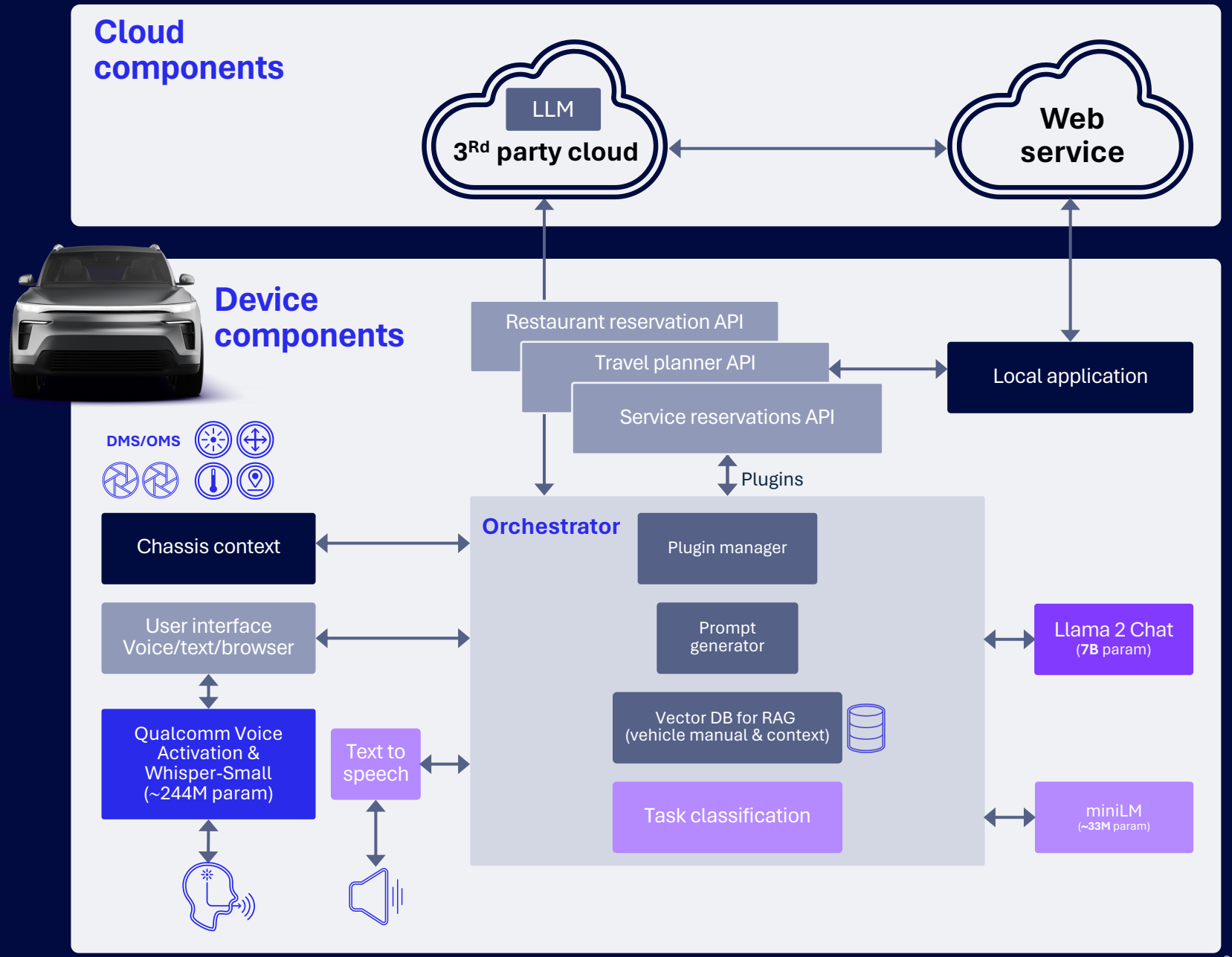
LM: language model

AI assistant can scale to use cases for digital cockpit

Orchestration across different tasks based on user queries **in automotive**

Powered by Llama 2 Chat (7B)

Voice UI with Qualcomm Voice Activation and Whisper-Small (244M)



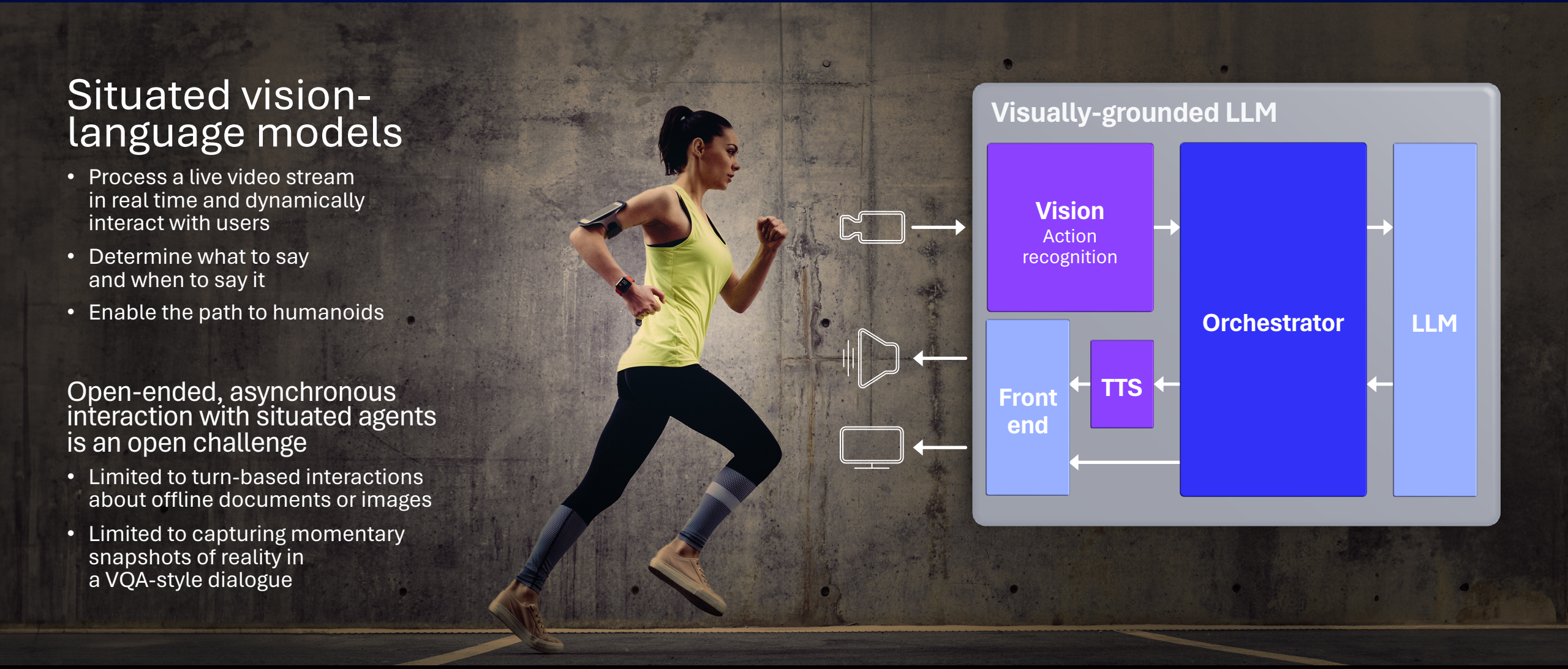
LM: language model; LLM: large language model

Situated vision-language models

- Process a live video stream in real time and dynamically interact with users
- Determine what to say and when to say it
- Enable the path to humanoids

Open-ended, asynchronous interaction with situated agents is an open challenge

- Limited to turn-based interactions about offline documents or images
- Limited to capturing momentary snapshots of reality in a VQA-style dialogue



Researching visually-grounded LLMs with the ability to reason and interact with the environment

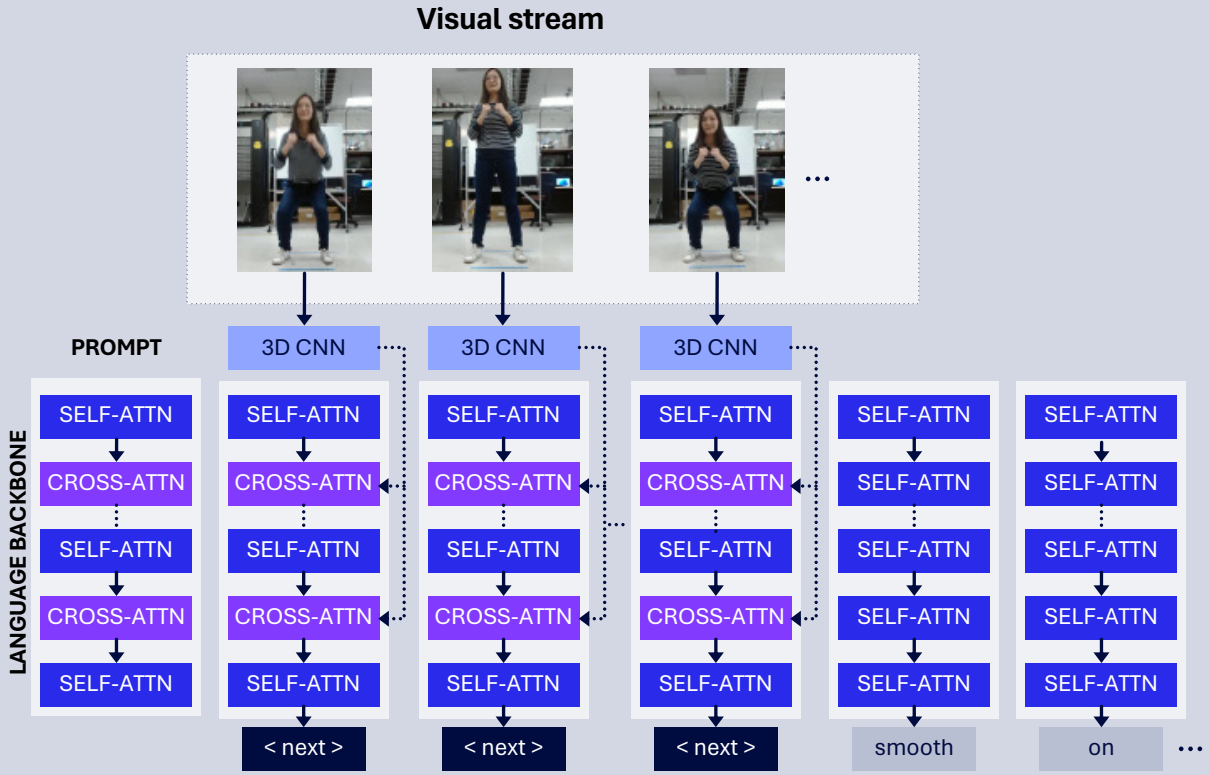
What to Say and When to Say it: Video-Language Model and Benchmark for Situated Interactions (2024); OpenEQA: Embodied Question Answering in the Era of Foundation Models (2024); VQA: visual question answering

Our situated vision-language model for fitness coaching

- 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

Key innovations

- 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)



Leading results

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 stools.

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)

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

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

Qualcomm

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

Efficient fine-tuning improves performance across use cases, and personalization is key to unlocking the power of generative AI

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

Questions

Connect with us

www.qualcomm.com/research/artificial-intelligence

www.qualcomm.com/news/ong

www.youtube.com/c/QualcommResearch

@QCOMResearch

https://assets.qualcomm.com/mobile-computing-newsletter-sign-up.html

www.slideshare.net/qualcommwirelessevolution

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:

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