Image Tokenization for Distributed Neural Cascades

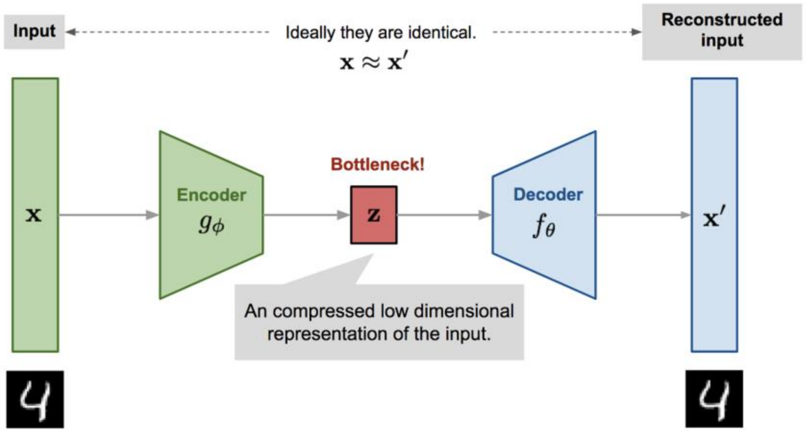
Derek Chow
Software Engineer, Google

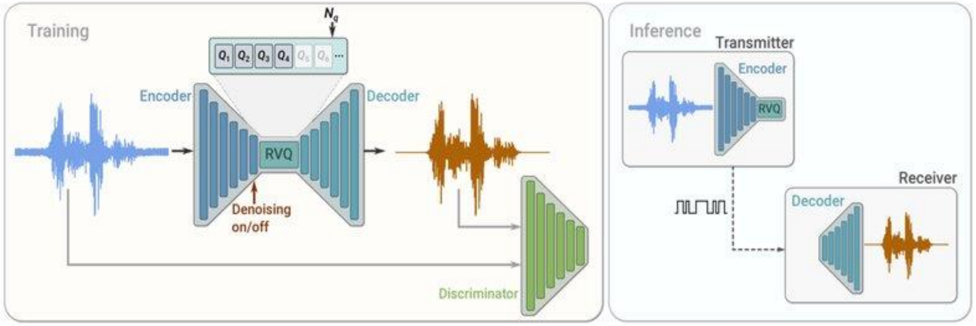
Shang-Hung Lin
Vice President of NPU Technology
VeriSilicon

What is Tokenization?

Tokenization is the process of converting a sensor modality into a neural encoding.

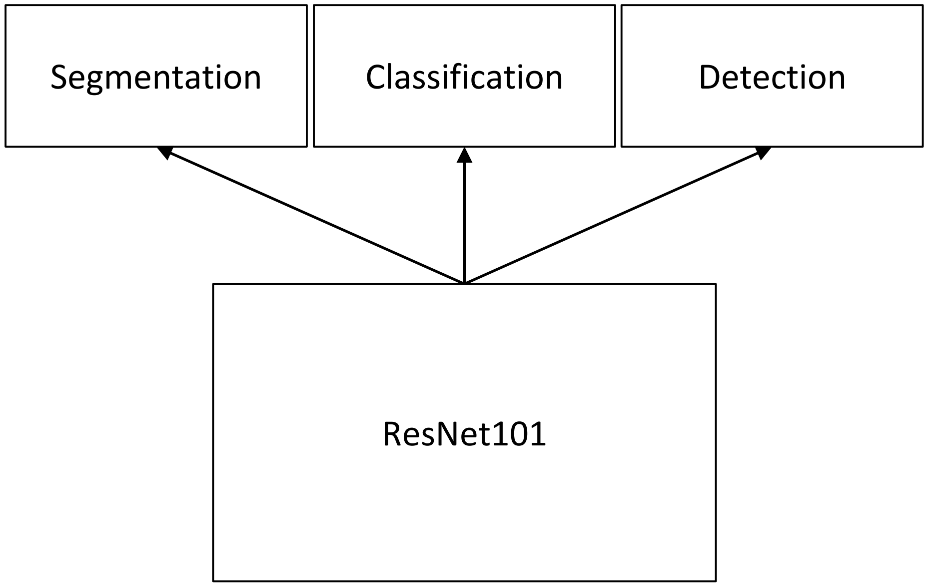
Examples of Tokenizers





Tokenizer is a Feature Extractor

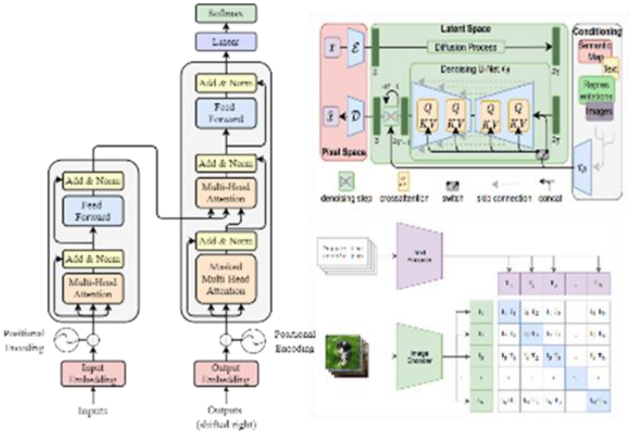
- Serves as a feature extractor for a neural network
- Enables features like classification, generation, RAG



Multimodal AI

- Text
- Audio
- Image
- Video
- Sensor
- 3D
- Graph
- ⋮

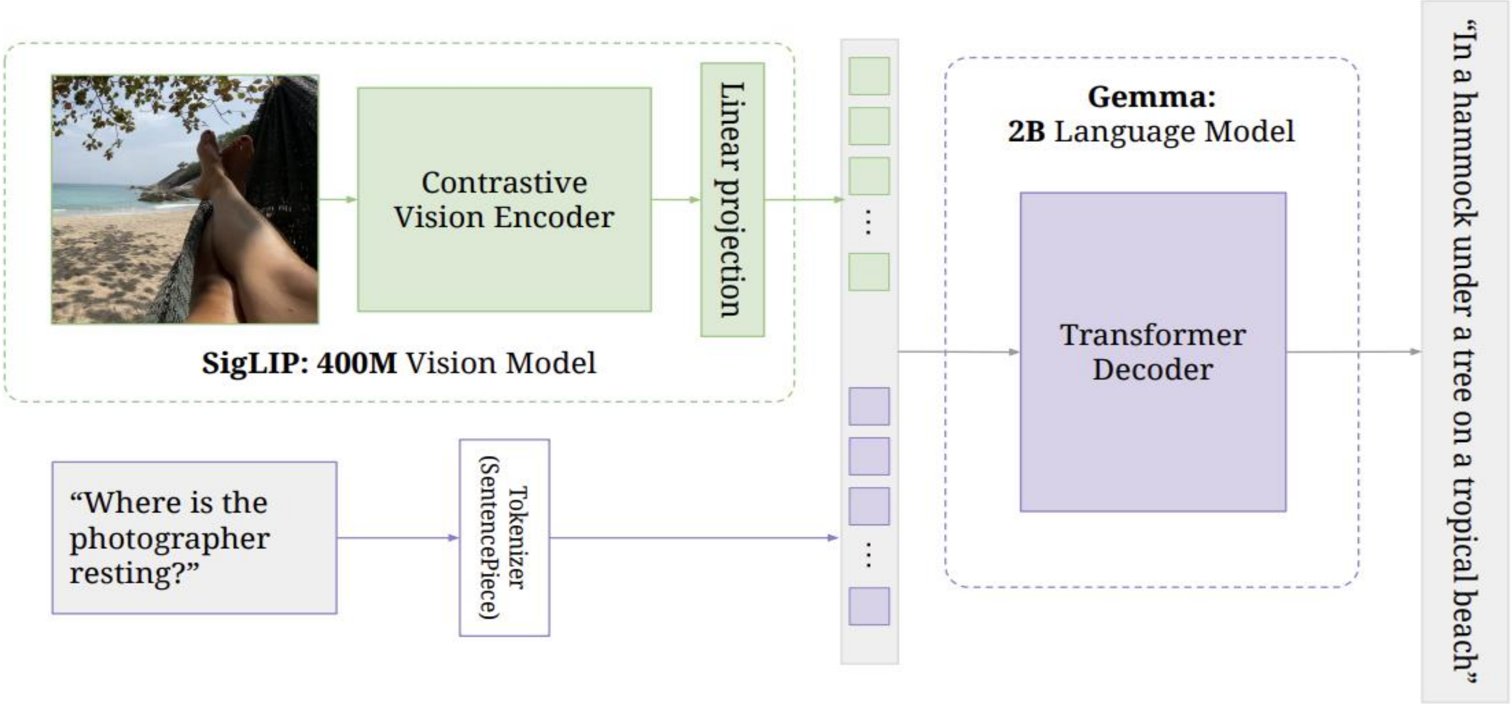
Foundation Model



Large Language Models | Vision Language Models | Diffusion Models …

- Question Answering
- Recall/Summarization
- Live Captions
- Content Creation
- Zero-Shot Learning
- Scene/Object Recognition
- Action Recognition
- ⋮

SigLIP / Gemma



Tokenization Creates a Form a Data Compression

- Tokenizer and detokenizer act as a Codec
- Saves power during transmission
- Saves capacity at rest

Diverse Hardware Ecosystem

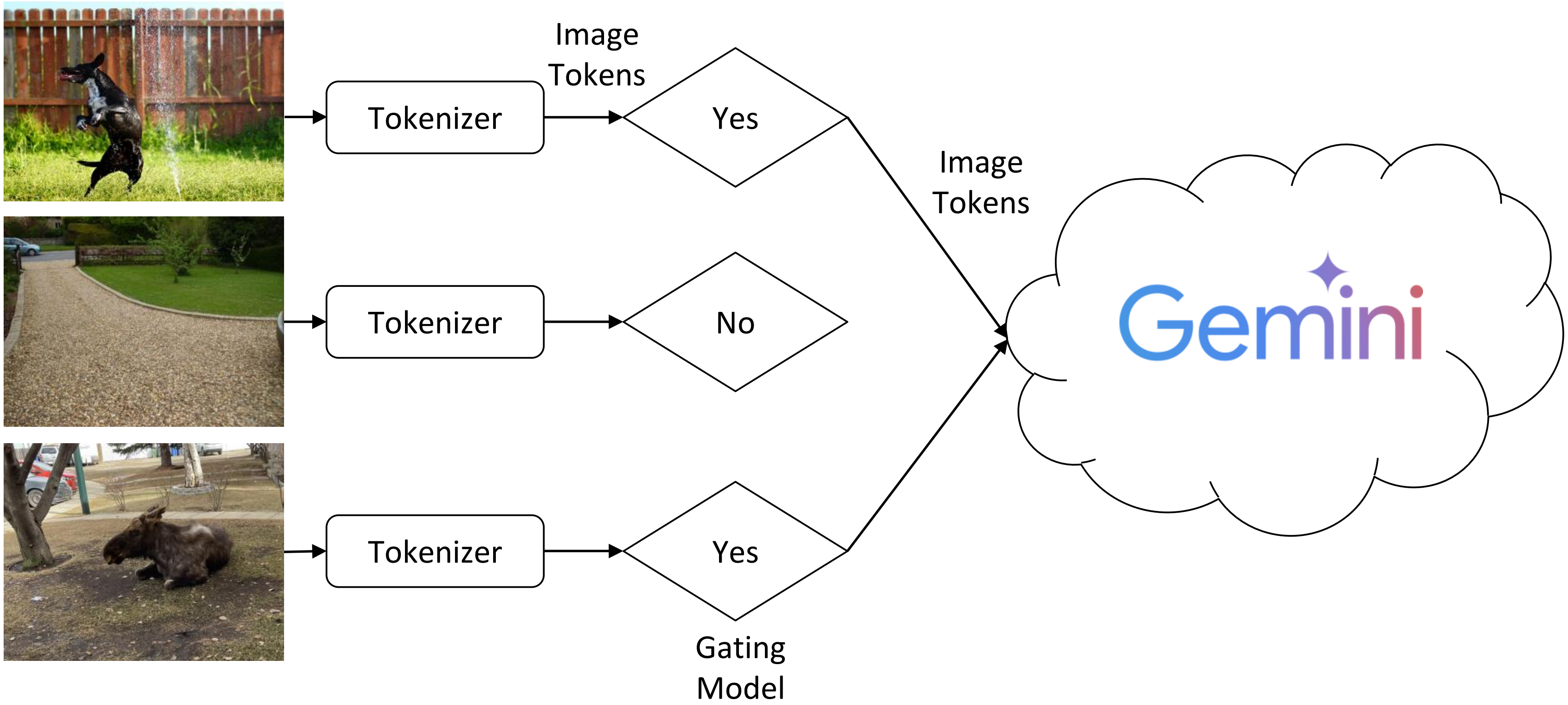
	Compute	Memory	Bandwidth
	High	High	High
	Medium	Medium	Medium
	Medium	Low	Low
	Low	Low	Low
	Low	Low	Low

World’s Leading Smart Home Products

Can we combine the strengths of multiple devices for GenAI experiences?

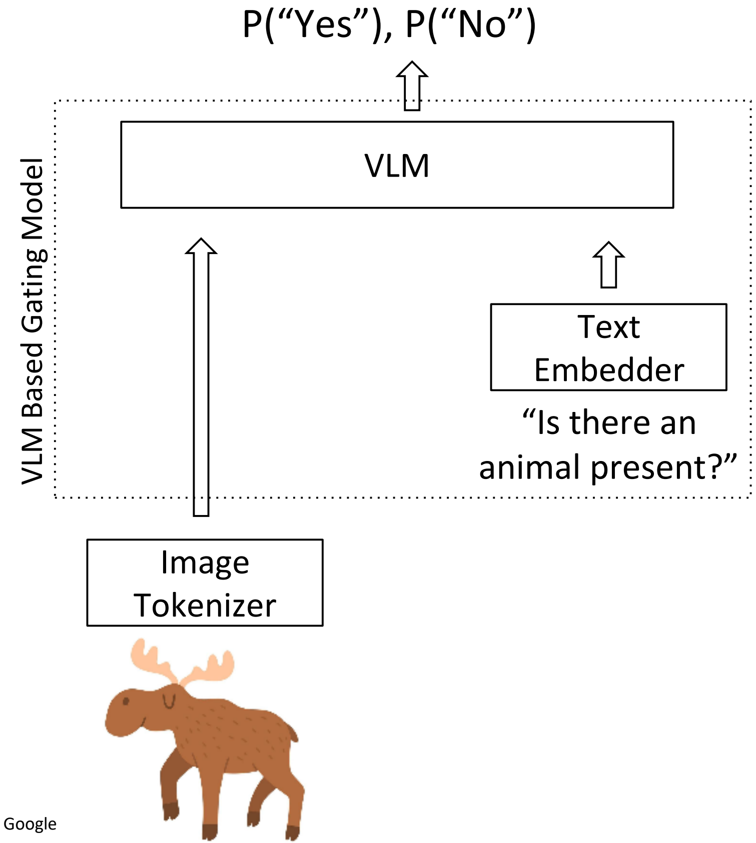
We think yes.

Anatomy of a Neural Cascade

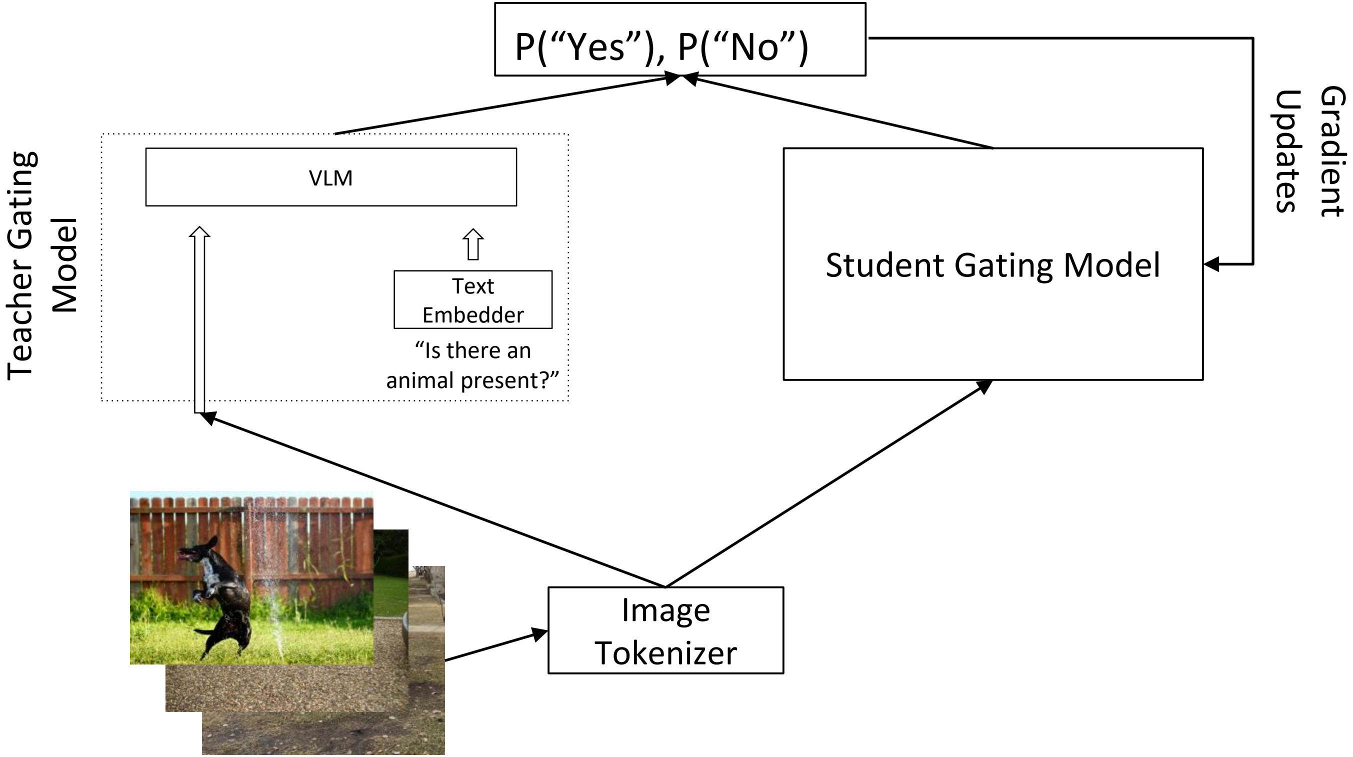


Building a Large Gating Model

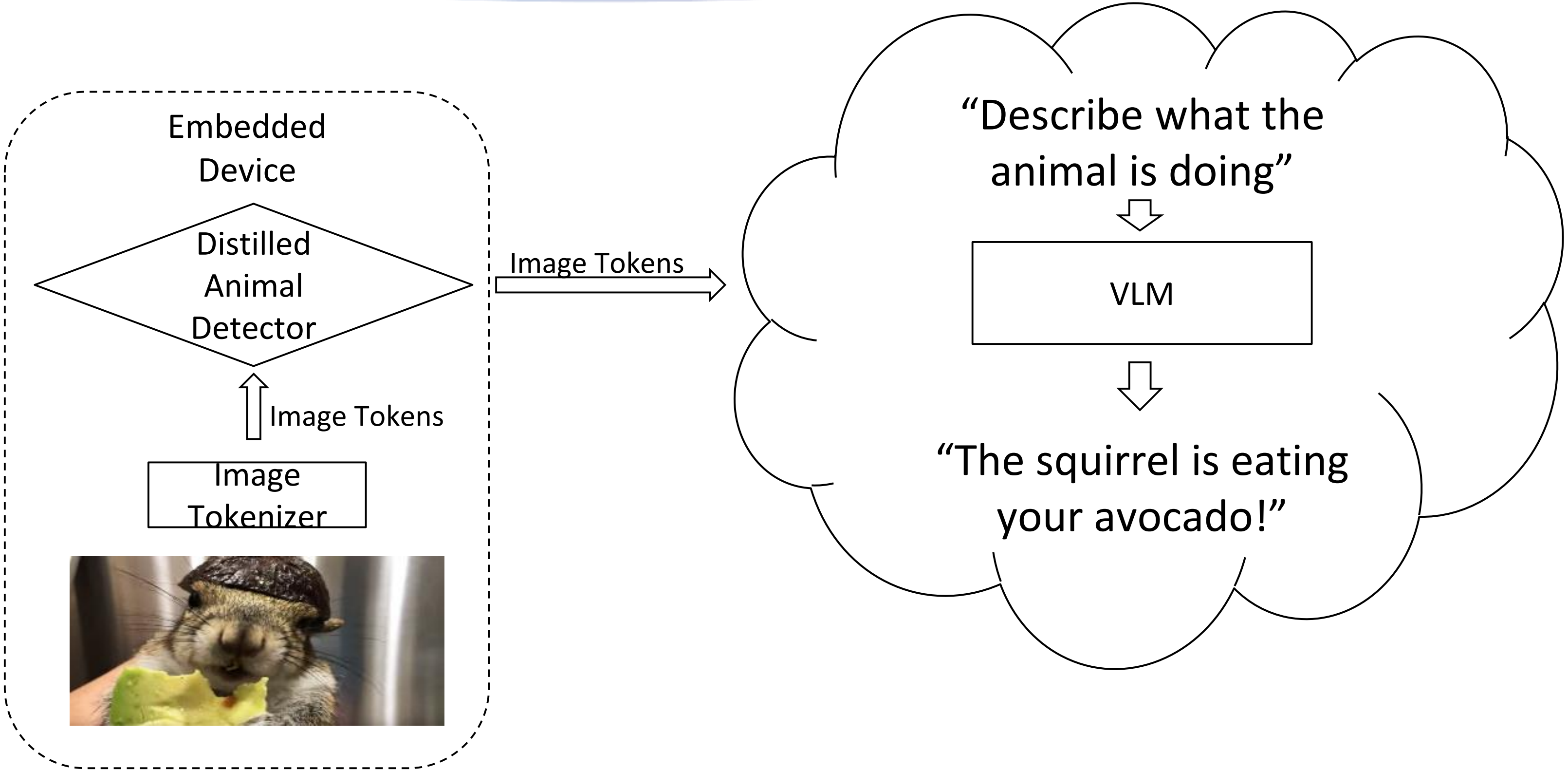
- We can build a gating model using a VLM
  - Provide a prompt to describe what you want to detect. i.e.: “Is there an animal present?”
  - Feed tokenized image into VLM
  - Check probability of emitting “Yes” or “No”



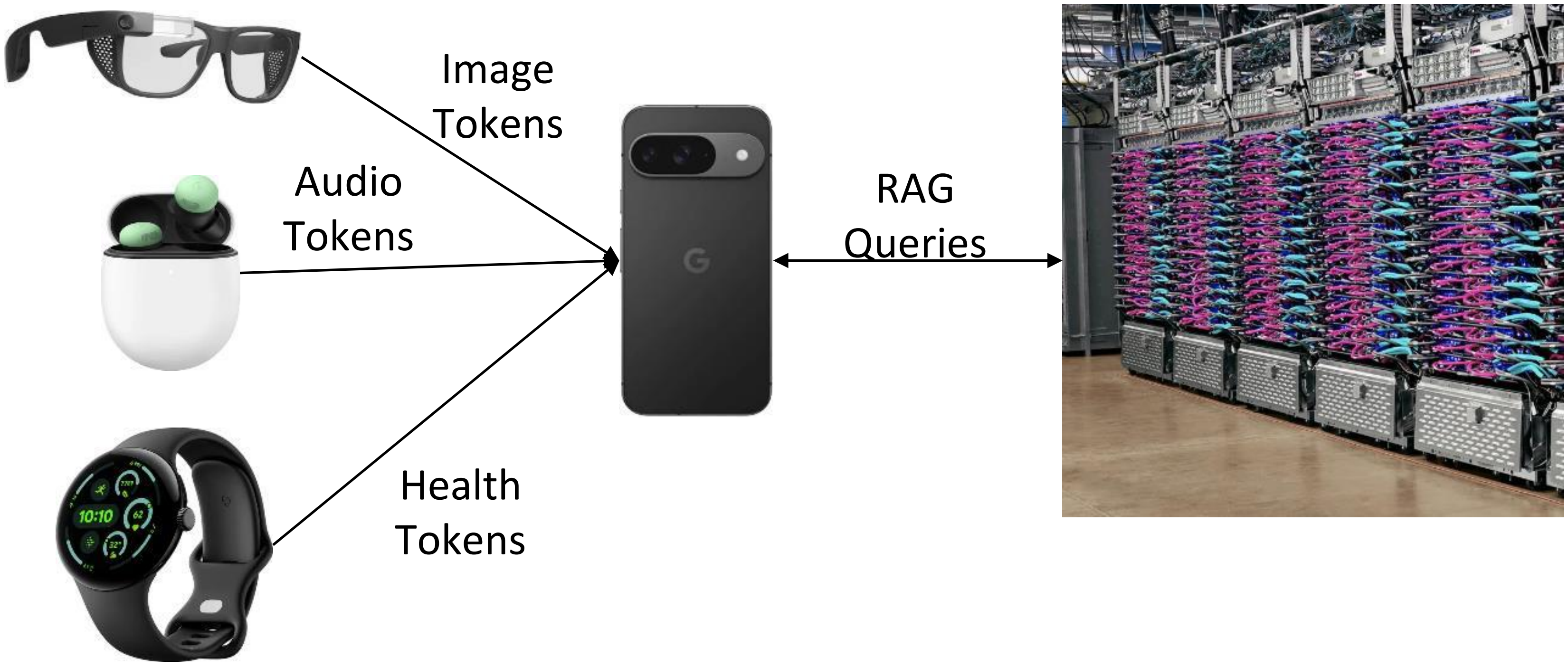
Distilling a Smaller Gating Model



Composing Models

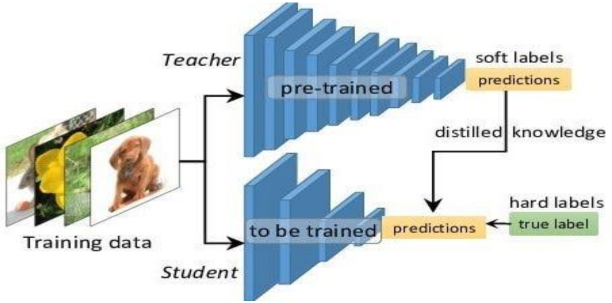


Cascades Beyond Two Devices

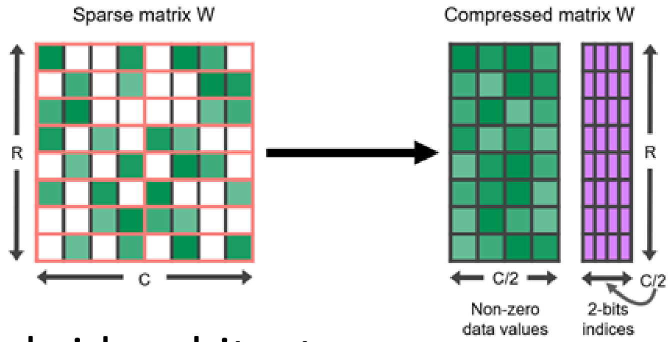


Squeezing Neural Cascade Frontend into Small Devices

- Knowledge distillation



- Sparsity, weight sharing



- Quantization

Format Name	Element Data Type	Element Bits (d)	Scaling Block Size (k)	Scale Data Type	Scale Bits (w)
MXFP8	FP8 (E5M2)	8	32	E8M0	8
	FP8 (E4M3)				
MXFP6	FP6 (E3M2)	6	32	E8M0	8
	FP6 (E2M3)				
MXFP4	FP4 (E2M1)	4	32	E8M0	8
MXINT8	INT8	8	32	E8M0	8

- Hybrid architecture

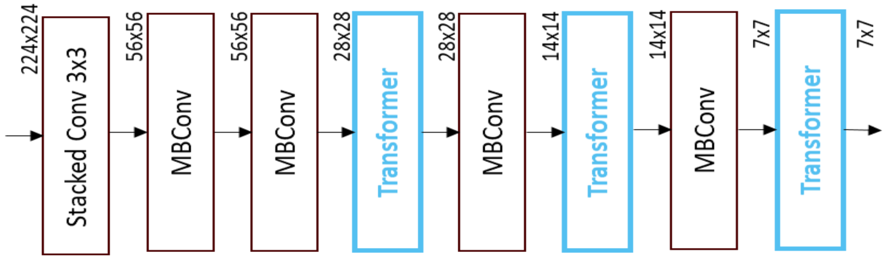
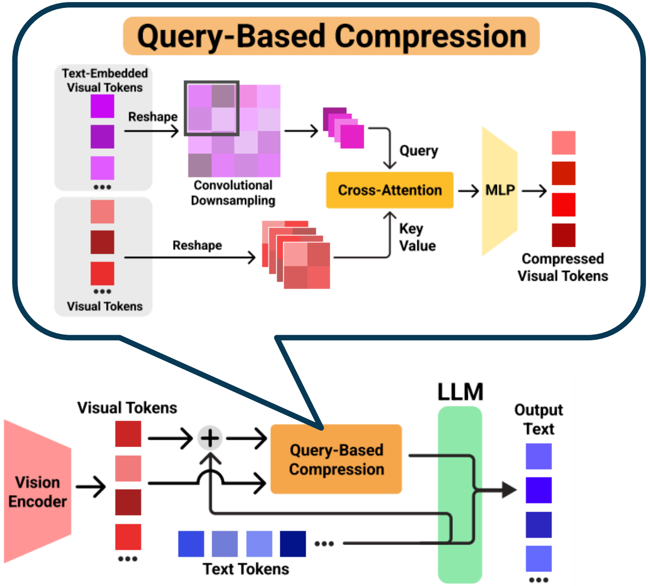


Image Token Compression

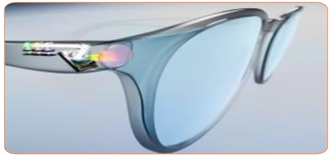
- Reducing image token numbers by text prompt

Compression Ratio	Method	# Token	GQA	MMB	MME	POPE	SQA	TextVQA	VizWiz	VQAv2
	LLaVA-1.5	576	62.0	64.3	1510.7	85.9	66.8	58.2	50.0	78.5
16x	PruMerge	∼32	57.2*	60.9	1350.3	76.3	68.5	**56.0**	45.2*	72.0
	TokenPacker	36	59.6	62.8	1440.9*	83.3*	**71.0***	53.2*	50.2	75.0
	Matryoshka Multi.	36	60.3	**64.8**	–	**85.5**	–	–	**52.8**	–
	Matryoshka Query	36	58.8	63.4	1416.3	81.9	66.8	–	51.0	73.7
	QueCC	36	**60.5**	62.5	**1442.0**	84.5	70.6	53.3	50.1	**75.8**
36x	TokenPacker	16	58.9*	**62.7***	1378.8*	**83.7***	68.1*	**52.5***	**50.5***	74.4*
	Matryoshka Query	16	57.6	61.9	**1408.5**	80.8	67.5	–	49.8	71.1
	QueCC	16	**59.0**	62.2	1408.0	83.4	**70.7**	51.3	47.7	**74.5**
144x	TokenPacker	4	56.2*	61.5*	1347.6*	81.7*	68.5*	**49.2***	45.7*	70.5*
	Matryoshka Query	4	53.0	56.5	1176.1	77.6	65.1	–	**49.4**	64.1
	QueCC	4	**56.5**	**62.1**	**1390.3**	**81.8**	**68.6**	48.7	45.0	**70.6**
576x	TokenPacker	1	53.4*	58.7*	1262.4*	80.7*	69.4*	46.2*	41.1*	66.9*
	Matryoshka Multi.	1	52.6	**59.5**	–	78.4	–	–	**49.4**	–
	Matryoshka Query	2	50.8	54.4	1144.0	74.5	65.0	–	48.5	61.0
	QueCC	1	**53.5**	59.4	**1269.1**	**81.3**	**69.9**	**46.8**	44.1	**67.3**



QueCC (ICLR 2025, arxiv:2411.03312)

Project Open Se Cura – Edge and Cloud Collaborative Computing



Cloud computing

Extremely low power consumption

- Always on
- Ambient computing

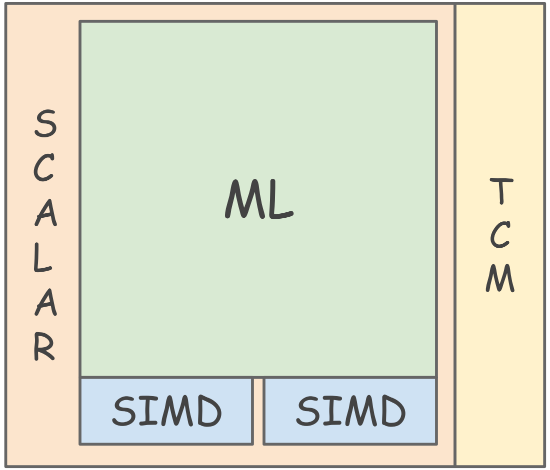
Realizing large models everywhere

- Responsiveness
- **Privacy (local & cloud)**
- Computational resources

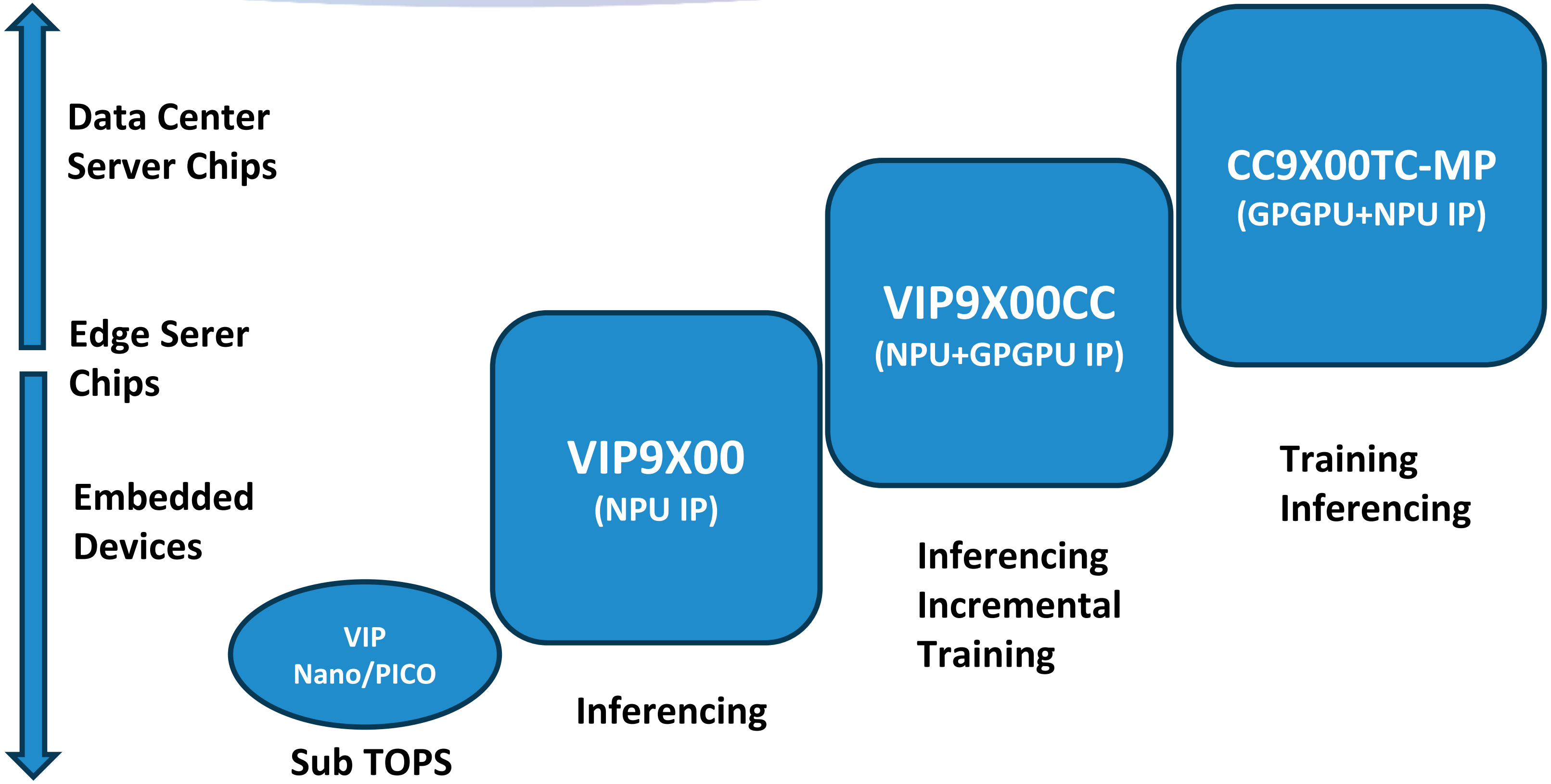
Kelvin: A RISC-V ML Accelerator for Edge

Kelvin is a RISC-V based ML Accelerator

- Open-source design as part of Open Se Cura
- Provides familiar framework for programming ML kernels to experts with SIMD/GPU experience
- Support for RISC-V Vector and Matrix extensions is in development, targeting 256+ MACs/cycle
- Security extensions via CHERI are on our roadmap



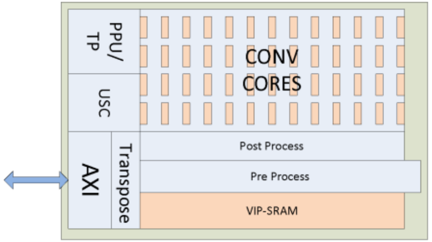
VeriSilicon AI-Computing IP Product Lineup



High Efficiency Inference NPU for VLMs & LLMs

VIP9000
4 TOPS
16 GB/s

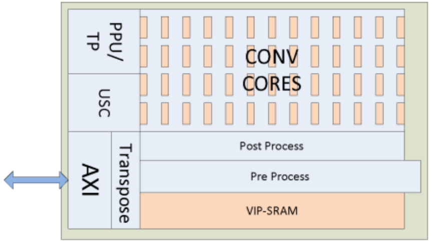
Qwen2
1.5B



Embedded Devices

VIP9000
40 TOPS
128 GB/s

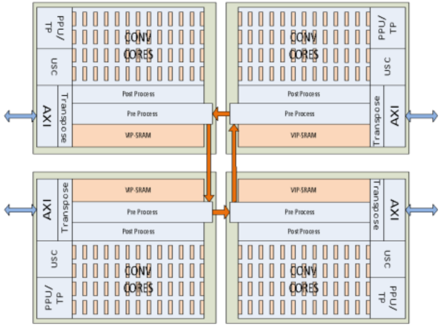
LLaMA2
7B



AI-PC, Mobile

VIP9400
160 TOPS
512 GB/s

LLaMA3
70B



Edge Server

Summary and Challenges

Summary

- Tokenizers provide a framework building multi-modal LLMs
- Distillation based training can create a gating mechanism to separate tokenizers from the LLM
- Once separated, compute can be distributed between embedded devices and the cloud

Challenges

- Technical
  - Memory and compute scaling for tokenizers and LLMs
  - Infrastructure for training distributed models
- Ecosystem
  - Changing model landscape
  - Diverse hardware landscape
  - Fostering community

Resources

Gemma

https://ai.google.dev/gemma

Project Open Se Cura

https://www.opensecura.googlesource.com

VeriSilicon NPU IP

https://www.verisilicon.com/en/IPPortfolio/VivanteNPUIP

2025 Embedded Vision Summit

Visit us at booth 508!