



Foundations of Edge AI

Lecture05

The Frontier of Edge AI

Lanyu (Lori) Xu

Email: lxu@oakland.edu

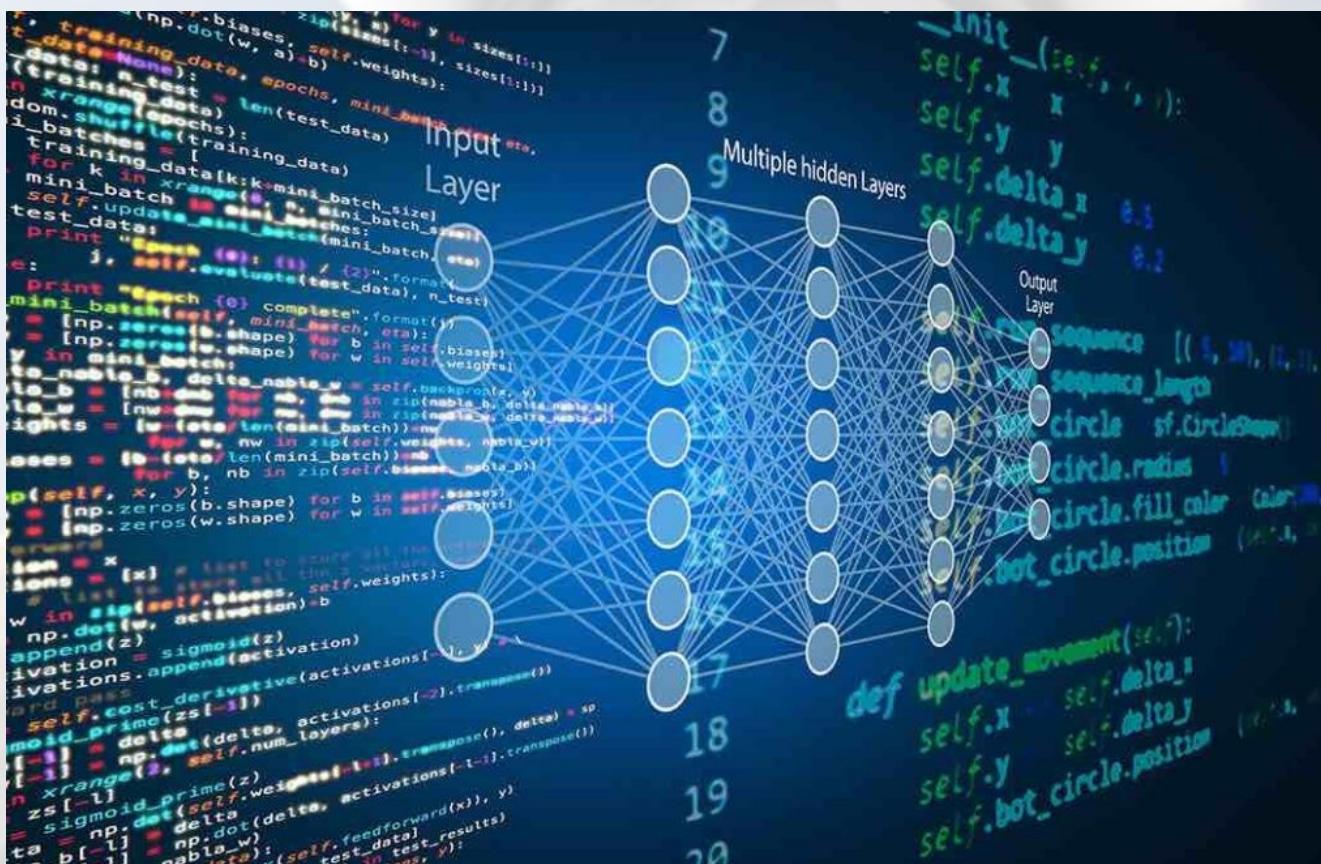
Homepage: <https://lori930.github.io/>

Office: EC 524



Deep Learning

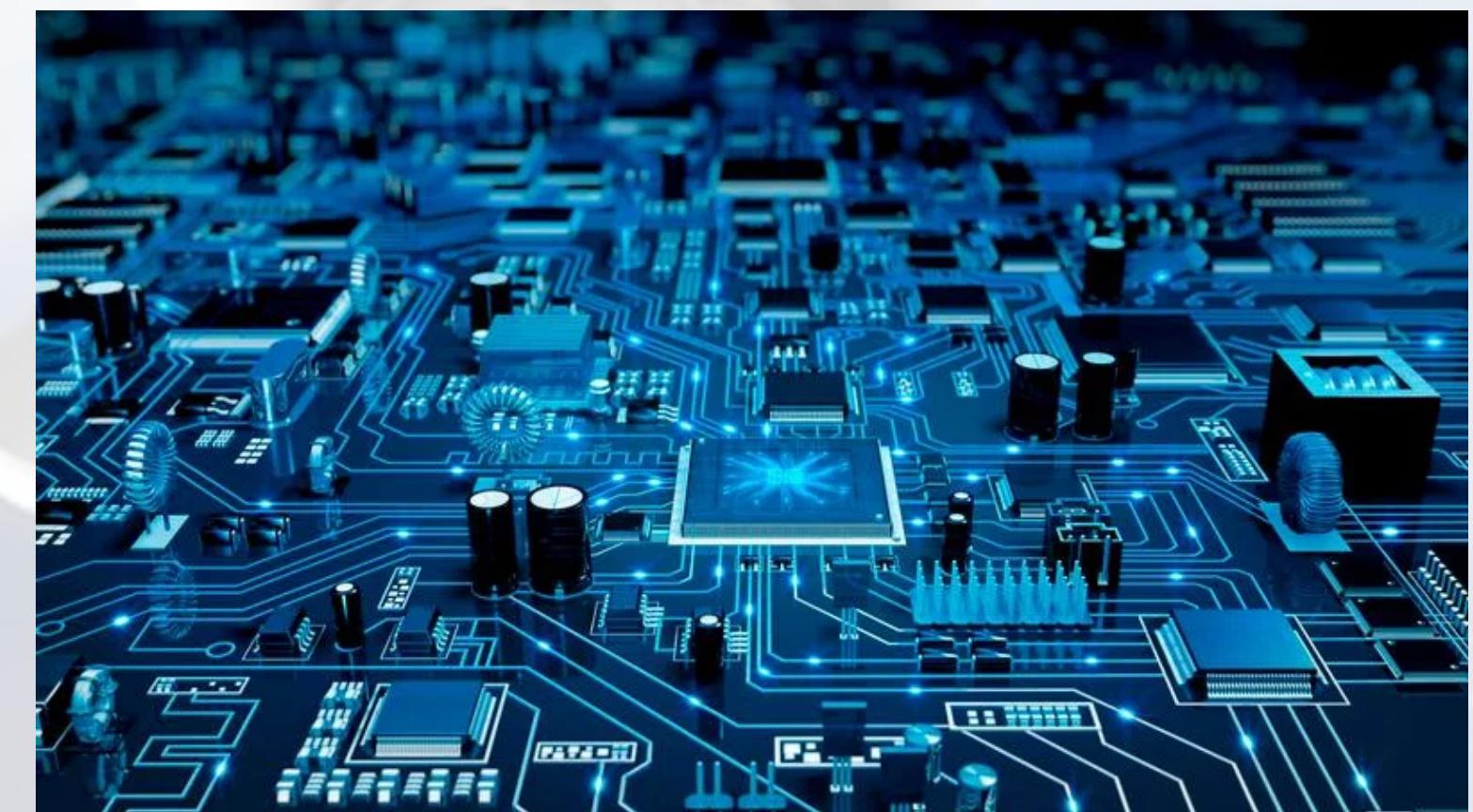
Three Pillars



Algorithm



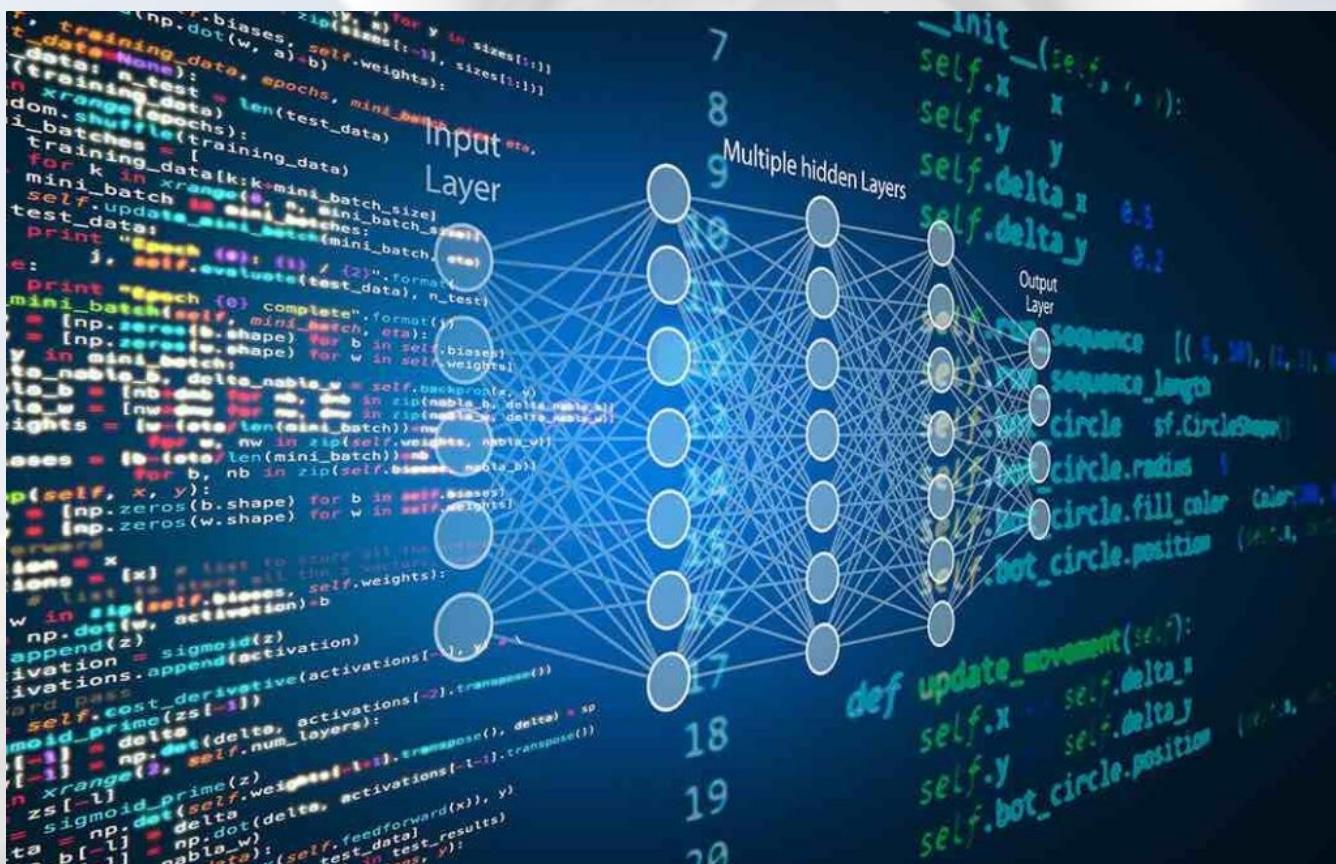
Data



Hardware

Deep Learning

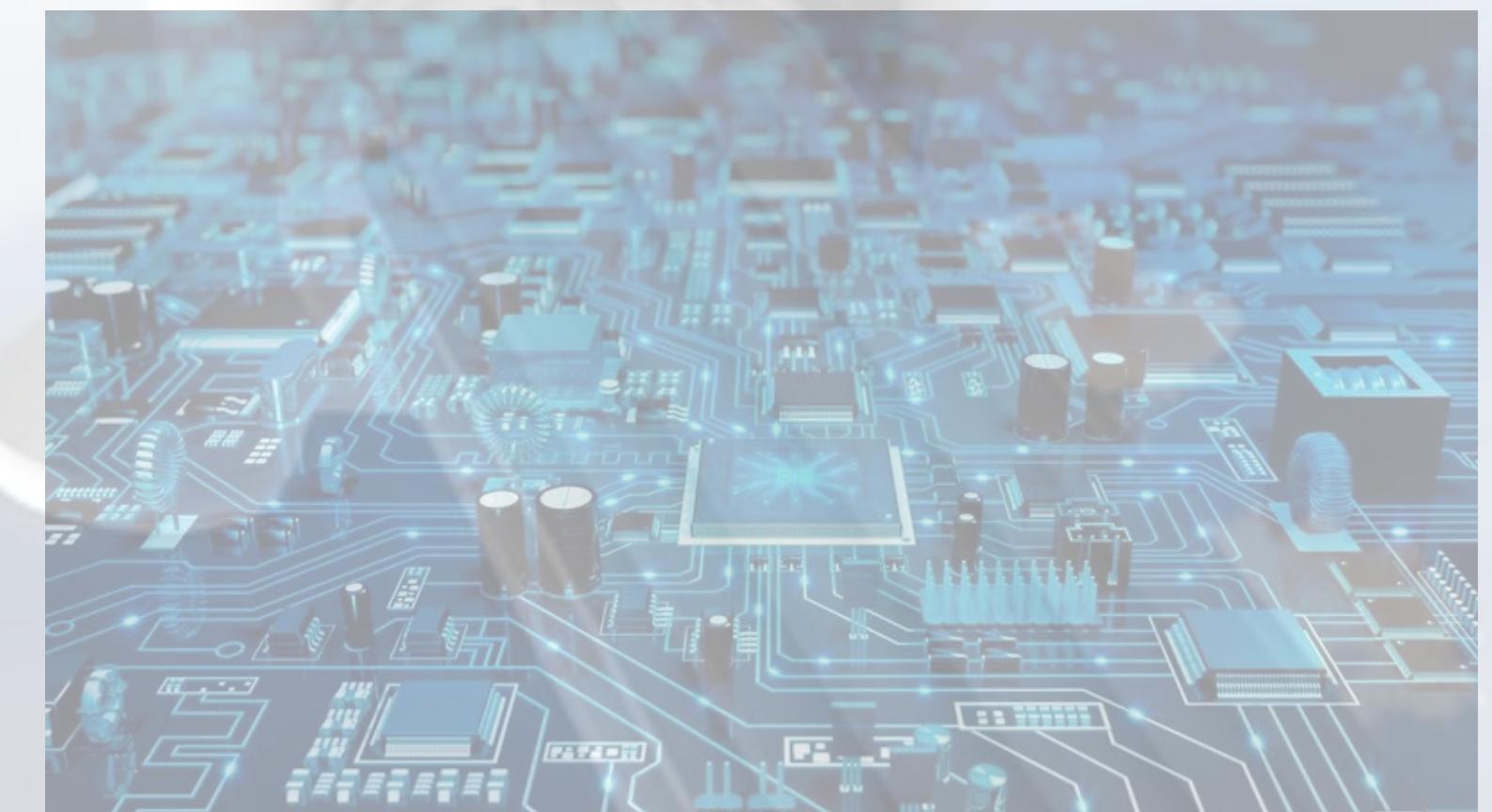
Three Pillars



Algorithm

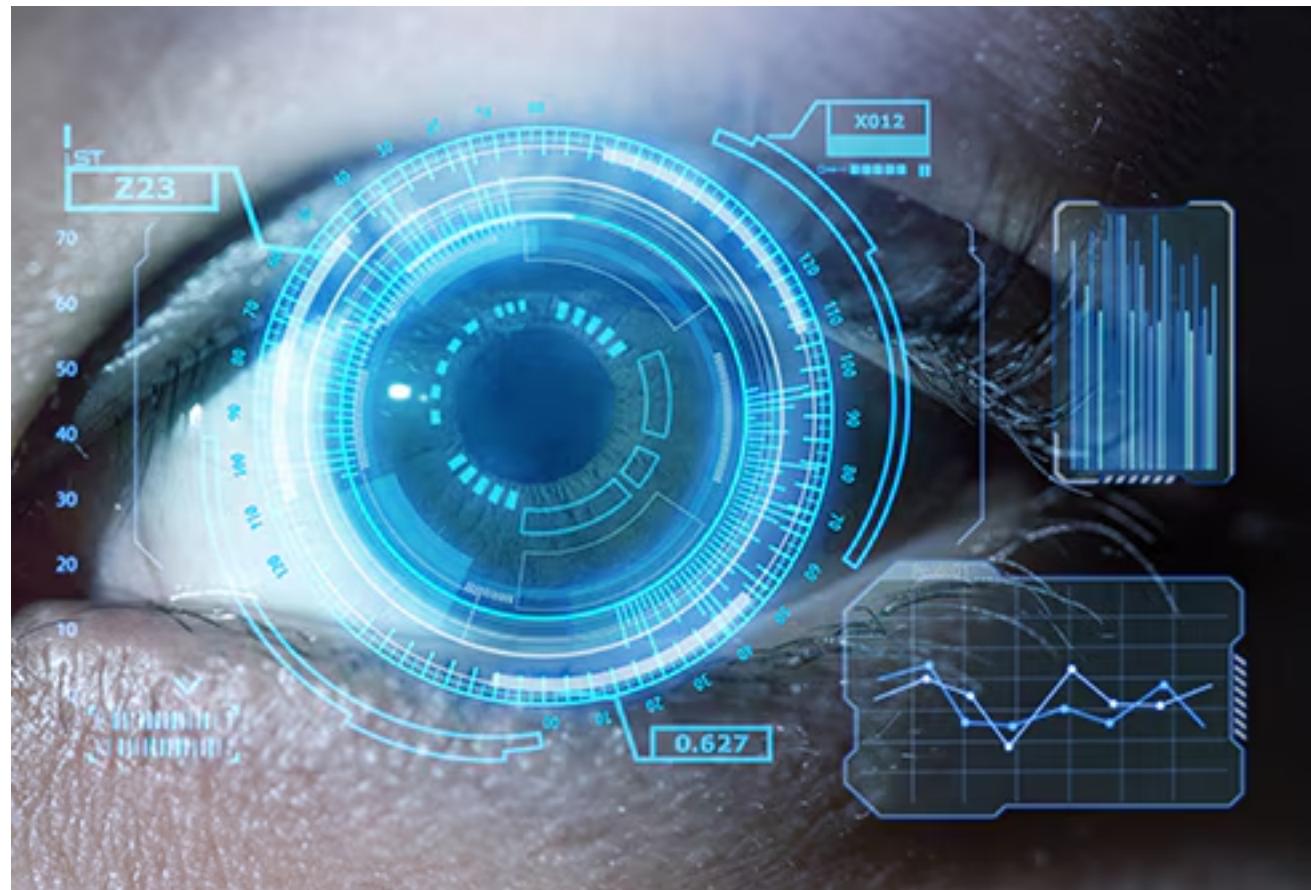


Data



Hardware

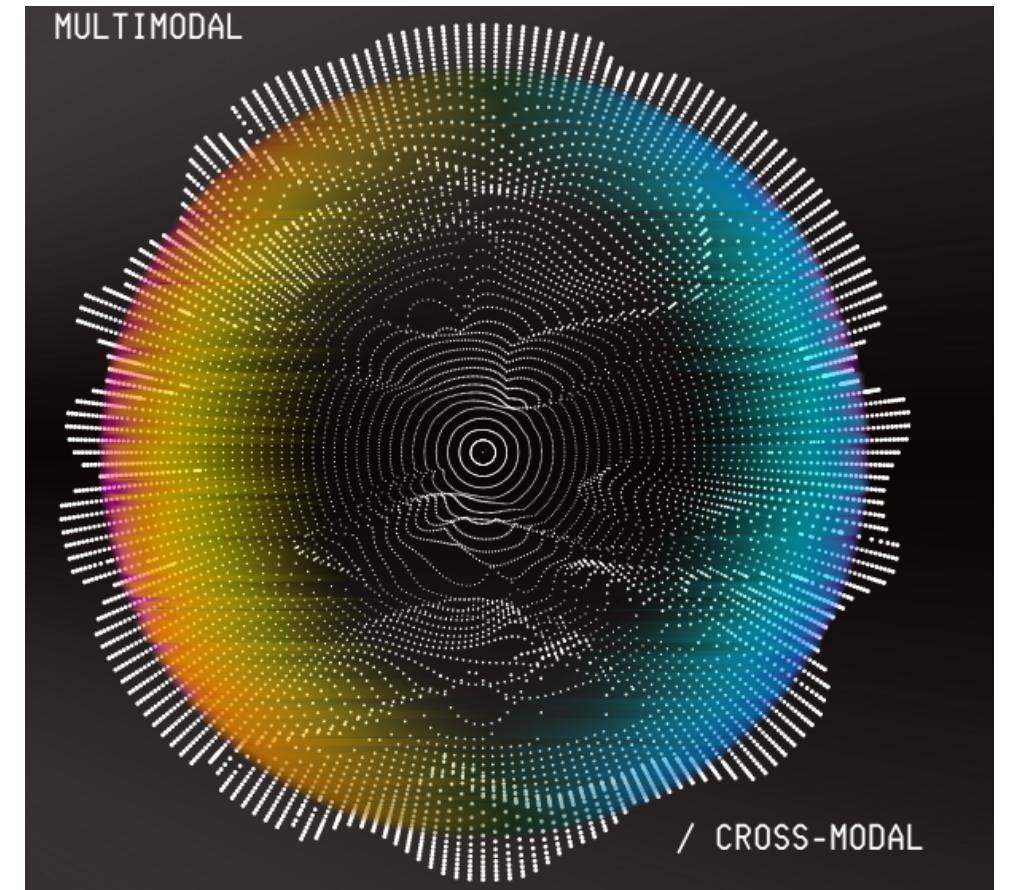
Deep Learning is Everywhere



Vision

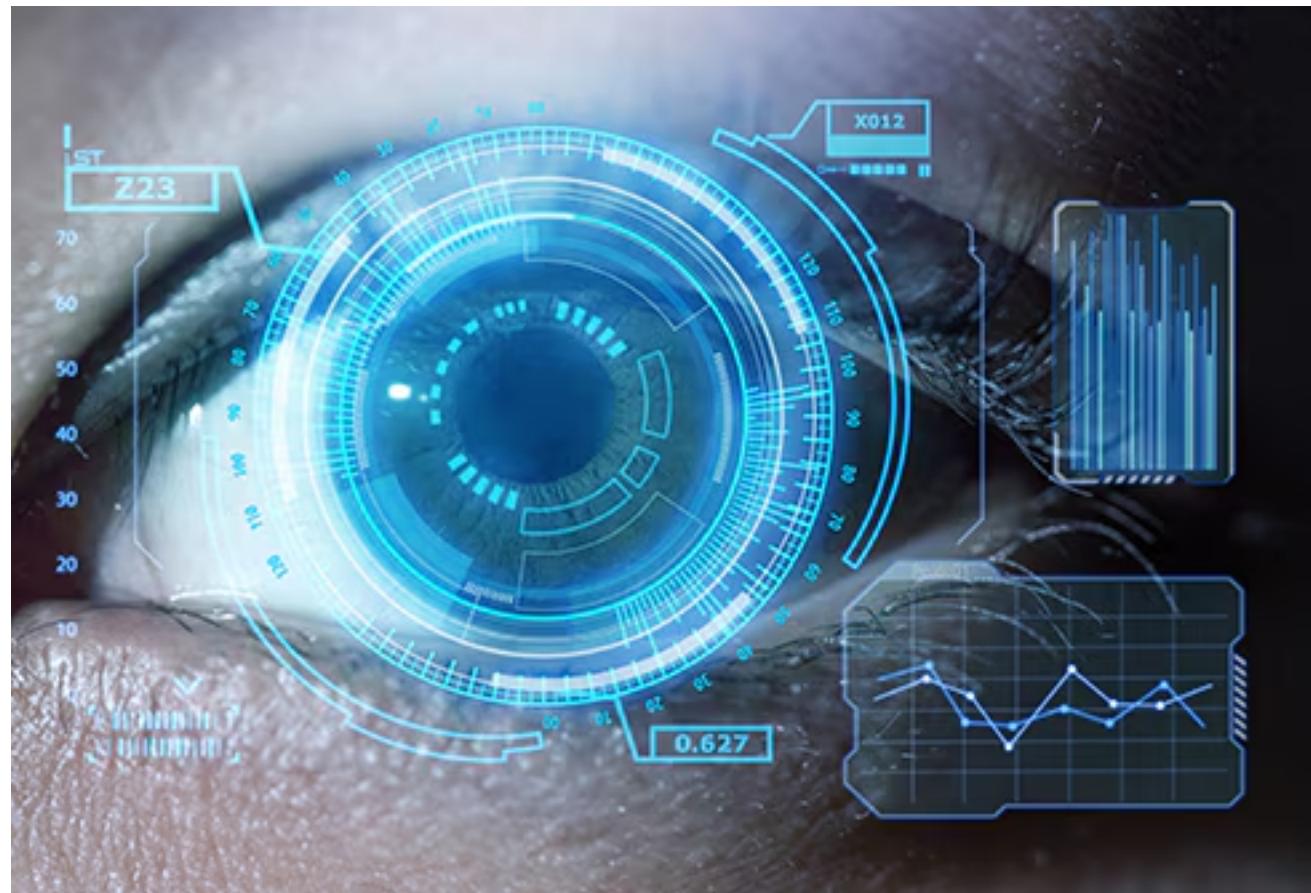


Language



Multimodal

Deep Learning is Everywhere



Vision



Language

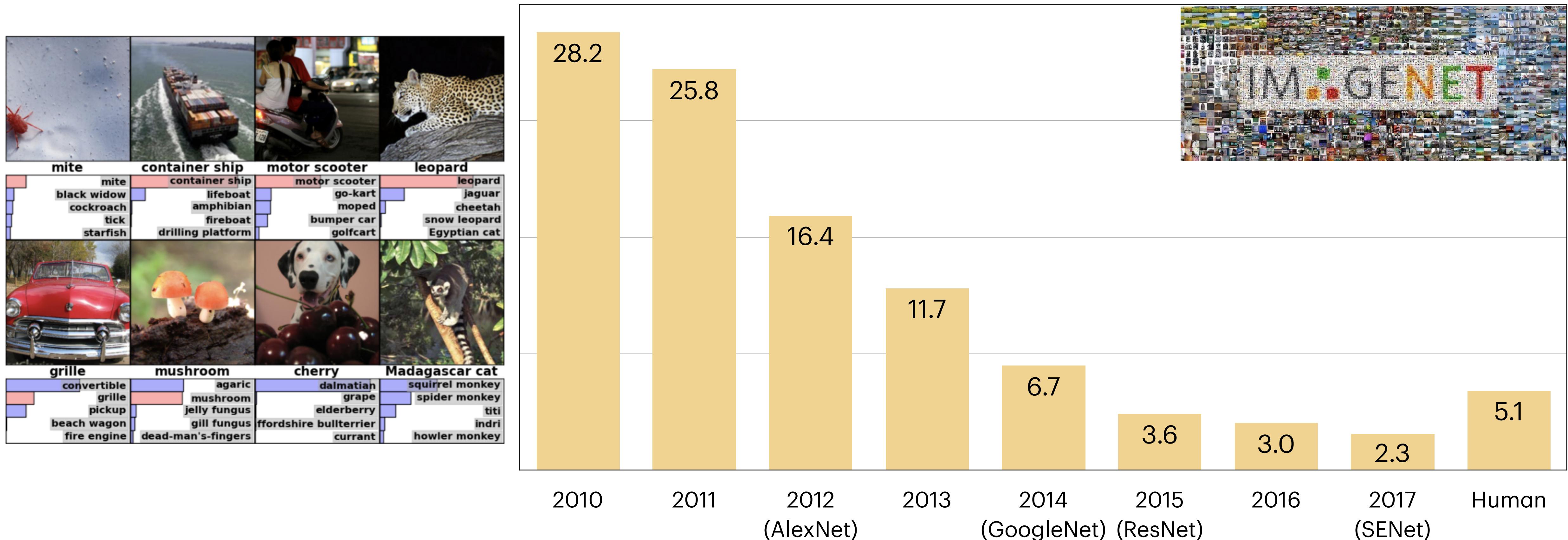


Multimodal

Deep Learning for Image Classification

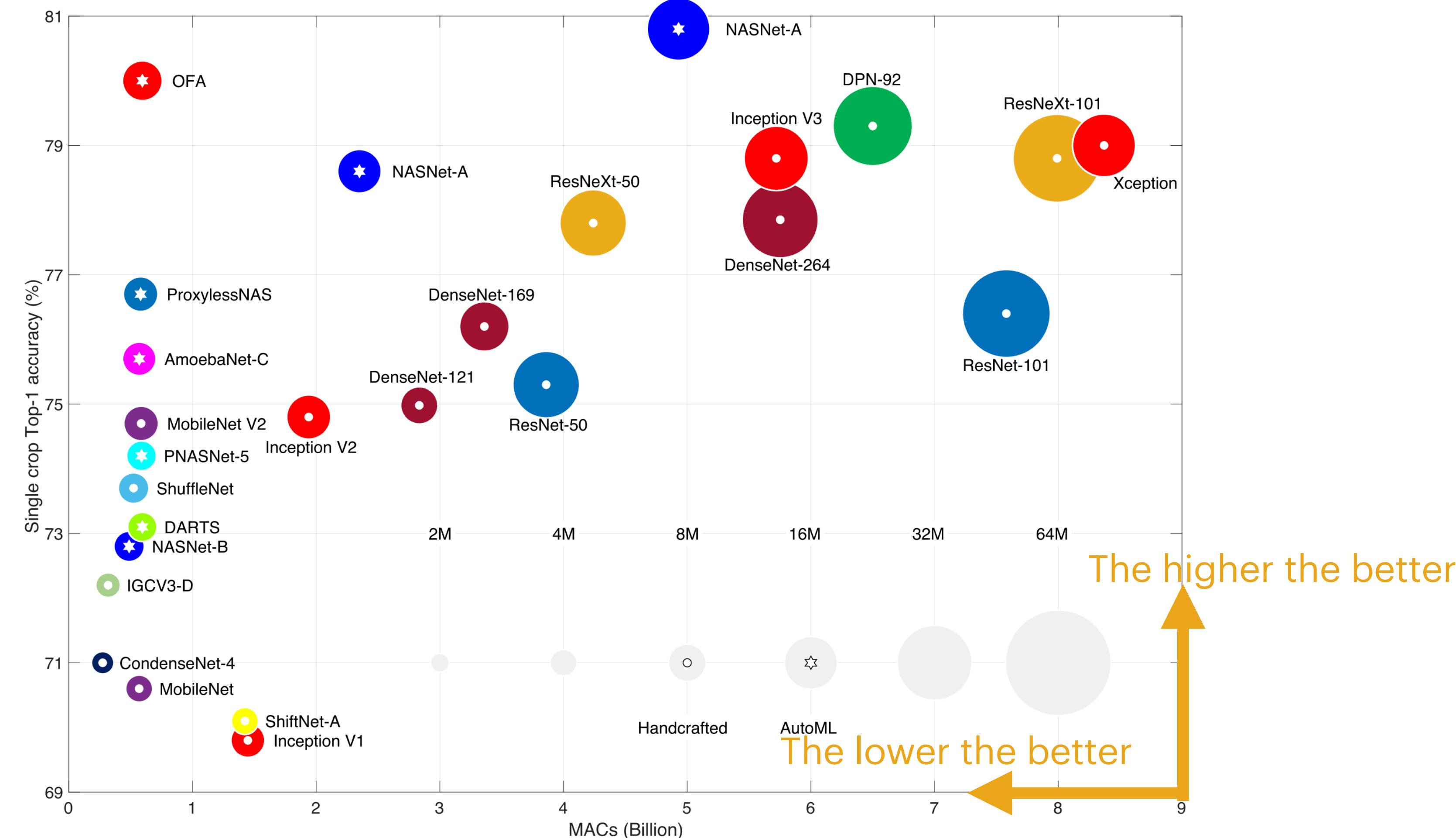
DNNs achieves super-human classification accuracy on ImageNet

ImageNet Contest Winning Entry: Top 5 Error Rate (%)



Efficient Image Classification

Neural architecture search reduces the computational cost



Deng, L., Li, G., Han, S., Shi, L., & Xie, Y. (2020). Model compression and hardware acceleration for neural networks: A comprehensive survey. Proceedings of the IEEE, 108(4), 485-532.

Efficient Image Classification

Efficient deep learning enables daily life application on mobile phones

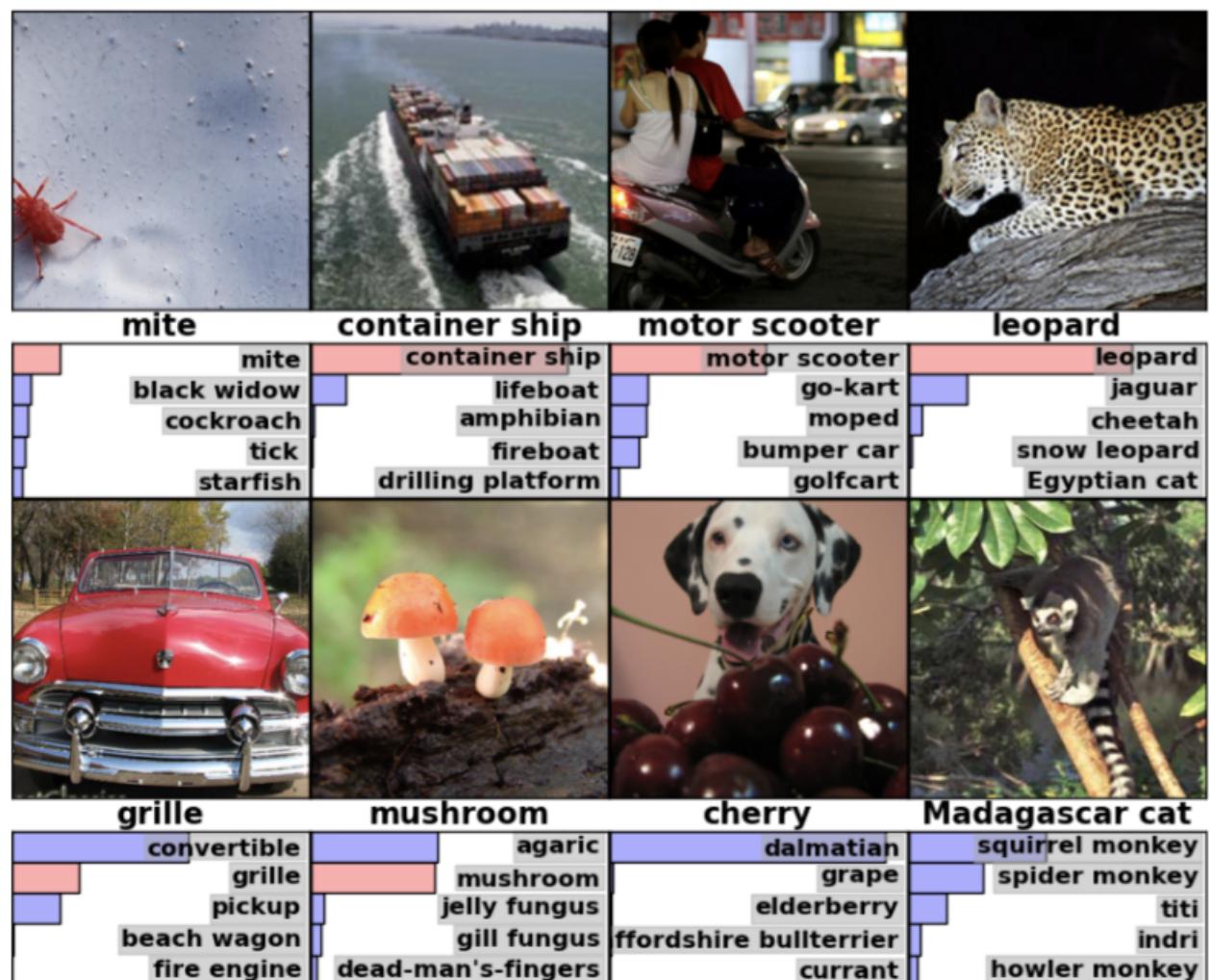


Image Classification

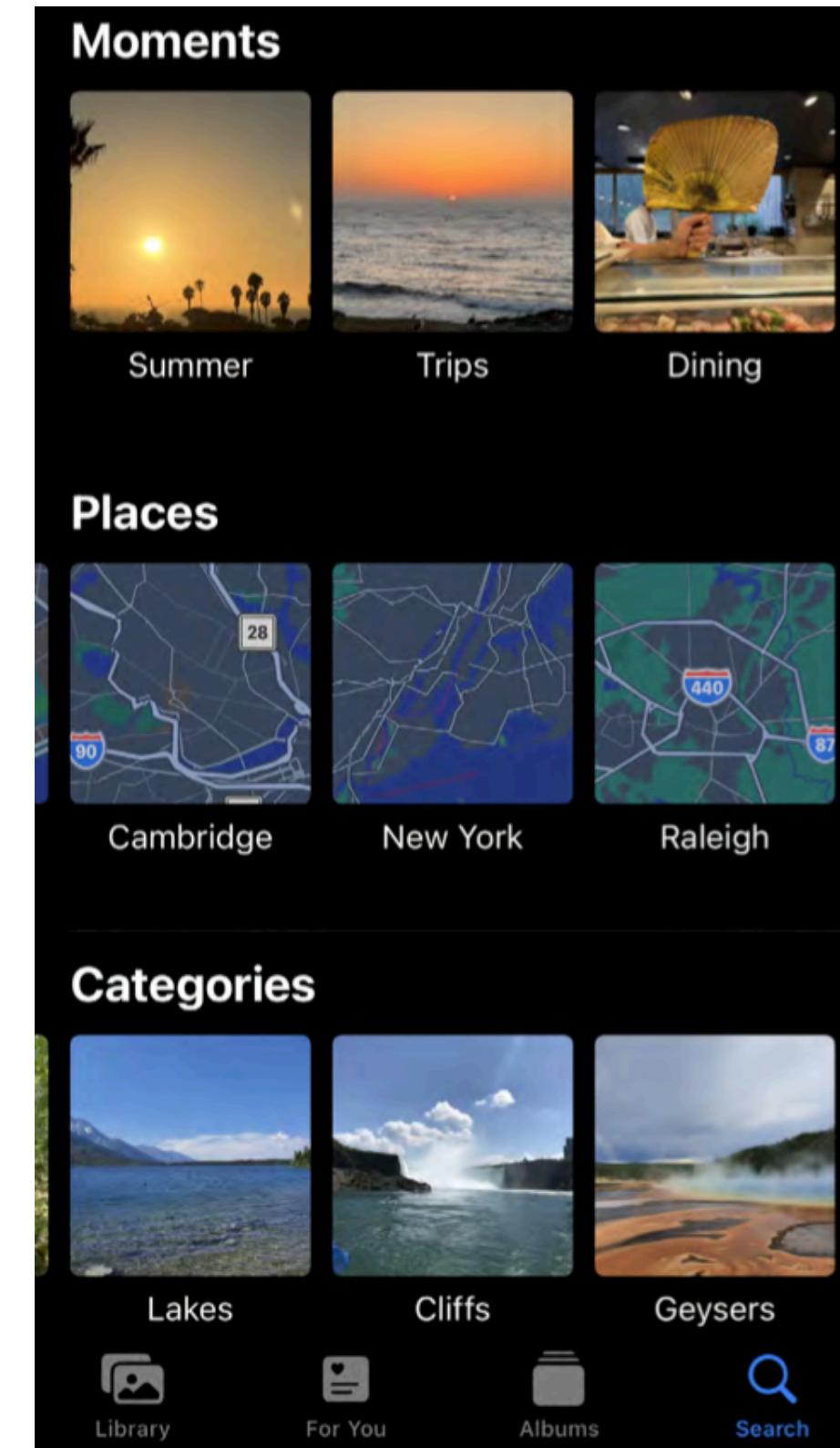


Photo Tags

Efficient Image Recognition

Efficient deep learning enables daily life application on mobile phones



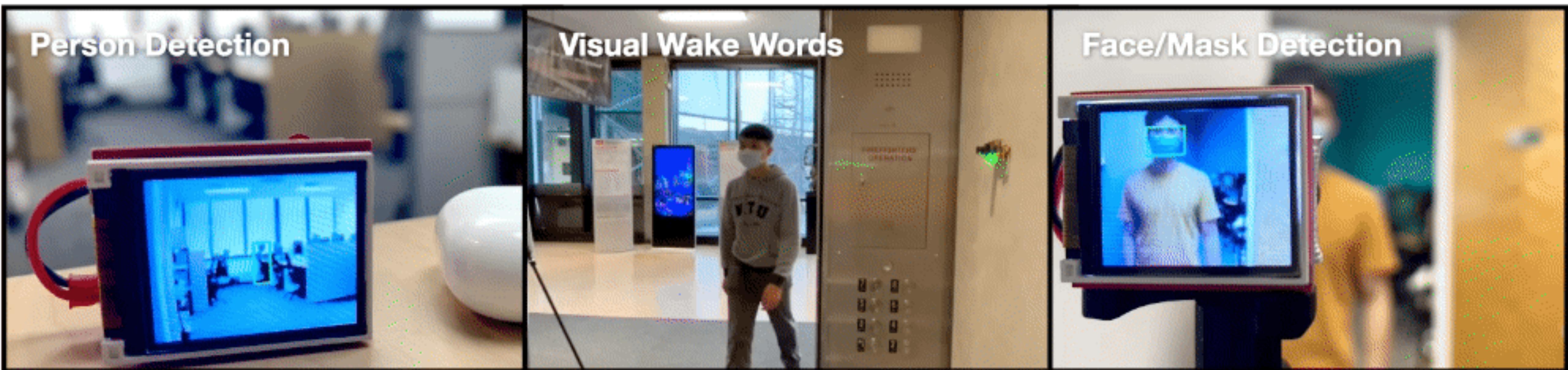
People Recognition

Recognizing People in Photos Through Private On-Device Machine Learning

Efficient Image Recognition on MCUs

MCUNet enables TinyML on IoT devices

- TinyNAS + TinyEngine
- ARM Cortex-M7 + OpenMV Cam



Lin, J., Chen, W. M., Lin, Y., Gan, C., & Han, S. (2020). Mcunet: Tiny deep learning on iot devices. Advances in neural information processing systems, 33, 11711-11722.

Segment Anything Model (SAM)

<https://segment-anything.com/>

- A new AI model from Meta AI that can "cut out" any object, in any image, with a single click



Prompt it with interactive points
and boxes

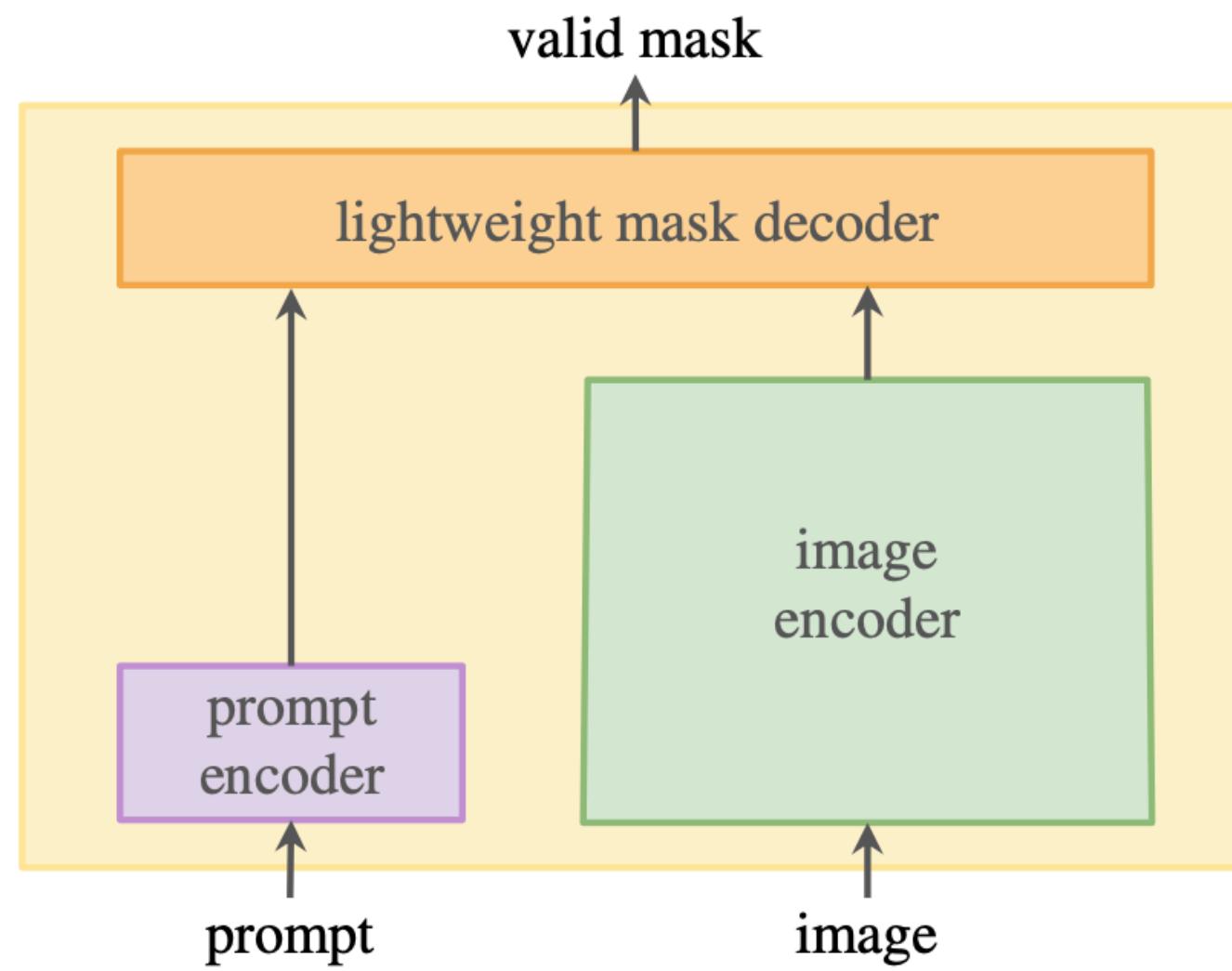


Automatically segment
everything in an image



Generate multiple valid masks
for ambiguous prompts

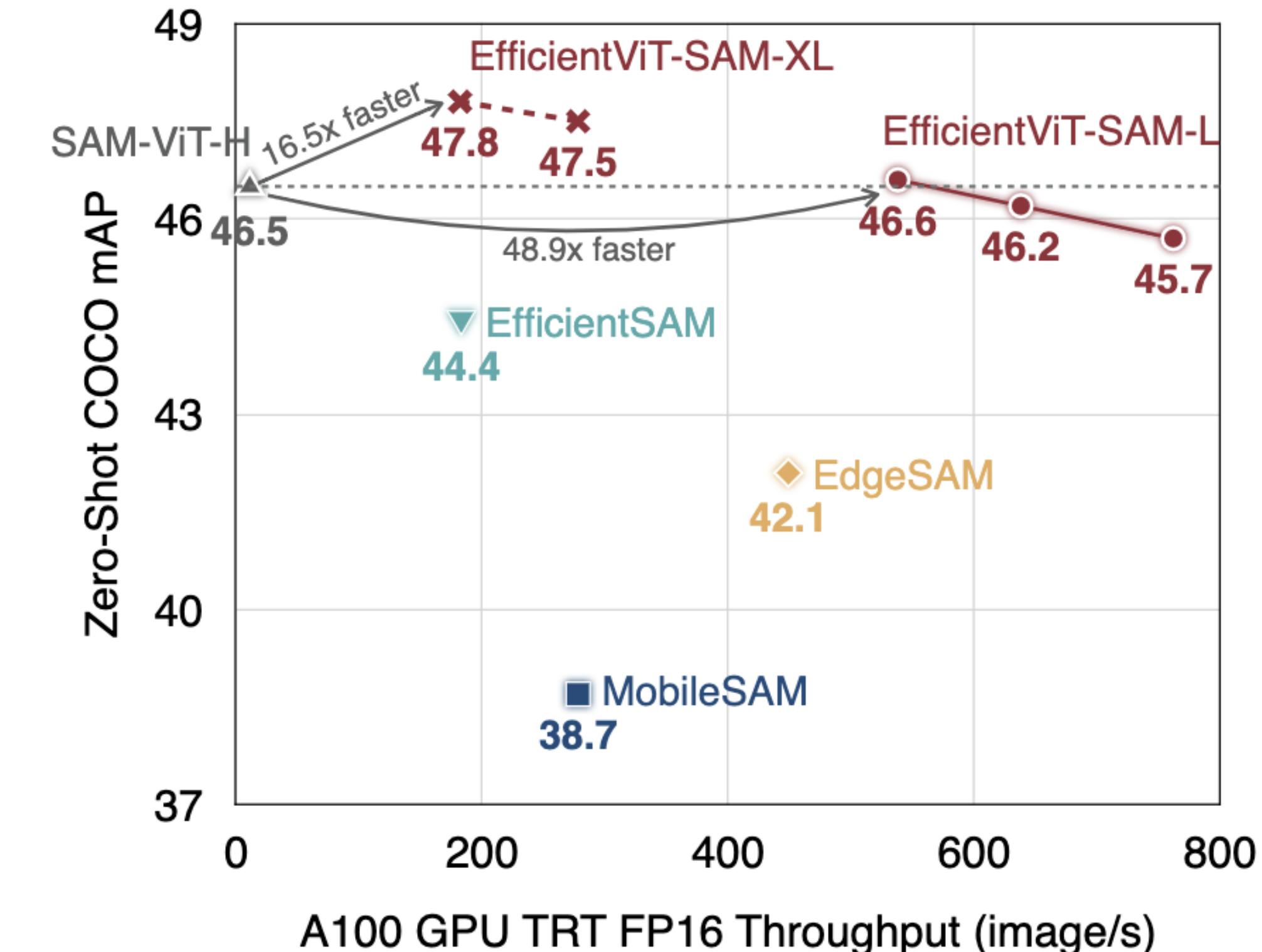
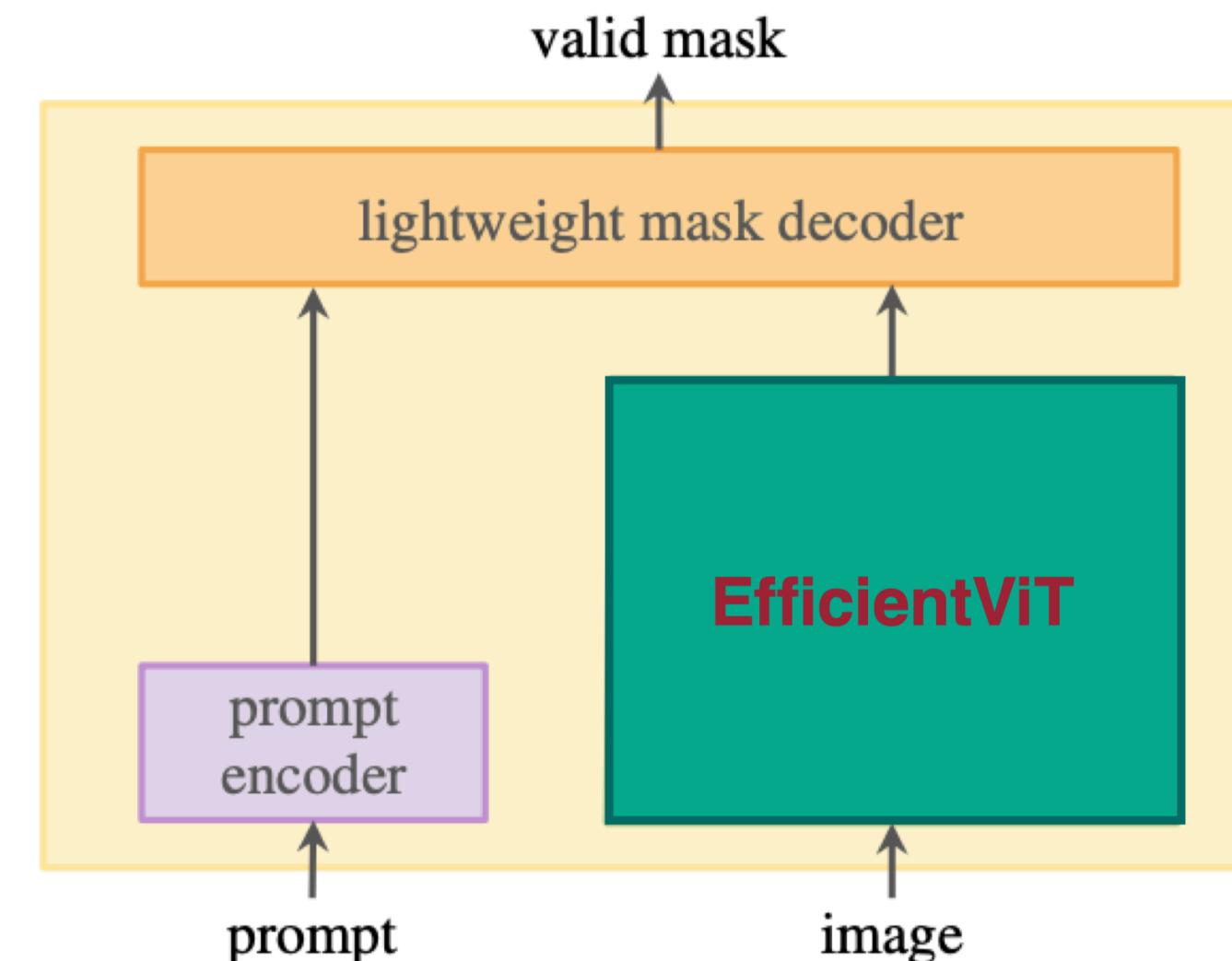
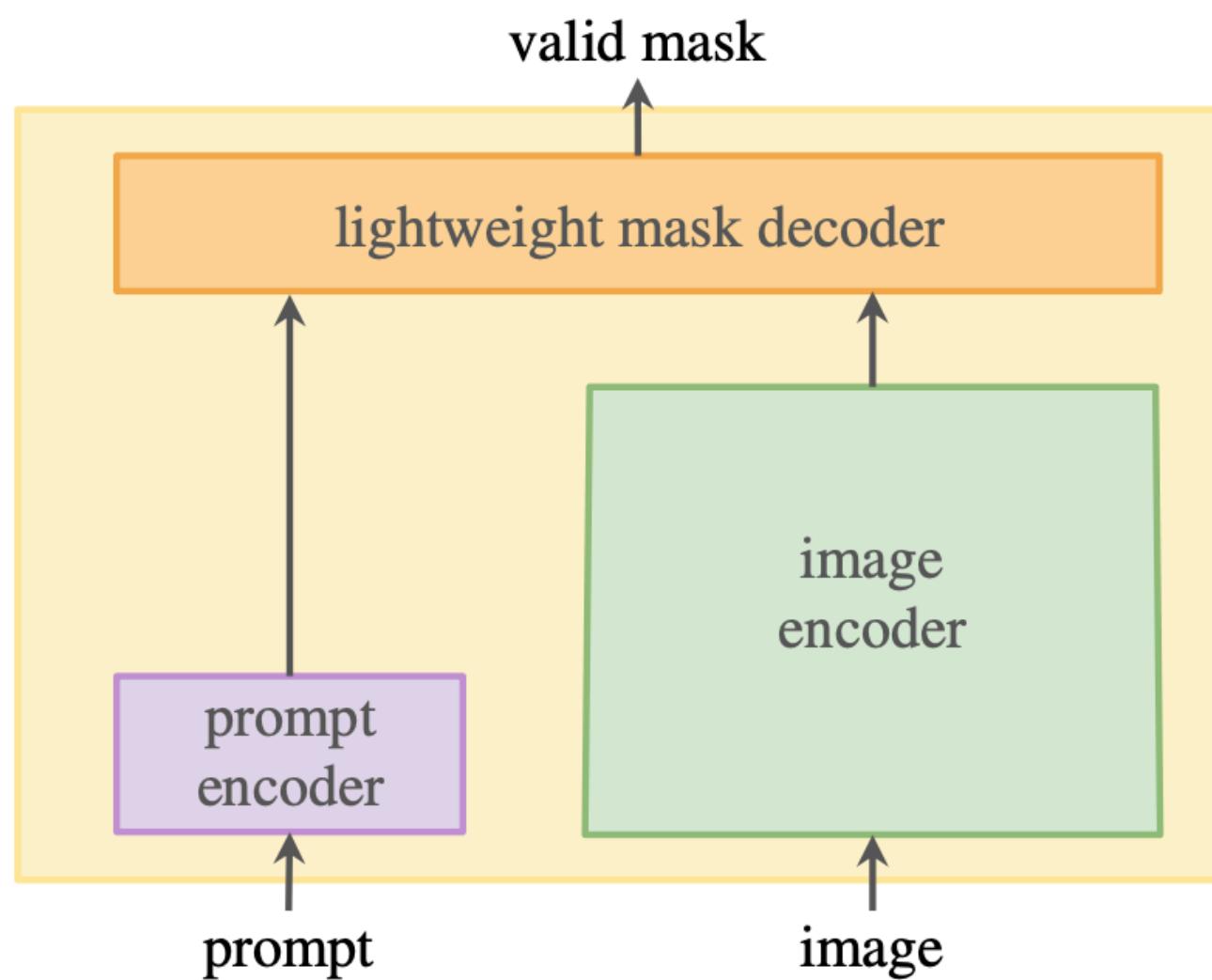
Promptable Image Segmentation



[Source](#)

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 4015-4026).

Accelerate SAM with EfficientViT



Source

- Match SAM-ViT-H's zero-shot segmentation performance with 48.9x speedup on A100 GPU.
- [EfficientViT-SAM Demo](#)

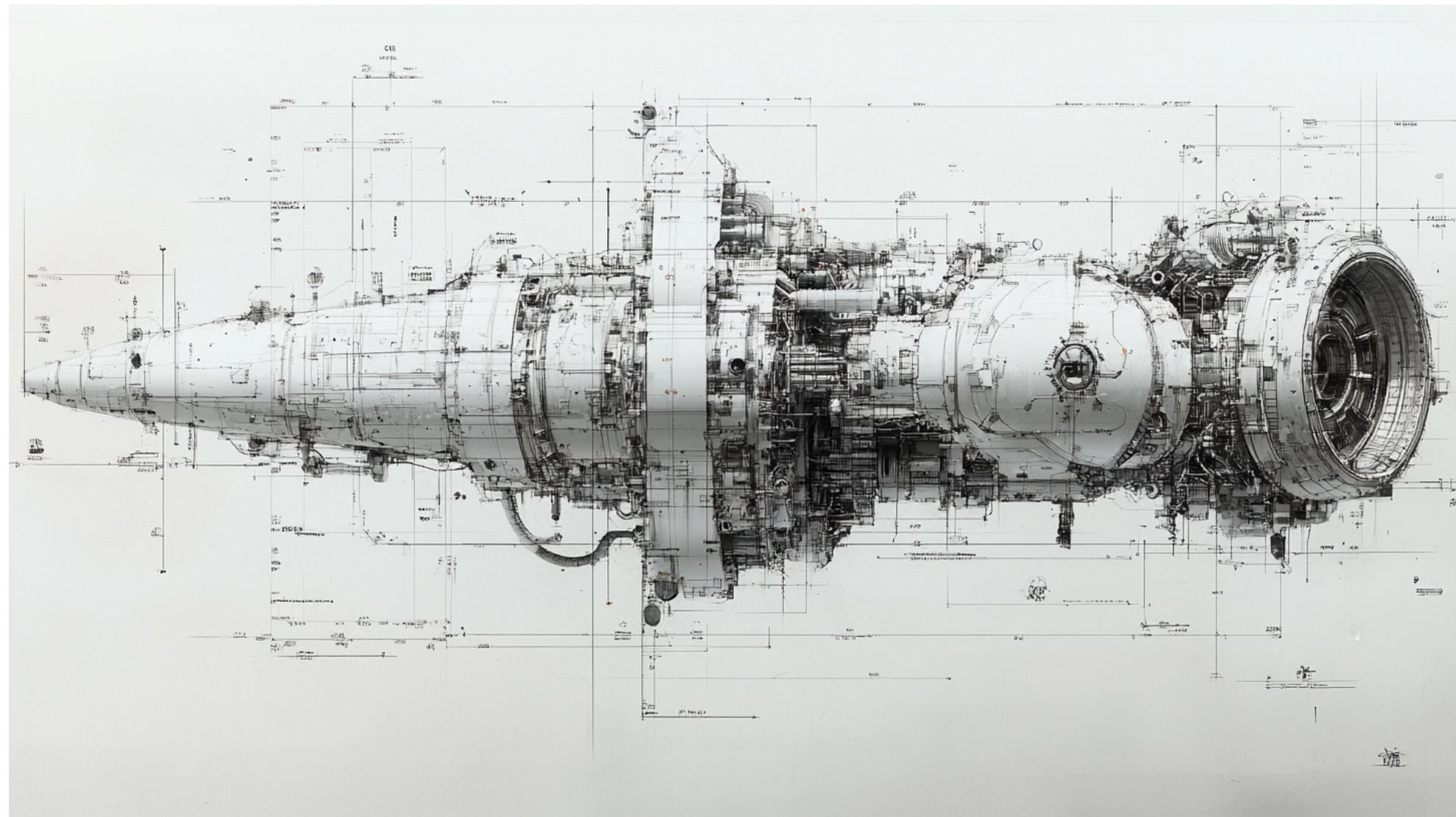
From Discrimitive Model to Generative Model

Diffusion models create realistic images from a natural language description

Healing Seven cats, stacked on top of each other, looking up, little a, childish and warm, Style cartoon, illustration, cute, healing



Exploded technical diagram of an enormous and mysterious futuristic machine. In the style of a blueprint.



Exploded technical diagram of an enormous and mysterious futuristic machine. In the style of a blueprint.

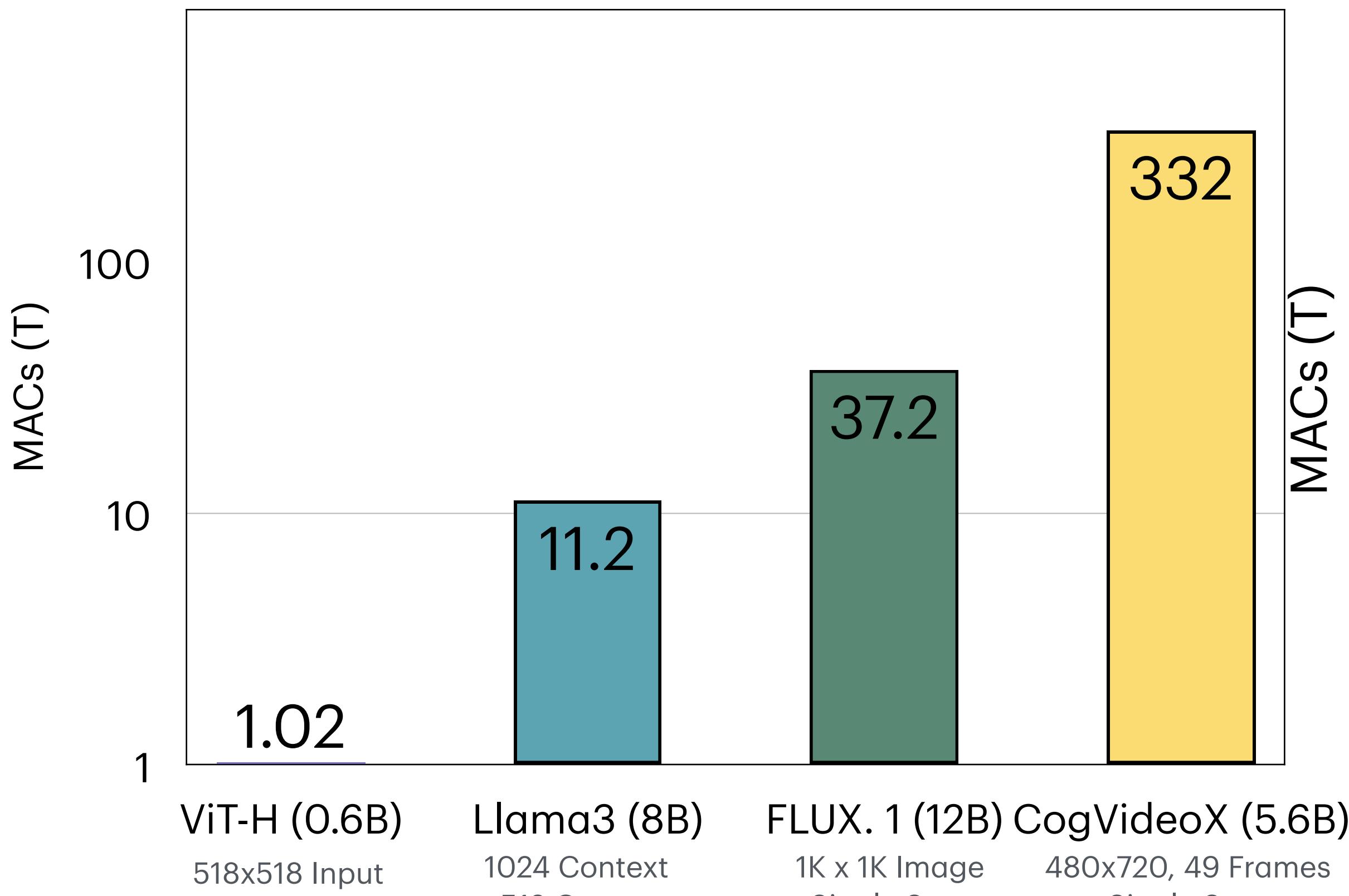


- Training stable diffusion costs **\$600,000 (256 A100s, 150k hours)**

Midjourney: <https://www.midjourney.com/>

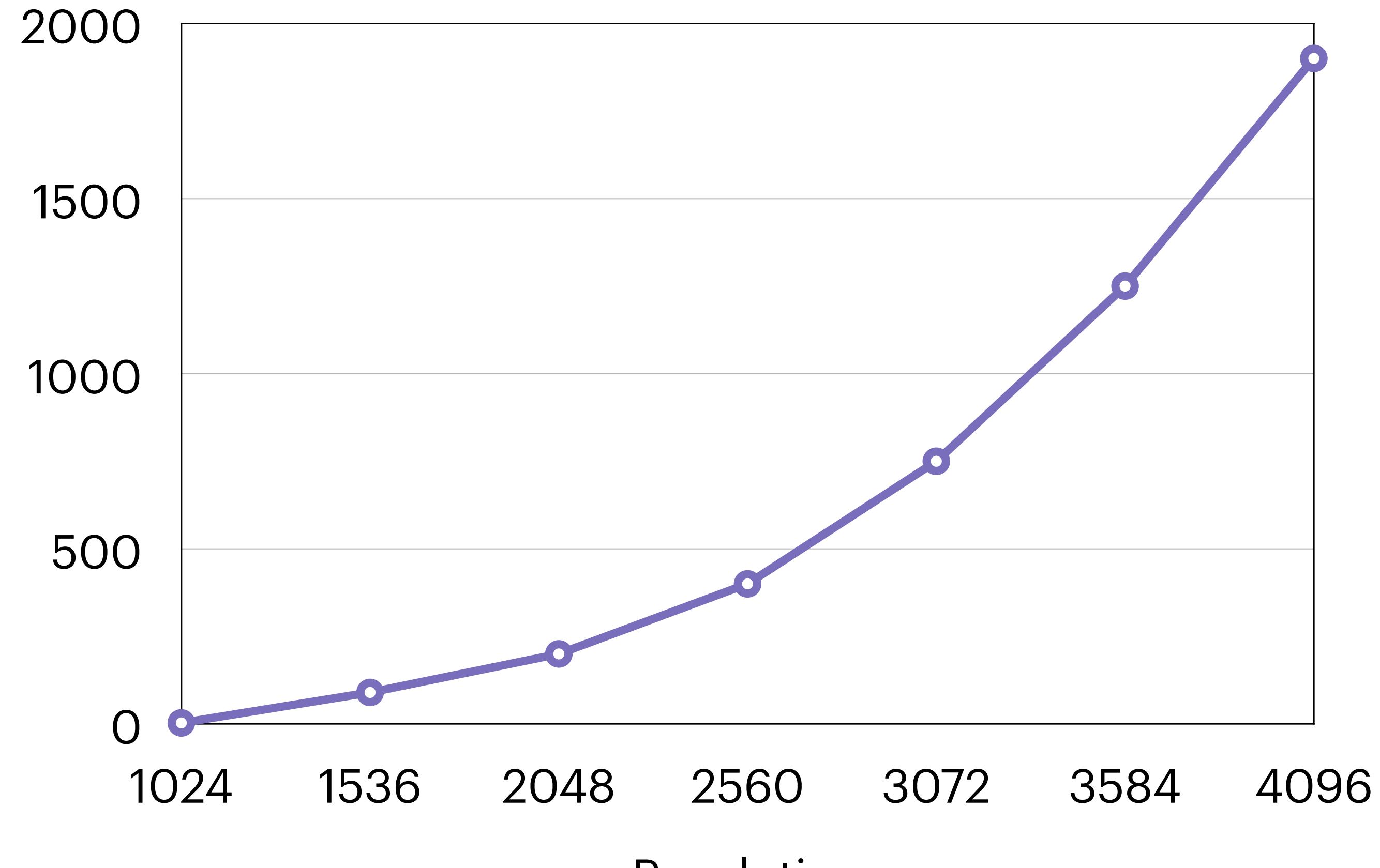
Generative Models are Computationally Exensive

Comparisons with other models



Diffusion models require more computation than other models

Comparisons across resolutions on FLUX.1



The computation scales super-quadratically with resolution.

Efficient Image Generation

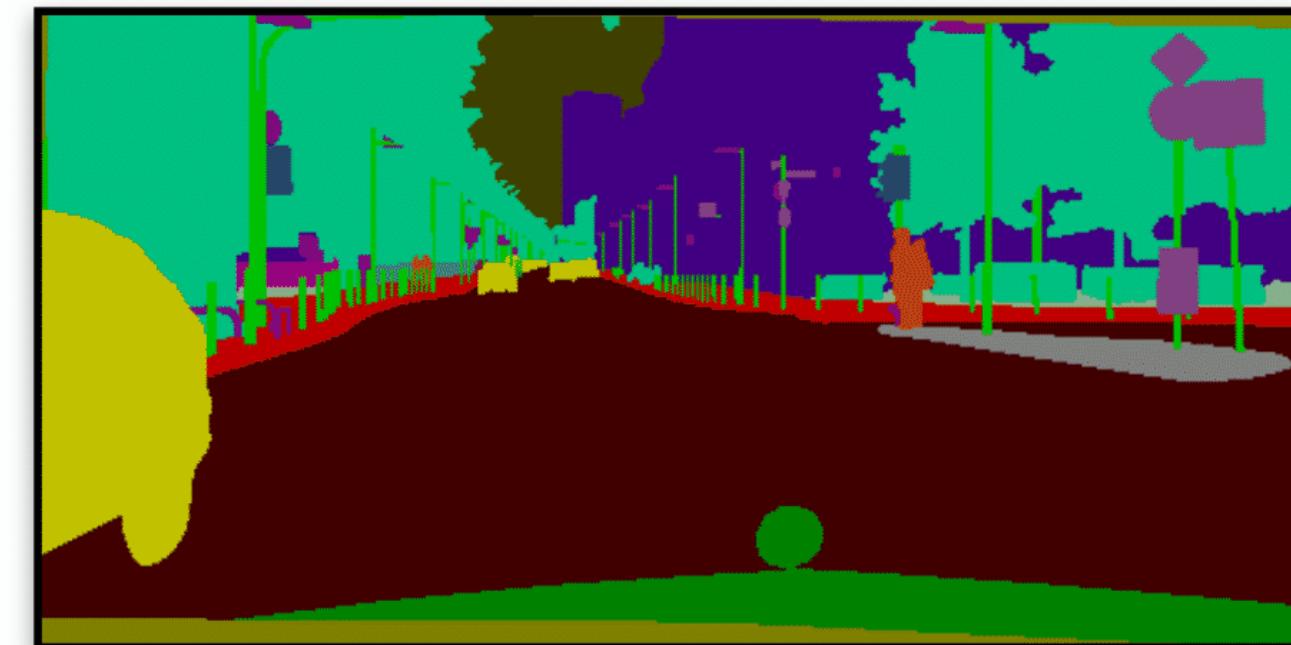
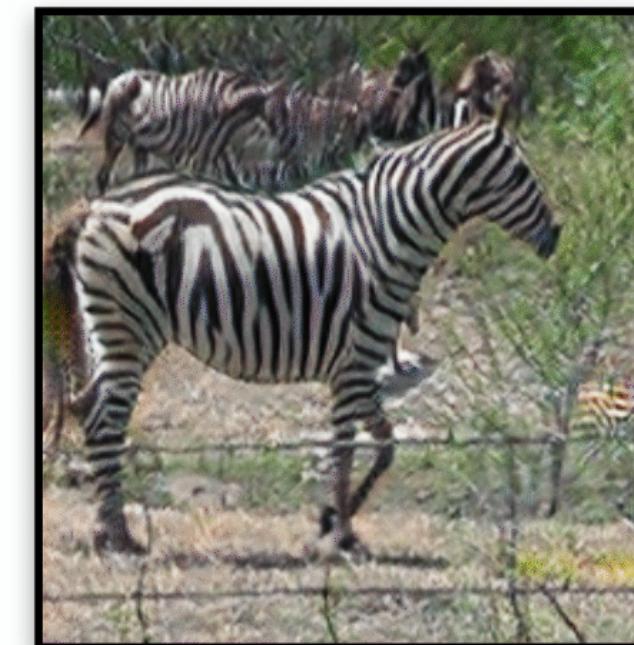
GAN Compression reduces the computation by 9-21X by pruning



Input



CycleGAN: 56.8G MACs Ours: 2.67G (**21.2×**)



Input



Ground Truth



Input



pix2pix: 56.8G MACs



Ours: 4.81G (**11.8×**)



GauGAN: 281G MACs

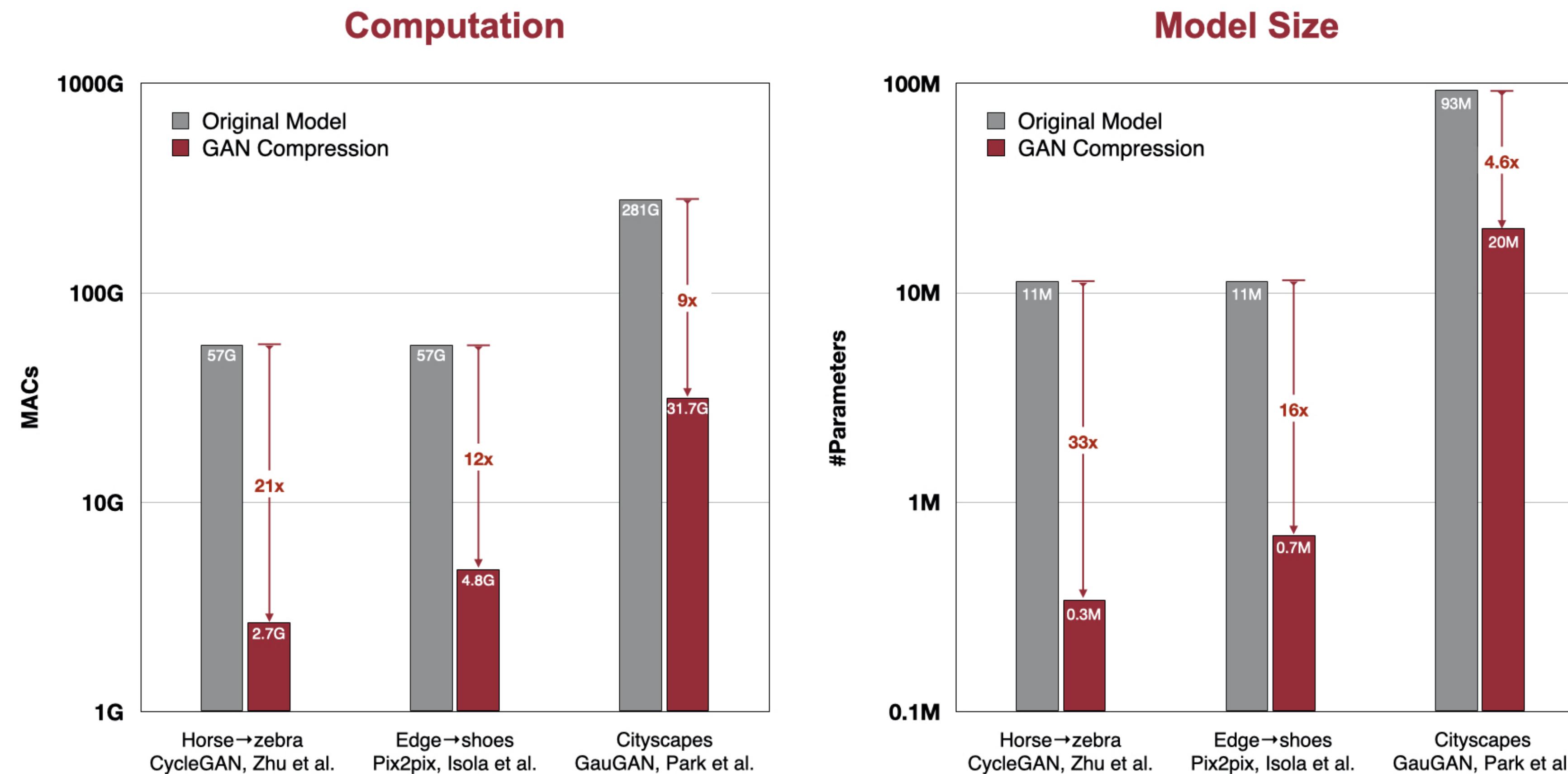


Ours: 31.7G (**8.8×**)

Li, M., Lin, J., Ding, Y., Liu, Z., Zhu, J. Y., & Han, S. (2020). GAN compression: Efficient architectures for interactive conditional gans. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5284-5294).

Efficient Image Generation

GAN Compression reduces the computation by 9-21X by pruning

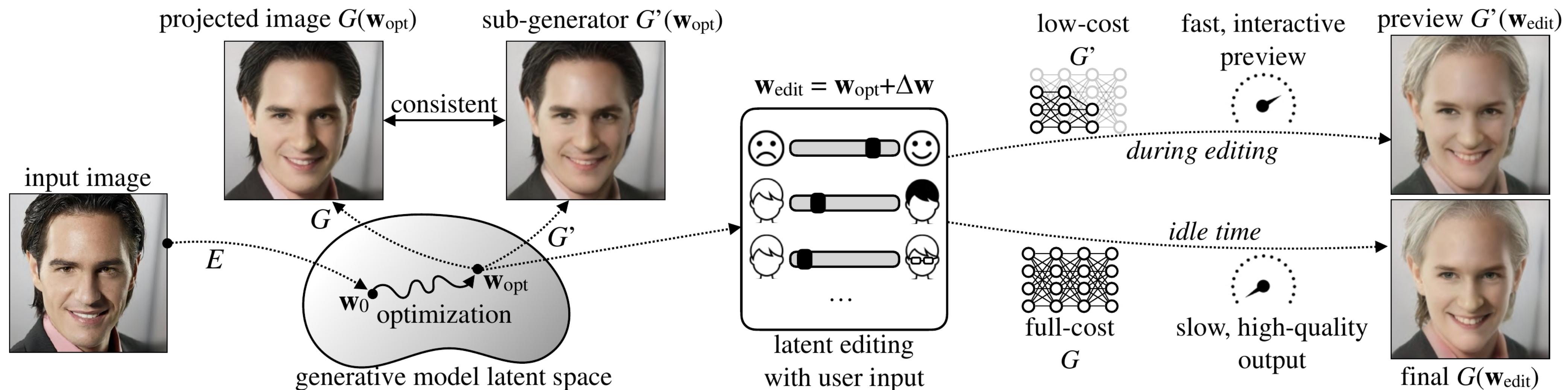


Li, M., Lin, J., Ding, Y., Liu, Z., Zhu, J. Y., & Han, S. (2020). GAN compression: Efficient architectures for interactive conditional gans. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5284-5294).

Efficient Image Generation

AnycostGAN achieves interactive image synthesis and editing on a laptop

- Generative Adversarial Network (GAN) is computationally heavy and slow
- Difficult for interactive photo editing on mobile device (iPad)
- Anycost GAN with once-for-all network: Train once, get (1) Small sub-net: low cost, fast prototyping; (2) Large sub-net: high-quality, finalization



Efficient Image Generation

AnycostGAN achieves interactive image synthesis and editing on a laptop



MACs:  100% 1.0x reduction

Anycost GAN

Compute Budget	1X	0.7X	0.5X	0.4X	0.2X	
Tiered Pricing	\$0.01	\$0.007	\$0.005	\$0.004	\$0.002	The quality is still reasonably good.

Efficient Image Generation

SIGE accelerates stable diffusion by >4X with spatial sparsity

Prompt: *A photograph of a horse on a grassland.*

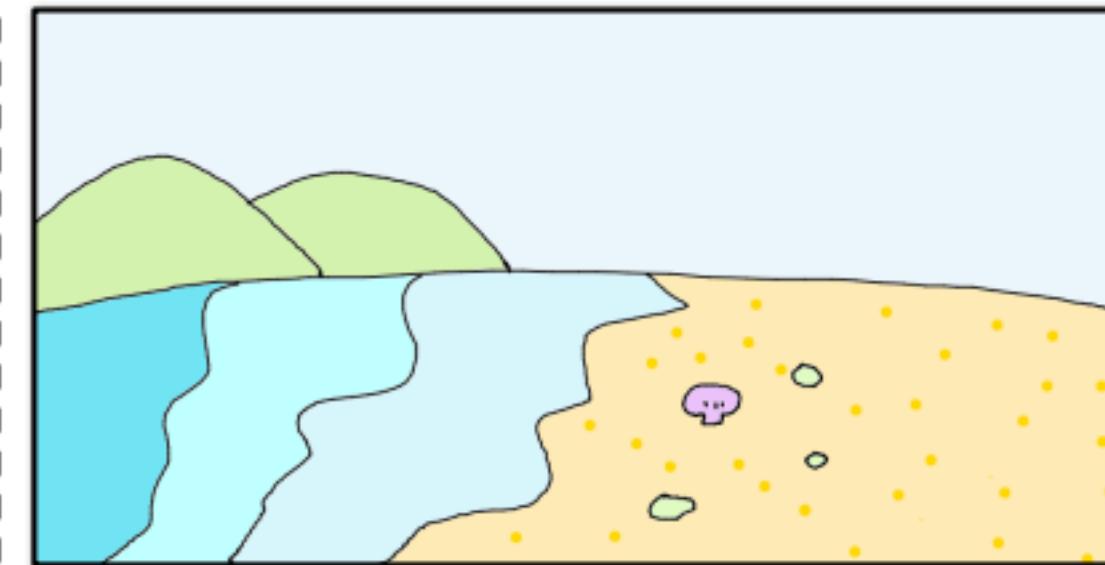


Original

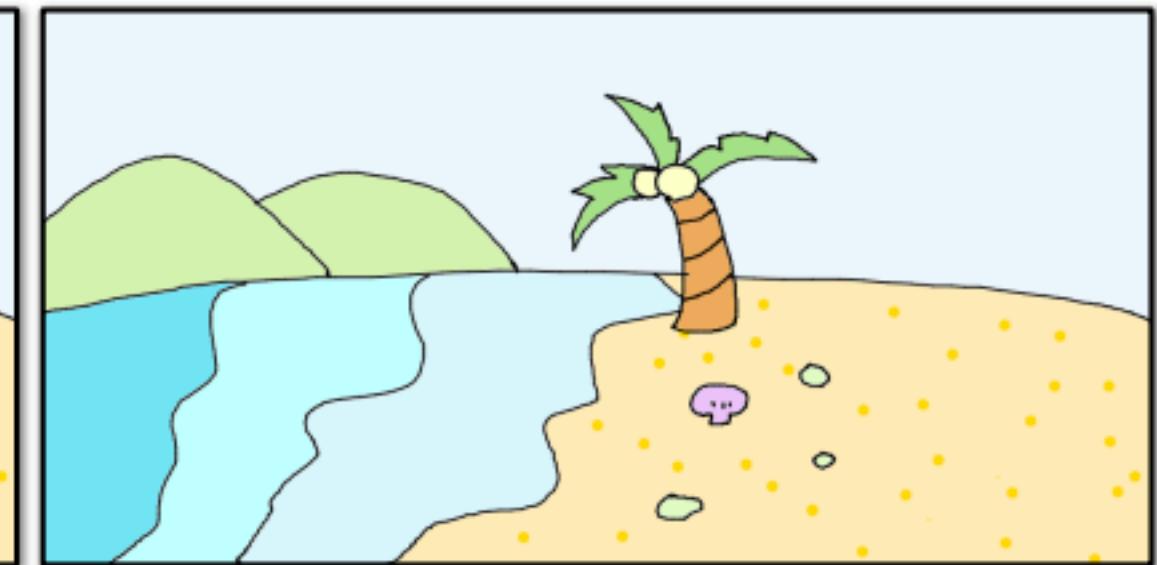


11.6% Masked

Prompt: *A fantasy beach landscape, trending on artstation.*



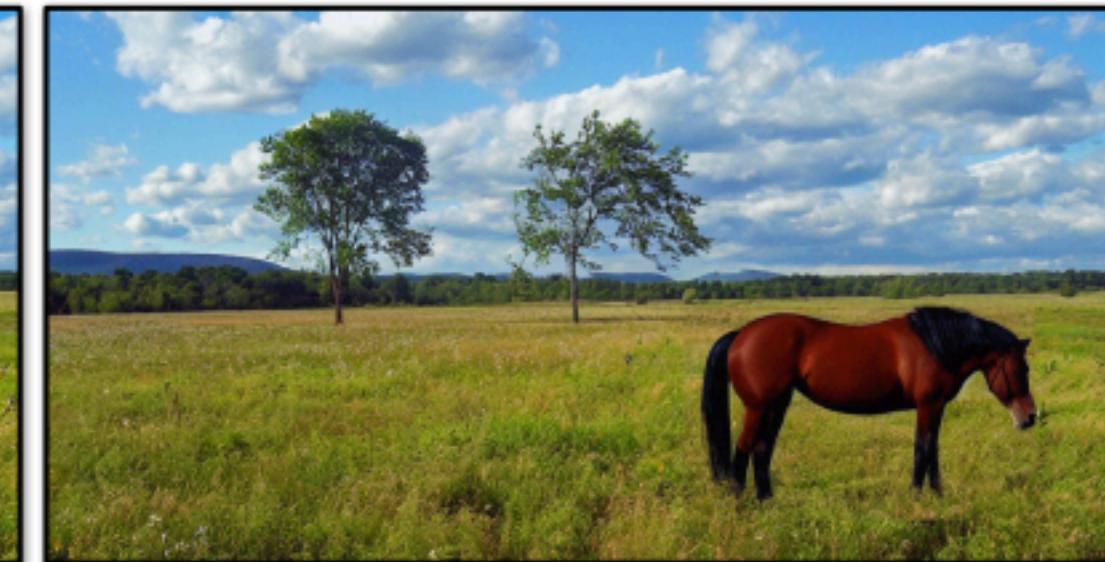
Original



2.8% Edited



Stable Diffusion:
1855GMACs 369ms



Ours:
514GMACs (3.6×) 95.0ms (3.9×)

(a) Image Inpainting



- Latency measured on NVIDIA RTX 3090

(b) Image Editing

Stable Diffusion+SDEdit:
1855GMACs 369ms

Ours:
353GMACs (5.3×) 76.4ms (4.8×)

Efficient Image Generation

FastComposer achieves tuning-free multi-subject image generation



Xiao, G., Yin, T., Freeman, W. T., Durand, F., & Han, S. (2024). Fastcomposer: Tuning-free multi-subject image generation with localized attention. International Journal of Computer Vision, 1-20.

Efficient Image Generation

DistriFusion accelerates high-resolution generation with distributed parallelism

Original
MACs: 907T
Latency: 12.3s



Naïve Patch (4 Devices)
MACs Per Device: 190T (4.8× Less)
Latency: 3.14s (3.9× Faster)



Ours (4 Devices)
MACs Per Device: 227T (4.0× Less)
Latency: 4.16s (3.0× Faster)



Ours (8 Devices)
MACs Per Device: 113T (8.0× Less)
Latency: 2.74s (4.5× Faster)



Prompt: *Ethereal fantasy concept art of an elf, magnificent, celestial, ethereal, painterly, epic, majestic, magical, fantasy art, cover art, dreamy.*



Prompt: *Romantic painting of a ship sailing in a stormy sea, with dramatic lighting and powerful waves.*

Li, M., Cai, T., Cao, J., Zhang, Q., Cai, H., Bai, J., ... & Han, S. (2024). Distrifusion: Distributed parallel inference for high-resolution diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7183-7193).

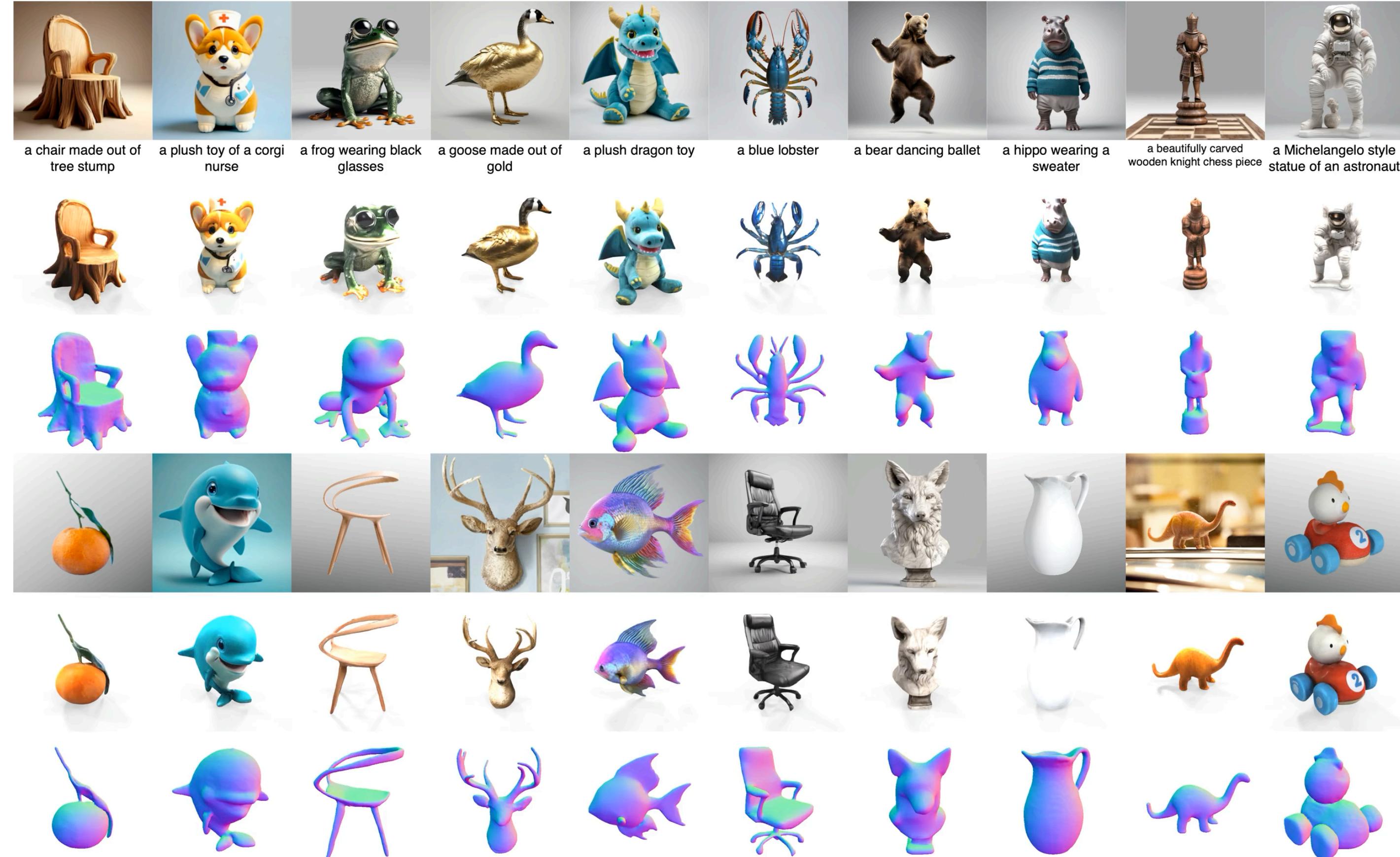
Efficient Image Generation

- Apple Intelligence



3D Generation

Diffusion models create 3D objects from a reference image

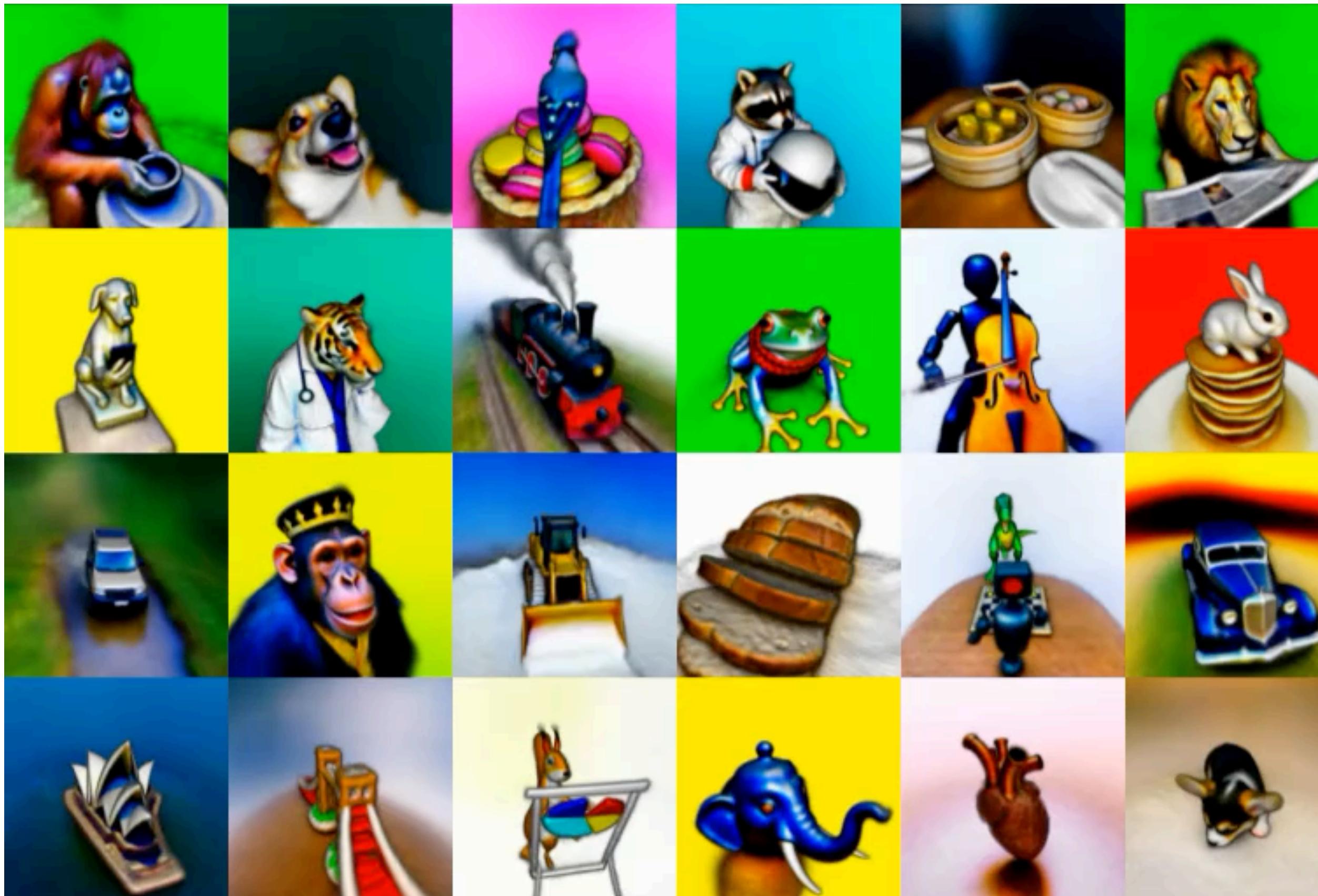


One-2-3-45++

Liu, M., Shi, R., Chen, L., Zhang, Z., Xu, C., Wei, X., ... & Su, H. (2024). One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10072-10083).

3D Generation

Diffusion models create 3D objects from a natural language description



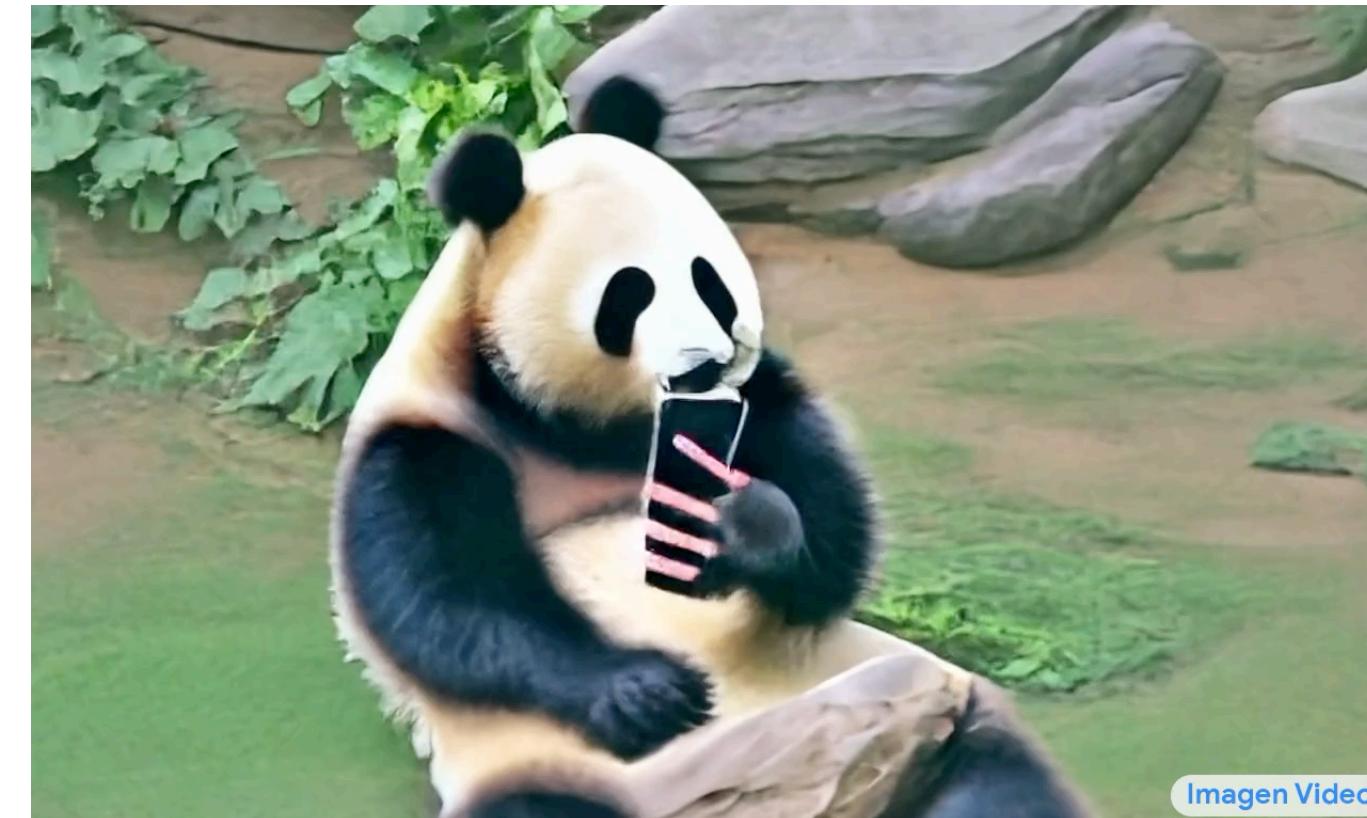
<https://dreamfusion3d.github.io/>

Video Generation

Diffusion models create realistic images from a natural language description



An umbrella on top of a spoon.



A panda taking a selfie



A cat eating food out of a bowl, in the style of Van Gogh

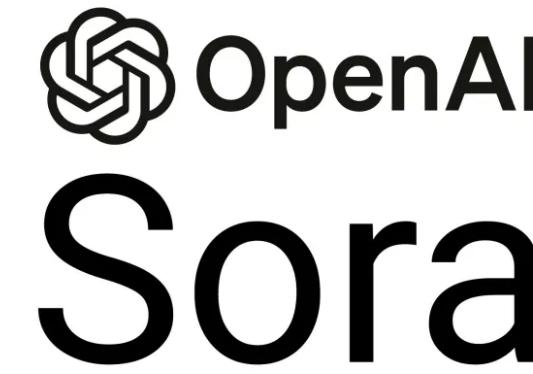
- Video modeling is a harder task for which performance is not yet saturated at **5.6B model size.**

Imagen Video: <https://imagen.research.google/video/>

Video Generation

Diffusion models create realistic videos from a natural language description

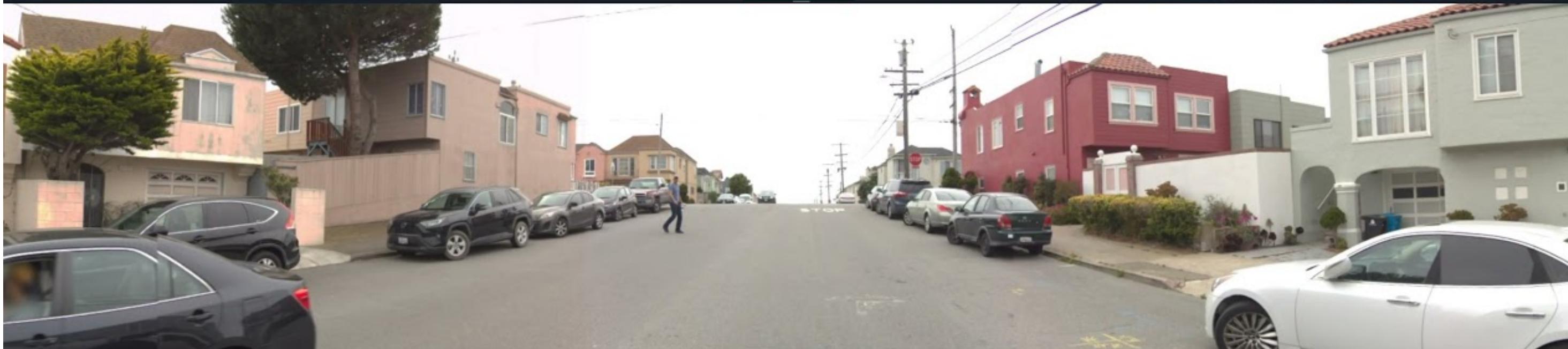
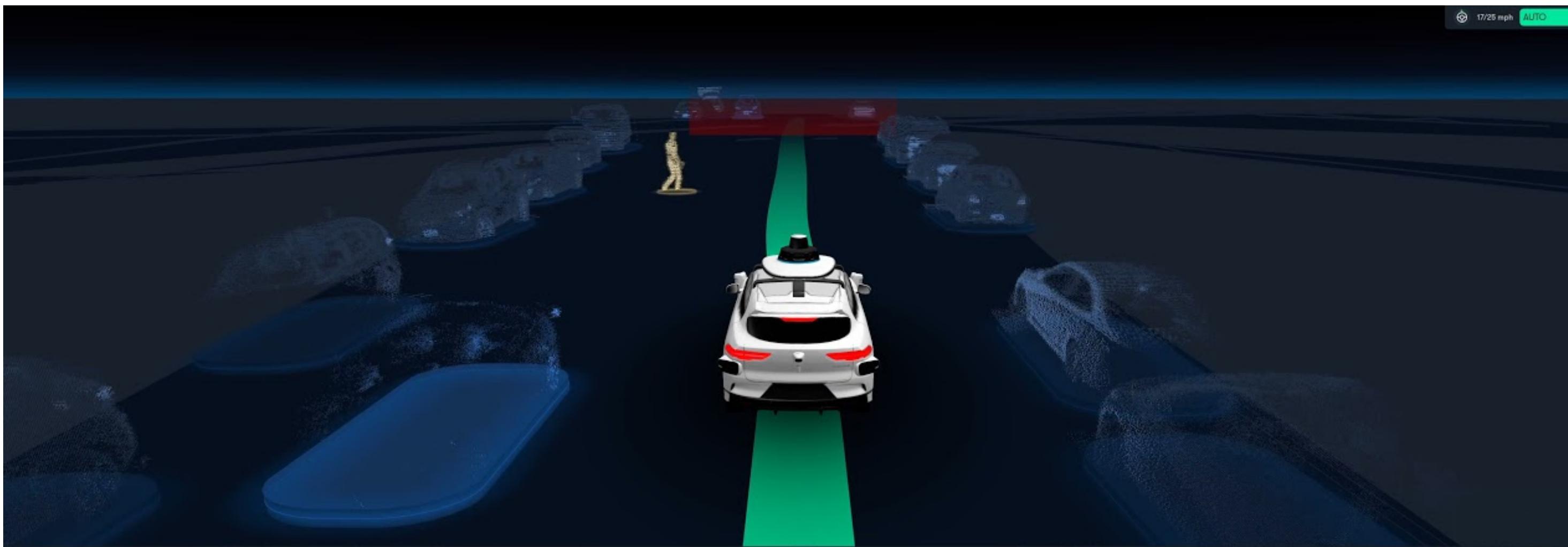
- Prompt: The camera follows behind a white vintage SUV with a black roof rack as it speeds up a steep dirt road surrounded by pine trees on a steep mountain slope, dust kicks up from its tires, the sunlight shines on the SUV as it speeds along the dirt road, casting a warm glow over the scene. The dirt road curves gently into the distance, with no other cars or vehicles in sight. The trees on either side of the road are redwoods, with patches of greenery scattered throughout. The car is seen from the rear following the curve with ease, making it seem as if it is on a rugged drive through the rugged terrain. The dirt road itself is surrounded by steep hills and mountains, with a clear blue sky above with wispy clouds.



[OpenAI Sora](#)

From 2D Vision to 3D Vision

Deep learning helps machine perceive the surrounding environment



How the Waymo Driver navigates San Francisco streets



A **whole trunk** of workstation

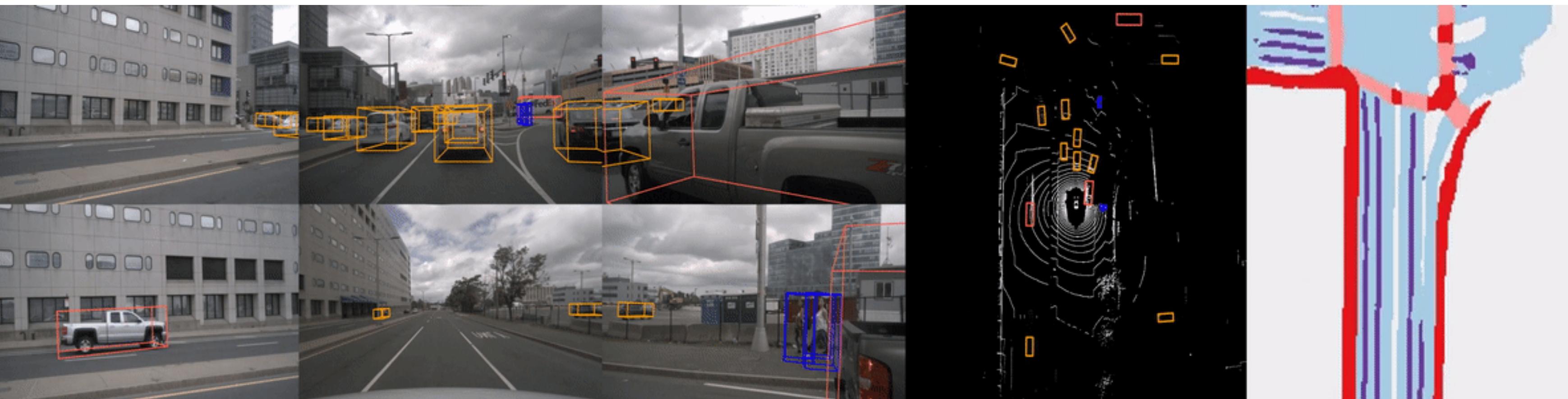
Efficient 3D Perception

- Fast-LiDARNet accelerates 3D perception with algorithm/system co-design

- 30FPS is the common frame rate



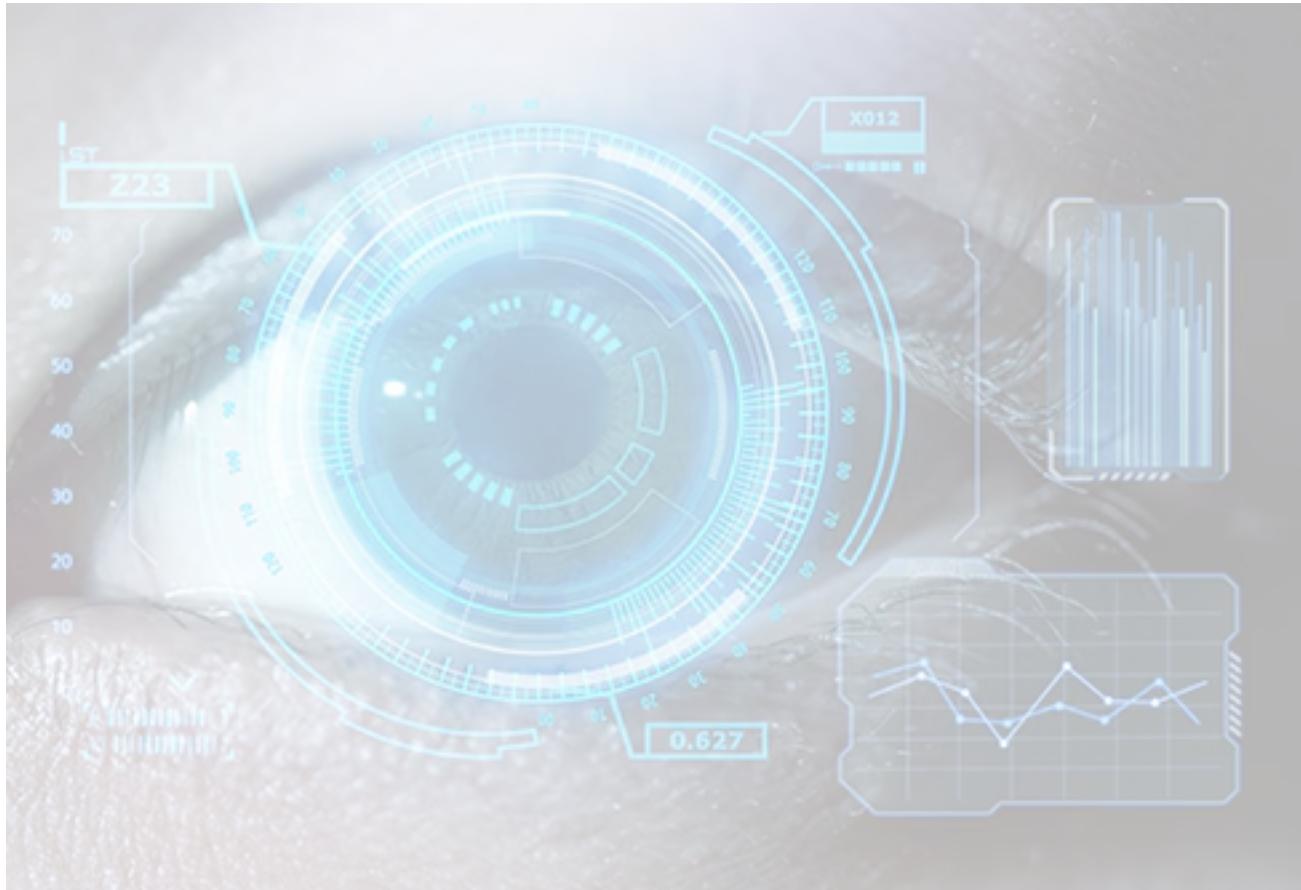
- BEVFusion supports efficient multi-task multi-sensor fusion



Liu, Z., Amini, A., Zhu, S., Karaman, S., Han, S., & Rus, D. L. (2021, May). Efficient and robust lidar-based end-to-end navigation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 13247-13254). IEEE.

Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D. L., & Han, S. (2023, May). Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE international conference on robotics and automation (ICRA)* (pp. 2774-2781). IEEE.

Deep Learning is Everywhere



Vision



Language



Multimodal

ChatGPT and Large Language Model

Large language models produce human-like text based on past conversations

Draft a short description for the course "Foundations of Edge AI"

 Course Title: Foundations of Edge AI

Course Description: This course provides an introduction to the principles and applications of Edge AI, where artificial intelligence is deployed at the edge of networks to enable real-time decision-making. Students will explore key concepts in edge computing, including architectures, system design, and differences between cloud and edge computing. The course will focus on hands-on projects involving machine learning and AI optimizations for resource-constrained environments. Topics will include edge AI frameworks, model compression techniques, and real-world applications in industries like autonomous vehicles, healthcare, and IoT. Suitable for both senior undergraduates and graduate students, the course emphasizes practical implementation and collaborative group work.

Code Generation

GitHub CoPilot can make meaningful coding suggestions based on context

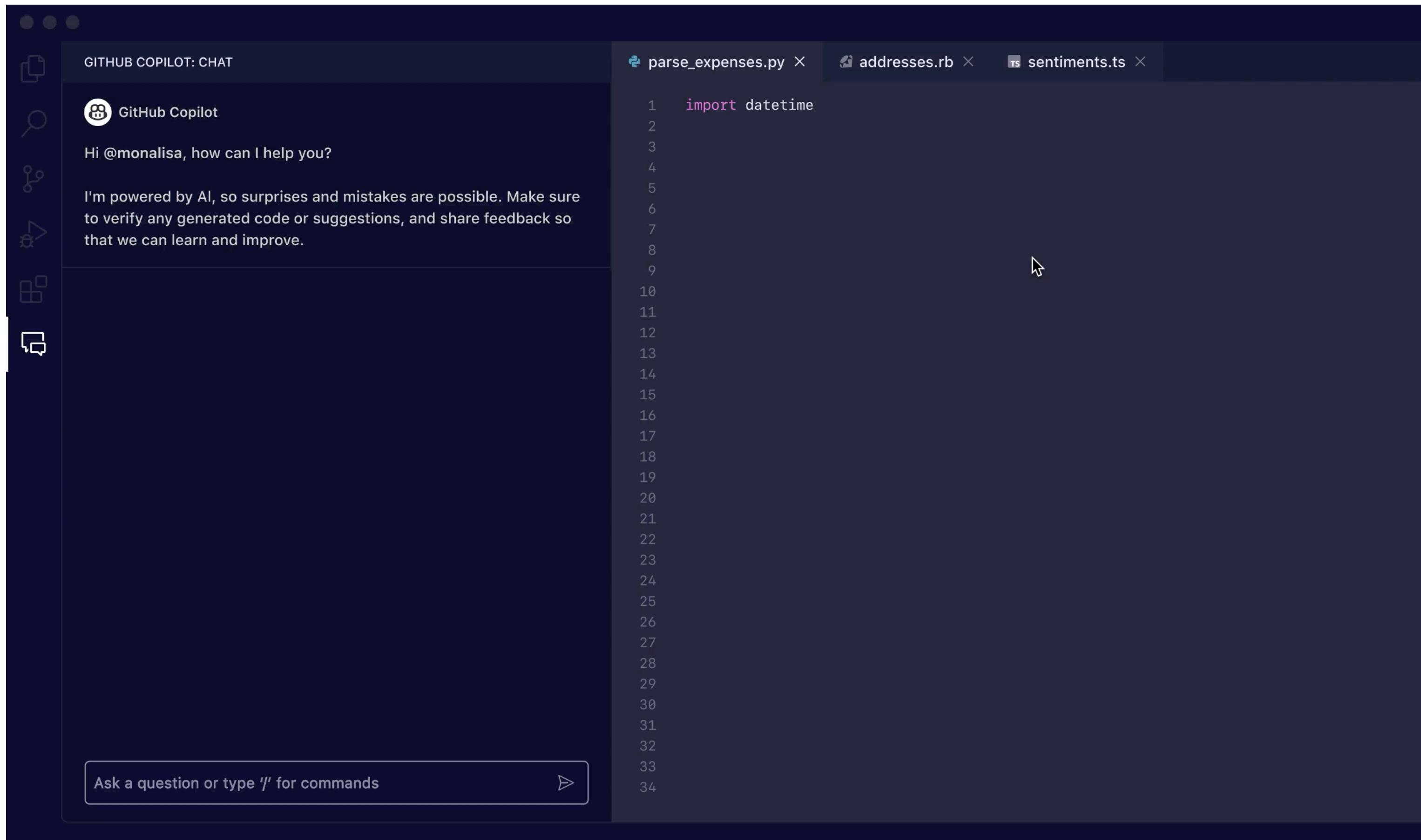


Image credit: <https://github.com/features/copilot/>

Neural Machine Translation

Neural machine translation bridges the language barrier

The screenshot shows the Google Translate mobile application. At the top, it displays language detection as "English - Detected" and offers other language options: English, Spanish, French, and Chinese (Simplified). Below this, there are two text boxes. The left text box contains an English paragraph about edge AI, and the right text box contains its Chinese (Simplified) translation. At the bottom of each text box are various interaction icons like microphone, speaker, and share.

In this course, we examine the foundations of edge artificial intelligence (AI). In the data explosion era, edge computing has become a significant computation paradigm, and AI techniques have promised to push the civilization to the next generation. However, current deep neural networks rely on extraordinary levels of computation and storage, hindering their deployment on edge devices. This course focuses on the edge computing applications and efficient AI techniques that enable powerful deep learning applications on resource-constrained edge devices, including efficient inference and domain-specific acceleration techniques. Students will get a good understanding of edge computing applications, and hands-on experience implementing model compression techniques.

在本课程中，我们将研究边缘人工智能 (AI) 的基础。在数据爆炸时代，边缘计算已成为重要的计算范例，AI 技术有望推动文明迈向下一代。然而，当前的深度神经网络依赖于非凡的计算和存储水平，阻碍了它们在边缘设备上的部署。本课程重点介绍边缘计算应用和高效的 AI 技术，这些技术可在资源受限的边缘设备上实现强大的深度学习应用，包括高效推理和特定领域的加速技术。学生将对边缘计算应用有很好的了解，并获得实施模型压缩技术的实践经验。

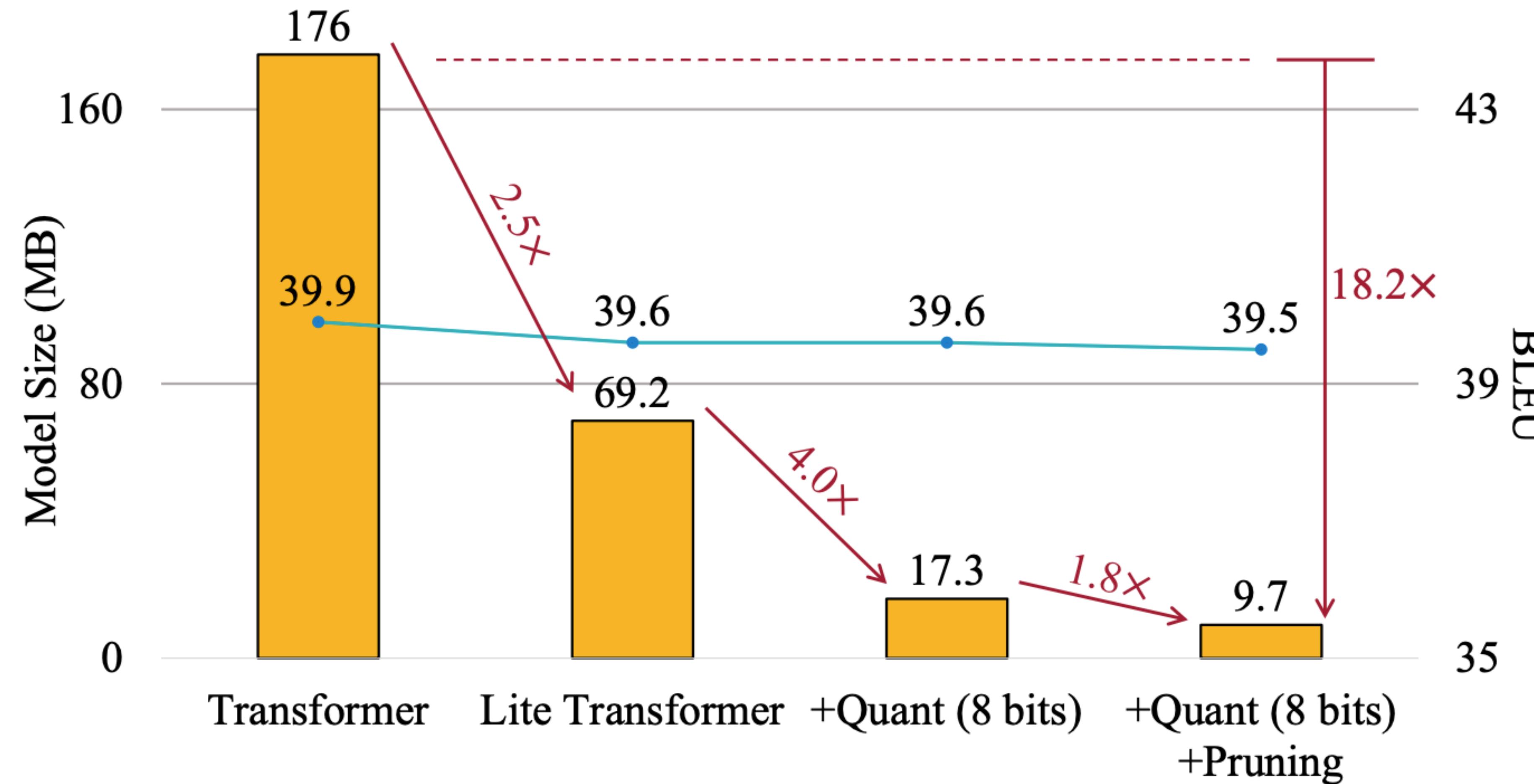
<https://translate.google.com/>



Conversation Translation

Efficient Neural Machine Translation

Lite Transformer reduces the model size with pruning and quantization



Wu, Z., Liu, Z., Lin, J., Lin, Y., & Han, S. (2020). Lite transformer with long-short range attention. arXiv preprint arXiv:2004.11886.

Large Language Models

LLMs show emergent behaviors: zero/few learning

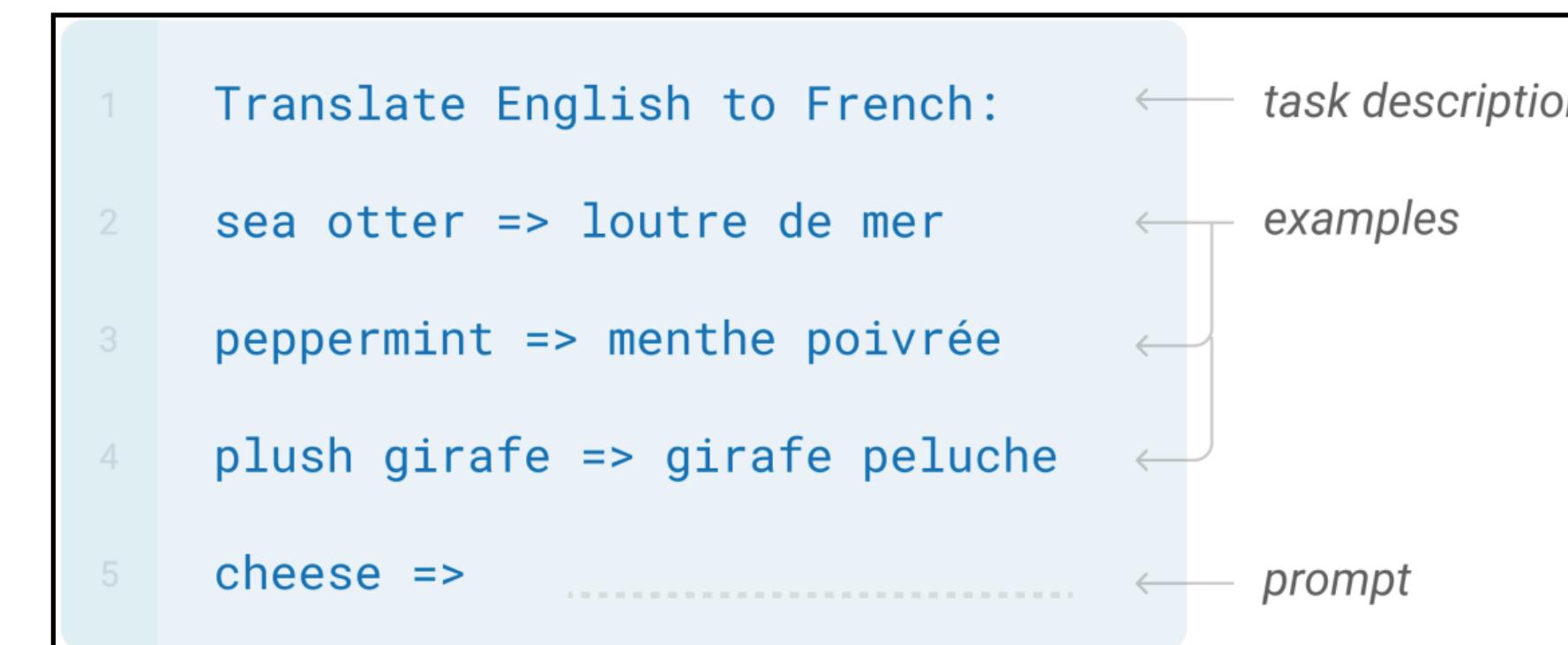
- **Zero-shot**

- The model predicts the answer given only a natural language description of the task.
No gradient updates are performed.



- **Few-shot**

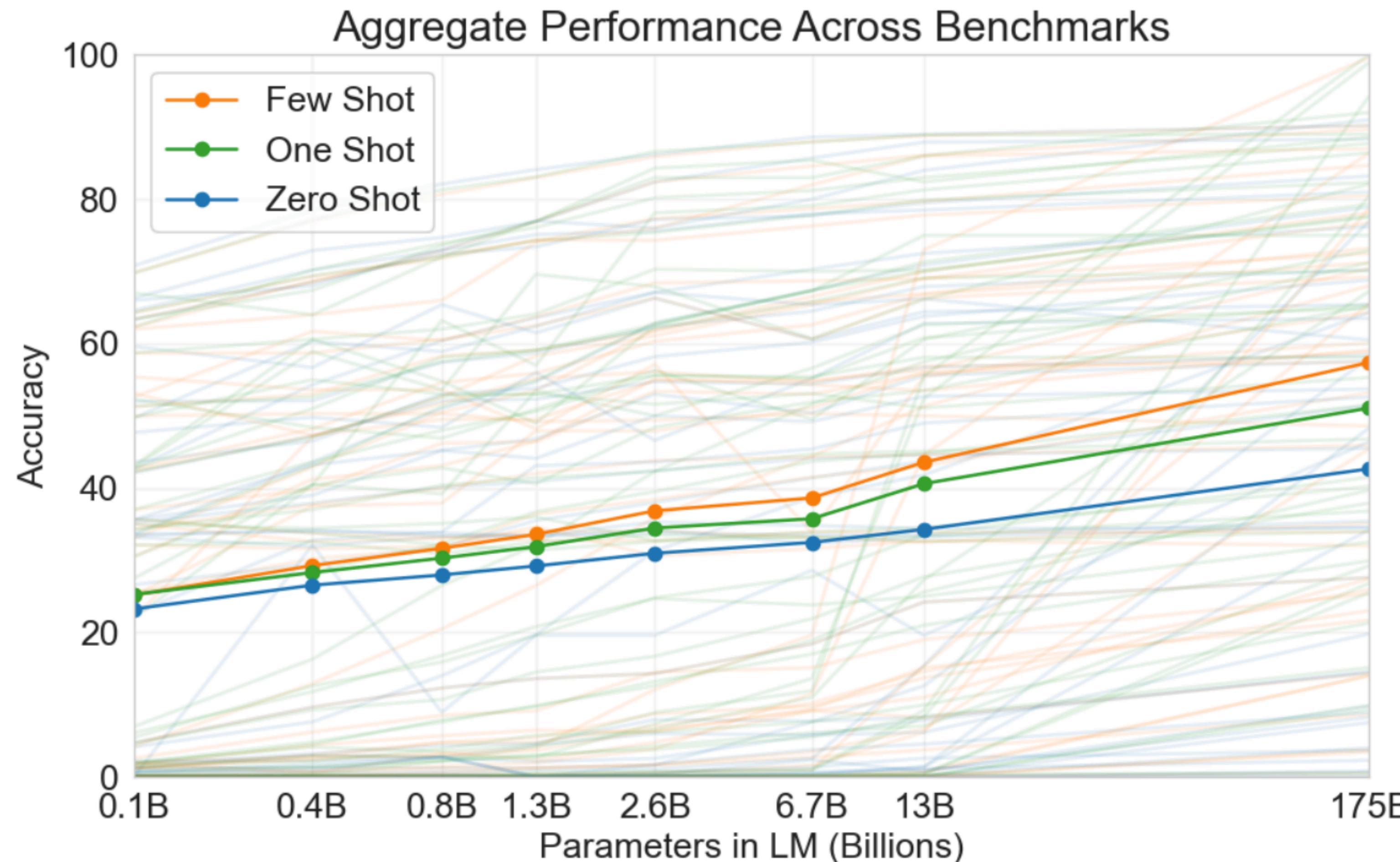
- In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



Brown, T. B. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Large Language Models

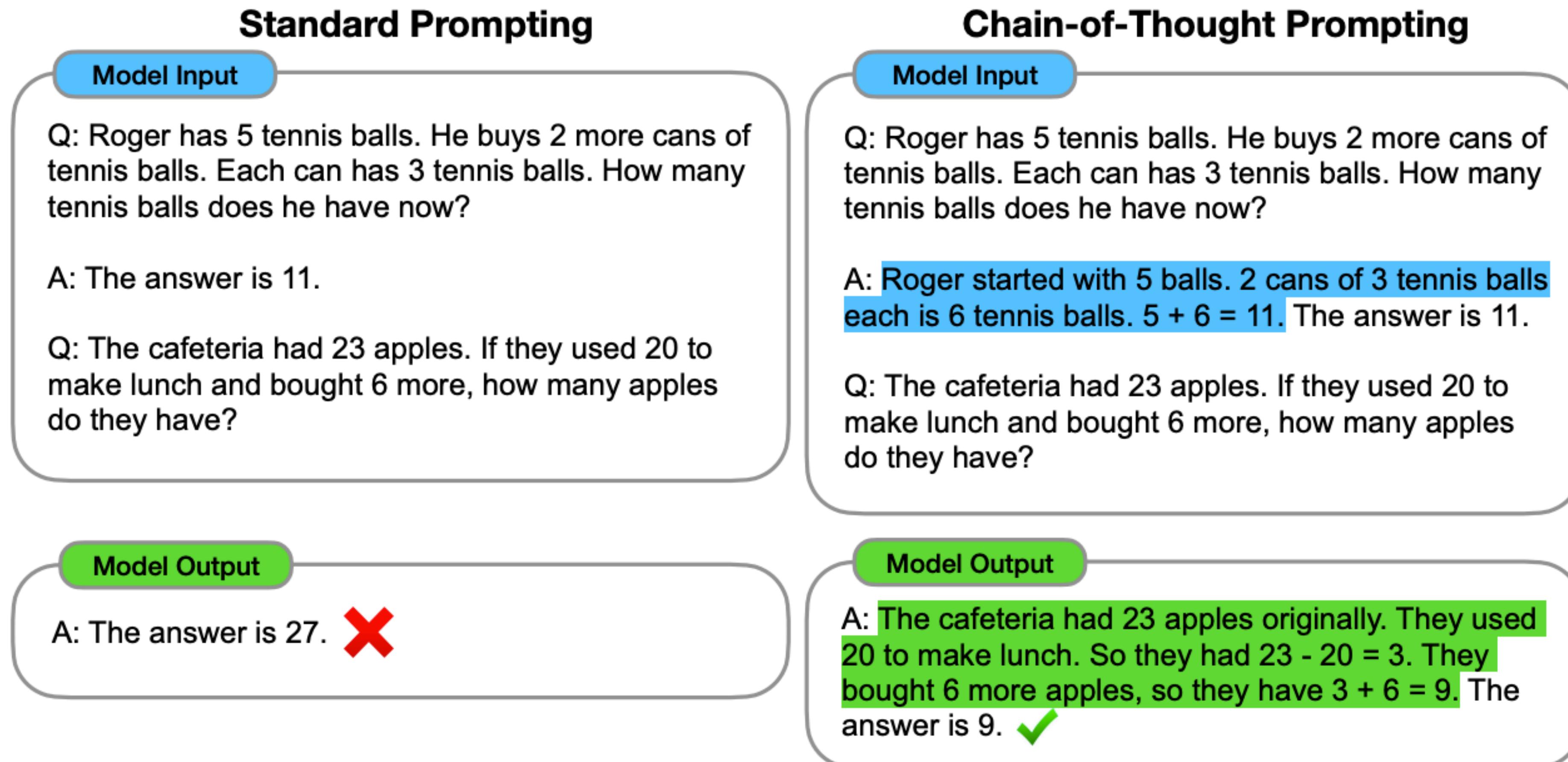
But it comes at the cost of large model size



Brown, T. B. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Large Language Models

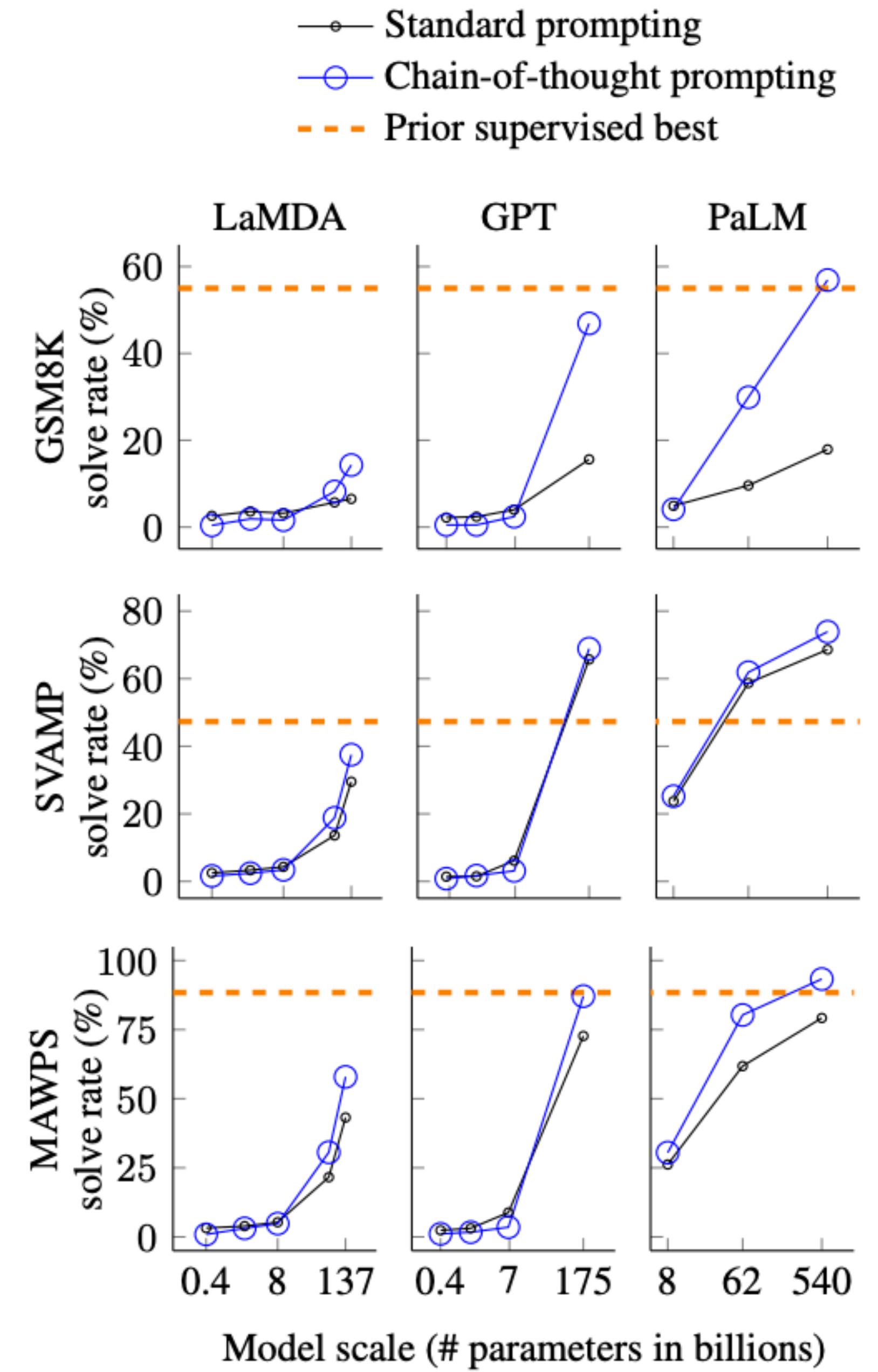
LLMs show emergent behaviors: chain-of-thought



Large Language Models

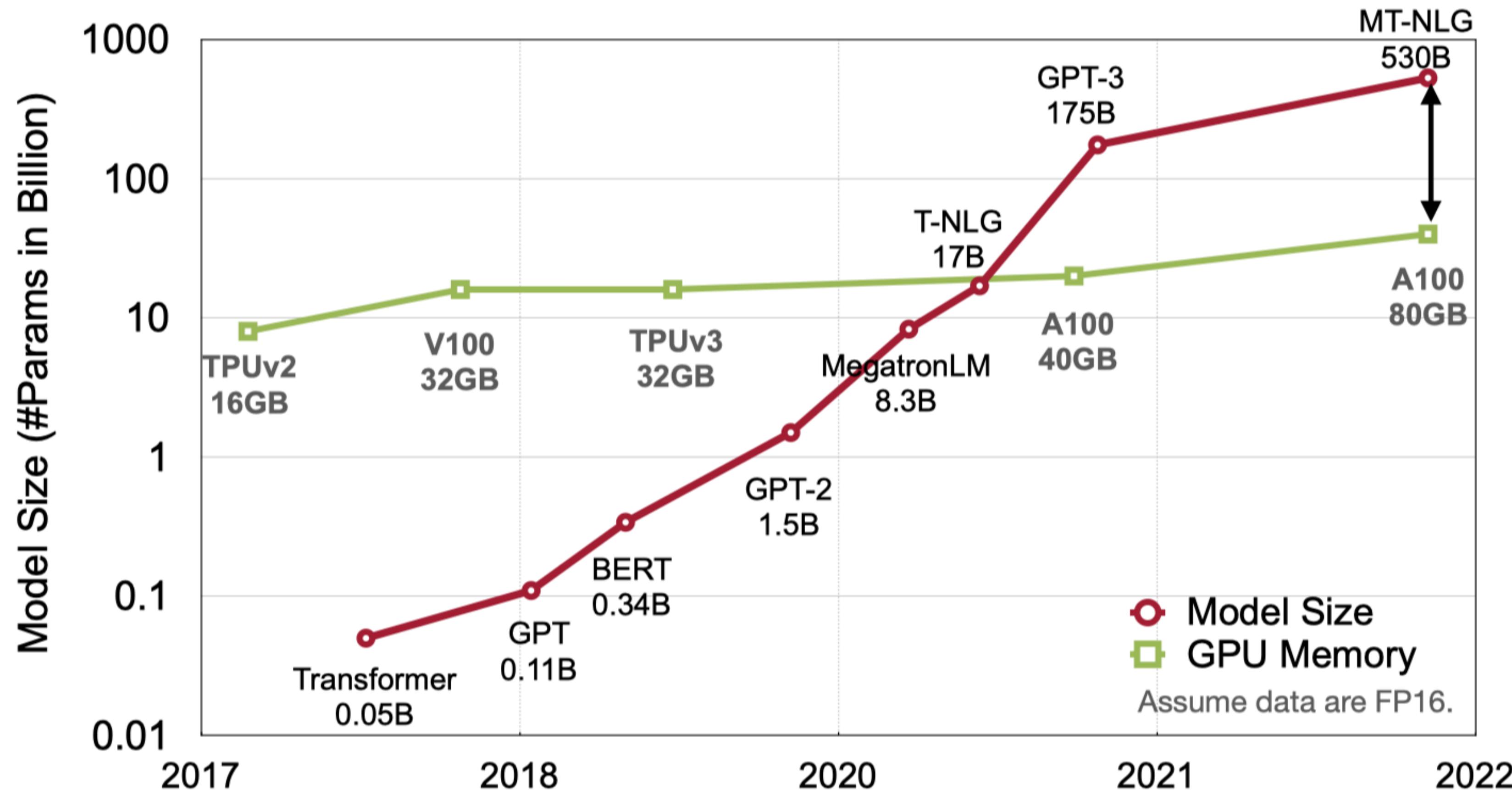
But it comes at the cost of large model size

- **GSM8K (Grade School Math 8K)**: A dataset for benchmarking arithmetic reasoning tasks in machine learning models, containing grade-school-level math word problems.
- **SVAMP (Simple Variations of Math Problems)**: A benchmark designed to test robustness and generalization in machine learning models, particularly focusing on variations of math word problems to assess how well models can handle small changes in problem phrasing.
- **MAWPS (Math Word Problem Solver)**: A dataset of annotated math word problems designed to improve the ability of machine learning models to understand and solve math word problems automatically.



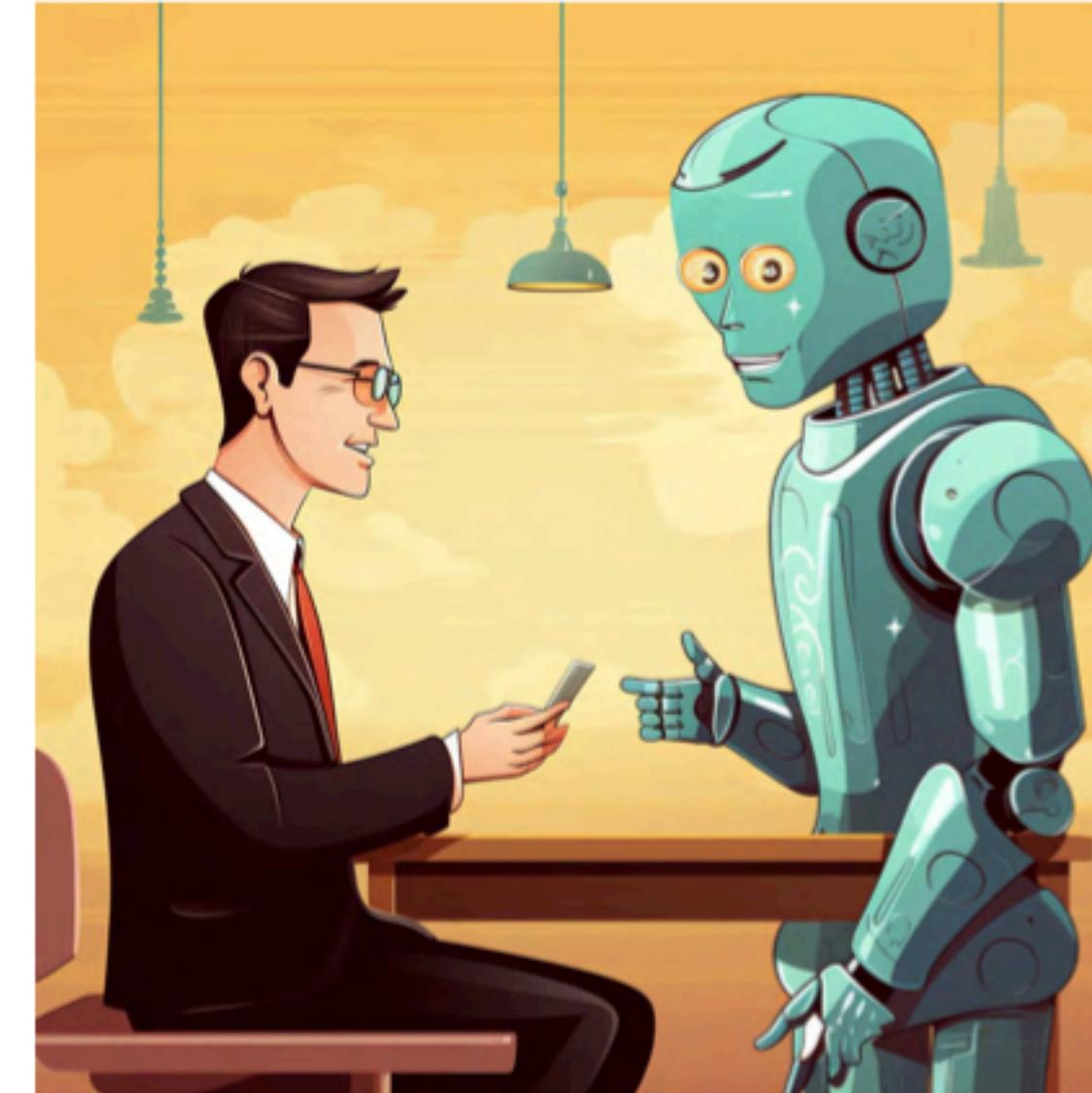
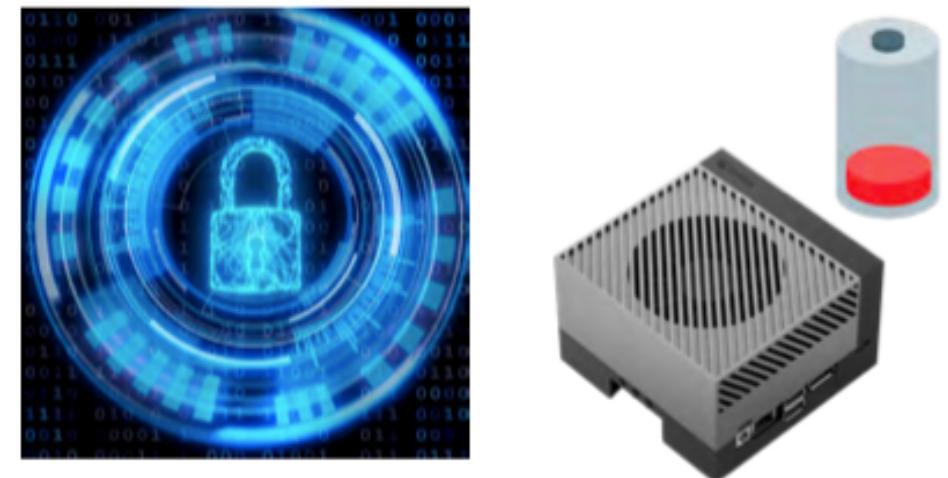
Large Language Models

Model size of language models is growing exponentially



Efficient Large Language Models

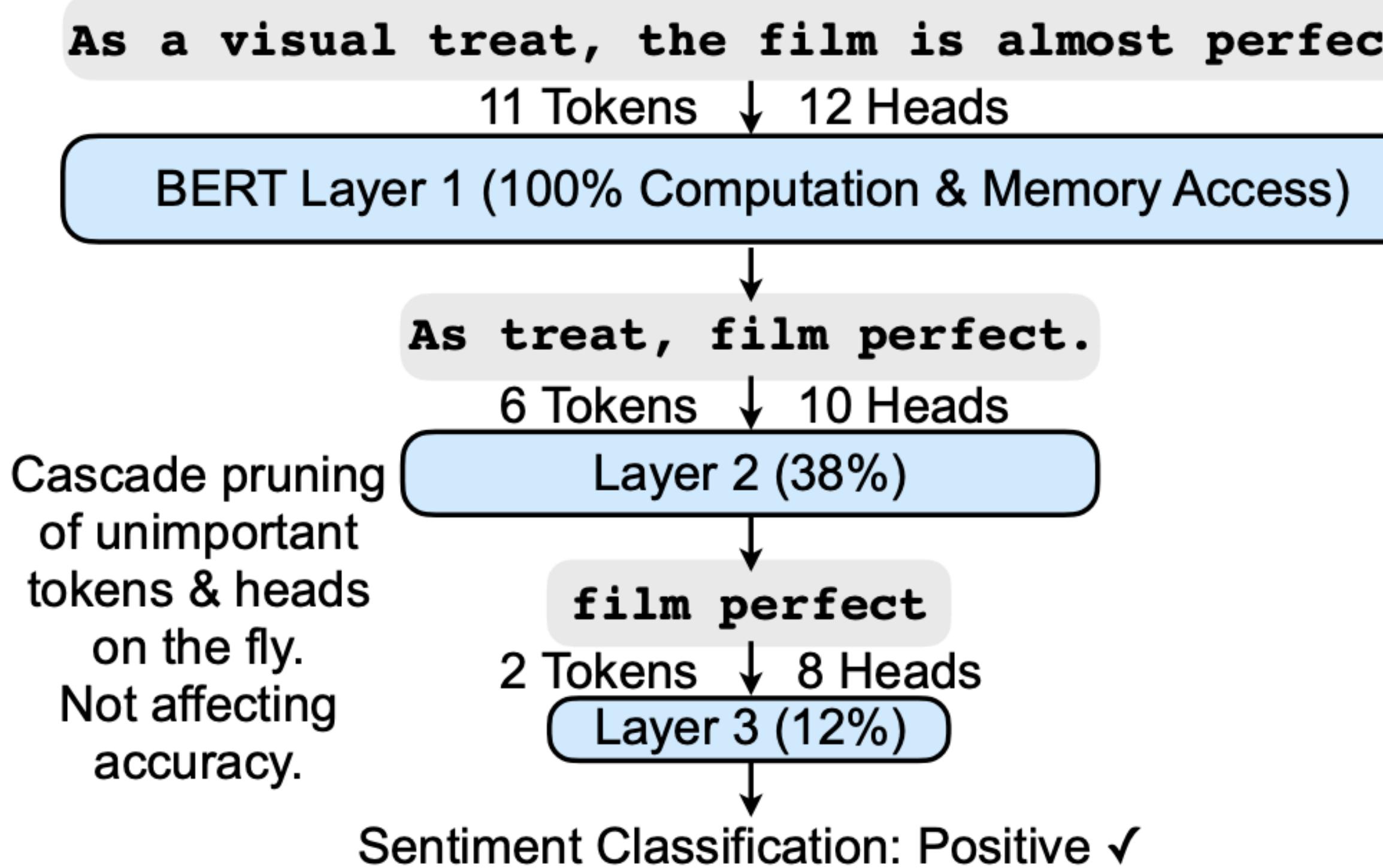
Running LLM on the edge is very important



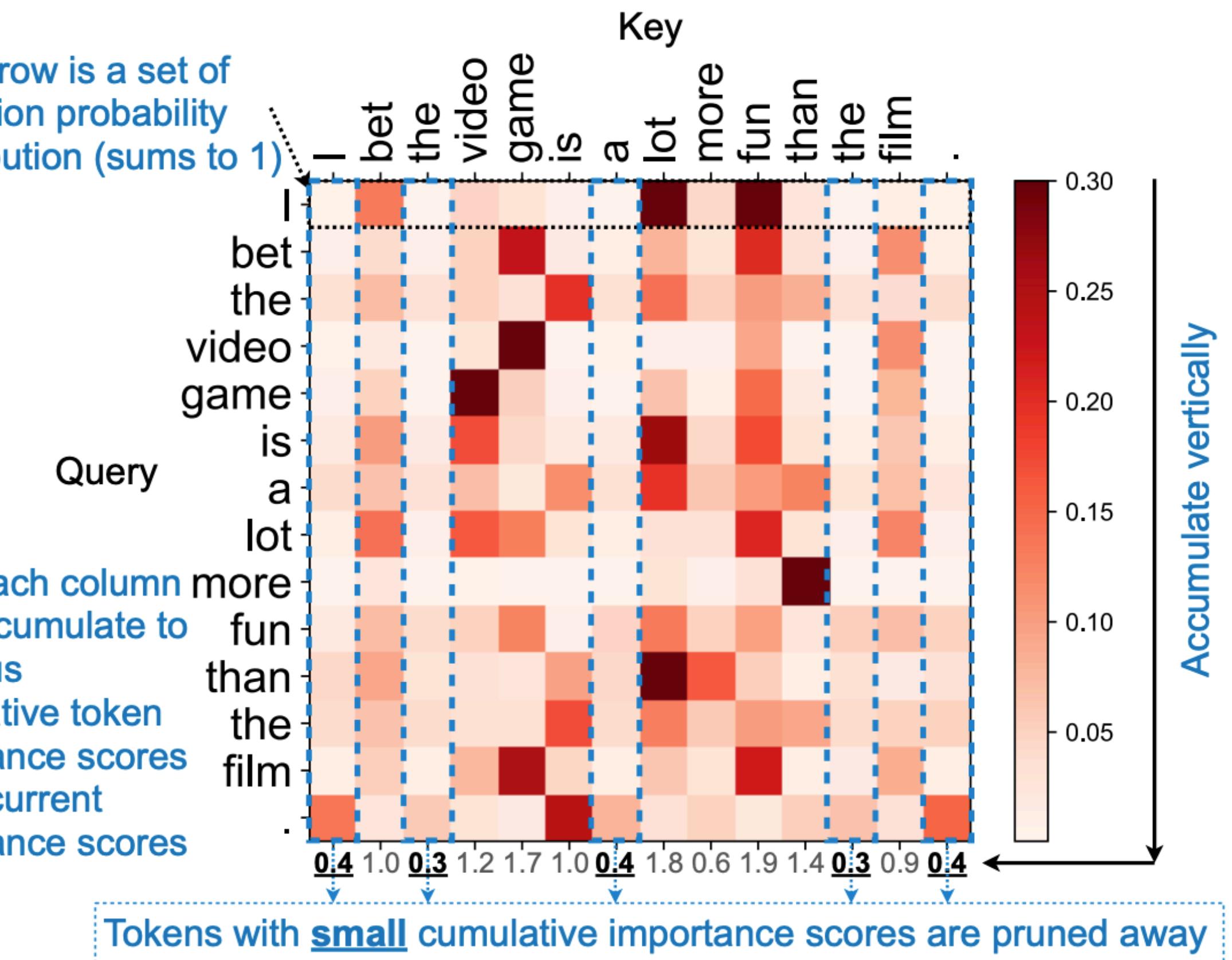
- Deploying LLM on the edge is useful: running copilot services (code completion, office, game chat) locally on laptops, cars, robots, and more. These devices are **resource-constrained, low-power** and sometimes **do not have access to the Internet**.
- **Data privacy** is important. Users do not want to share personal data with large companies.

Efficient Large Language Models

SpAtten accelerates language models by pruning redundant tokens



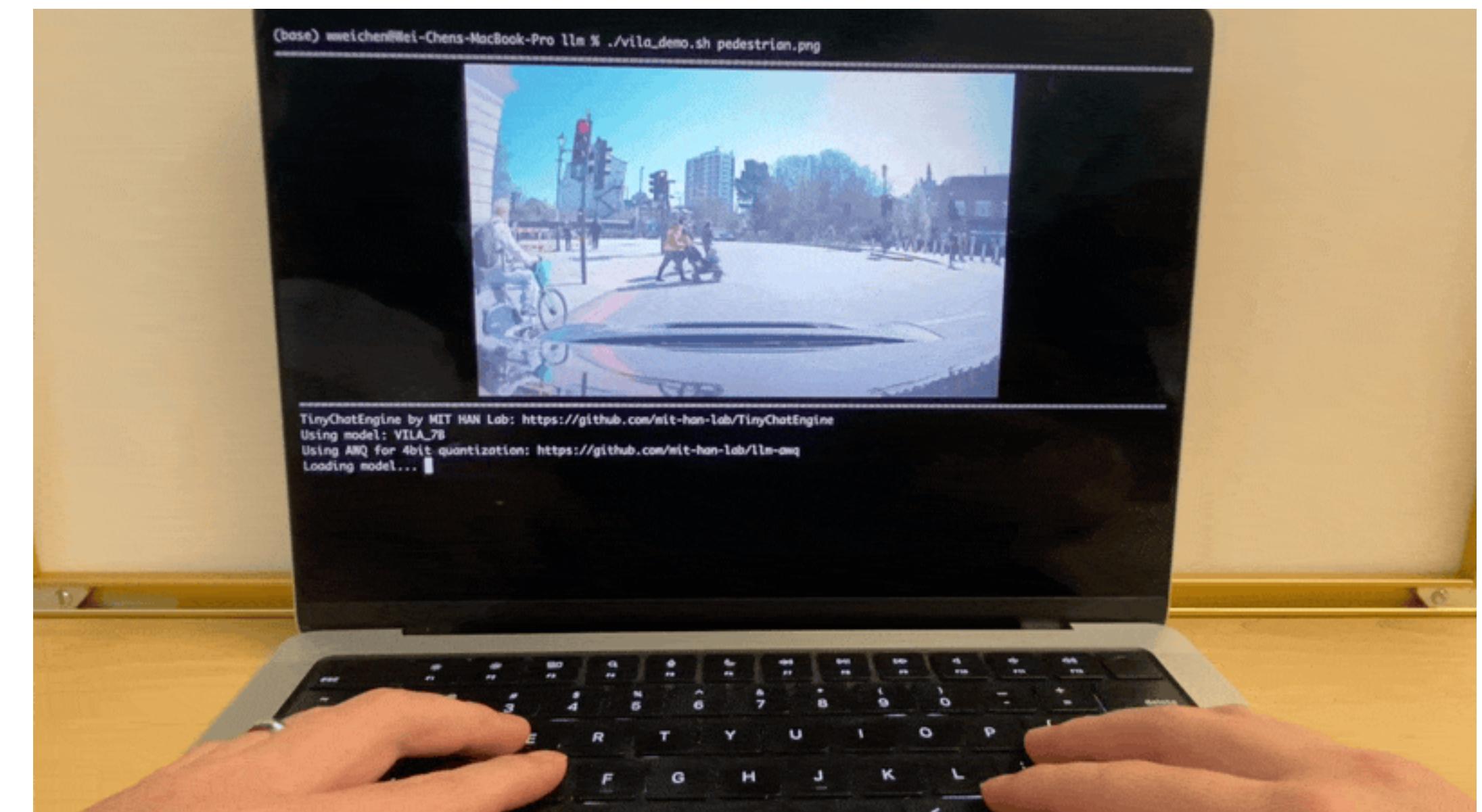
Each row is a set of attention probability distribution (sums to 1)



LLM Quantization

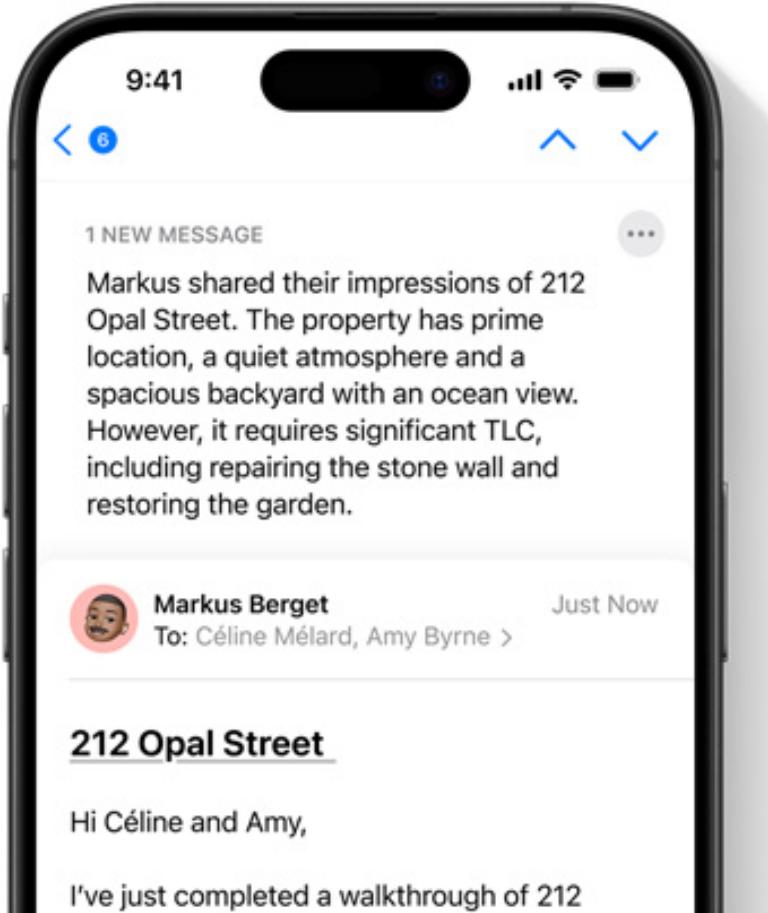
TinyChatEngine: On-Device LLM/VLM Inference Library

- LLMs are too large to fit into the memory of edge devices. We will talk about the edge deployment of LLMs through quantization.
- TinyChat is a lightweight LLM inference engine, which implement the compressed inference, built from C/C++ from scratch, easy to install and migrate to edge platforms.

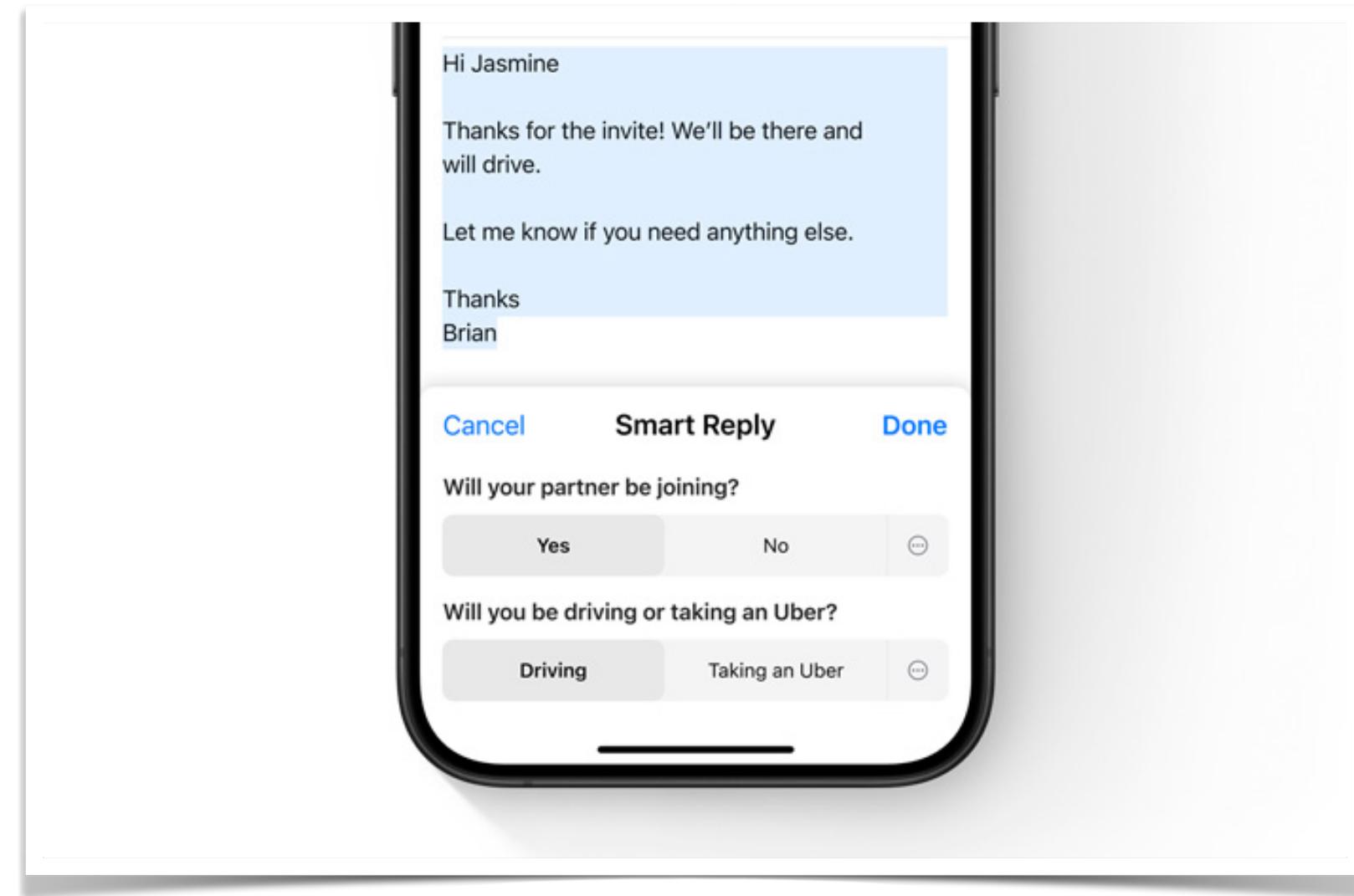


LLM on Phones

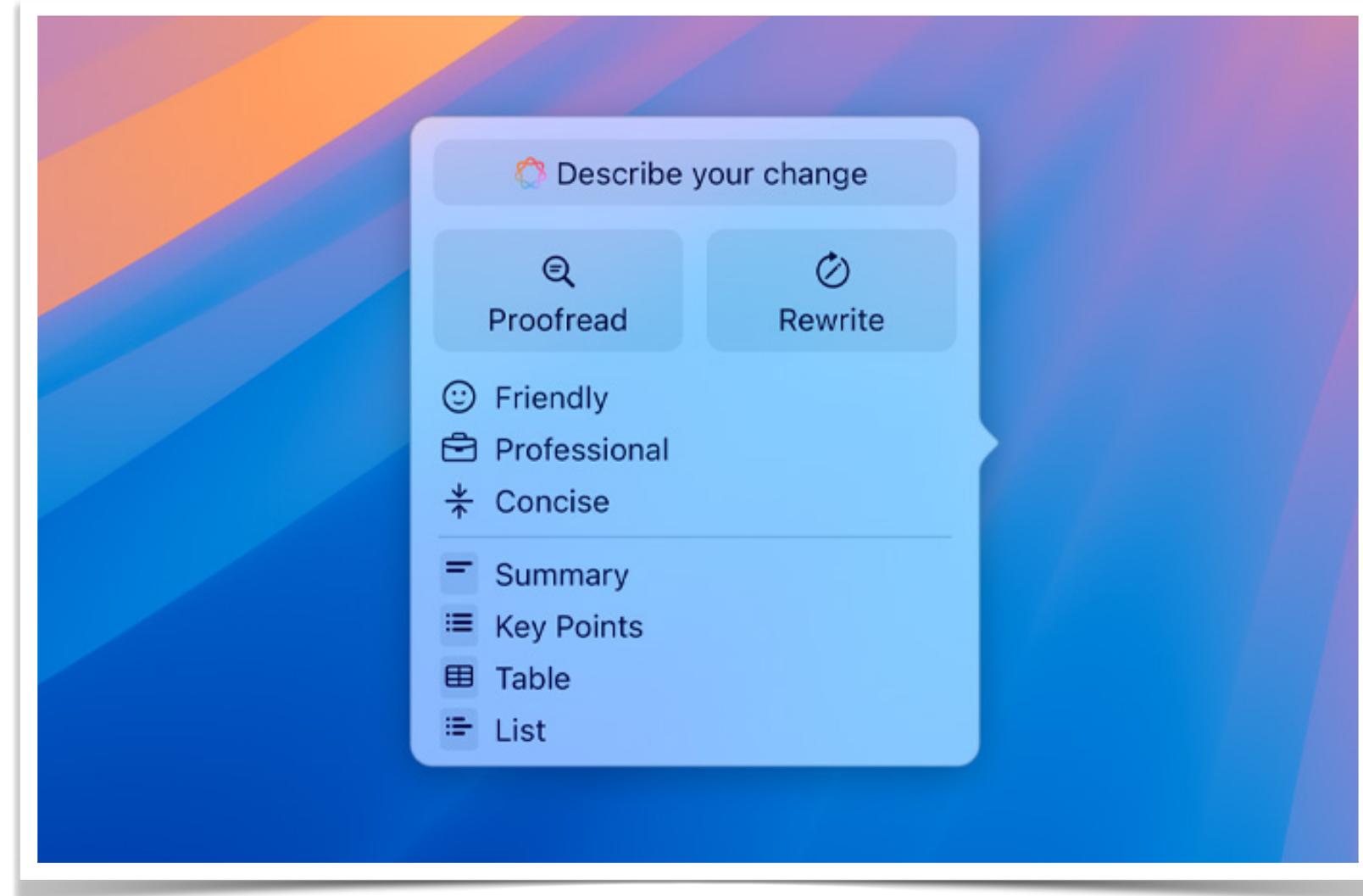
LLM empowers writing tools on phones



Summarization



Smart Reply



Writing Tools

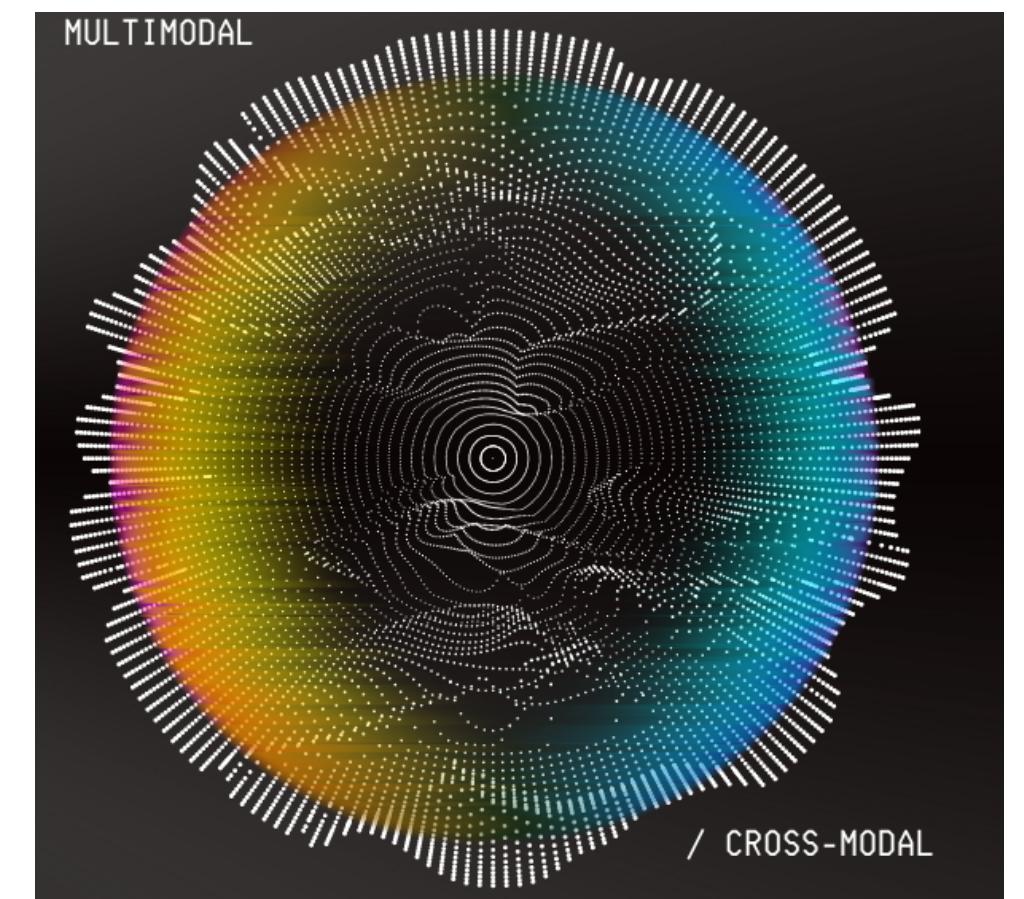
<https://www.apple.com/apple-intelligence/>

Deep Learning is Everywhere

Vision



Language



Multimodal

Vision-Language Models

LLaVA achieves general-purpose visual and language understanding

- LLaVA uses a 13B LLaMA for language understanding

User

User

Do you know who drew this painting?

 LLaVA

The painting depicts a woman, commonly believed to be Mona Lisa, the famous artwork by Leonardo da Vinci. It is a portrait painting that showcases the woman's enigmatic smile and has become one of the most famous and iconic art pieces in the world. The original work is displayed in the Louvre Museum in Paris, and it is known for its intricate details, use of oil paint, and the artist's innovative techniques that contributed to its enduring appeal and mystery.

Efficient Vision-Language Models

AWQ quantizes vision-language models to 4 bits with high quality

Sometimes I just look at pictures of the earth from space and I marvel at how beautiful it all is.



Question: Can you explain this meme in detail?

INT4-g128 RTN: The meme in the image features a plate of food, likely fried chicken, with a unique twist. **There are small pictures of the Earth and other planets placed on top of the food**, creating an interesting and amusing visual. Alongside the plate, a caption reads, "Sometimes I look at pictures of the Earth from space and I marvel at how beautiful it all is." This combination of food and planet images adds a touch of humor and curiosity, making the viewer wonder about the context and intent behind the meme.

INT4-g128 AWQ: The meme in the image is a **lighthearted and humorous take on the concept of looking at pictures of the Earth from space**. A plate of fried food, specifically chicken nuggets, is presented with the caption, "Sometimes I look at pictures of the Earth from space and I marvel at how beautiful it all is." The playful combination of a food item and an Earth-related context creates a fun and entertaining visual.



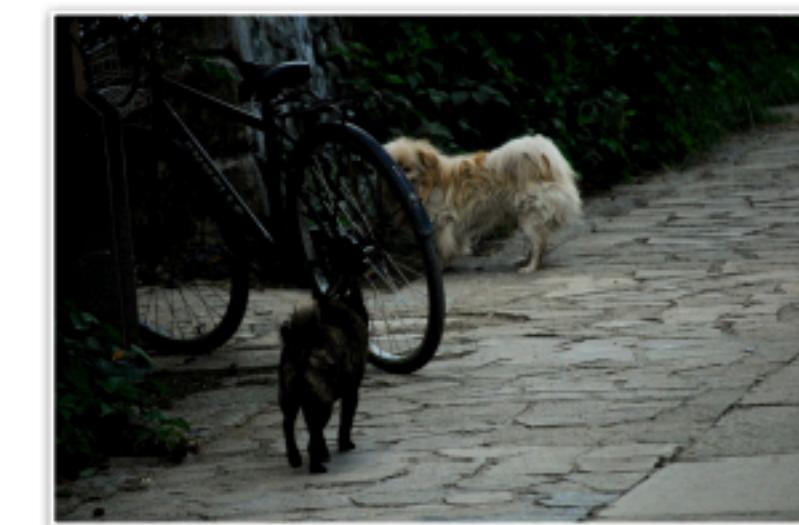
W4-RTN: A model airplane **flying in the sky**.

W4-AWQ: Two toy airplanes **sit on a grass field**.



W4-RTN: A man is **holding a baby elephant** in his arms.

W4-AWQ: A man and his daughter **pose with an elephant**.

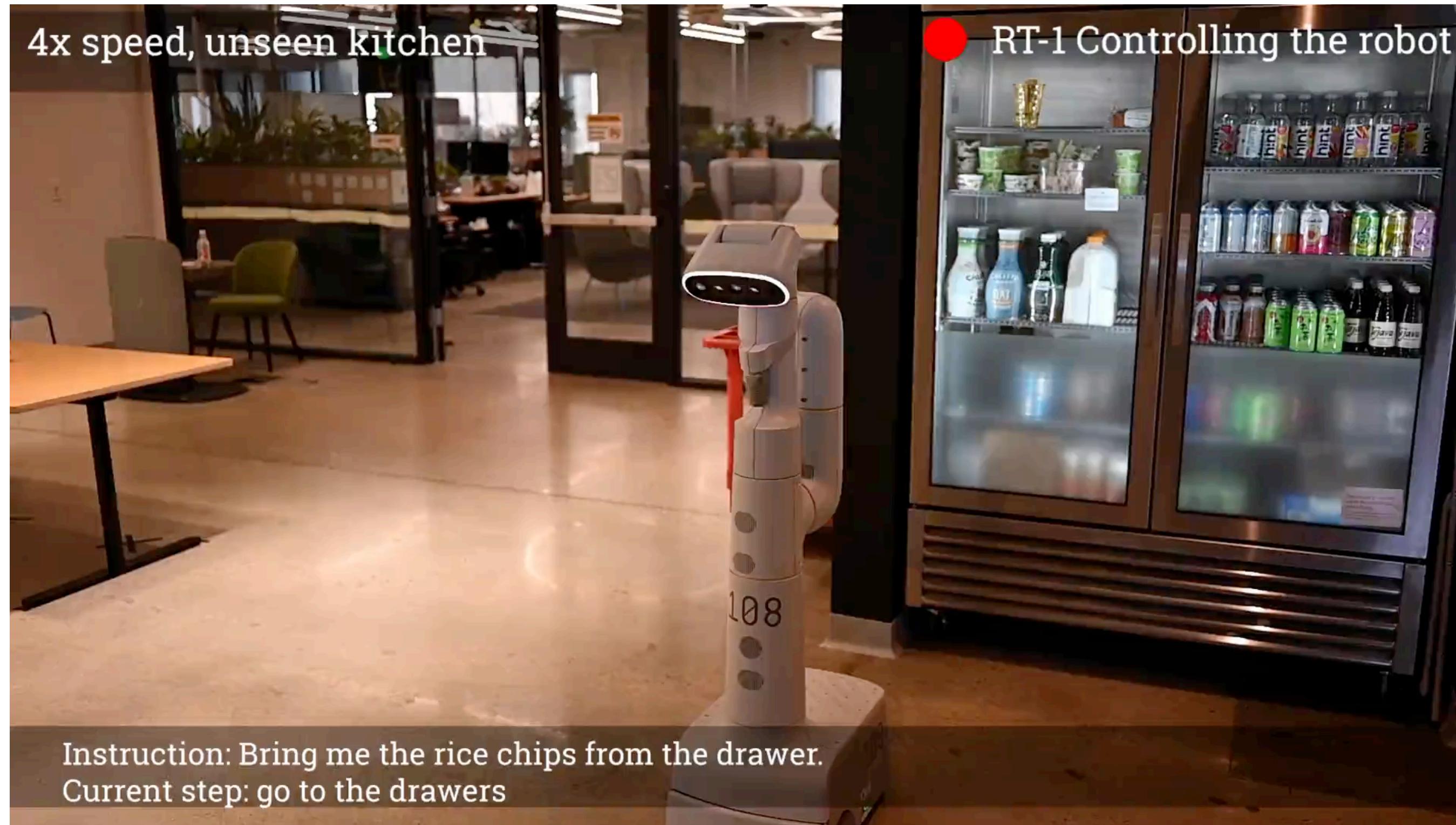


W4-RTN: **A man and a dog** walking past some bushes.

W4-AWQ: **Two dogs** are walking on the street.

Vision-Language-Action Models

Robotics transformers control robots based on language instructions



[Video source](#)

- Run at only **3Hz** due to the high computational cost and networking latency.

Deep Learning for Games

AlphaGo, AlphaZero, and AlphaMaster



AlphaGo



AlphaZero

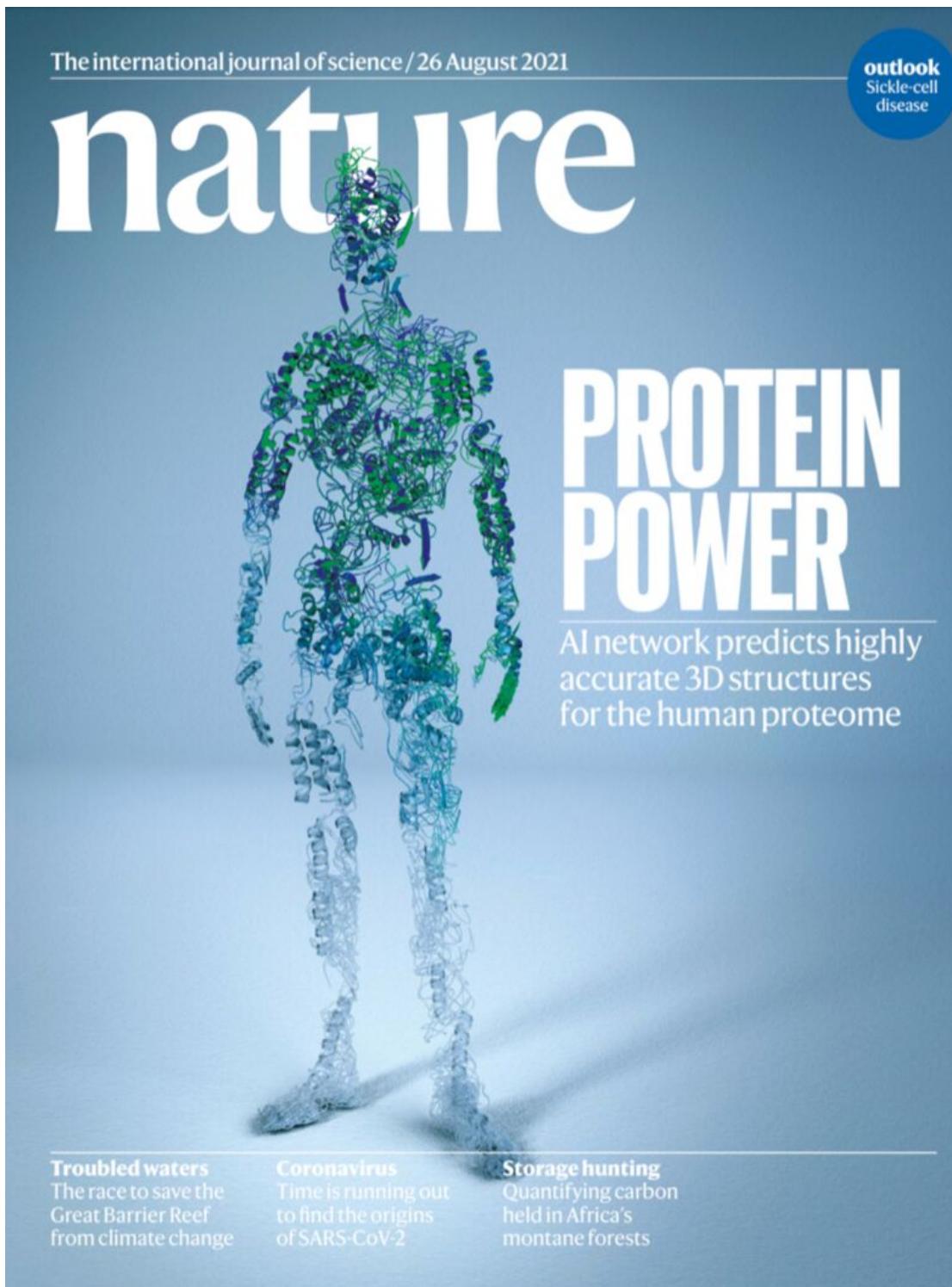


AlphaMaster

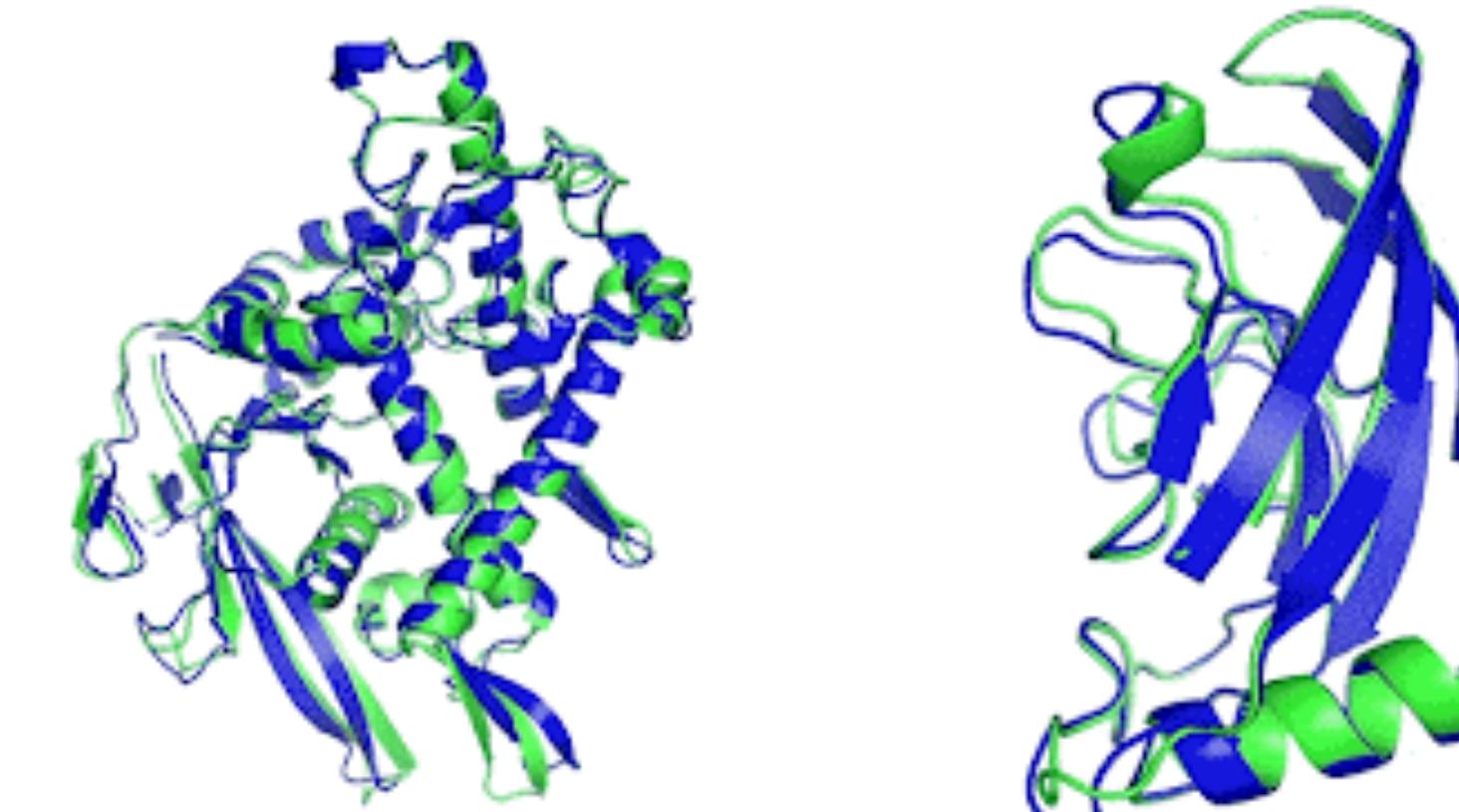
- AlphaGo: 1,920 CPUs and 2,80 GPUs (**\$3,000 electric bill** per game)
- AlphaZero: 5,000 TPUs for training over several days.

Deep Learning for Scientific Discovery

AlphaFold reveals the structure of the protein universe



AlphaFold (Nature 2021)



T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

T1049 / 6y4f
93.3 GDT
(adhesin tip)

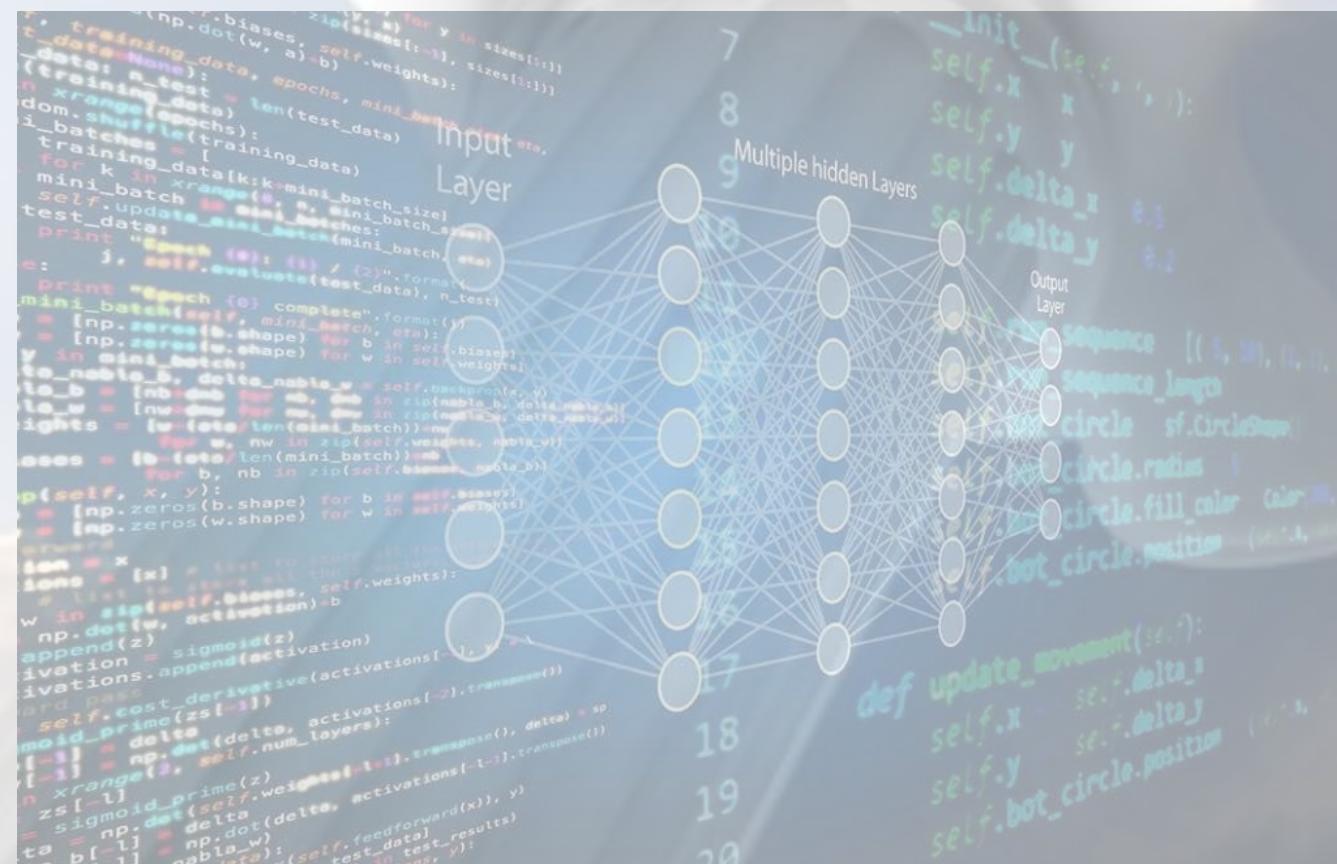
- Experimental result
- Computational prediction

Image source

- 16 TPUv3s (**128 TPUv3 cores**) run over a few weeks

Deep Learning

Three Pillars



Algorithm



Data

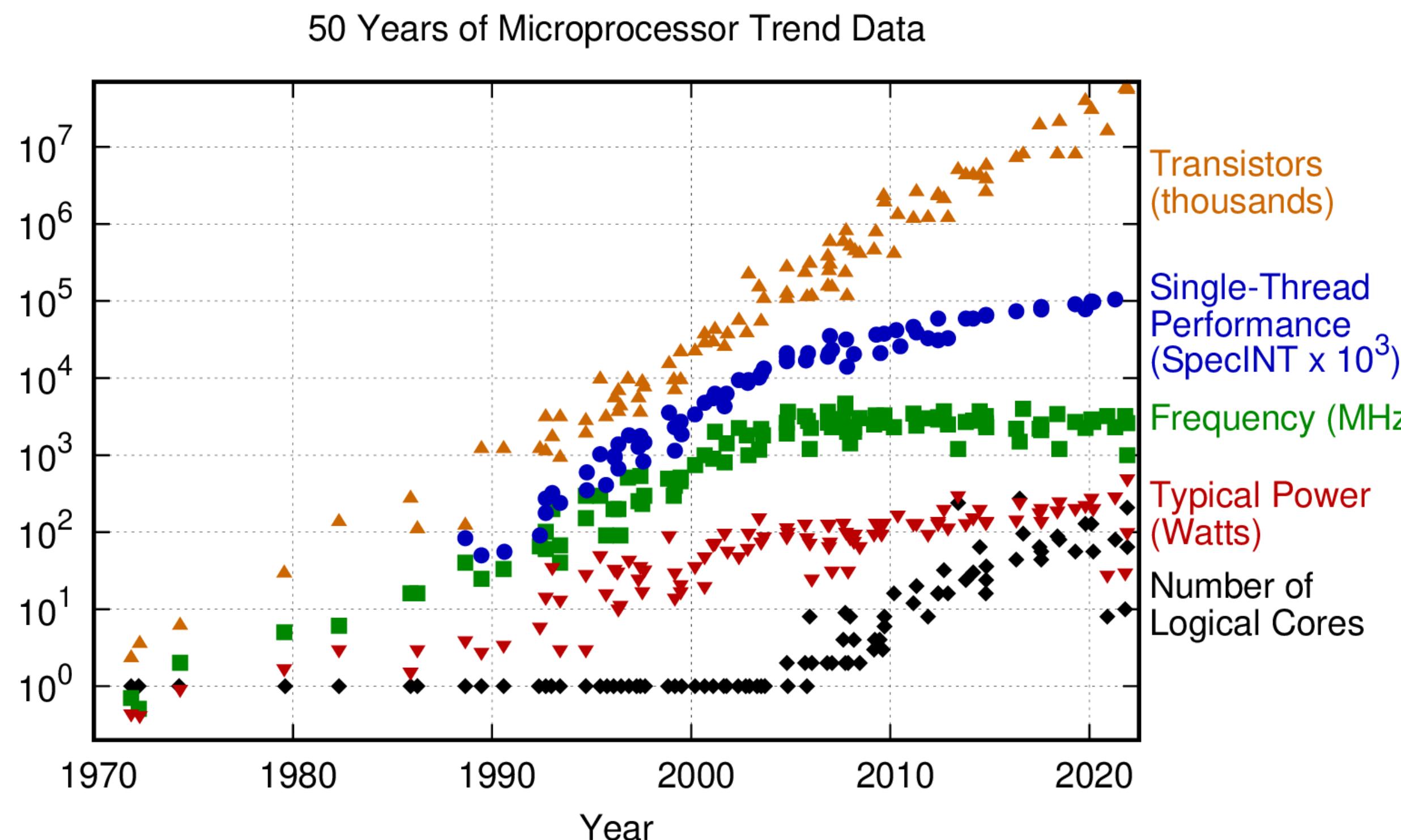


Hardware

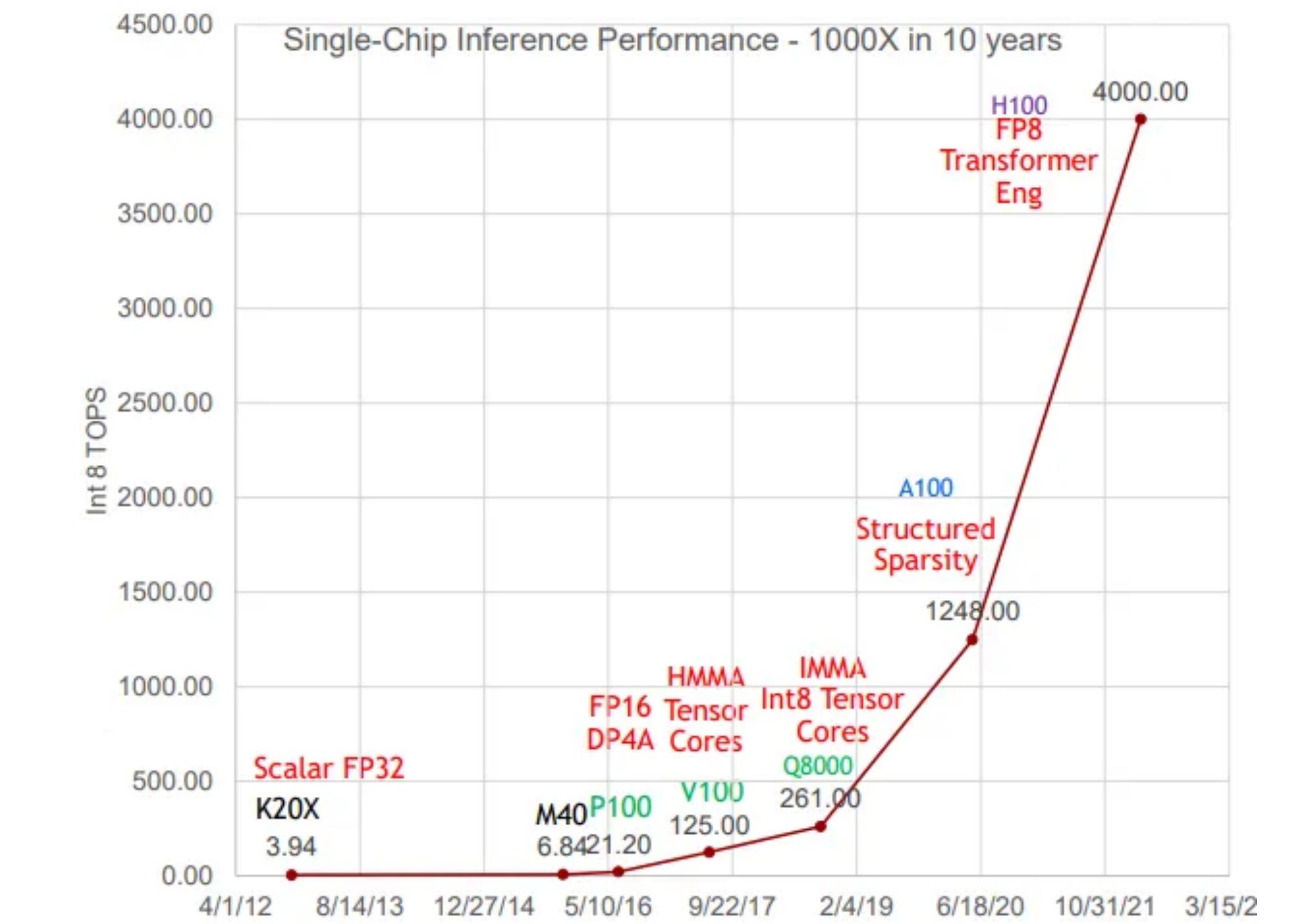
Deep Learning Hardware

Architectural support for quantization/pruning brings tremendous improvement

- FP32 → FP16 → Int8; dense → sparse



[Image source](#)



[AI for Science — Bill Dally](#)

Design Cost

Software innovation is important in advanced technology node

- The **software** cost dominates the cost breakdown of advanced technology nodes.

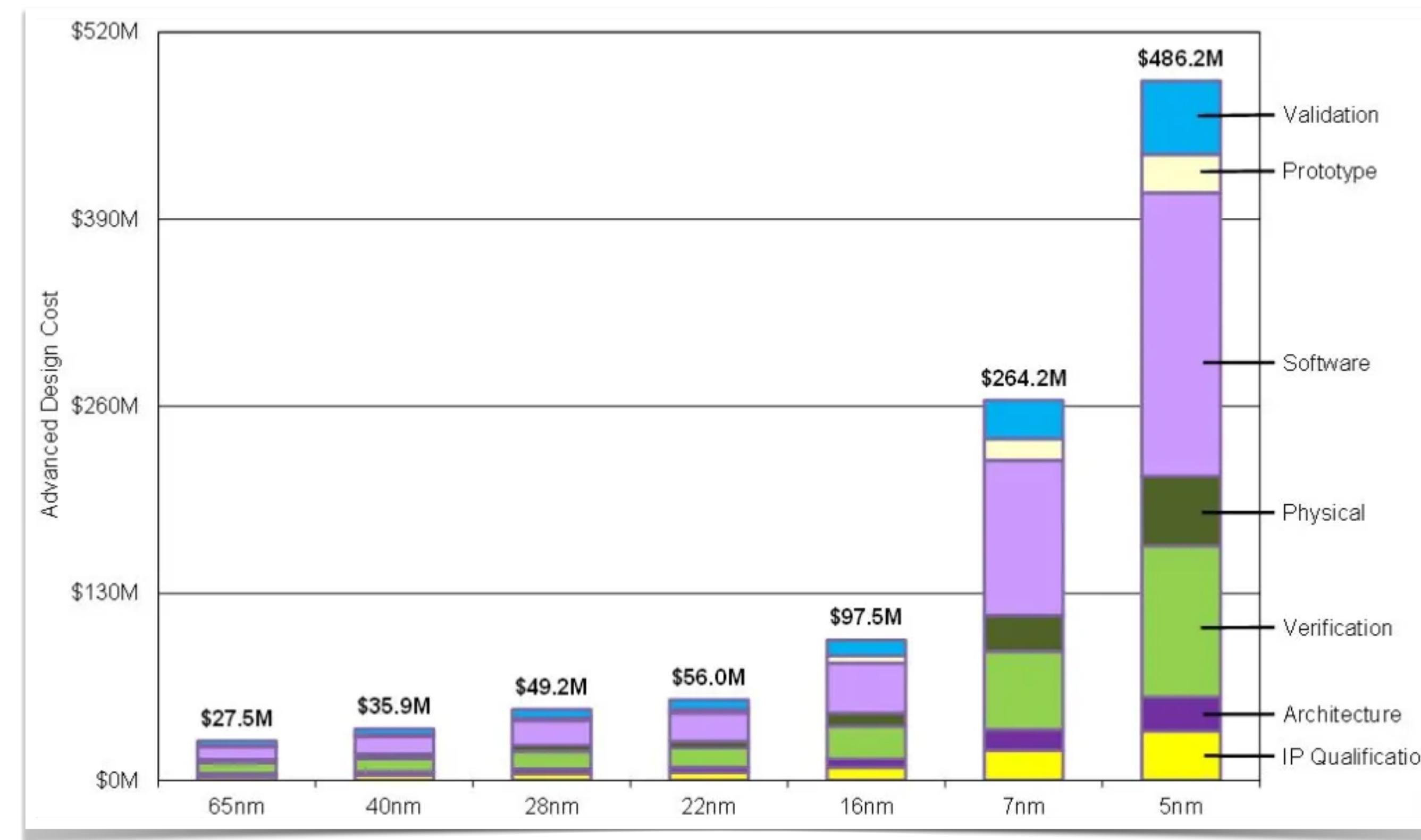
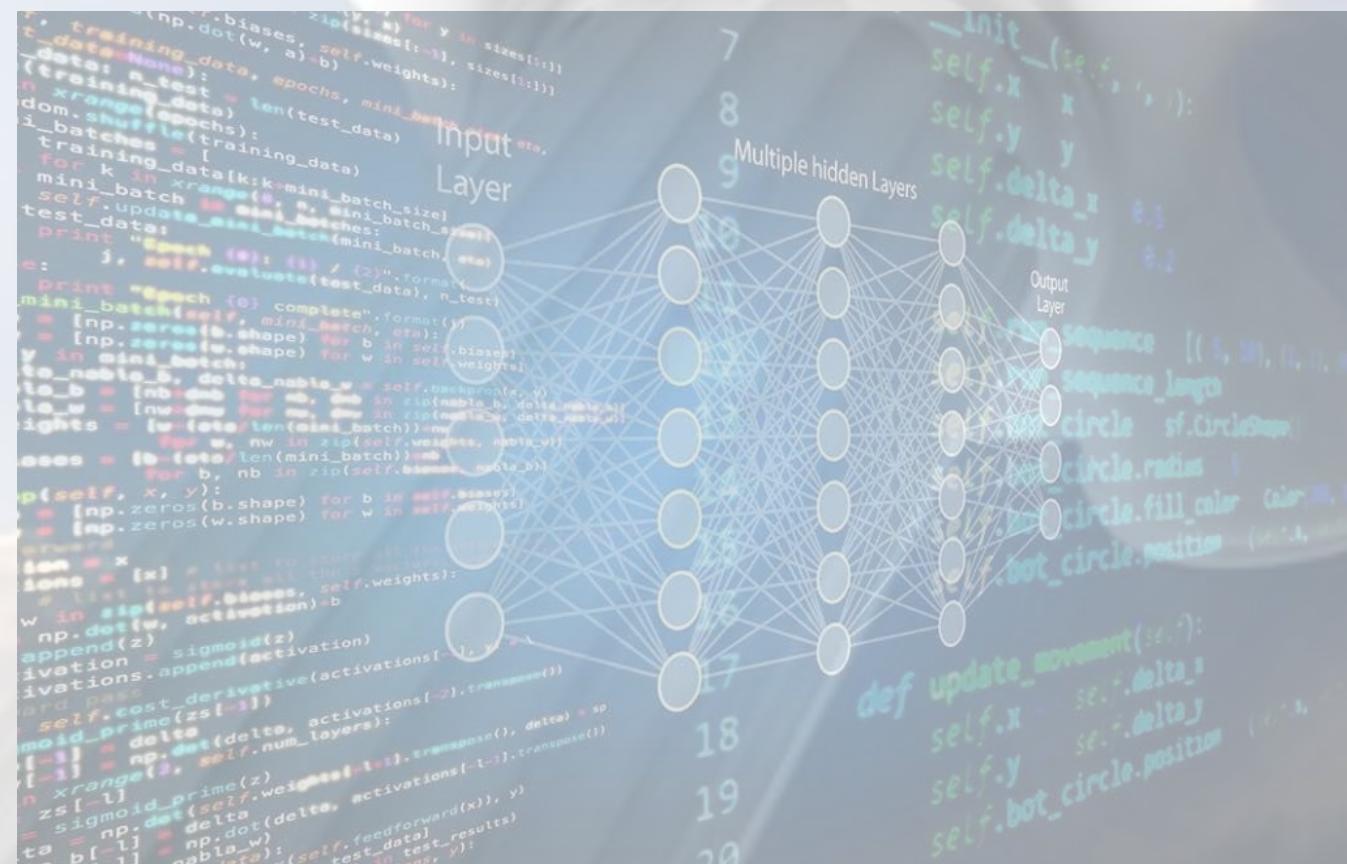


Image source

Deep Learning

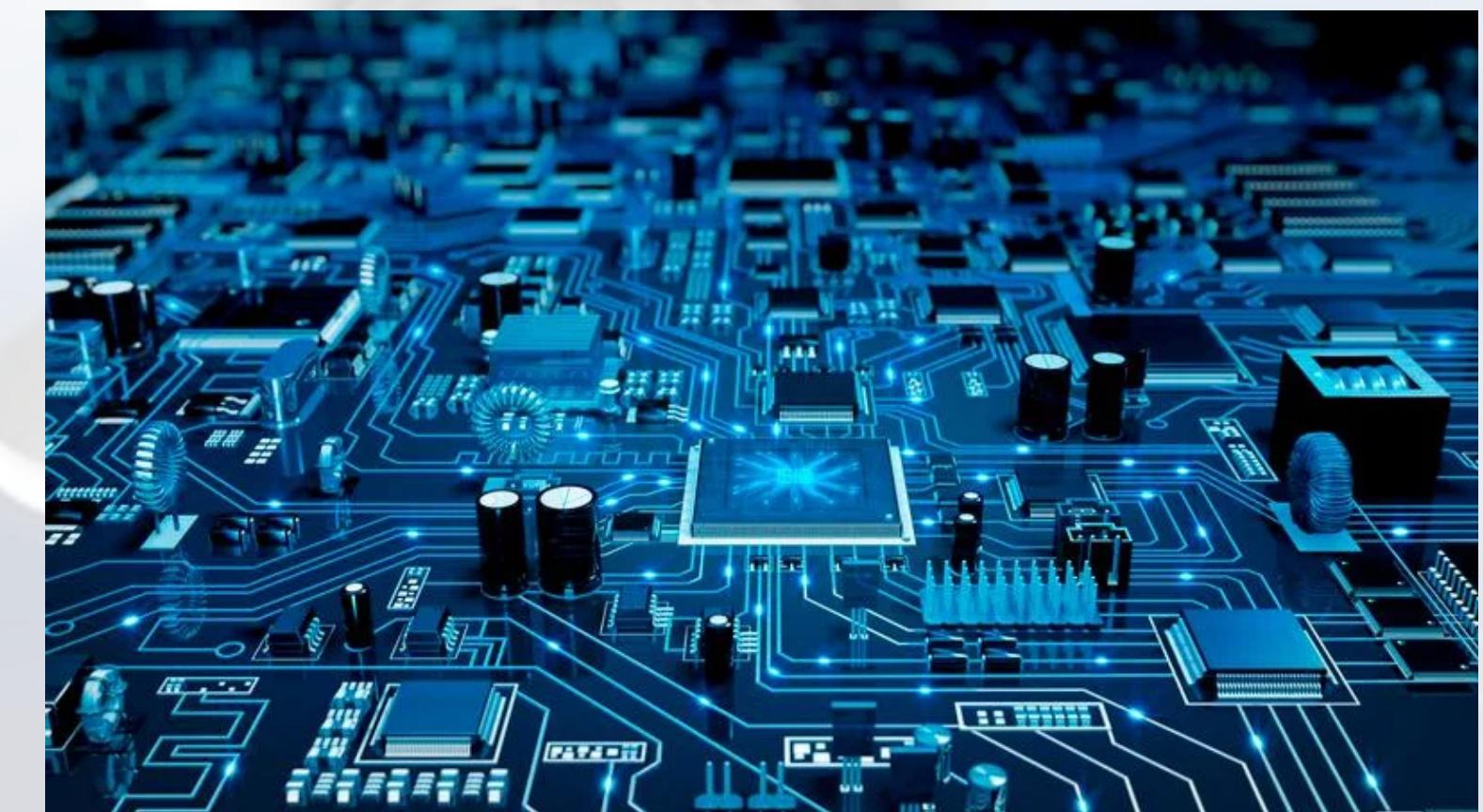
Three Pillars



Algorithm



Data



Hardware

Cloud AI Hardware



P100 (2016)



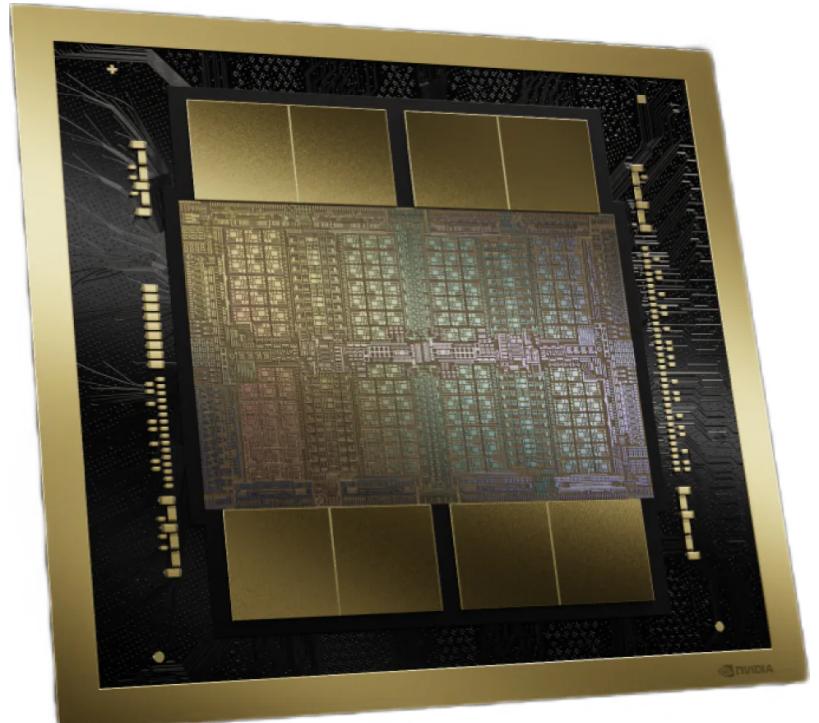
V100 (2017)



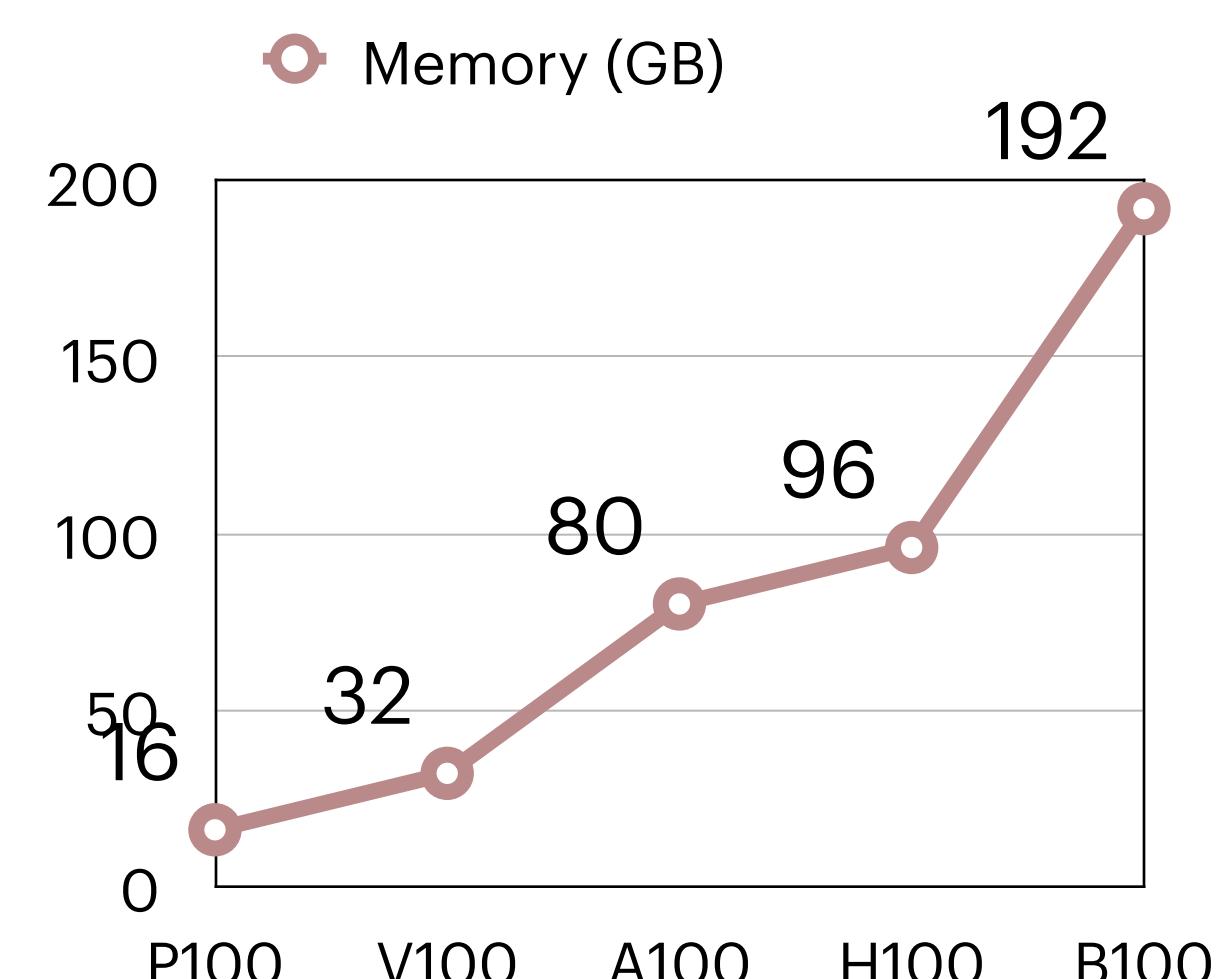
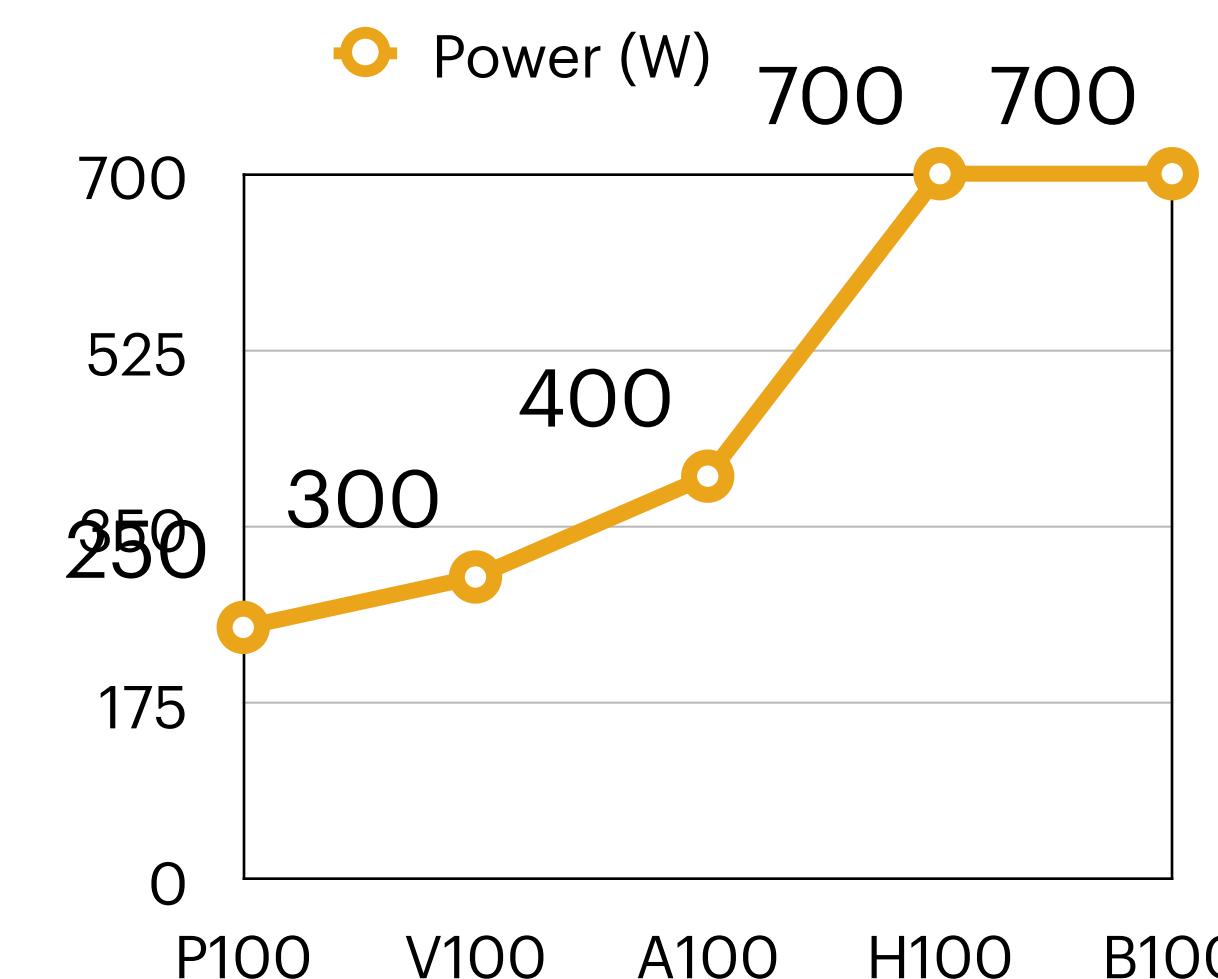
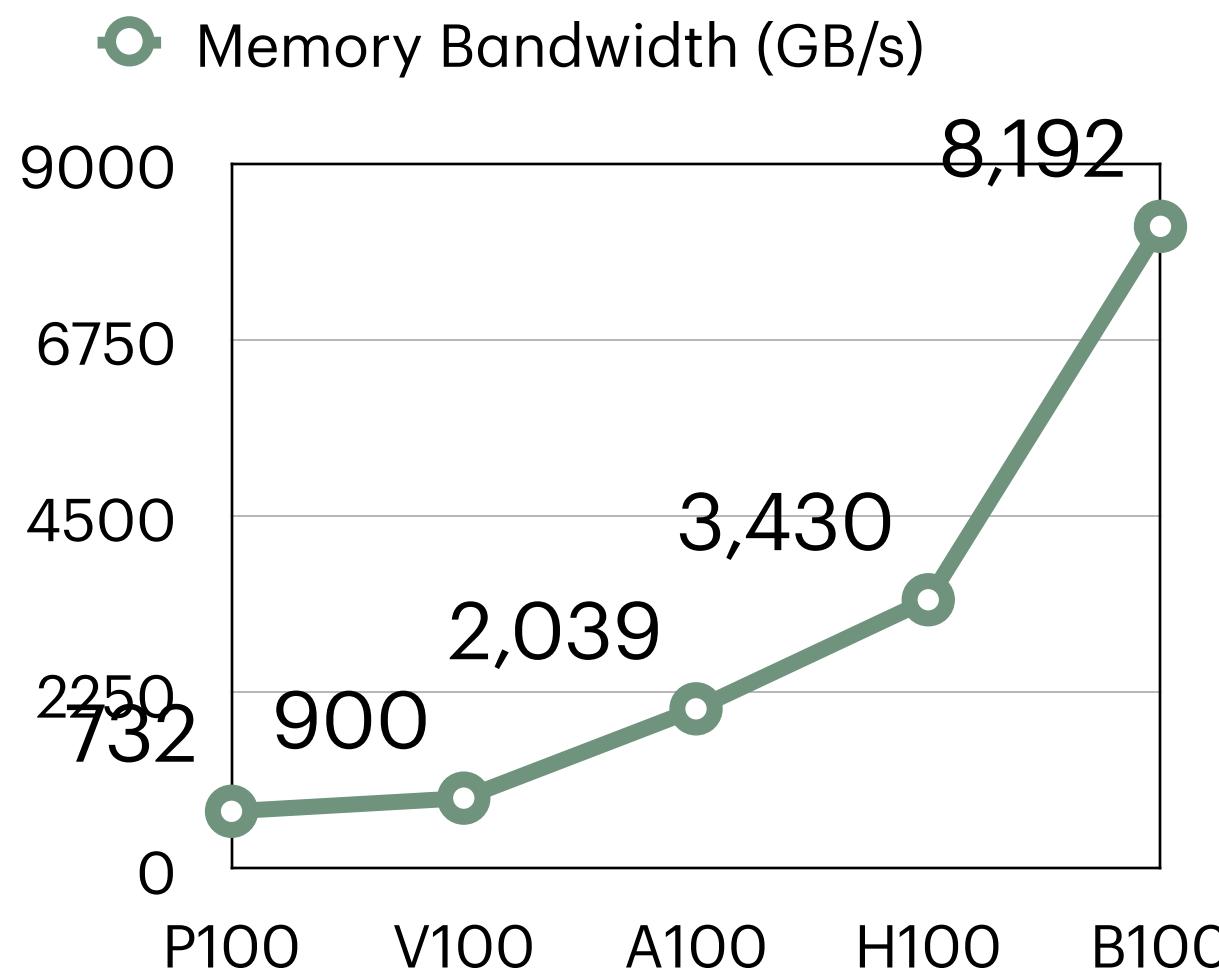
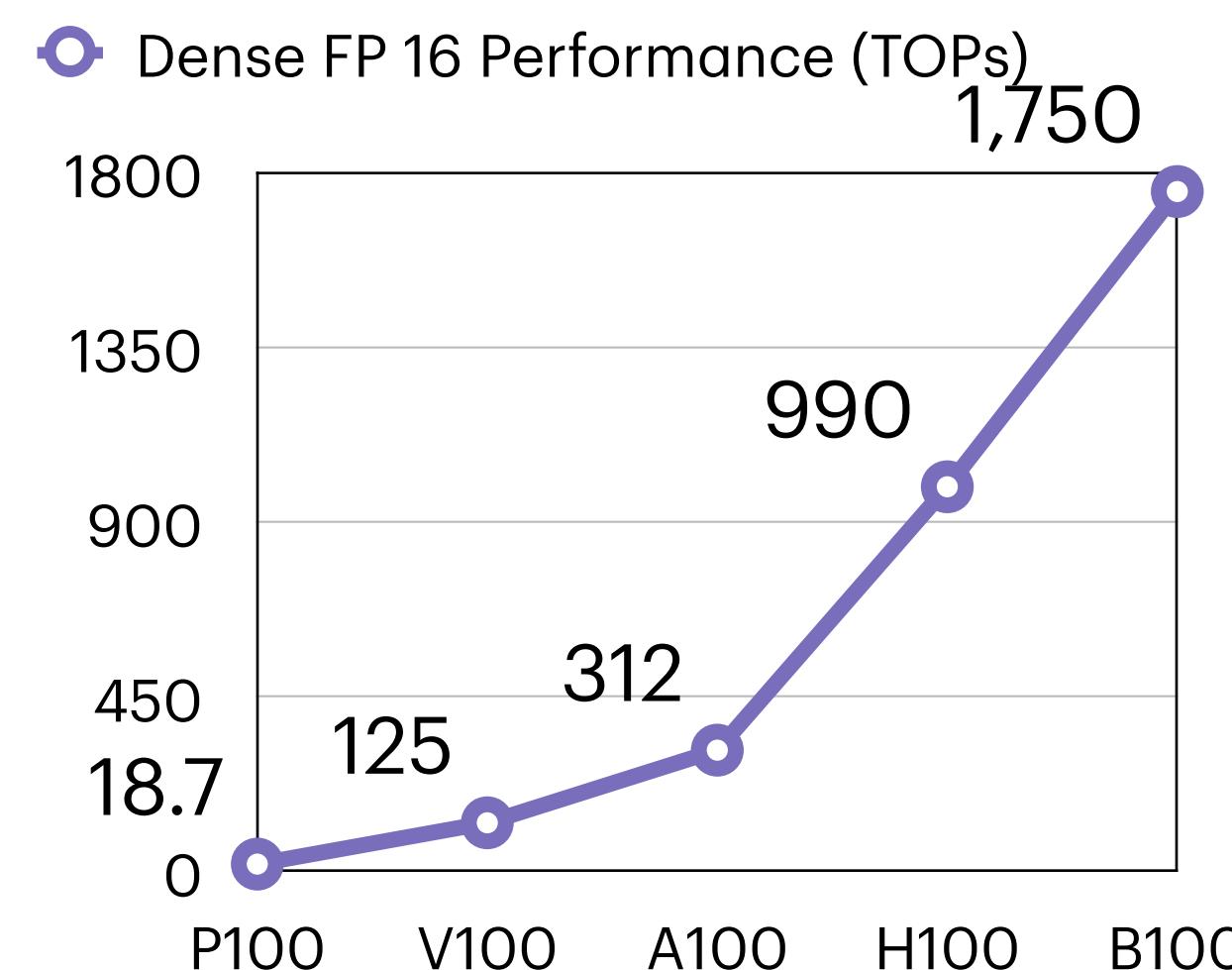
A100 (2020)



H100 (2022)



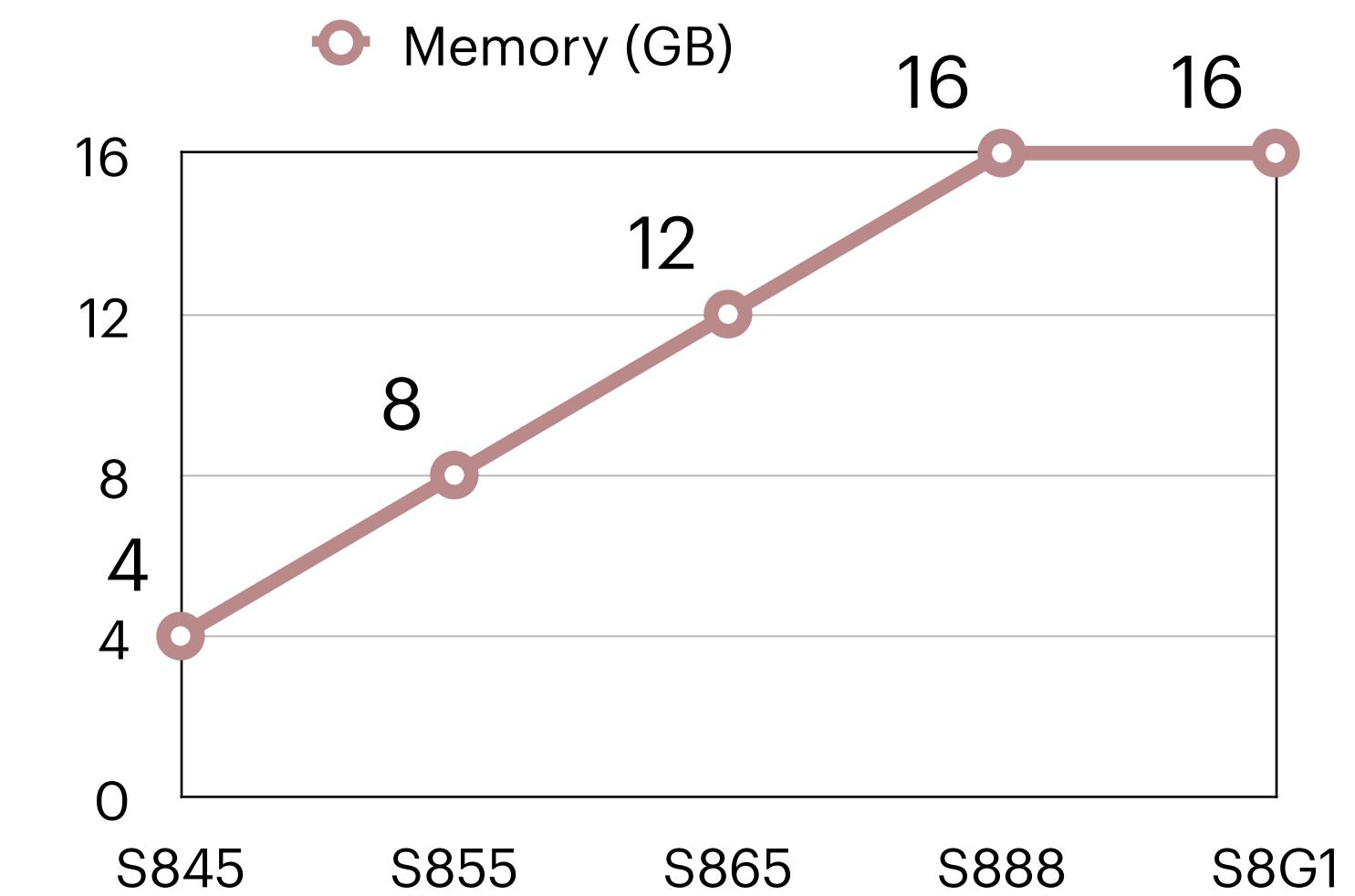
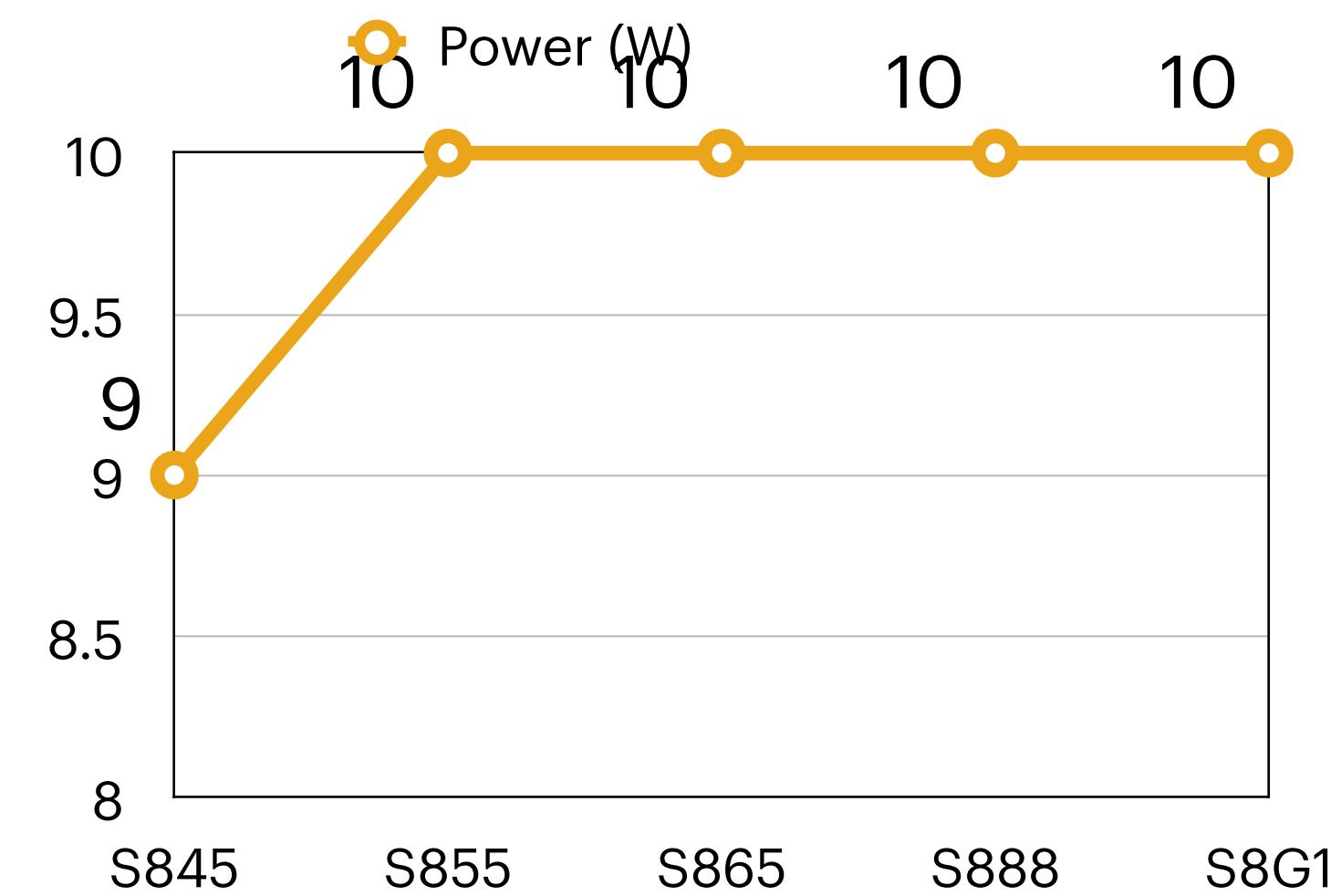
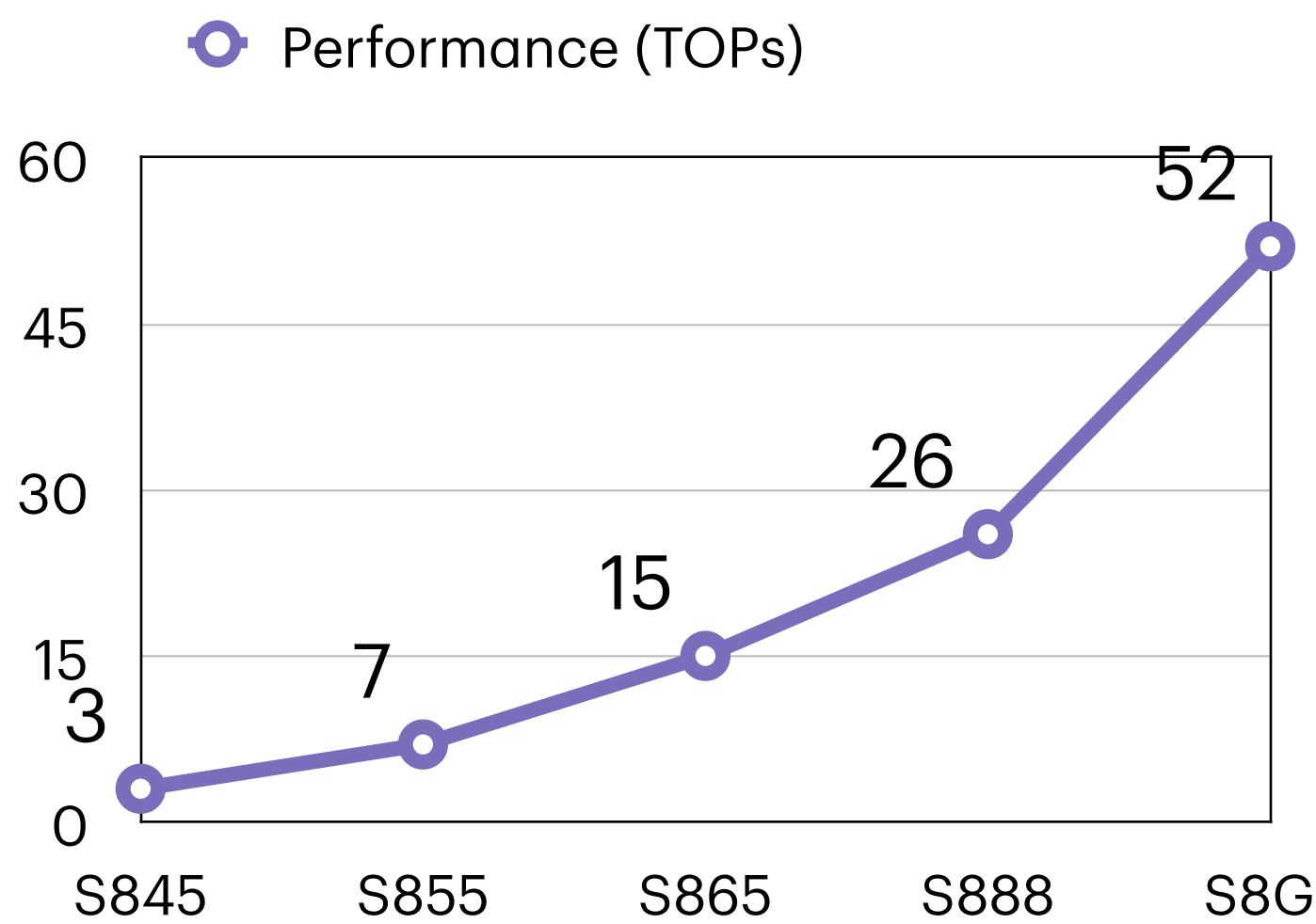
B100 (2024)



Edge AI Hardware

Qualcomm Hexagon DSP

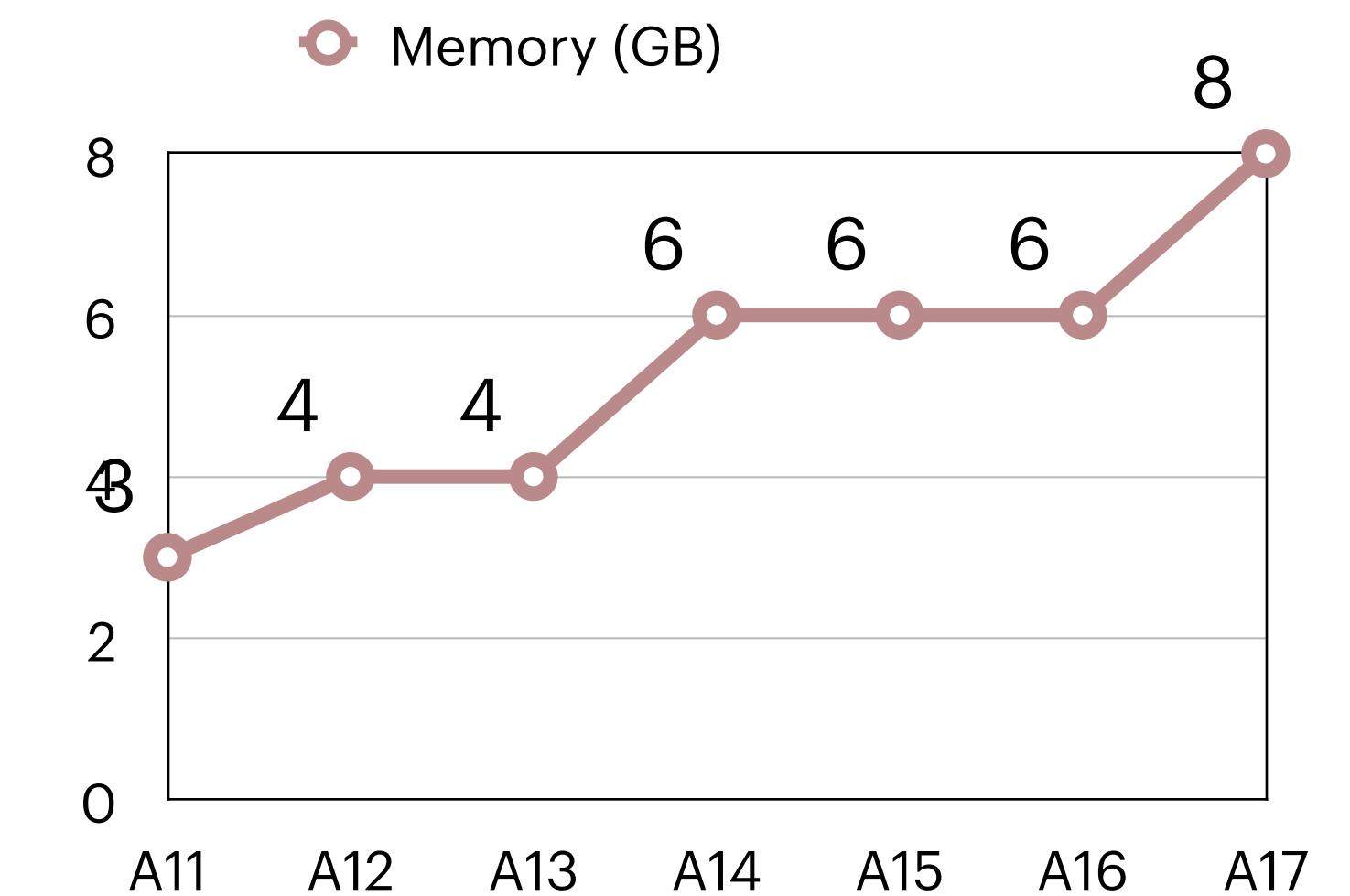
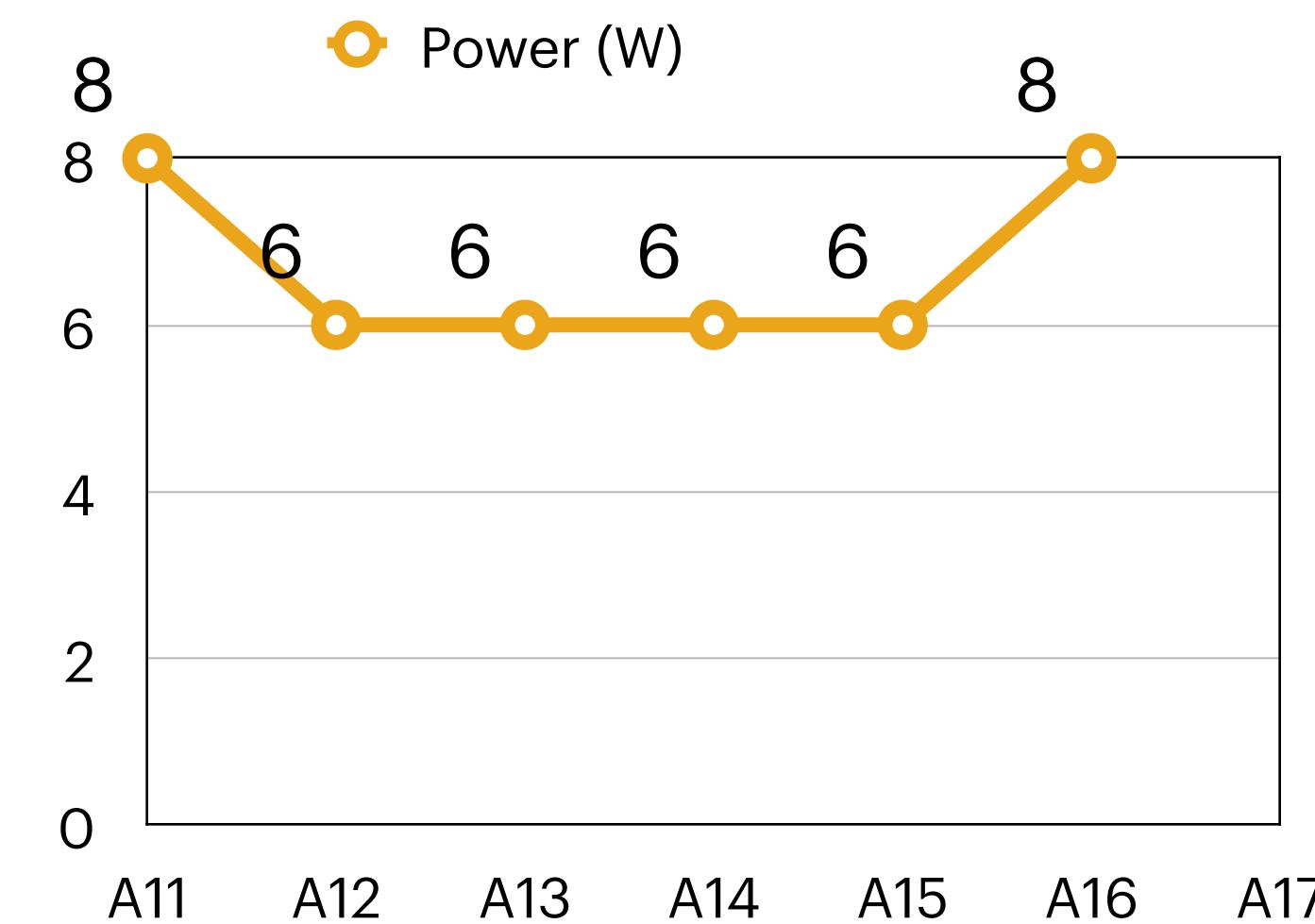
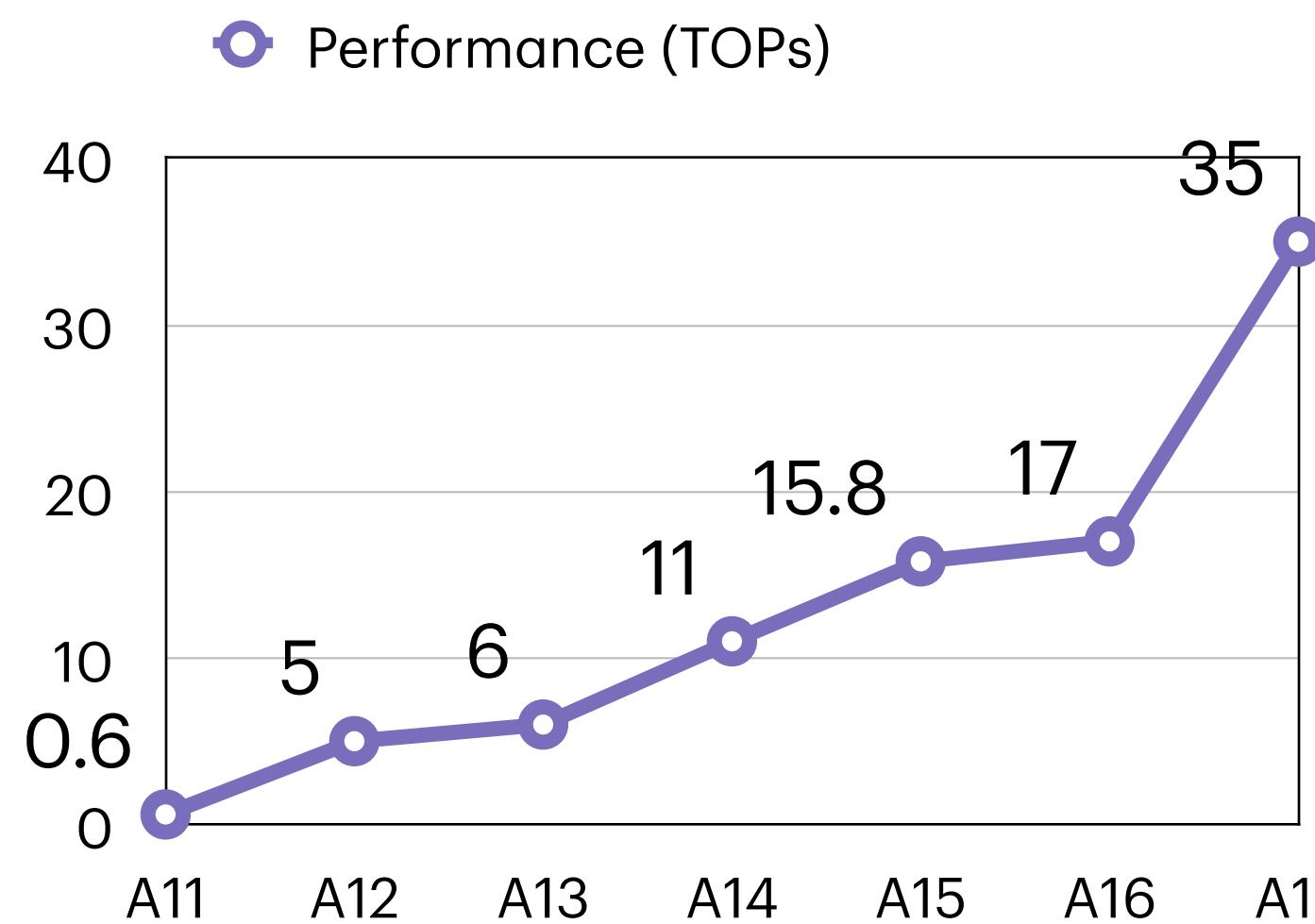
- Qualcomm Hexagon is a family of digital signal processor (DSP) products by Qualcomm. It is designed to deliver performance with low power over a variety of applications.



Edge AI Hardware

Apple Neural Engine

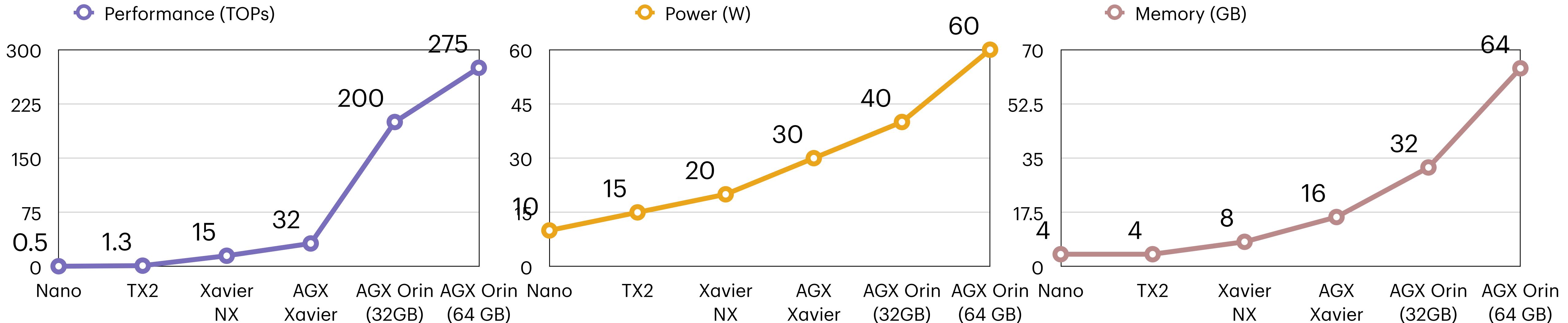
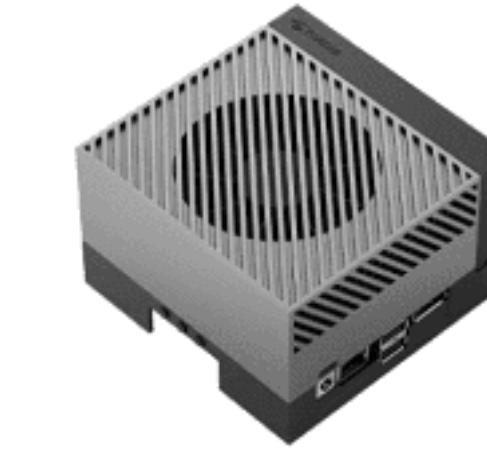
- The Apple Neural Engine (ANE) is an energy-efficient and high-throughput engine for ML inference on Apple silicon.



Edge AI Hardware

Nvidia Jetson

- NVIDIA Jetson is a complete System on Module (SOM) that includes a GPU, CPU, memory, power management, high-speed interfaces, and more.

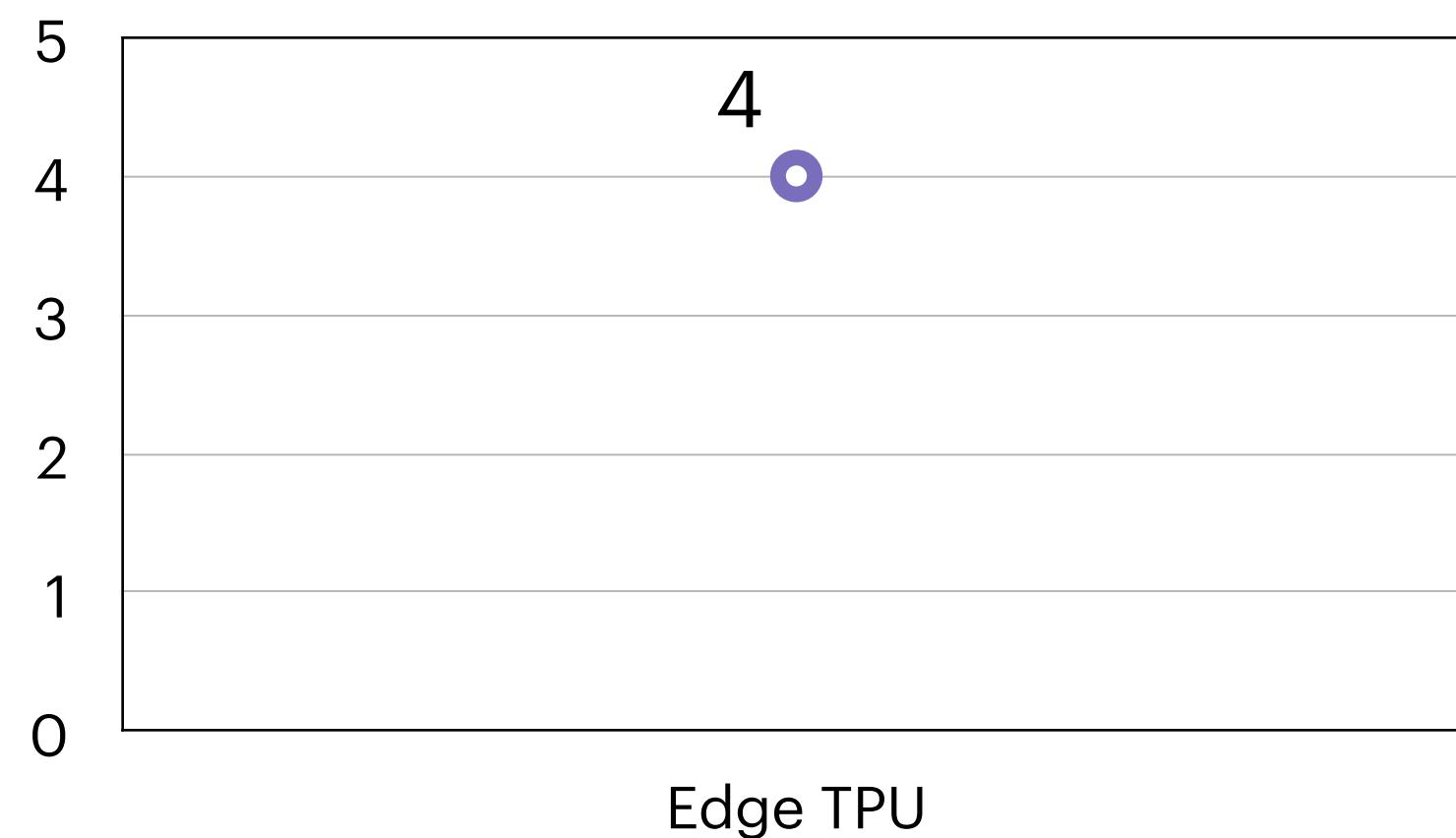


Edge AI Hardware

Tensor Processing Unit

- Tensor Processing Unit (TPU) is an AI accelerator application-specific integrated circuit (ASIC) developed by Google for neural network machine learning, using Google's own TensorFlow software

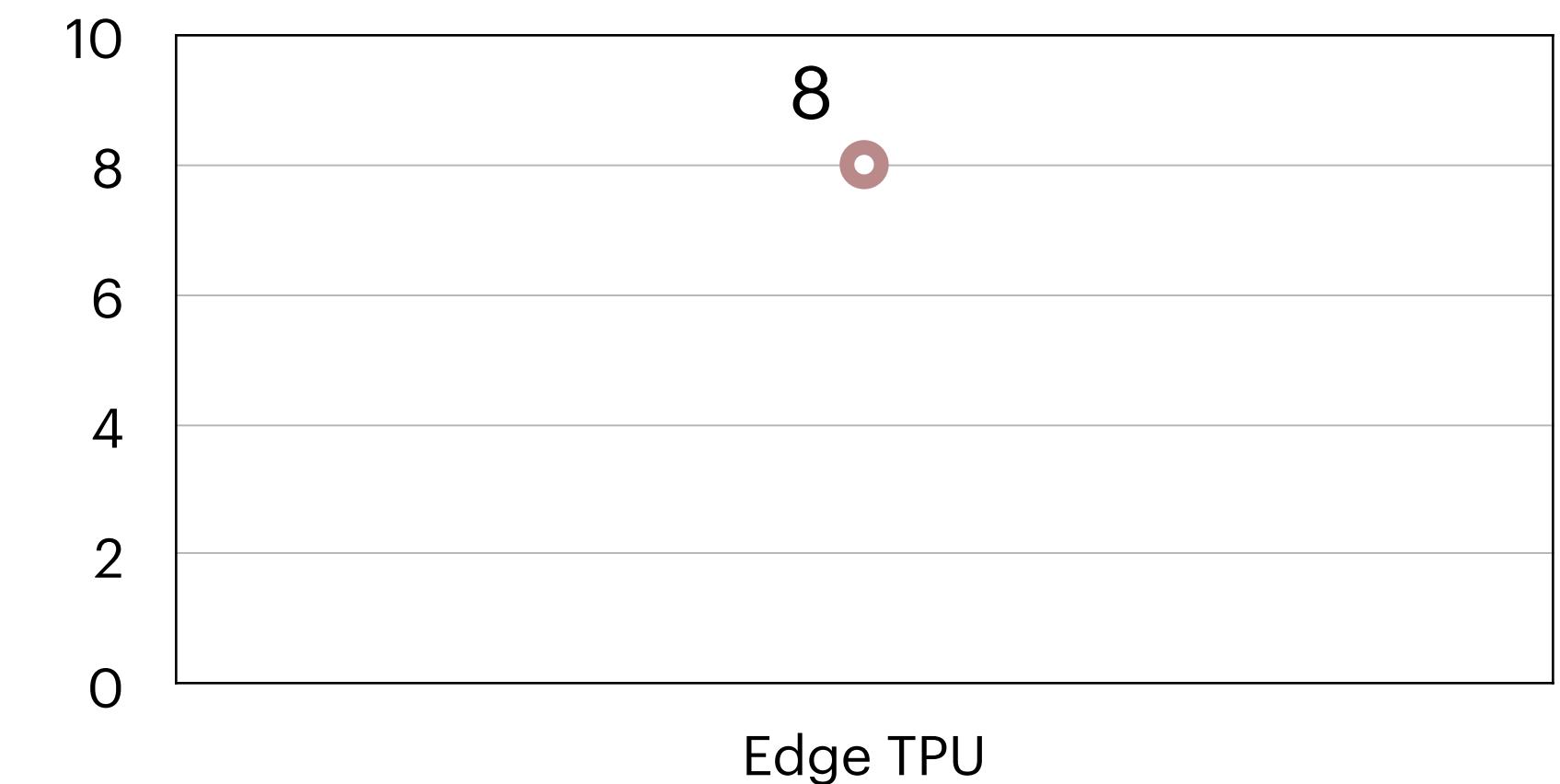
● Performance (TOPs)



● Power (W)



● Memory (GB)

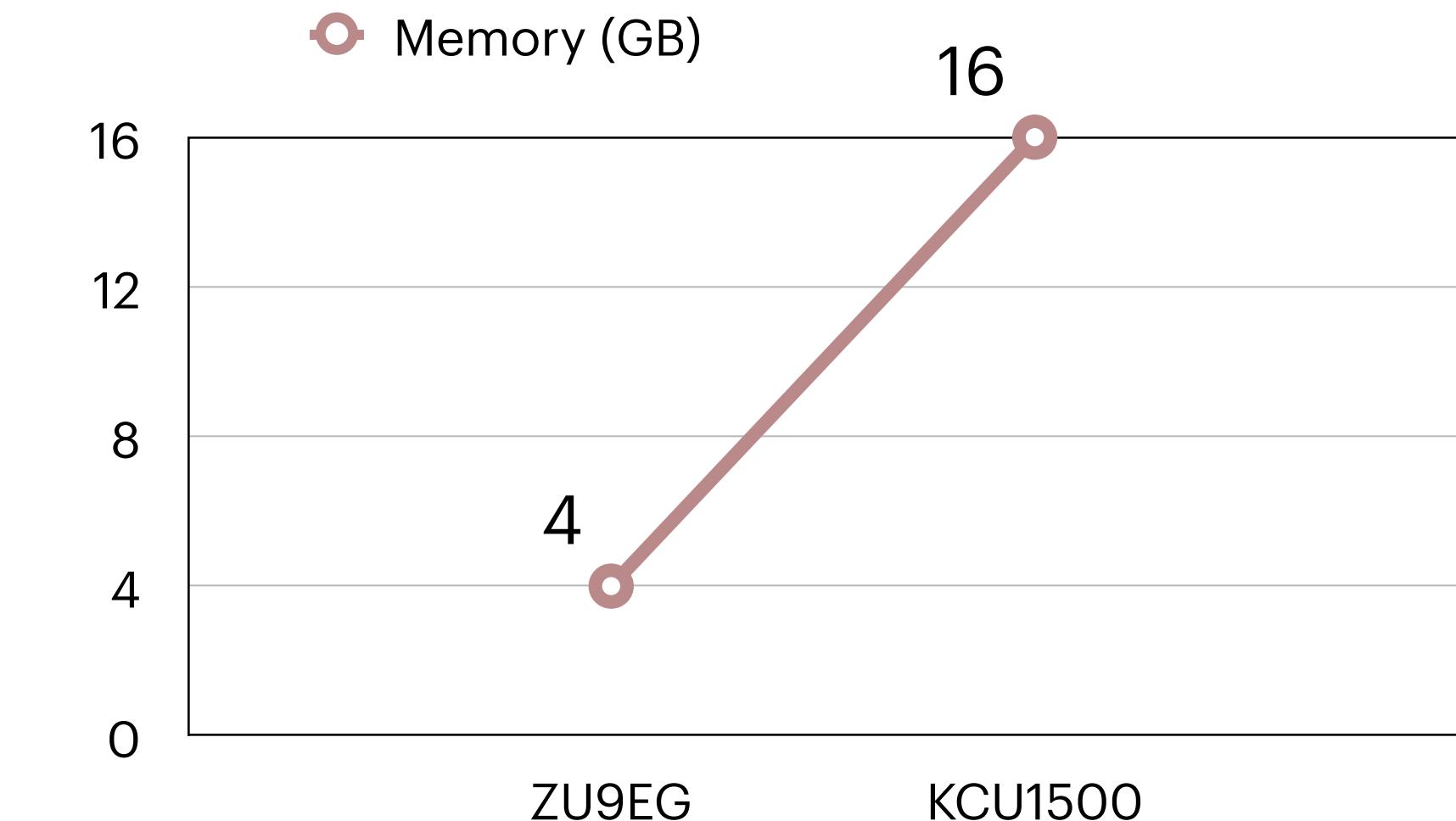
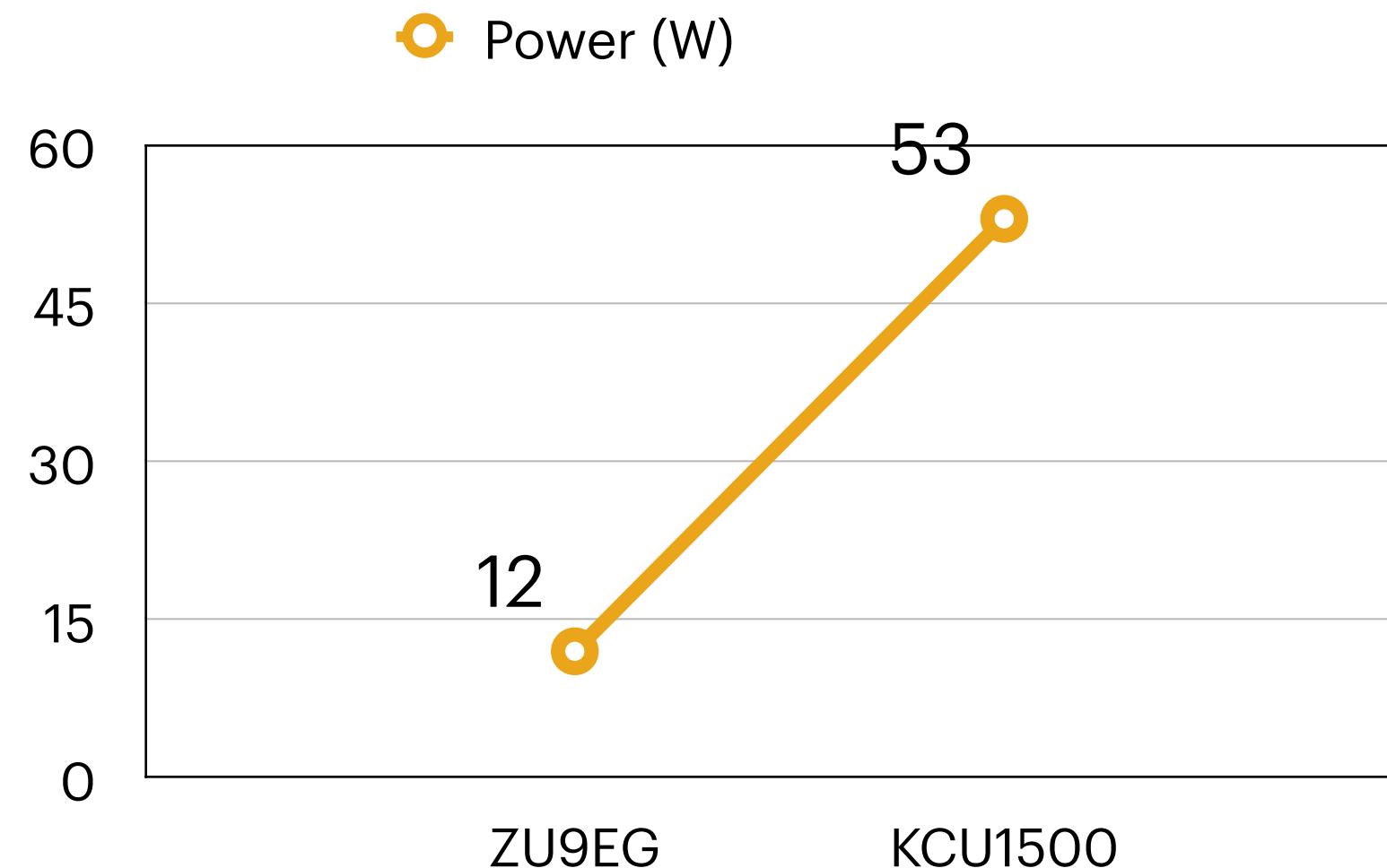
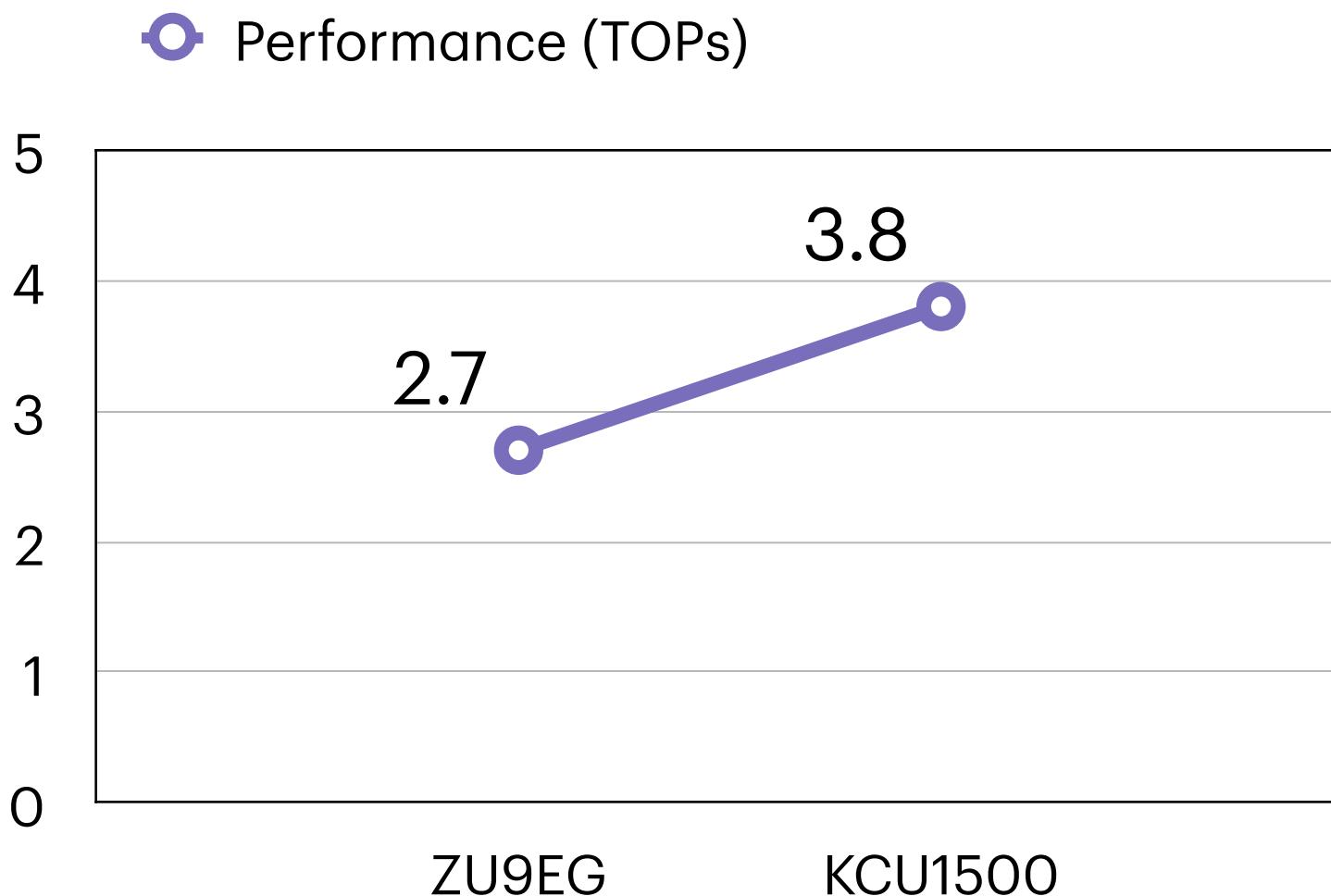
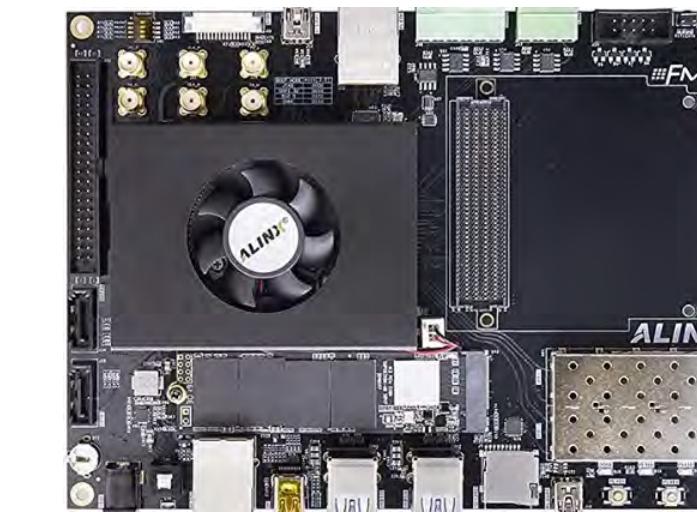


Tensor Processing Unit

Edge AI Hardware

FPGA-based accelerators

- Field Programmable Gate Arrays (FPGA) delivers higher performance compared to a fixed-architecture AI accelerator like a GPU due to efficiency of custom hardware acceleration

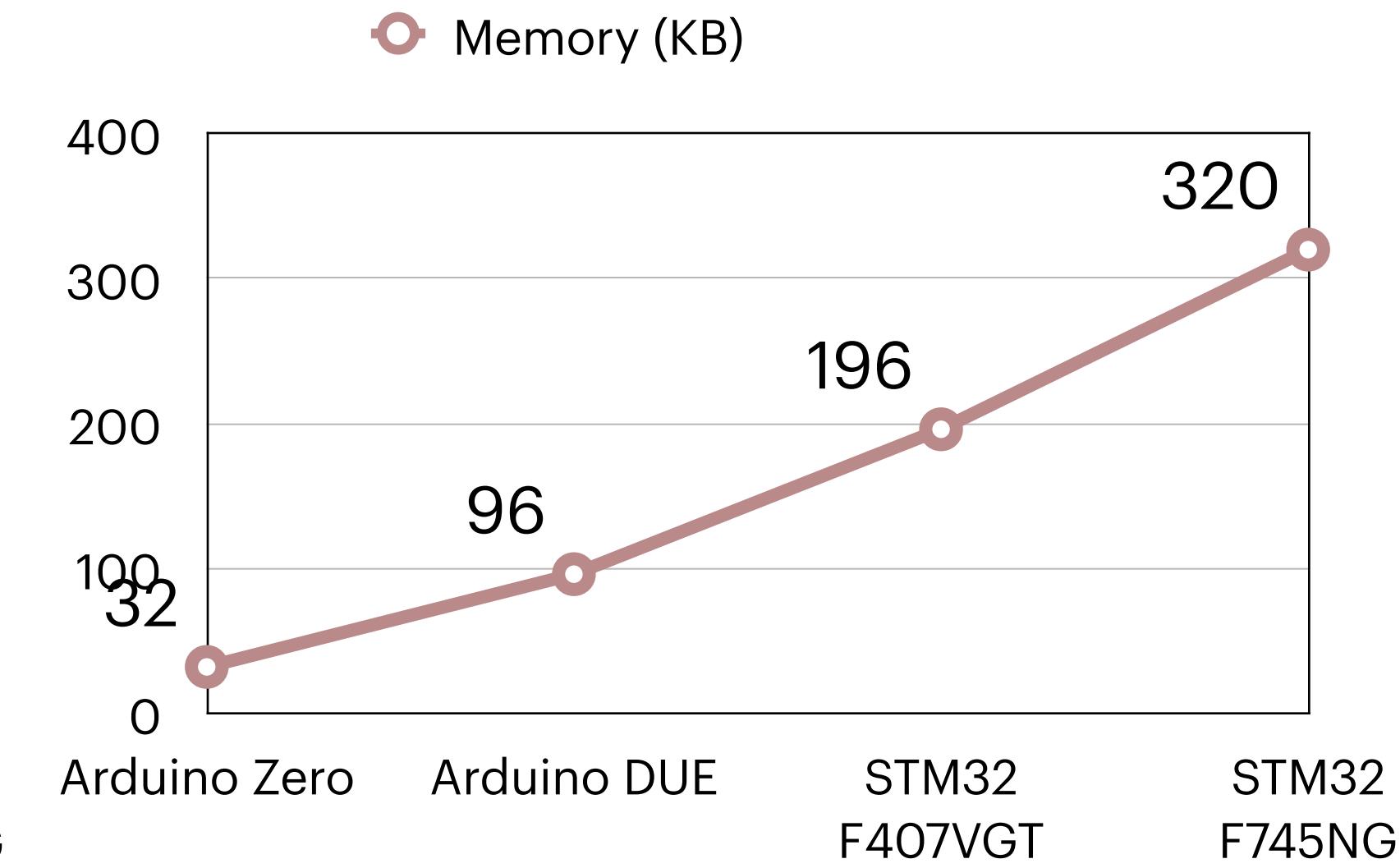
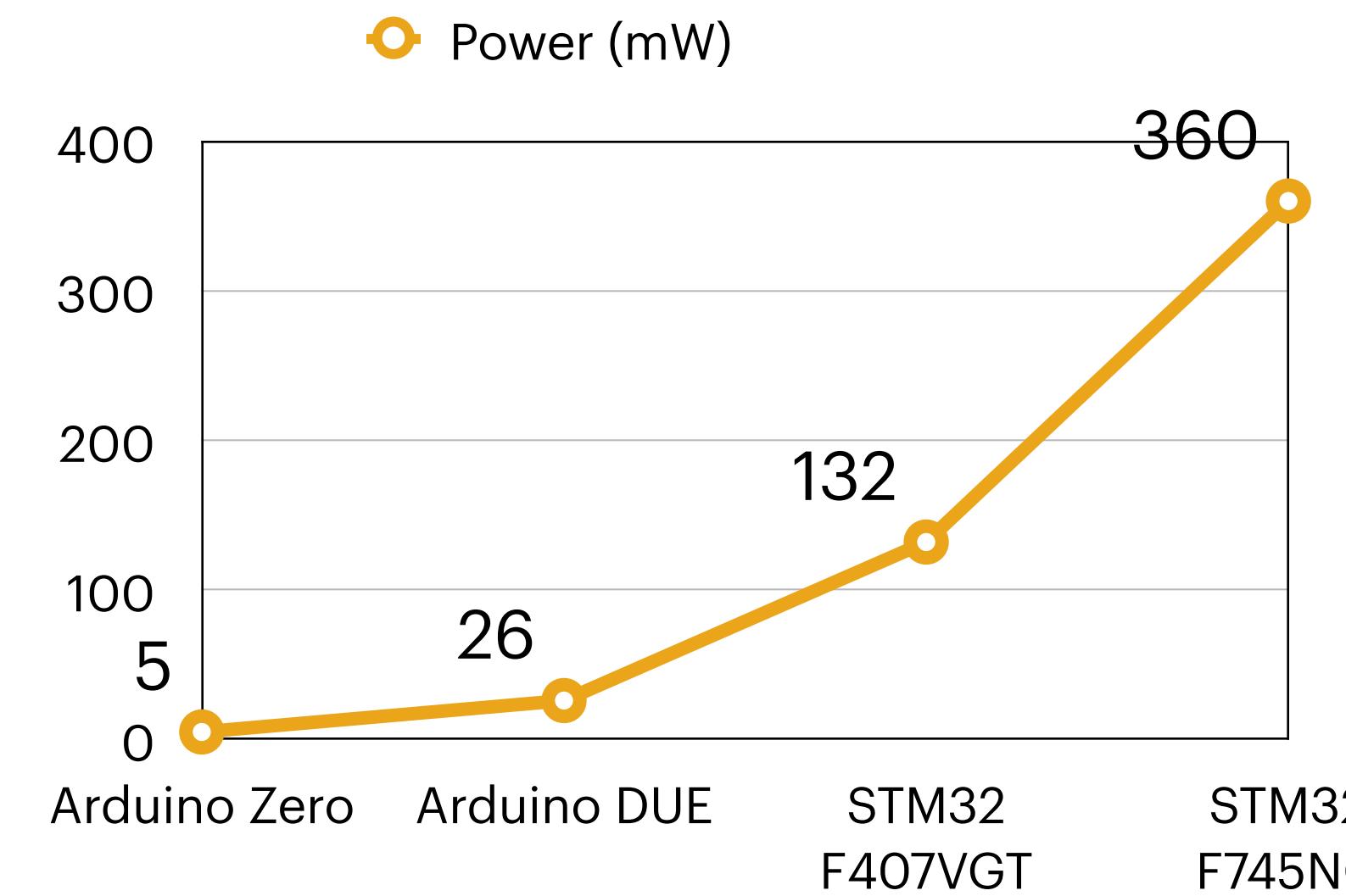
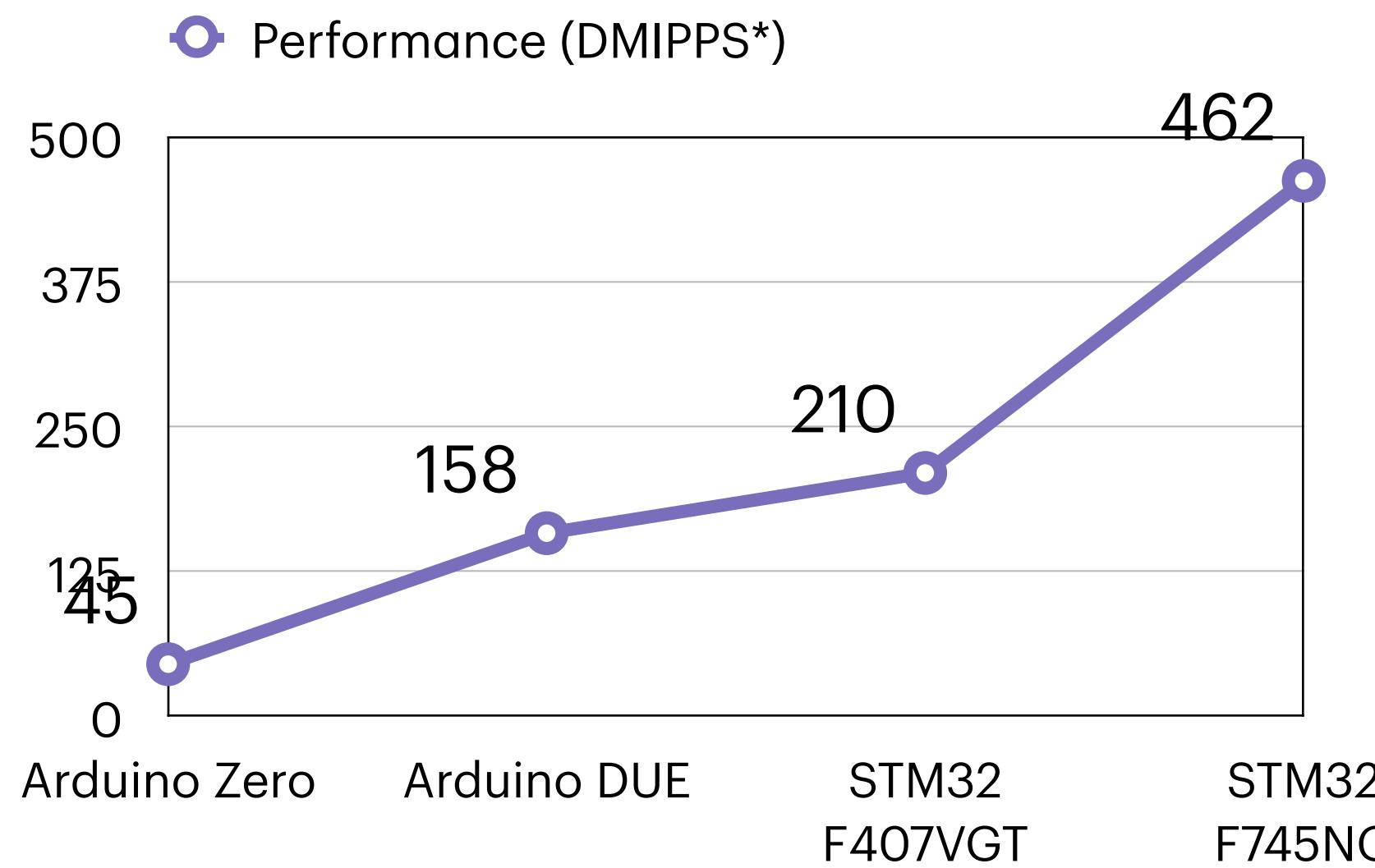


[Neural Network Accelerator Comparison](#)

Edge AI Hardware

Microcontrollers (MCU)

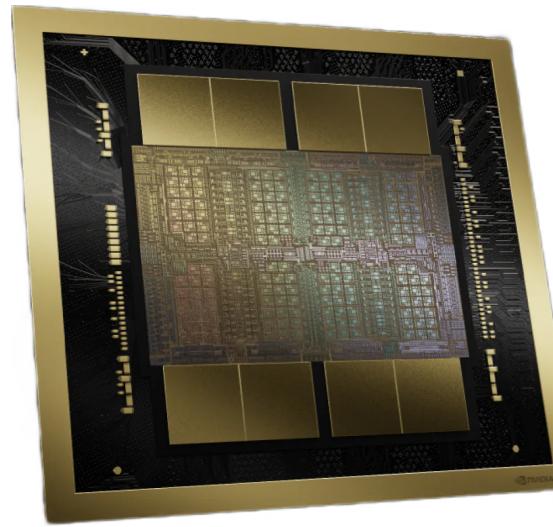
- A microcontroller is a compact integrated circuit designed for embedded systems. A typical microcontroller includes a processor, memory and input/output (I/O) peripherals on a single chip.



* Dhrystone Million Instructions Per Second (DMIPS) is an index for integer computation

Edge AI Hardware

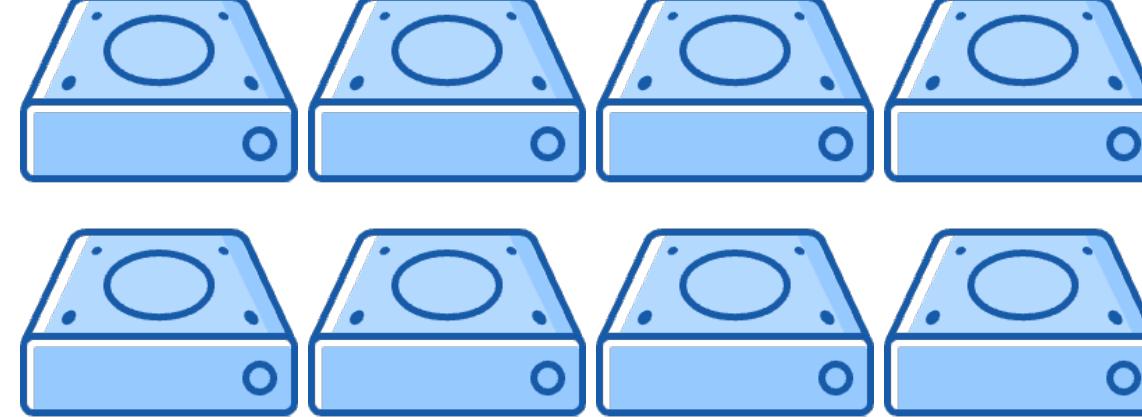
Edge AI devices still have huge gap to cloud processors



	Cloud AI	Mobile AI	Tiny AI
Memory (Activation)	80GB	4GB	320KB
Storage (Weights)	~TB/PB	256GB	1MB

Current Landscape of AI

Big Computation, engineer and data



A lot of computation
A lot of carbon



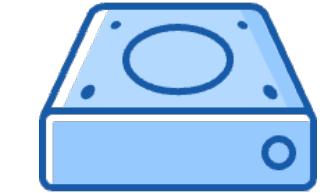
Many engineers



A lot of data

Edge AI

Lightweight computation, engineer and data

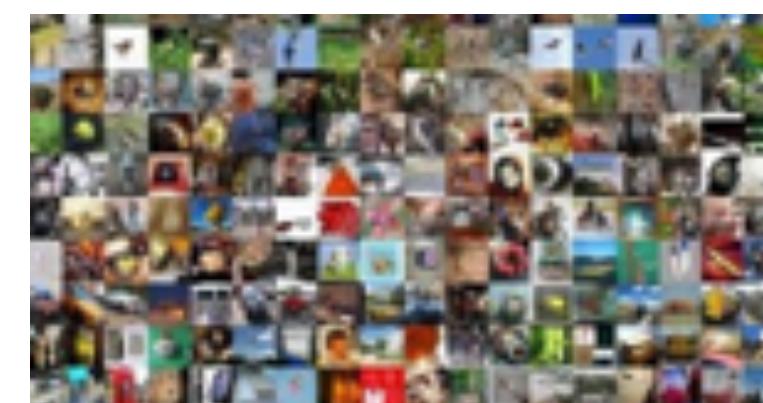


**Less computation
Less carbon**

Edge AI



Fewer engineers



Less data

Reference

- Deng, L., Li, G., Han, S., Shi, L., & Xie, Y. (2020). Model compression and hardware acceleration for neural networks: A comprehensive survey. *Proceedings of the IEEE*, 108(4), 485-532.
- Lin, J., Chen, W. M., Lin, Y., Gan, C., & Han, S. (2020). Mcunet: Tiny deep learning on iot devices. *Advances in neural information processing systems*, 33, 11711-11722.
- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 4015-4026).
- Cai, H., Li, J., Hu, M., Gan, C., & Han, S. (2022). Efficientvit: Multi-scale linear attention for high-resolution dense prediction. *arXiv preprint arXiv:2205.14756*.
- Li, M., Lin, J., Ding, Y., Liu, Z., Zhu, J. Y., & Han, S. (2020). GAN compression: Efficient architectures for interactive conditional gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5284-5294).
- Lin, J., Zhang, R., Ganz, F., Han, S., & Zhu, J. Y. (2021). Anycost gans for interactive image synthesis and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14986-14996).
- Li, M., Lin, J., Meng, C., Ermon, S., Han, S., & Zhu, J. Y. (2022). Efficient spatially sparse inference for conditional gans and diffusion models. *Advances in neural information processing systems*, 35, 28858-28873.
- Xiao, G., Yin, T., Freeman, W. T., Durand, F., & Han, S. (2024). Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, 1-20.
- Li, M., Cai, T., Cao, J., Zhang, Q., Cai, H., Bai, J., ... & Han, S. (2024). Distrifusion: Distributed parallel inference for high-resolution diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7183-7193).
- Liu, M., Shi, R., Chen, L., Zhang, Z., Xu, C., Wei, X., ... & Su, H. (2024). One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10072-10083).
- Liu, Z., Amini, A., Zhu, S., Karaman, S., Han, S., & Rus, D. L. (2021, May). Efficient and robust lidar-based end-to-end navigation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 13247-13254). IEEE.
- Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D. L., & Han, S. (2023, May). Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE international conference on robotics and automation (ICRA)* (pp. 2774-2781). IEEE.
- Wu, Z., Liu, Z., Lin, J., Lin, Y., & Han, S. (2020). Lite transformer with long-short range attention. *arXiv preprint arXiv:2004.11886*.
- Brown, T. B. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35, 24824-24837.
- Wang, H., Zhang, Z., & Han, S. (2021, February). SpAtten: Efficient sparse attention architecture with cascade token and head pruning. In *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)* (pp. 97-110). IEEE.
- Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2024). Visual instruction tuning. *Advances in neural information processing systems*, 36.
- Lin, J., Tang, J., Tang, H., Yang, S., Chen, W. M., Wang, W. C., ... & Han, S. (2024). AWQ: Activation-aware Weight Quantization for On-Device LLM Compression and Acceleration. *Proceedings of Machine Learning and Systems*, 6, 87-100.