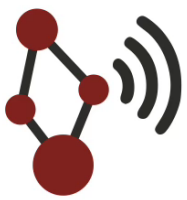
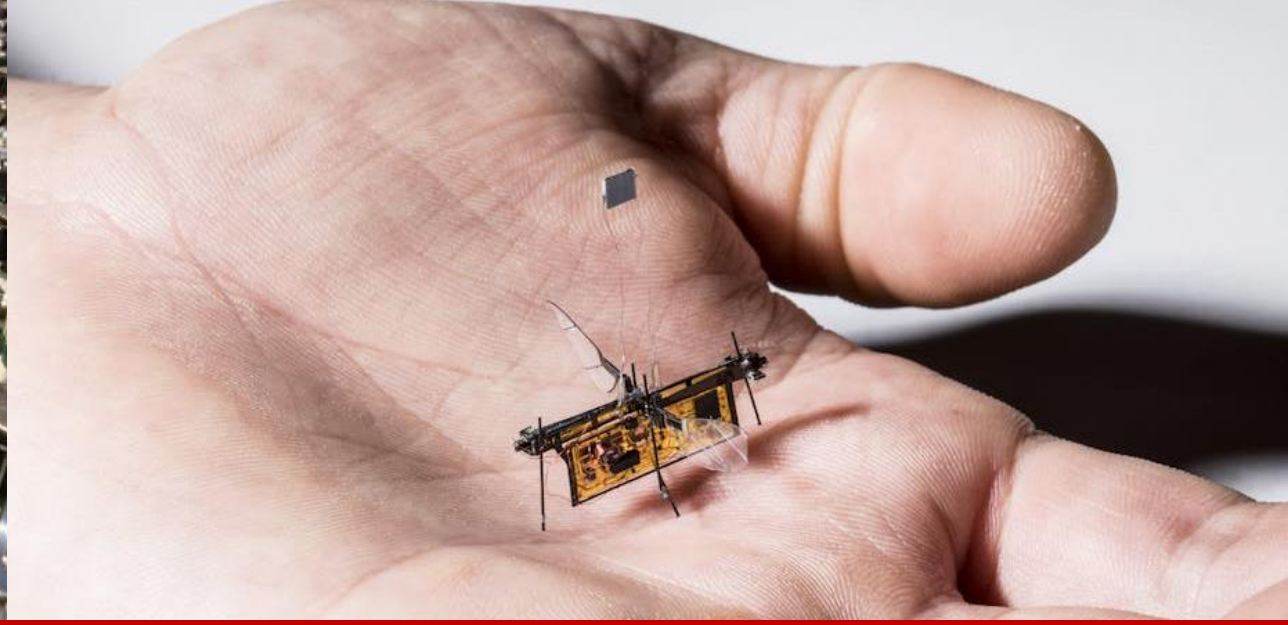
Ultra-Lightweight Edge Intelligence Using Vector Symbolic Architecture

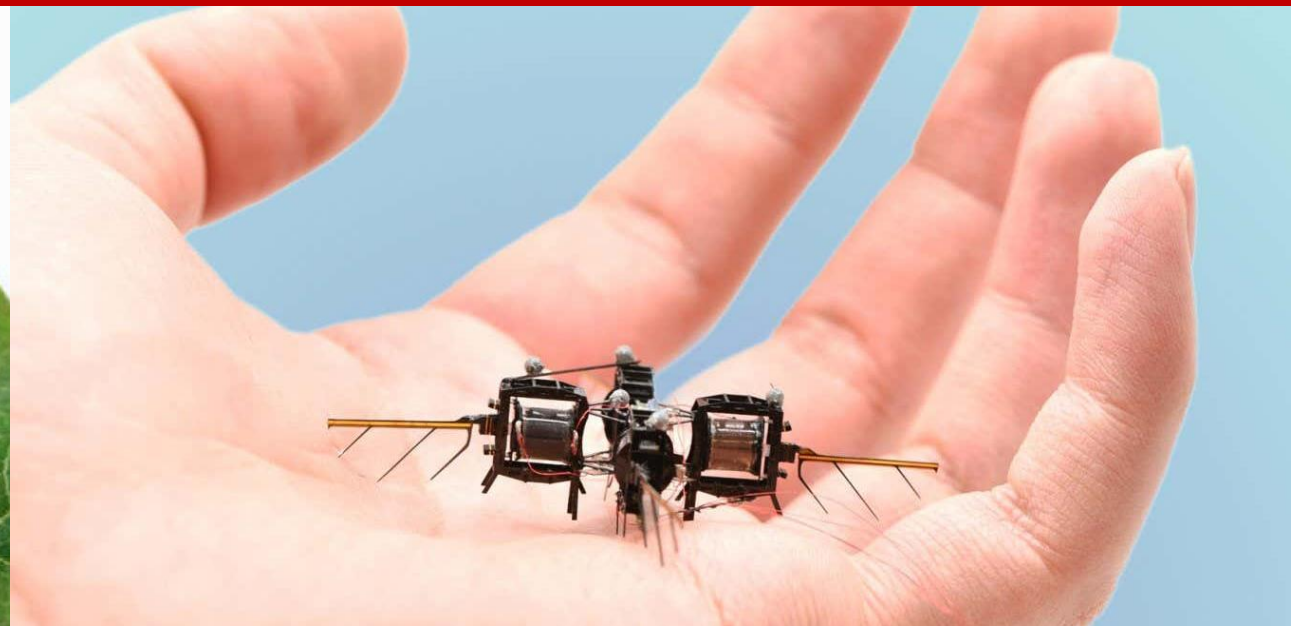
Chae Young Lee
Stanford University
chae@stanford.edu

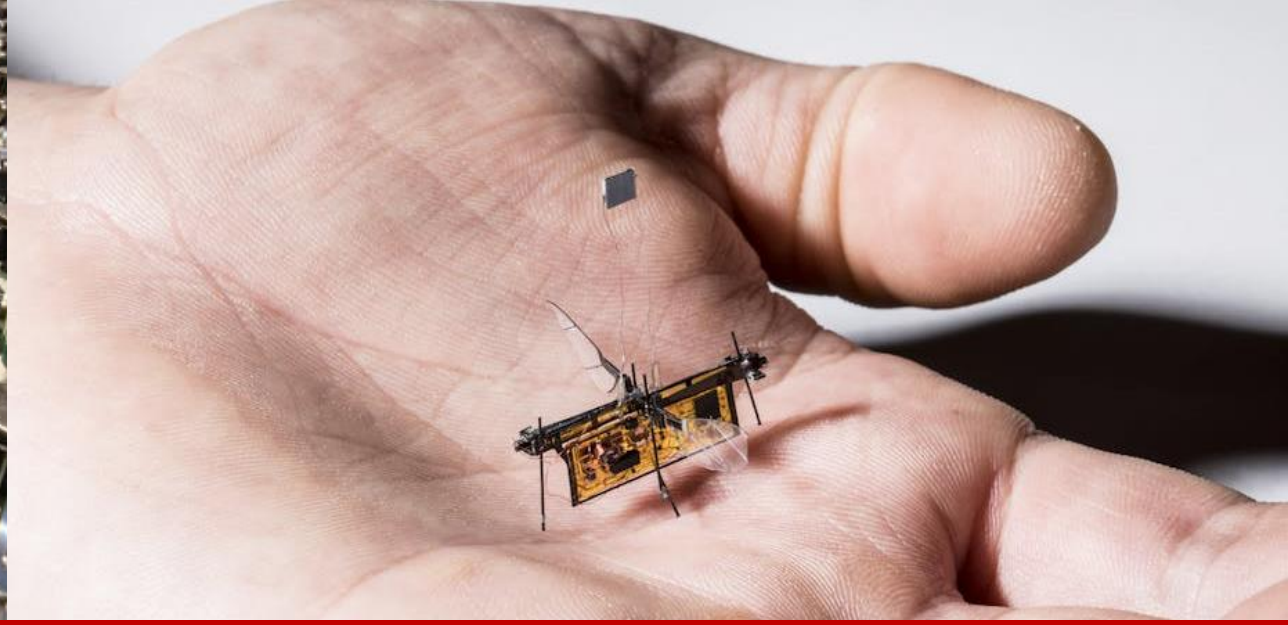


Stanford
SMART SENSING
SYSTEMS LAB

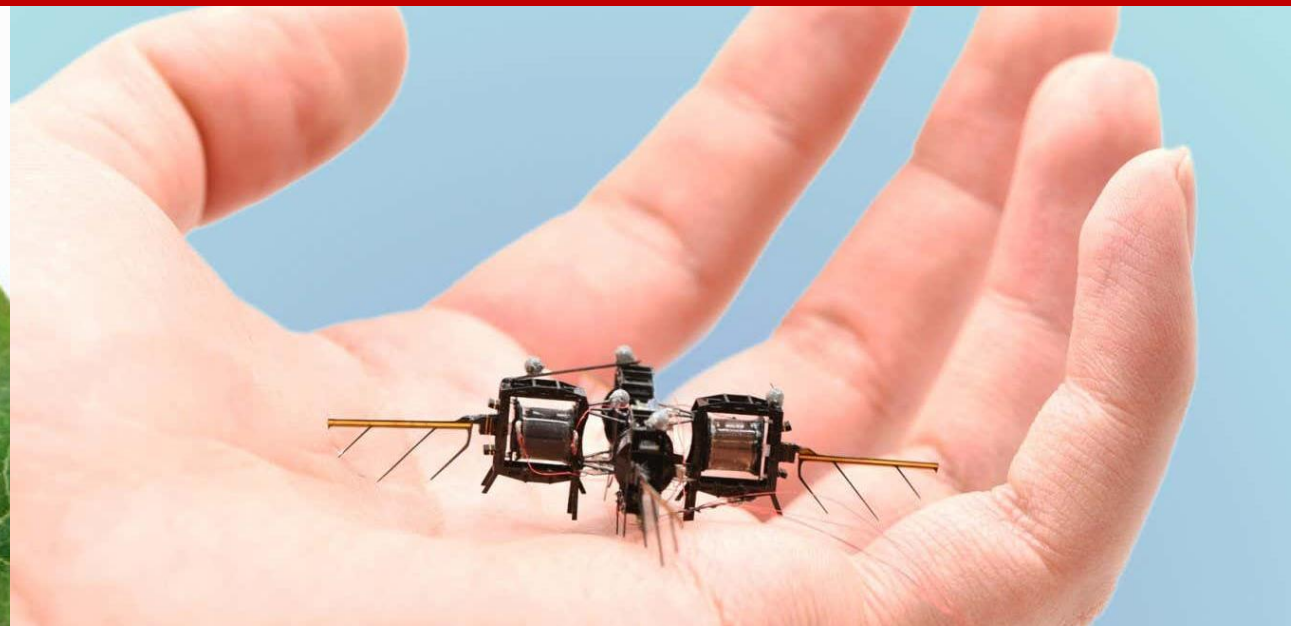


Tiny robots are emerging with many applications.





They require onboard intelligence for navigation.



IoT cameras are energy and bandwidth constrained.

They can greatly benefit from onboard intelligence.

Defining “ultra-lightweight” edge intelligence

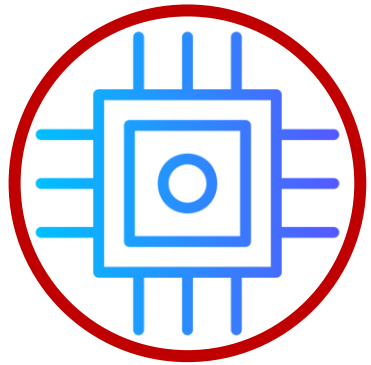
Providing intelligence with **ultra-light**

Power Consumption

<10mW

Memory Footprint

<100KB flash & RAM

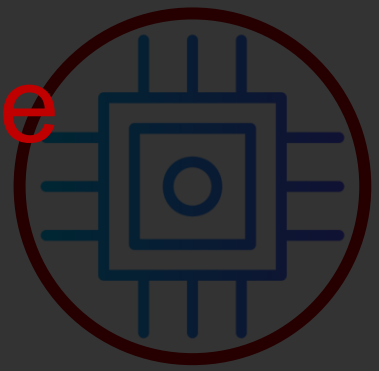


Compute Load

<100k CPU cycle

Defining “ultra-lightweight” edge intelligence

Providing intelligence with **ultra-light**



DNNs tradeoff performance with resource efficiency!

Power Consumption	Memory Footprint	Compute Load
<10mW	<100KB flash & RAM	<100k CPU cycle

Defining “ultra-lightweight” edge intelligence

Providing intelligence with **ultra-light**

Is there an alternative framework that is inherently more resource-efficient?

Power Consumption	Memory Footprint	Compute Load
<10mW	<100KB flash & RAM	<100k CPU cycle

Introducing Vector Symbolic Architecture (VSA)

1. Represents data as fixed-length vectors (typically binary).

Encoder

[0, 1, 0, 1, 1, …, 1]

Introducing Vector Symbolic Architecture (VSA)

2. Uses simple operations.

Encoder

[0, 1, 0, 1, 1, …, 1]

+, x, ^, >>

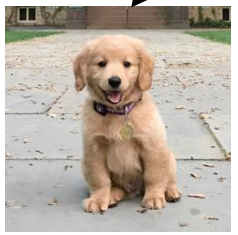
Introducing Vector Symbolic Architecture (VSA)

3. Performs classification using distances.

Encoder

[0, 1, 0, 1, 1, …, 1]

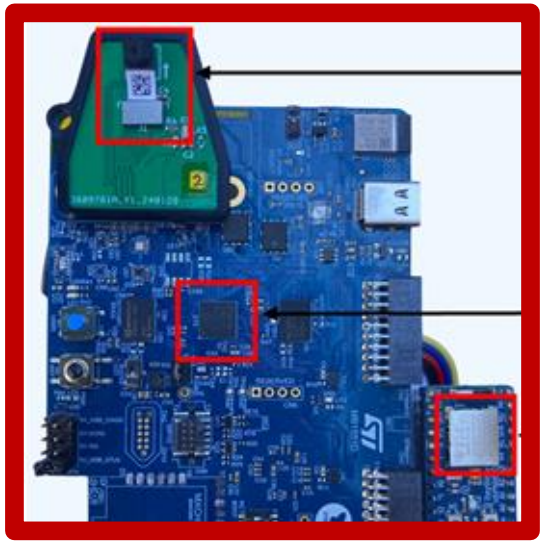
Smaller Distance!



Dog
[0, …, 1]

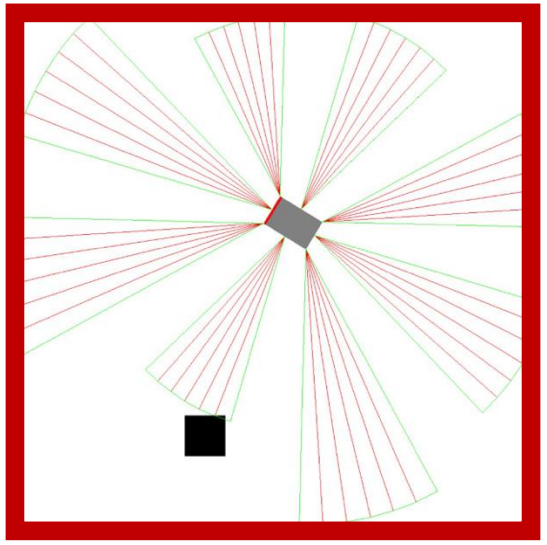
Cat
[1, …, 1]

My research builds ultra-lightweight VSA



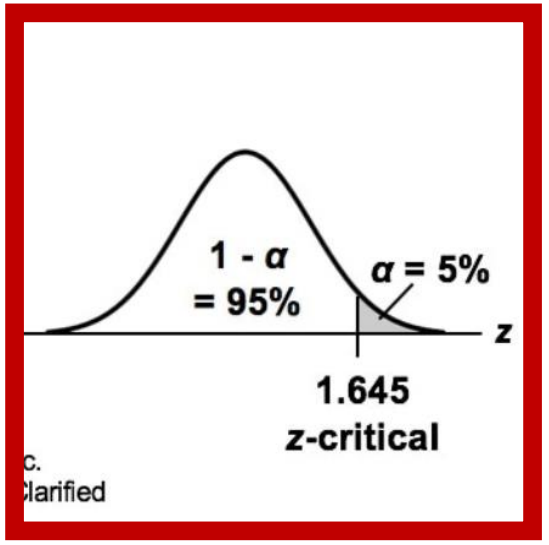
HyperCam
(MobiCom '25)

Image Classification
~43 kB of flash



NavHD
(IROS '25)

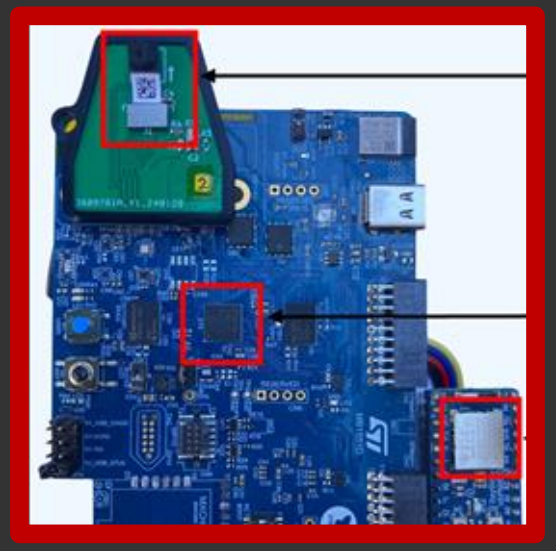
Tiny Robot Navigation
~10 kB of flash



Omen
(ASPLOS '25)

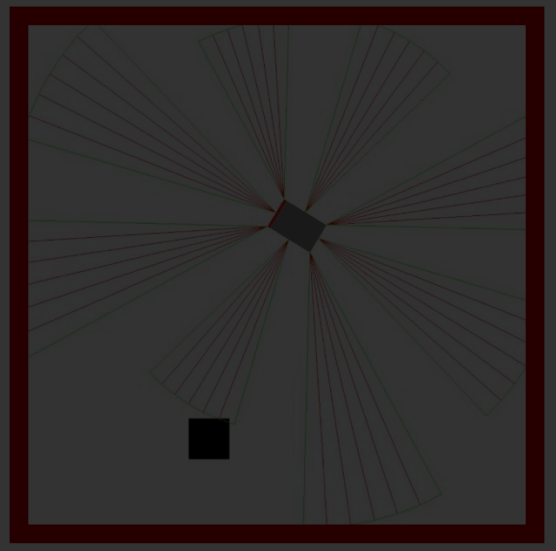
Inference Time
7-12x speedup

My research builds ultra-lightweight VSA



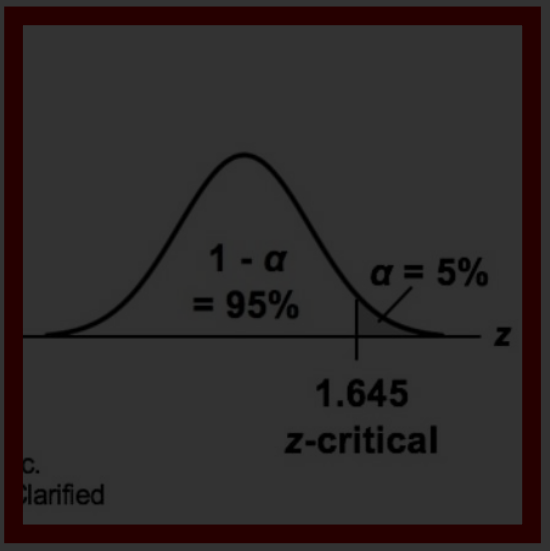
HyperCam
(MobiCom ’25)

Image Classification
~43 kB of flash



NavHD
(IROS ’25)

Tiny Robot Navigation
~10 kB of flash



Omen
(ASPLOS ’25)

Inference Time
7-12x speedup

HyperCam: onboard image classification

Naïve VSA encoding is memory-, time- consuming.

Encoder

[0, 1, 0, 1, 1, …, 1]

	Naïve VSA Encoder	HyperCam Encoder
Stored Vectors	284	6
Binding Operations	38400	256
Bundling Operations	19200	256 + 38*

*equivalent to 19200 sparse bundling operations

HyperCam: onboard image classification

Key idea is sparsity: $O(whn) \rightarrow O(whd)$,

where w, h are image width and height and $n = 10000, d = 20$.

$$\text{Bundle}(\mathbf{X}) = \text{MajorityVote}(\underbrace{\{1, \dots, 0\}}_{n}, \dots, \underbrace{\{1, \dots, 1\}}_{n})$$

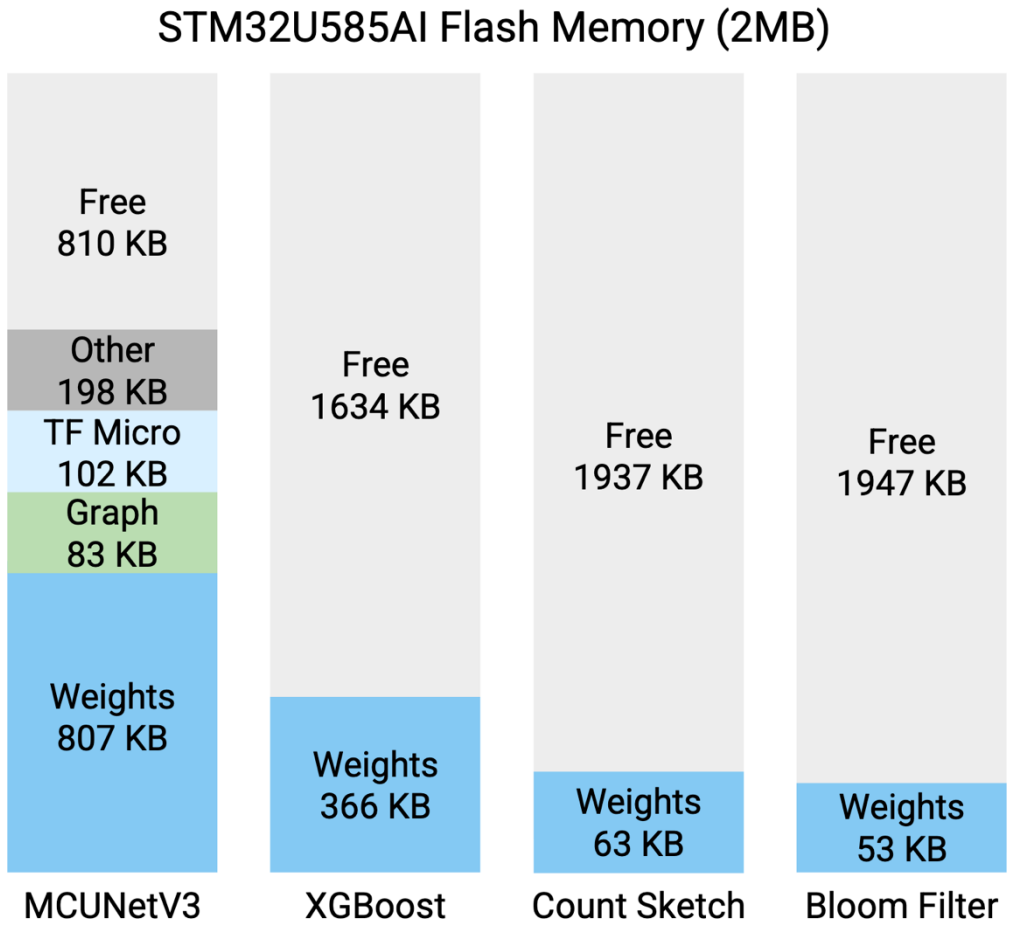
is approximated by $\text{SparseBundle}(\mathbf{S}) = \text{CountSketch}(\sum_{\mathbf{s} \in \mathbf{S}} p^{\mathbf{s}}[\mathbf{v}])$

$$= \text{BloomFilter}(\bigcup_{\mathbf{s} \in \mathbf{S}} p^{\mathbf{s}}[\mathbf{v}])$$

$\boldsymbol{s} \in \boldsymbol{S}$ is size n with $\boldsymbol{d}$ **non-zero bits**

HyperCam: onboard image classification

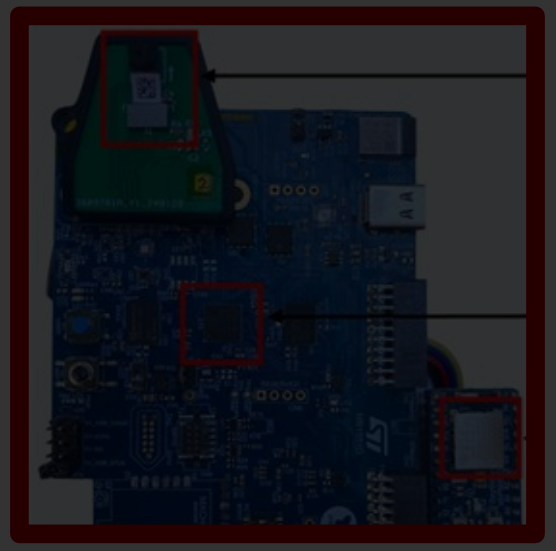
Face Detection Dataset



Achieves **+4%**, **-7%** test accuracy than MobileNet V3, MCUNet V3.

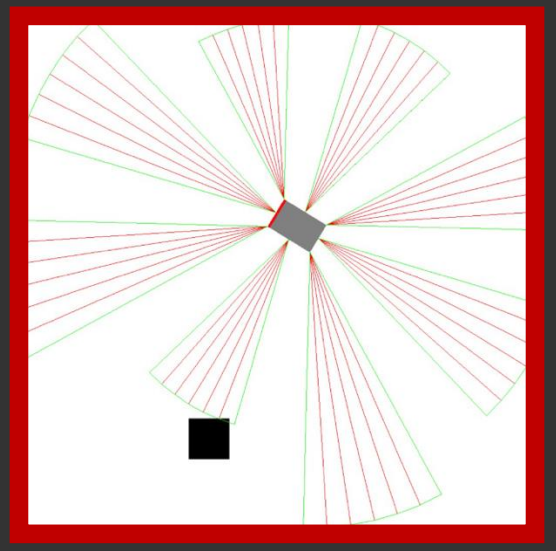
Uses **38x**, **27x** less memory than MobileNet V3, MCUNet V3.

My research builds ultra-lightweight VSA



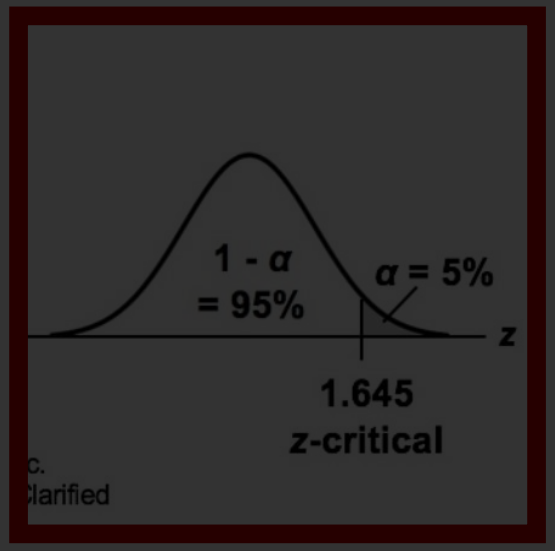
HyperCam
(MobiCom ’25)

Image Classification
~43 kB of flash



NavHD
(IROS ’25)

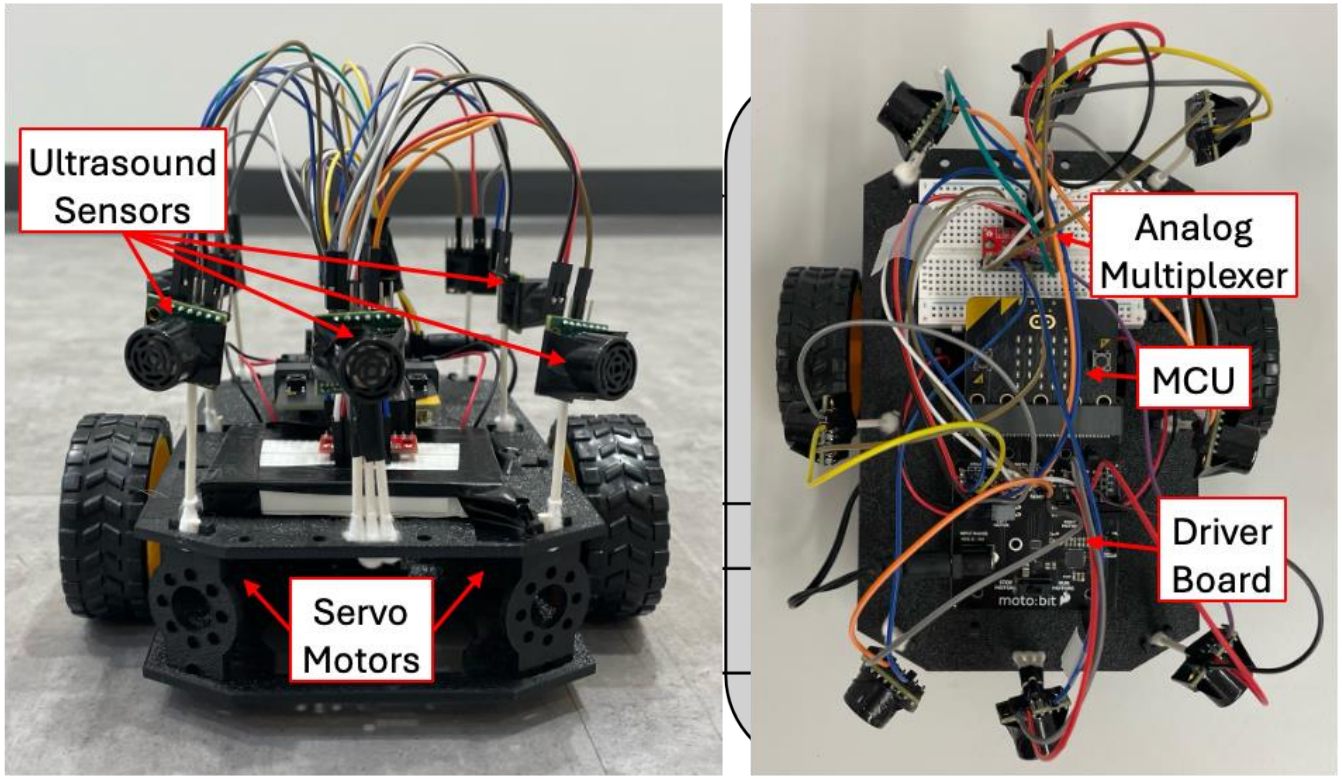
Tiny Robot Navigation
~10 kB of flash



Omen
(ASPLOS ’25)

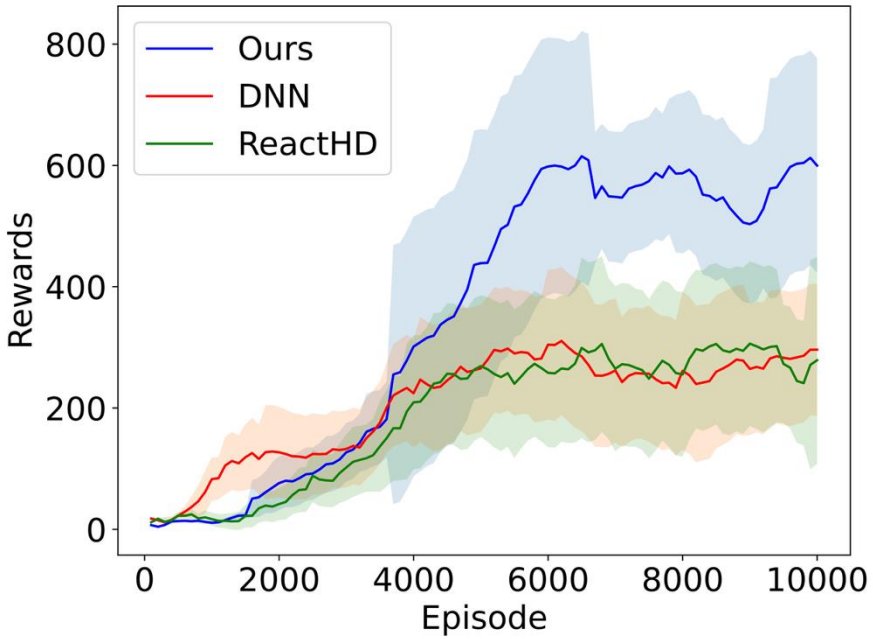
Inference Time
7-12x speedup

NavHD: off-policy RL for micro-robot navigation

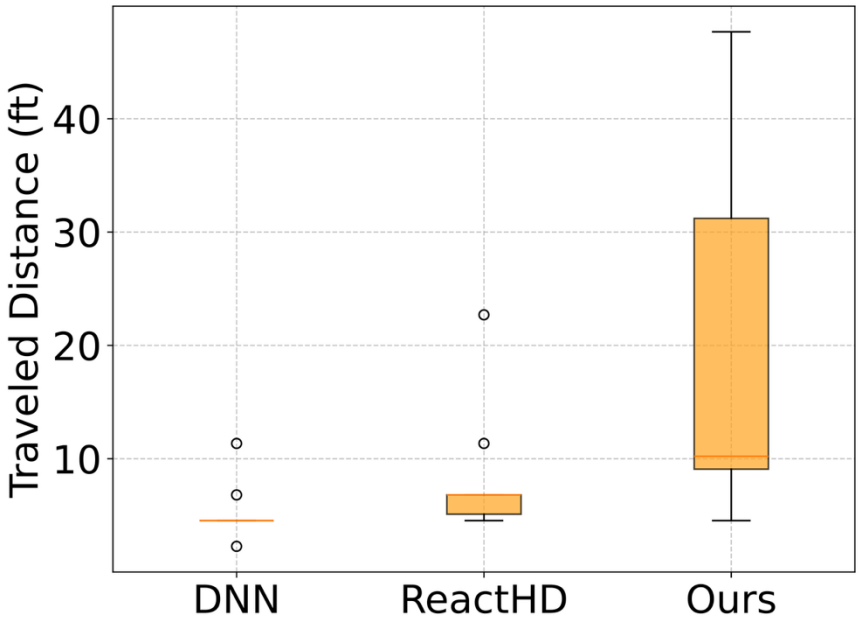


Simulation - Chase (1 object)

NavHD: off-policy RL for micro-robot navigation



Simulation results



Real world results

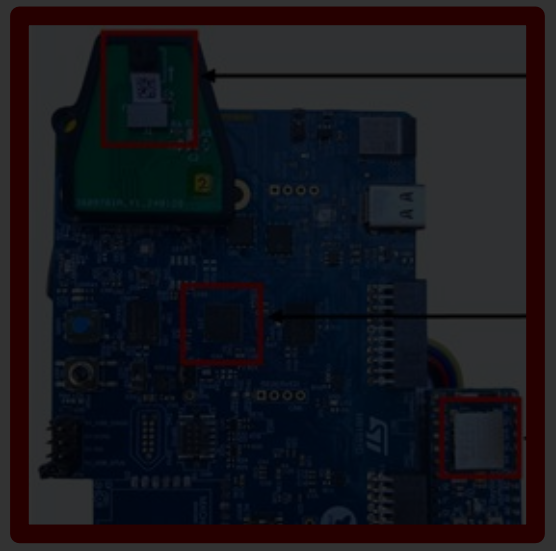
NavHD achieves 2x rewards and distance traveled.

NavHD: off-policy RL for micro-robot navigation

	Memory (kB)	Clock Cycle (k)	Energy (mJ)
DNN	263.8	2.9	3.4
VSA	19.5	8.4	9.9
NavHD (Ours)	**10.2**	**0.9**	**1.1**

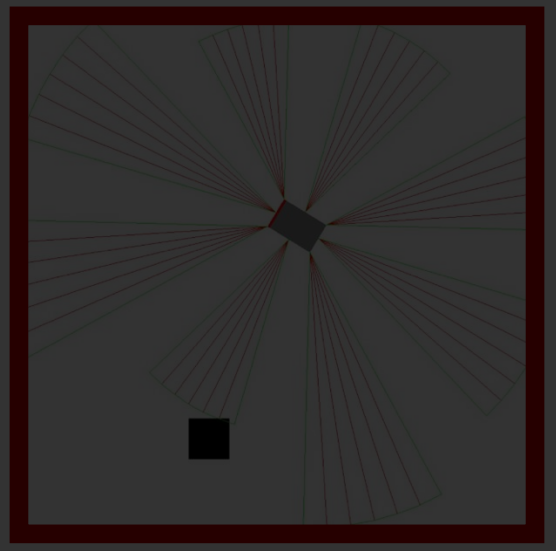
NavHD is 2-26x more resource-efficient than baseline!

My research builds ultra-lightweight VSA



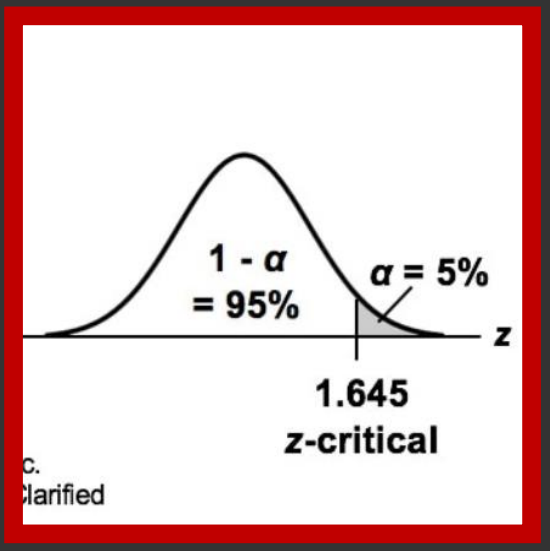
HyperCam
(MobiCom '25)

Image Classification
~43 kB of flash



NavHD
(IROS '25)

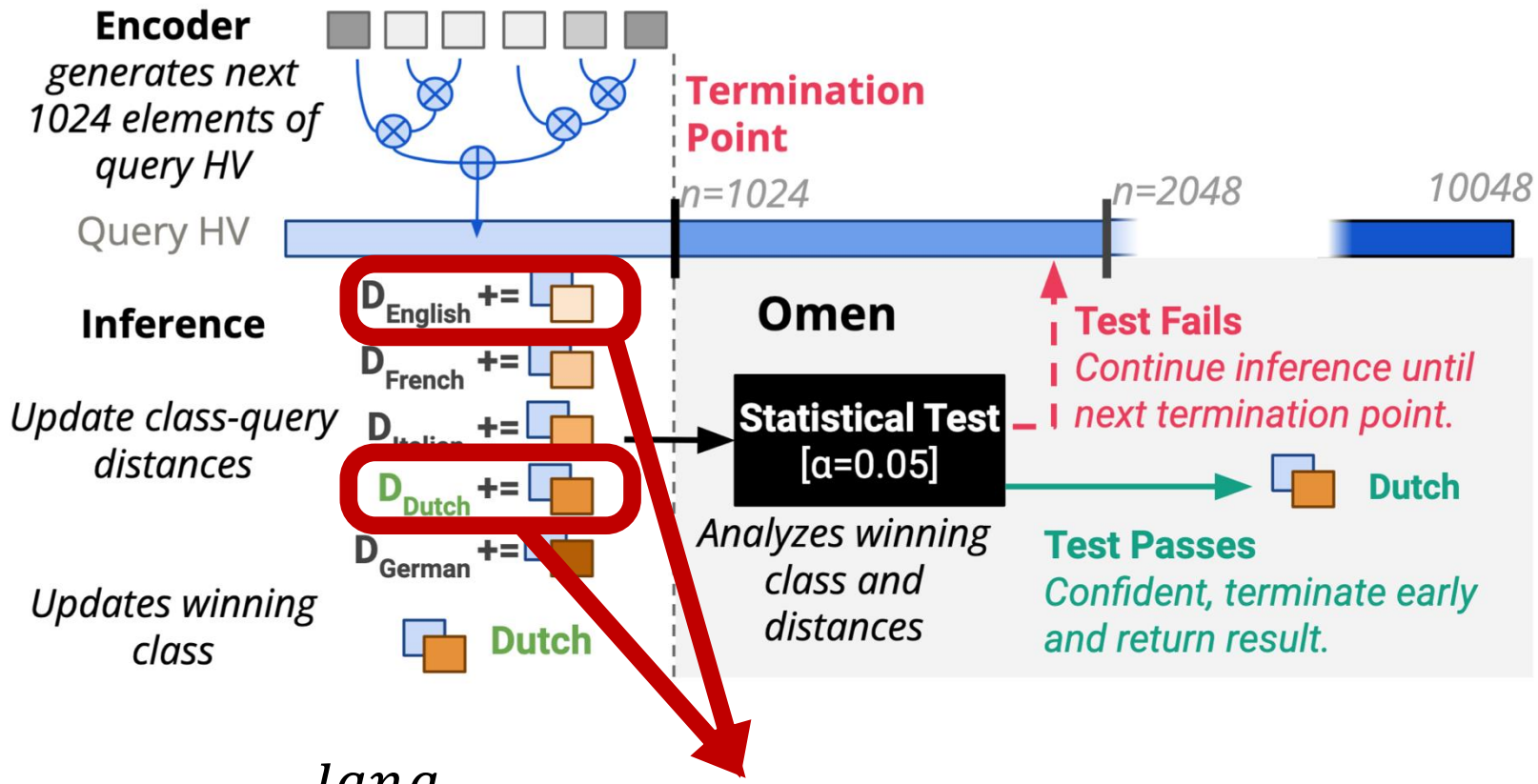
Tiny Robot Navigation
~10 kB of flash



Omen
(ASPLOS '25)

Inference Time
7-12x speedup

Omen: early termination of inference



Null hypothesis: $H_0^{lang}: E[D_{Dutch}] \geq E[D_{lang}]$

If p-value < α, reject H_0^{lang} and conclude $E[D_{Dutch}] < E[D_{lang}]$

Omen: early termination of inference

dataset	train/test	classes	description	NN acc
lang	42855/7145	5	language classification [21]	-
ucihar	7352/2947	6	activity recognition [3]	94.02 [12]
isolet	62371/560	26	voice recognition [4]	95.50 [69]
mnist	60000/10000	10	image classification [52]	95.83 [26]

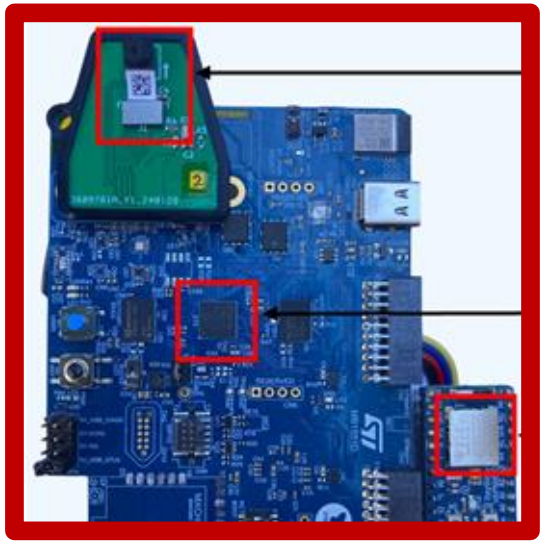
Table 3. Dataset configurations.

Name	Algorithm	Encoder	Class HV	Size	Initial/Step Terms (it/ct)
OHD	OnlineHD [26]	random	weighted $\oplus$	10048	512/64
LeH	LeHDC [14]	random	learned	5056	512/64
LDC	LDC [15]	learned	learned	256	16/4

Table 4. Hypervector (HV) Training algorithm descriptions.

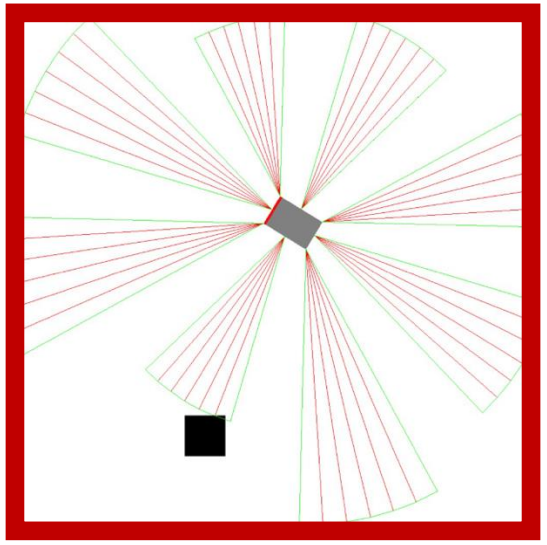
Omen achieves **7-12x speedup** while causing **0.0-0.7% accuracy drop** in 19 benchmarks, tested on 4 classification tasks.

My research builds ultra-lightweight VSA



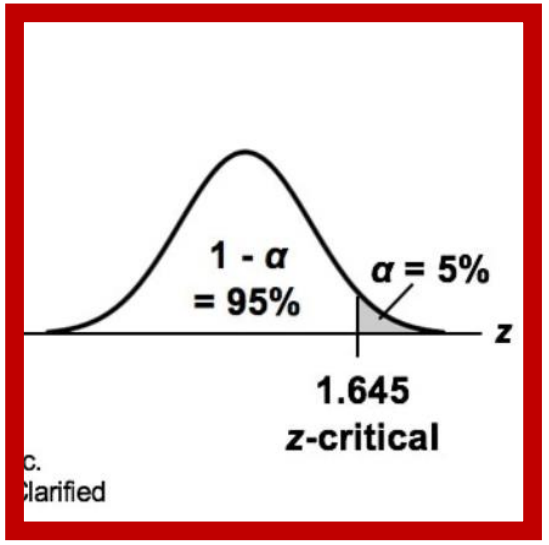
HyperCam
(MobiCom '25)

Image Classification
~43 kB of flash



NavHD
(IROS '25)

Tiny Robot Navigation
~10 kB of flash



Omen
(ASPLOS '25)

Inference Time
7-12x speedup

Read my work her

Thank you for listening!

Stanford
SMART SENSING
SYSTEMS LAB

Chae Young Lee
Stanford University
chae@stanford.edu