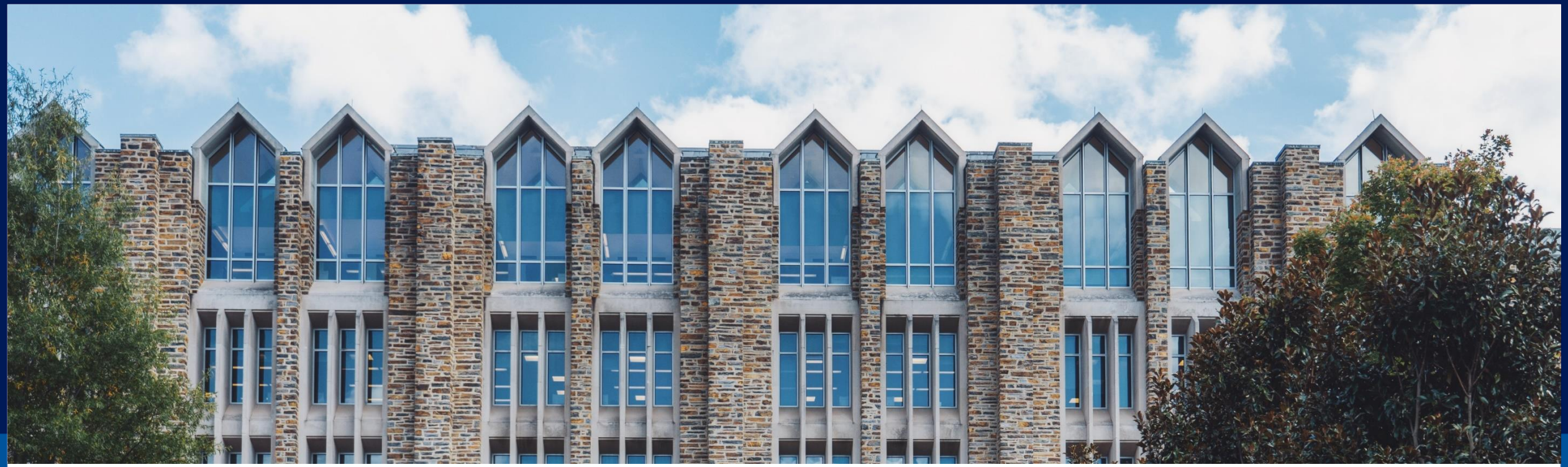


Focus : A Streaming Concentration Architecture for Efficient Vision-Language Models

Presenter: **Bowen Duan**, Duke University

Chiyue Wei*, Cong Guo*, Junyao Zhang, Haoxuan Shan, Yifan Xu, Ziyue Zhang, Yudong Liu,
Qinsi Wang, Changchun Zhou, Hai “Helen” Li, Yiran Chen Duke University

Feb. 2rd, 2026

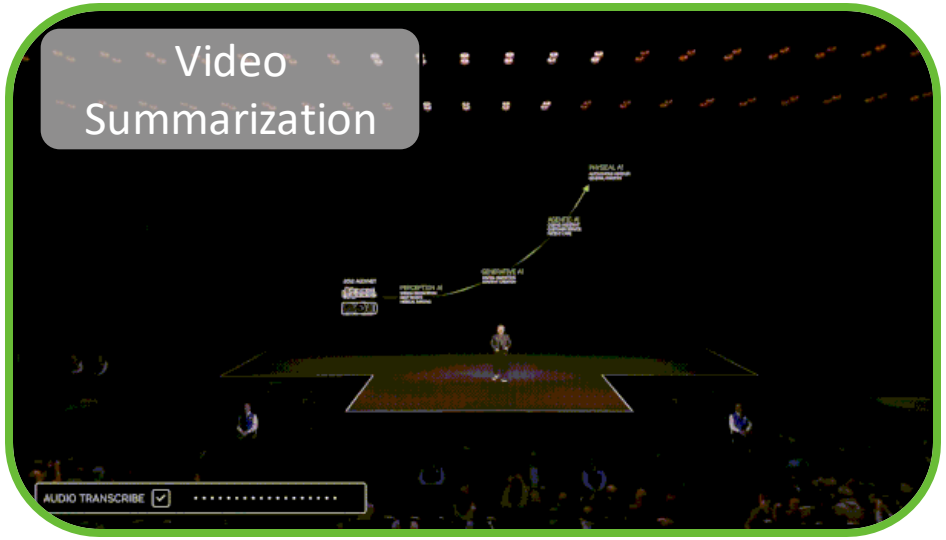
Duke

Contents

Duke

Vision–Language Models: Foundation of Multimodal AI

Video Summarization

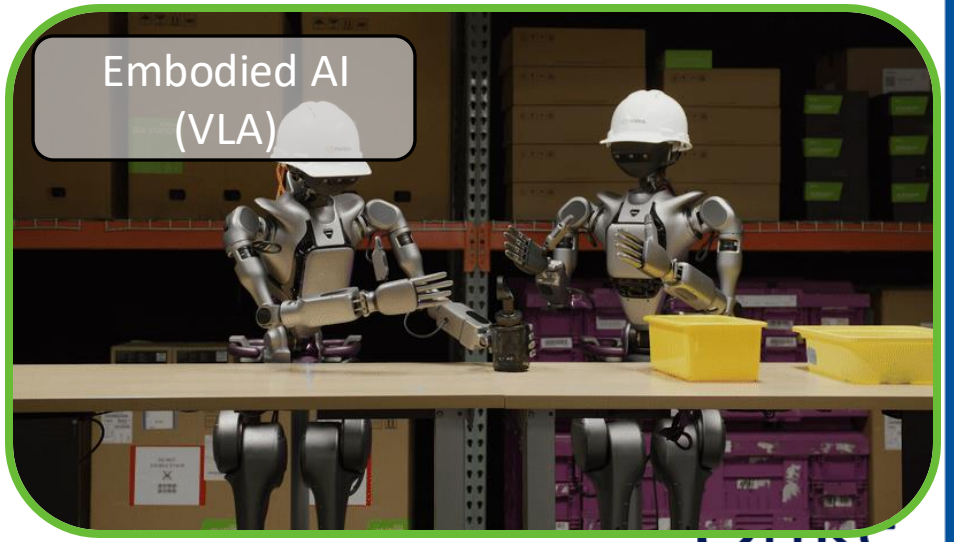


Multimodal Chatbot

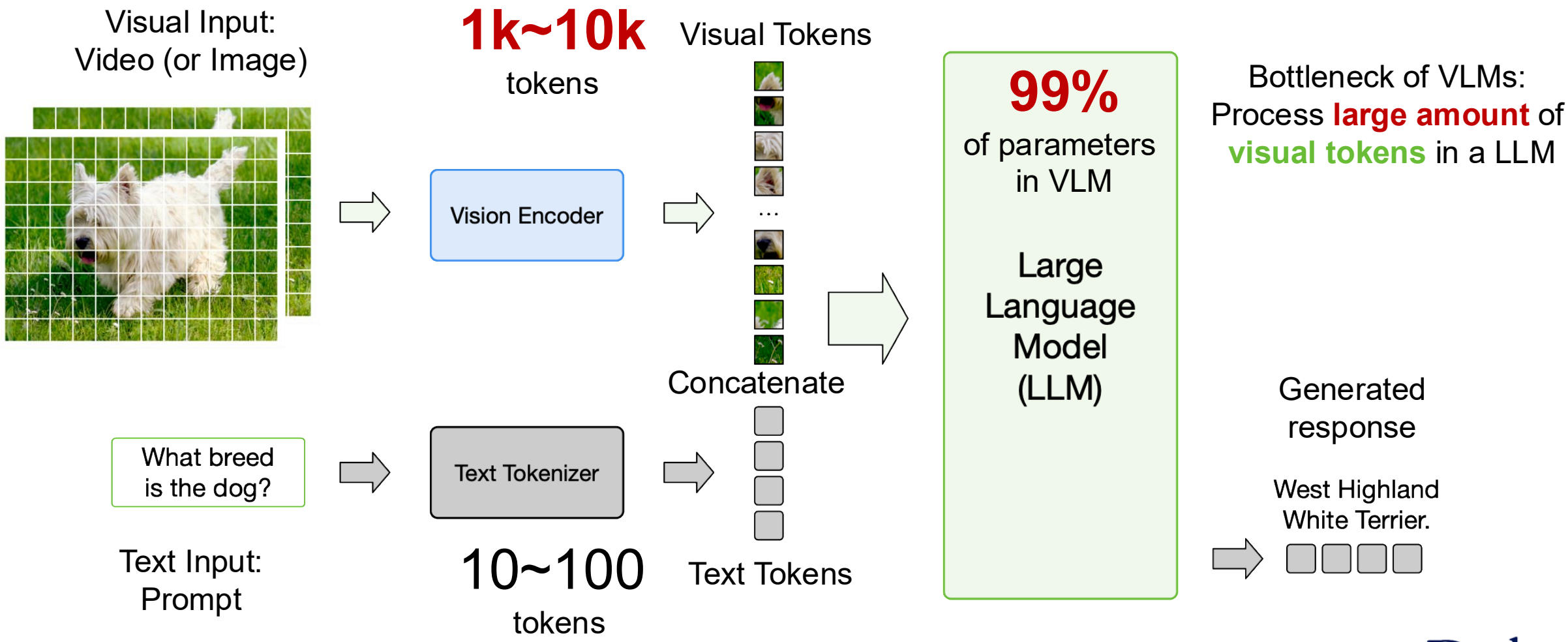
Augmented Reality Glasses

Autonomous Driving

Embodied AI (VLA)



Core Inference Bottleneck in Vision–Language Models



Contents

- Background
- **Motivation**
  - Semantic Redundancy
  - Visual Redundancy
  - Hardware-friendly Design
- Focus Architecture
- Evaluation

Duke

Algorithm Level: Semantic Redundancy in VLMs

- Semantic redundancy
- Attention to visual tokens shift according to prompt

Q: What breed is the dog?
A: West highland white terrier.

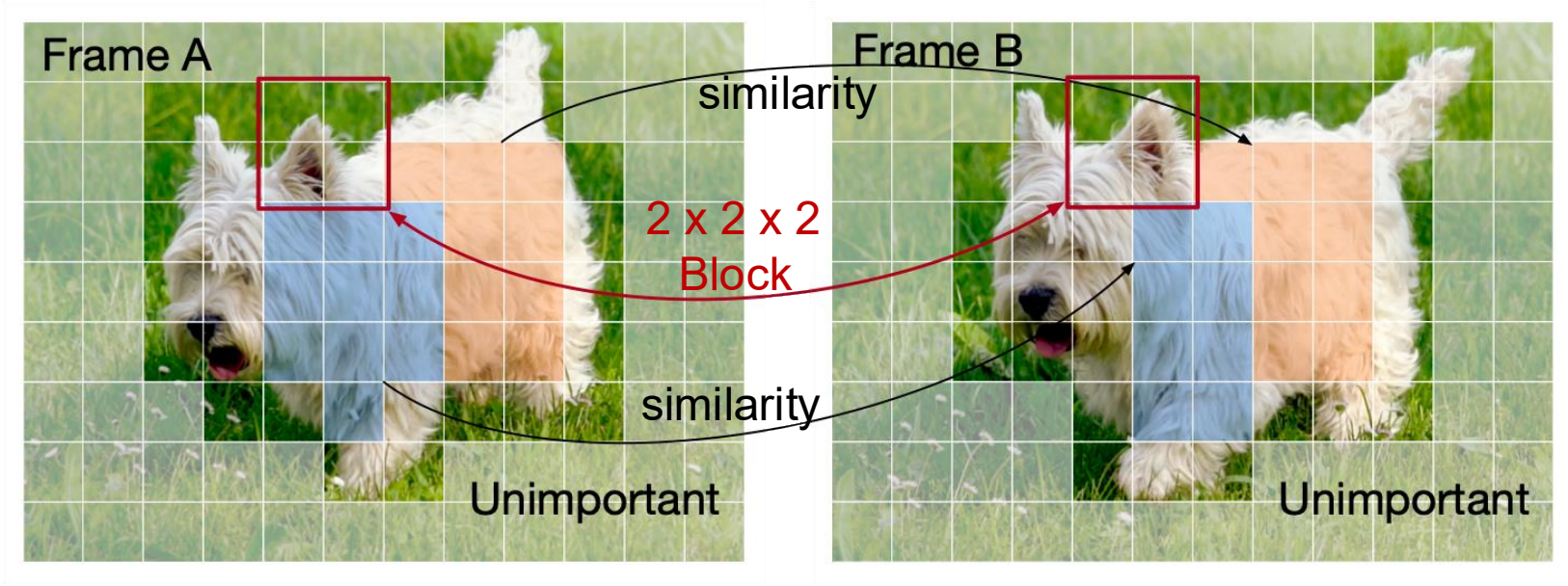
Q: What is the color of the flower?
A: White.

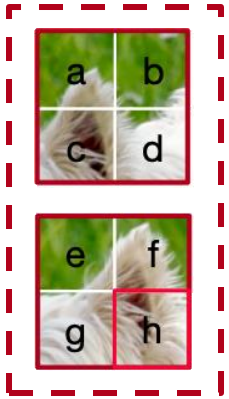
Cross-modal attention dynamically focuses on query-relevant regions.

Duke

Algorithm Level: Visual Redundancy in VLMs

- Visual redundancy: Similarity
- Similarity among adjacent visual tokens: temporal and spatial





2 x 2 x 2
Block

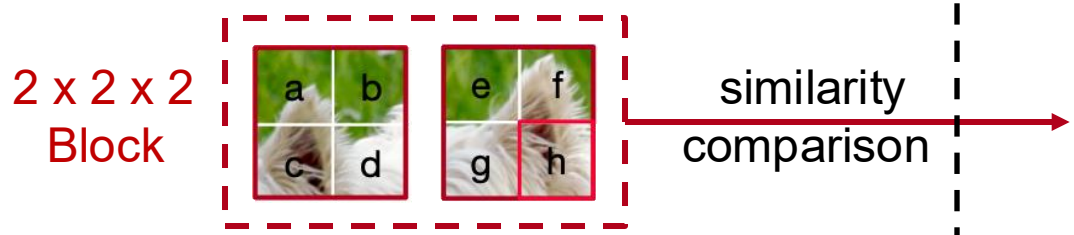
Local similarity comparison
within a block

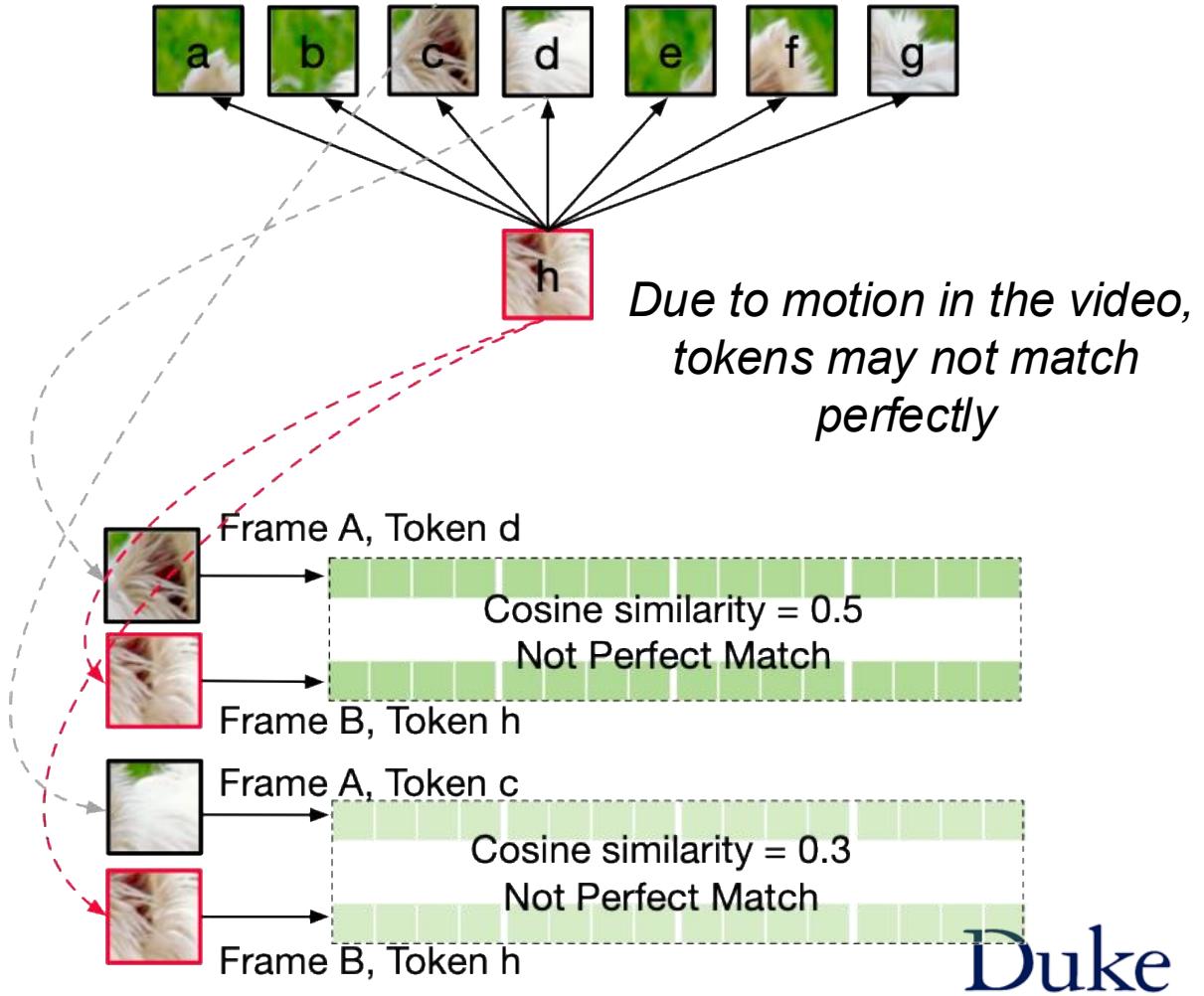
Visual redundancy arises from strong spatial and temporal similarity among adjacent visual tokens.

Duke

Algorithm Level: Visual Redundancy in VLMs

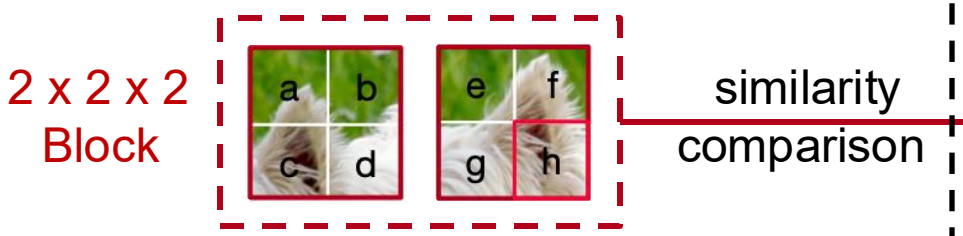
- Token-level visual similarity comparison works but is **coarse**

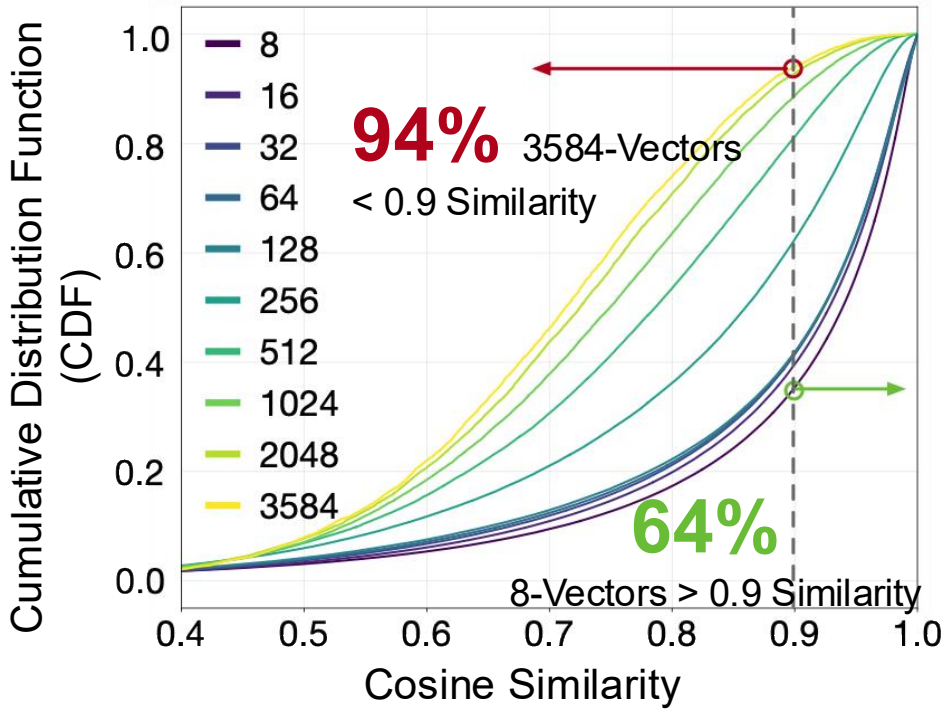


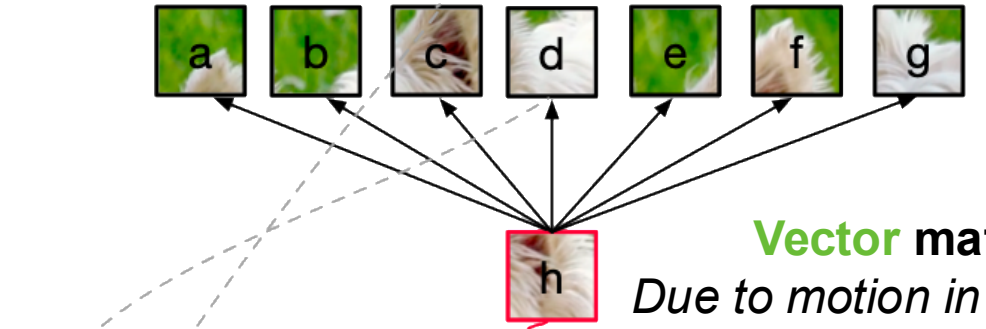


Algorithm Level: Visual Redundancy in VLMs

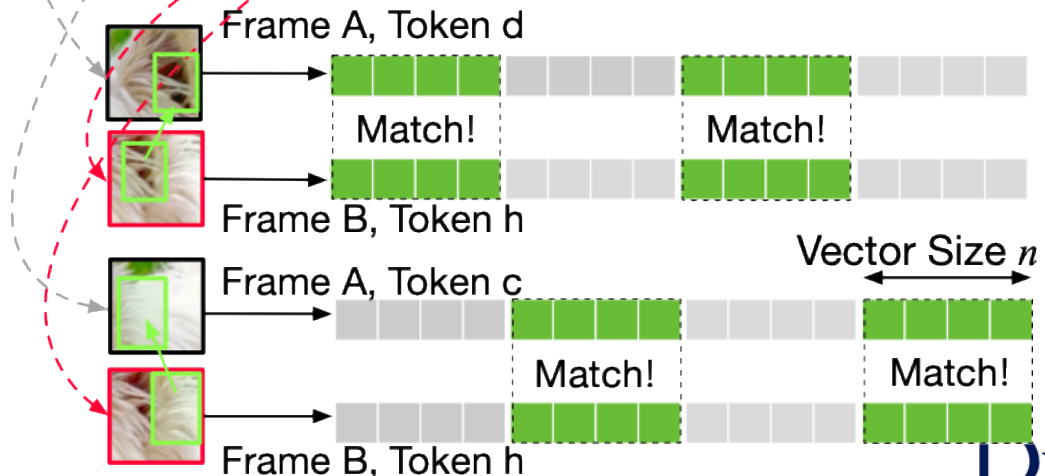
- Visual similarity comparison should be **fine-grained**





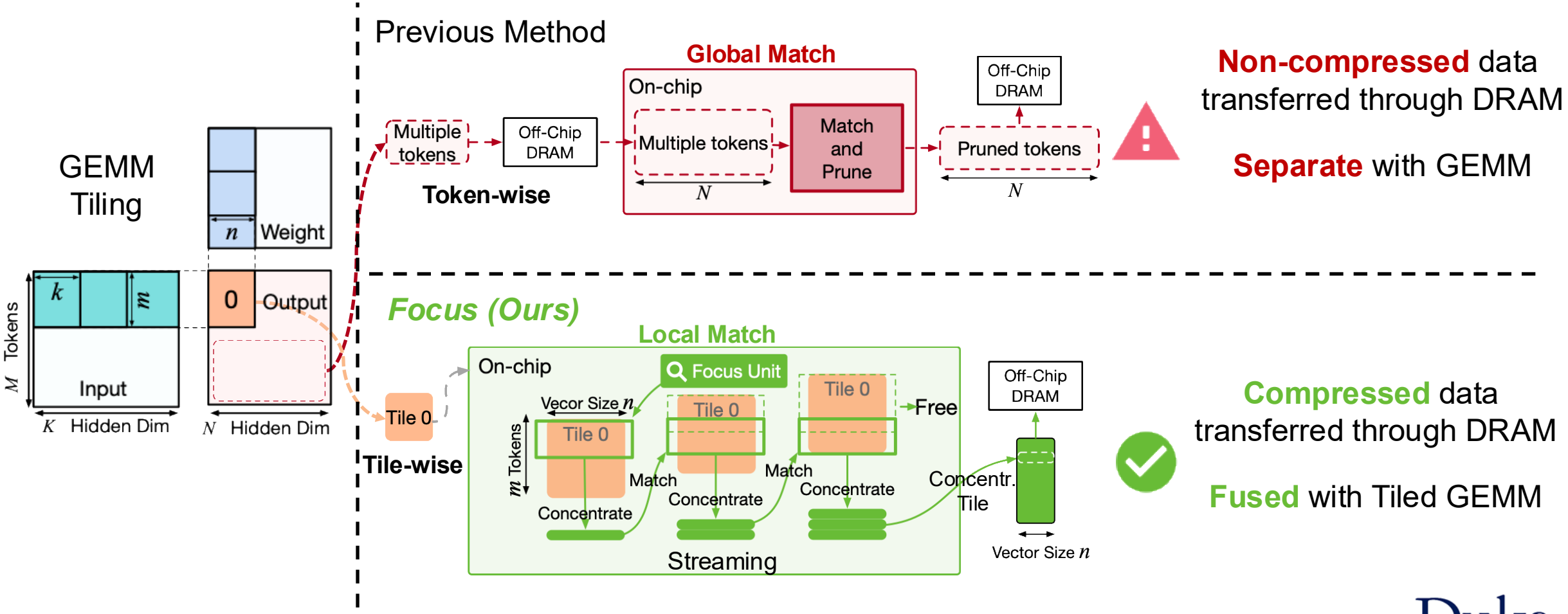


Vector match:
Due to motion in the video, different parts of a token may match with different tokens



Architecture Level: Hardware-friendly Similarity Removal

- Efficient GEMM considering fine-grained similarity is possible



Duke

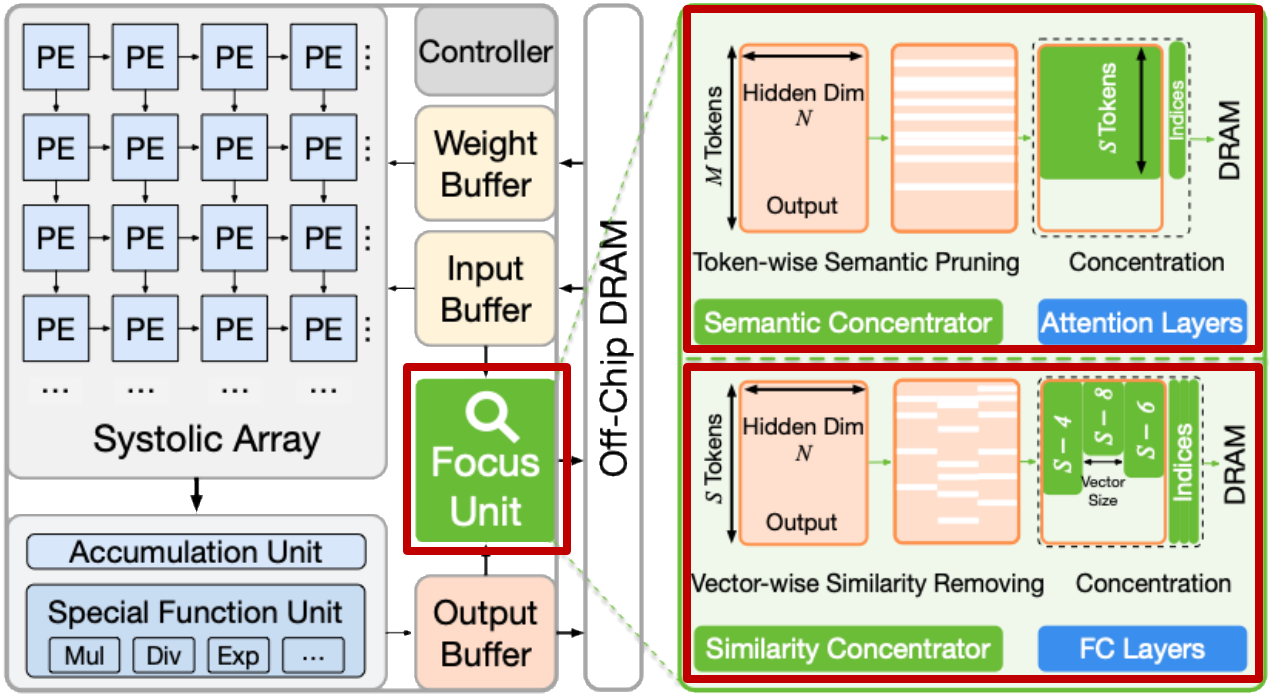
Contents

- Background
- Motivation
- **Focus Architecture**
  - Semantic Concentrator
  - Similarity Concentrator
- Evaluation

Duke

A Streaming Concentration Architecture - *Focus*

- A **modular** unit in systolic-array-based accelerator
- Intercept and process GEMM input & output in **streaming** manner
- Two types of **concentration** on visual tokens



2.7% additional area.

0.9% additional power.

Duke

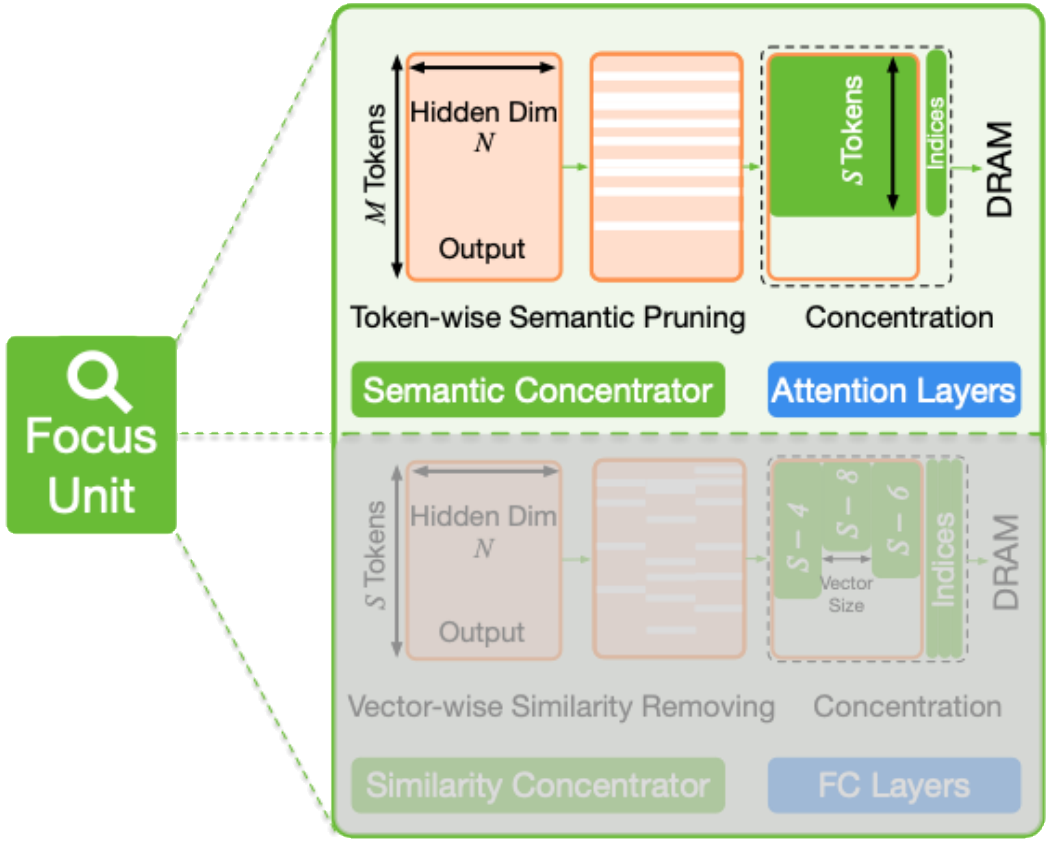
Contents

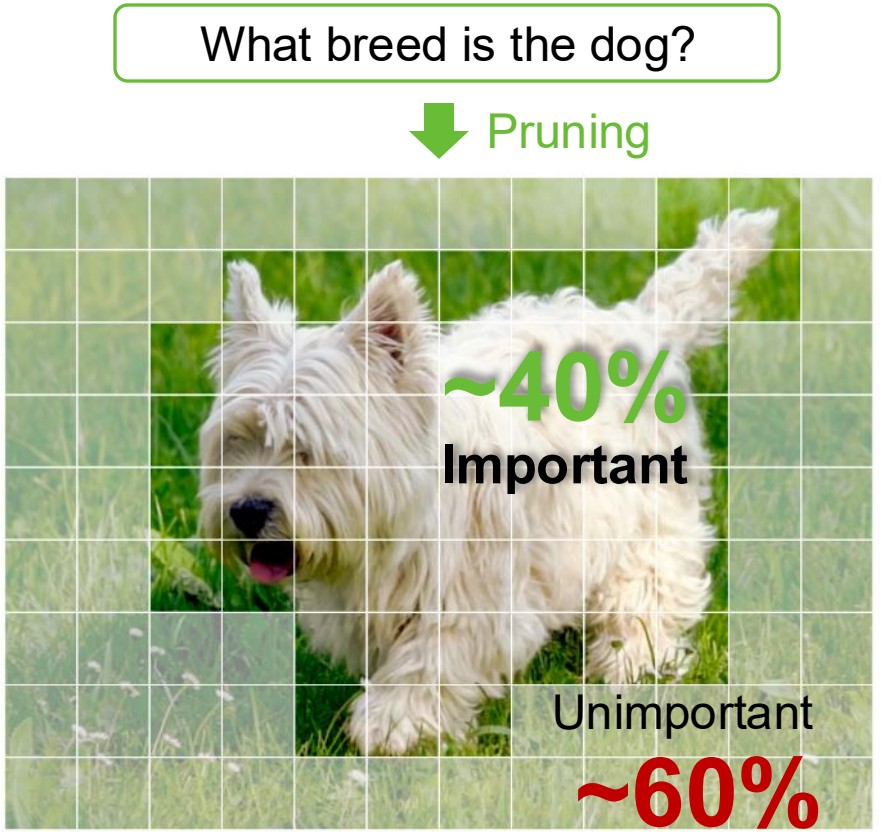
- Background
- Motivation
  - Semantic Redundancy
  - Visual Redundancy
  - Hardware-friendly Design
- **Focus Architecture**
  - Semantic Concentrator
  - Similarity Concentrator
- Evaluation

Duke

Outline of Semantic Concentrator

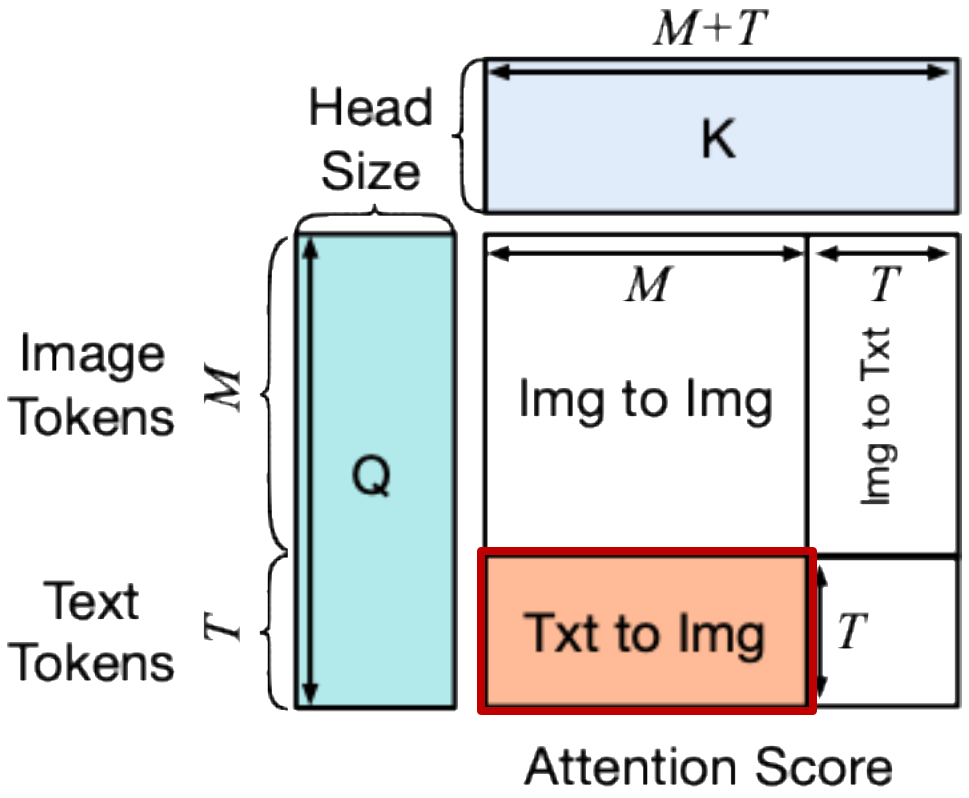
- **Token-wise** pruning according to **semantic**
- Pruning happen in **Attention** layers

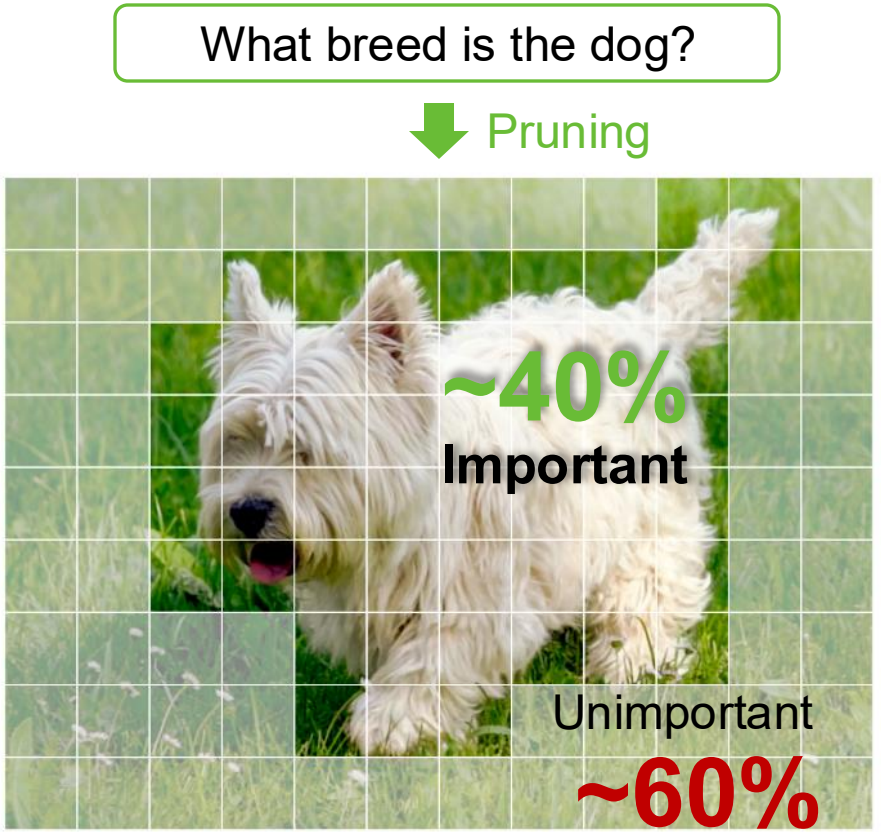




Outline of Semantic Concentrator

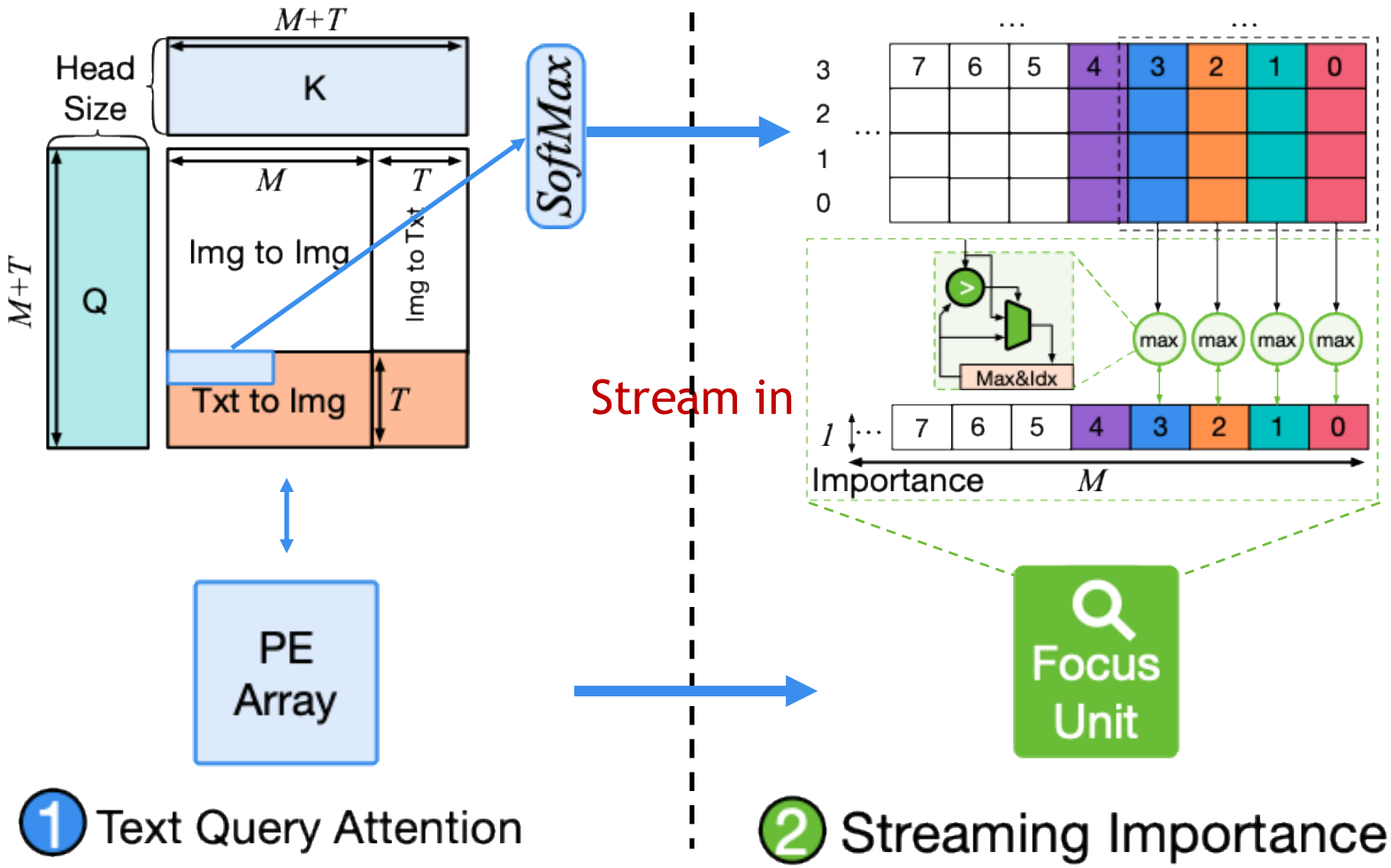
- **Attention score** reflect importance of tokens
- Use **Text to Image** attention to find visual tokens related to text





Workflow of Semantic Concentrator - 1

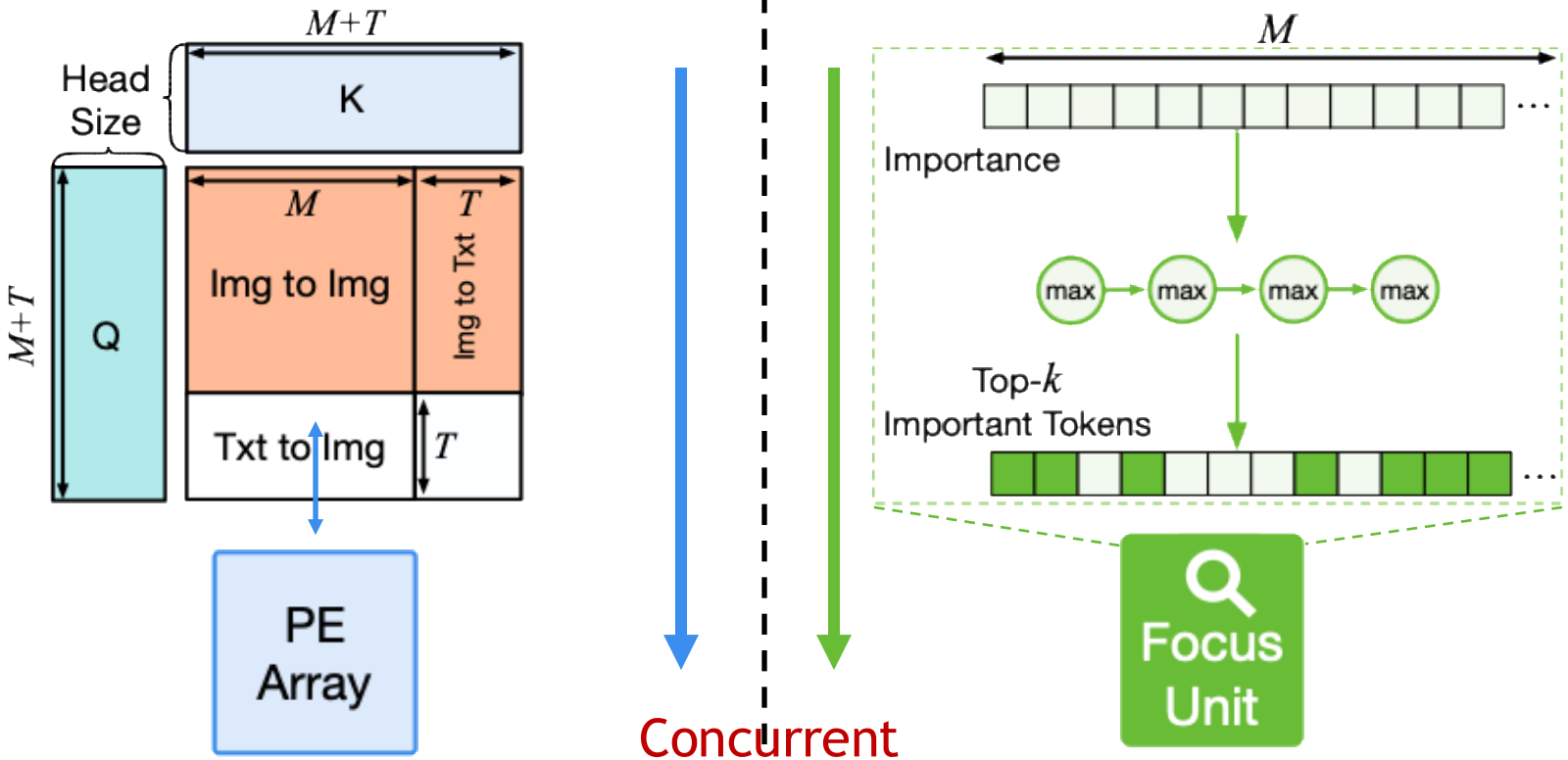
- Get visual **token importance** using text to image attention score



Within-token aggregation:
*Aggregate attention scores across **heads** and **text tokens** to obtain the maximum value*

Workflow of Semantic Concentrator - 2

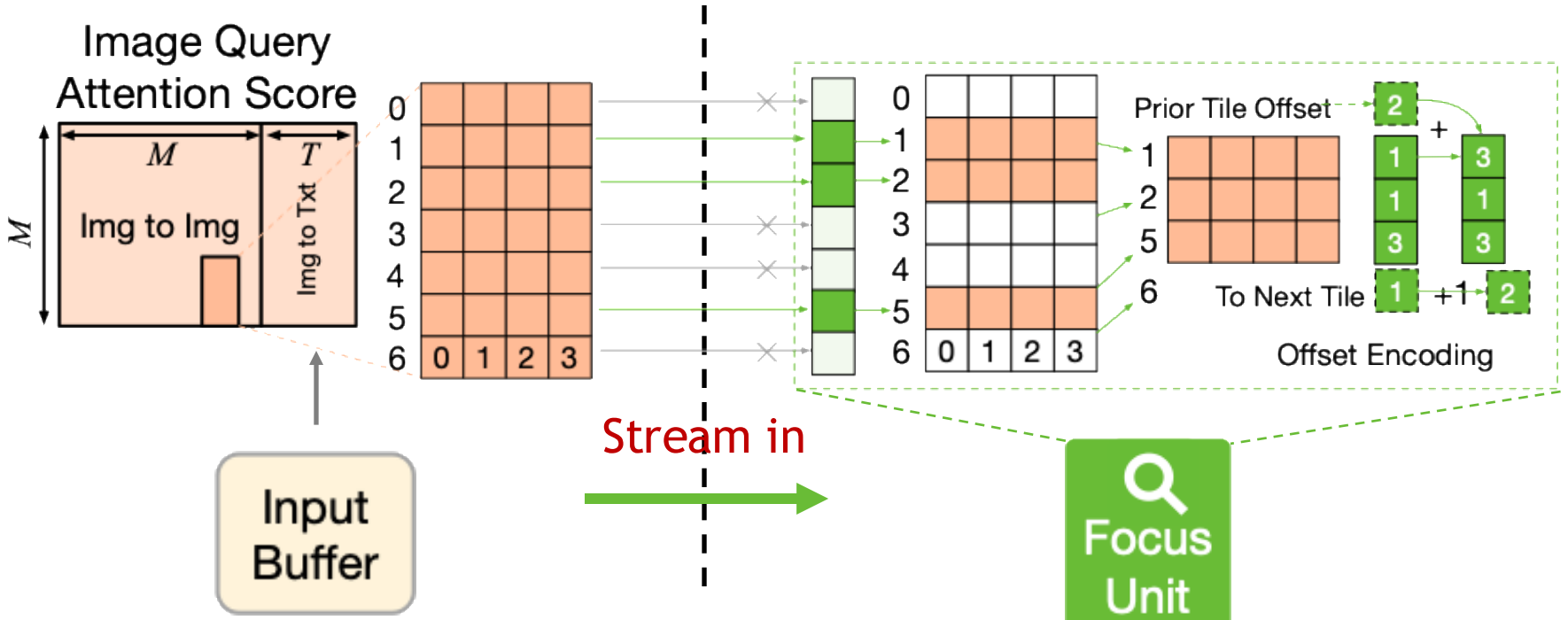
- Get visual token importance using text to image attention score
- Get **top-k** important tokens during attention computation



Across-token selection:
Select top-k tokens based on token importance.

Workflow of Semantic Concentrator - 3

- Get visual token importance using text to image attention score
- Get top-k important tokens during attention computation
- **Pruning** during data loading right before $AttnScore \times V$



Pruning:
Only the top-k tokens are kept during data loading.

5 Semantic Pruning and Localized Offset Encoding

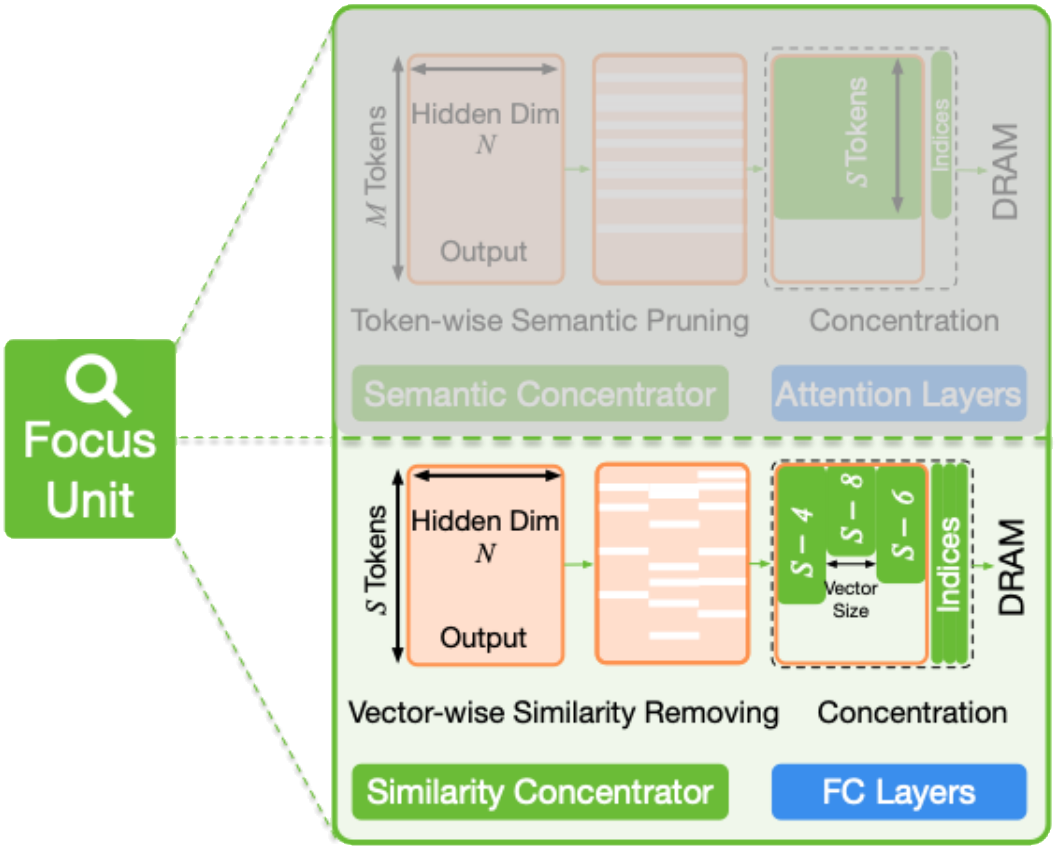
Contents

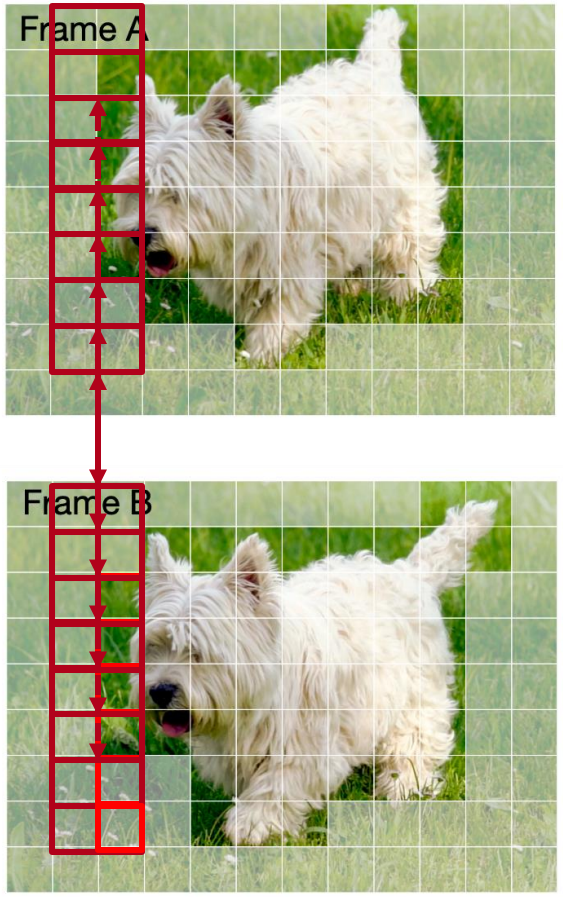
- Background
- Motivation
  - Semantic Redundancy
  - Visual Redundancy
  - Hardware-friendly Design
- **Focus Architecture**
  - Semantic Concentrator
  - Similarity Concentrator
- Evaluation

Duke

Outline of Similarity Concentrator

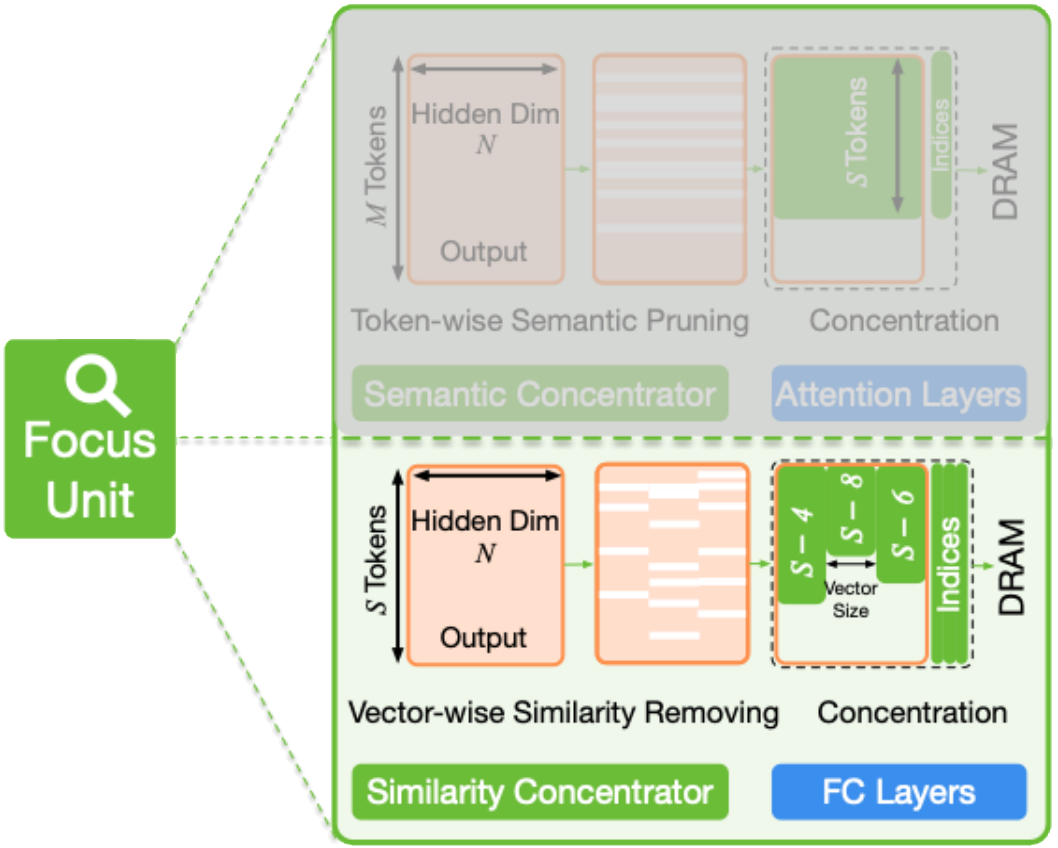
- **Vector-wise** concentration using **spatial and temporal similarity**
- Concentration for **GEMM** operations in FC layers

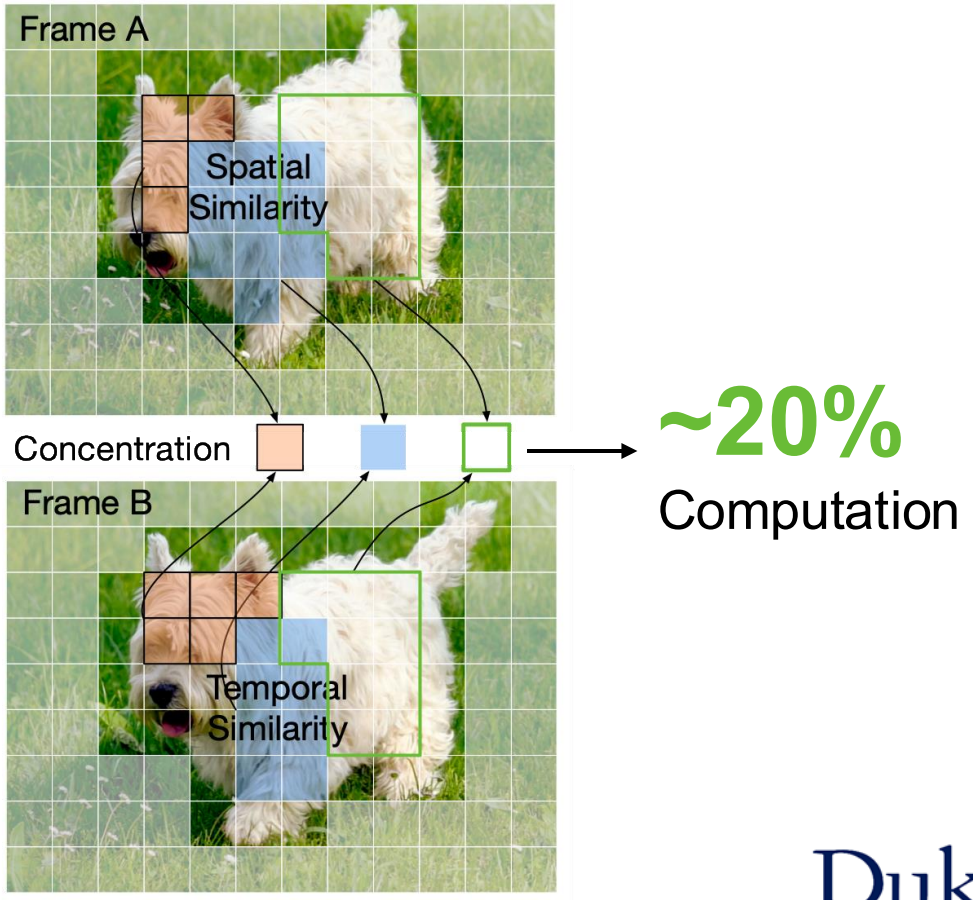




Outline of Similarity Concentrator

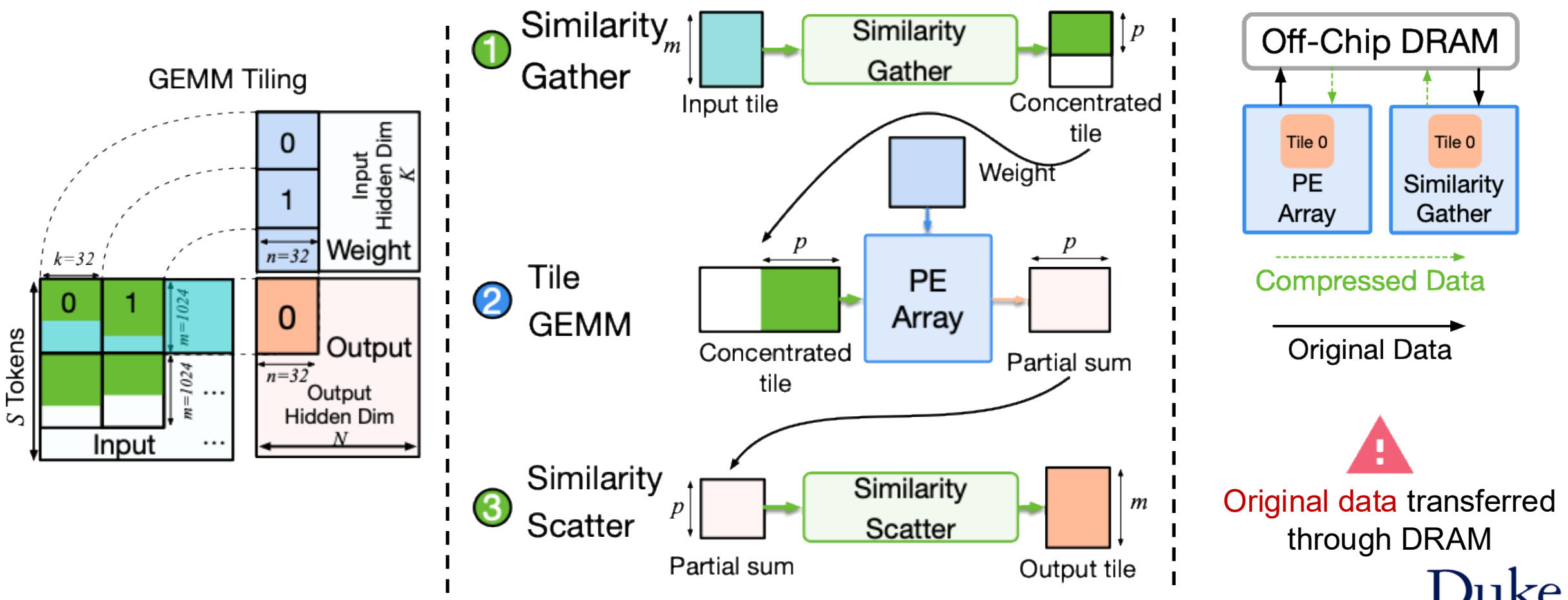
- **Vector-wise** concentration using **spatial and temporal similarity**
- Concentration for **GEMM** operations in FC layers





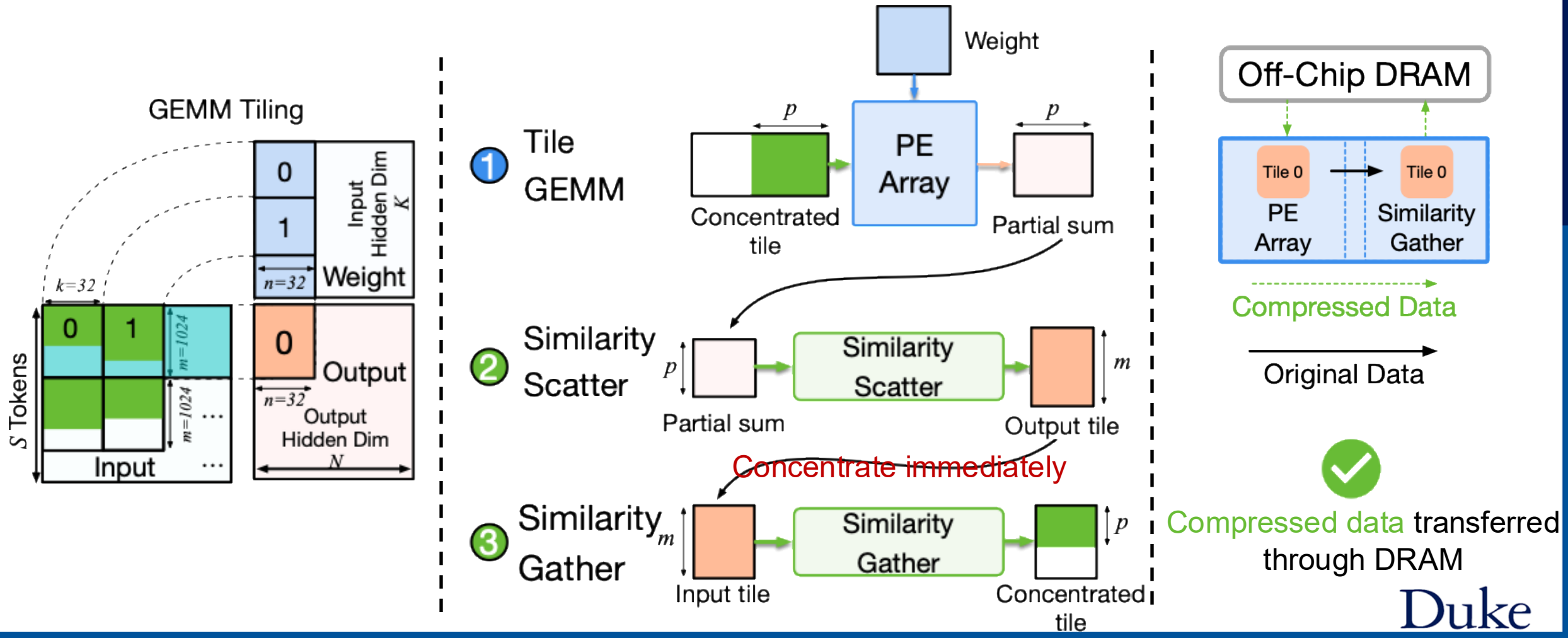
Workflow of Similarity Concentrator

- Concentration seamlessly **integrated with GEMM Tiling**



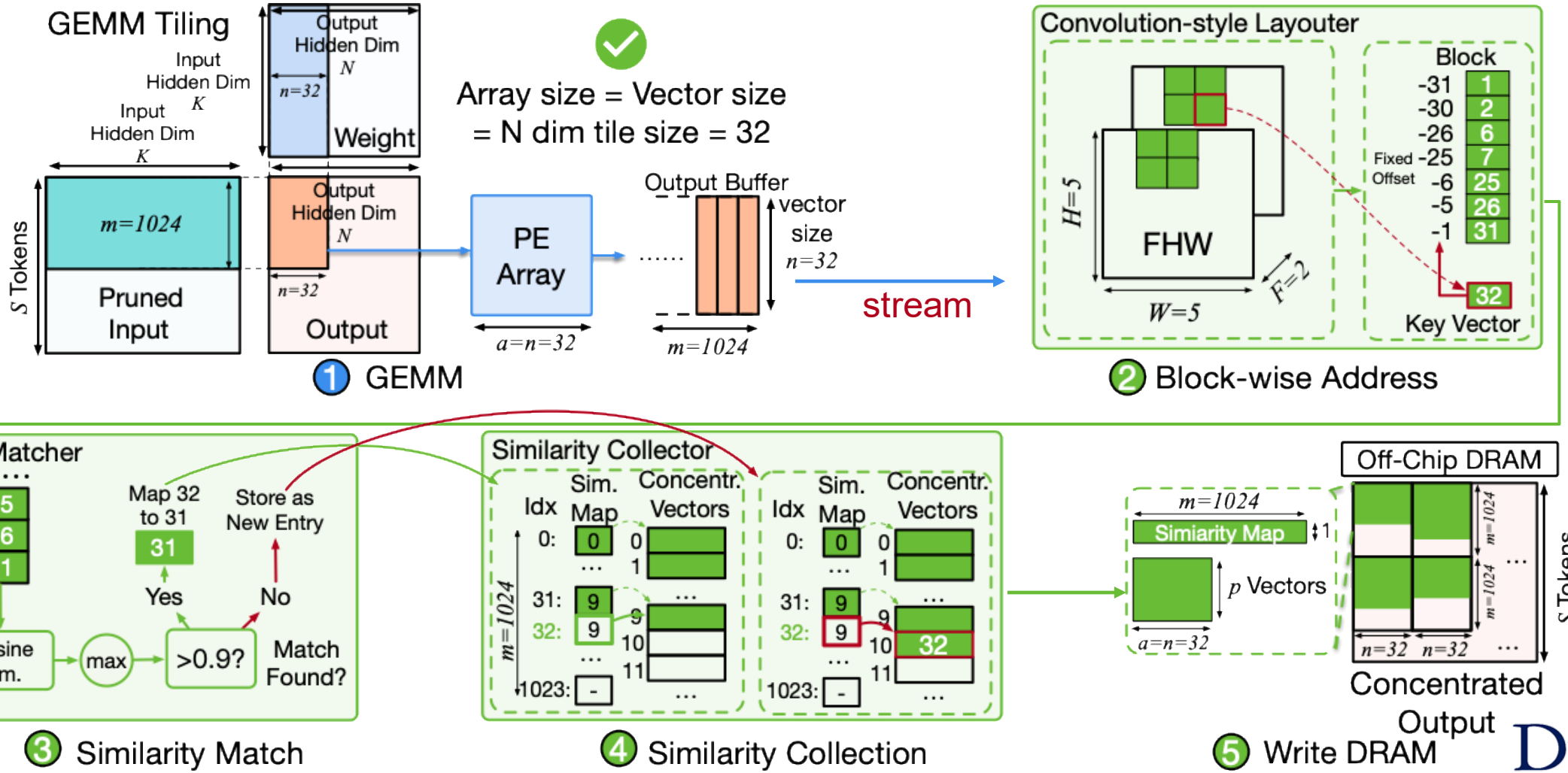
Workflow of Similarity Concentrator

- Concentration seamlessly **integrated with GEMM Tiling**



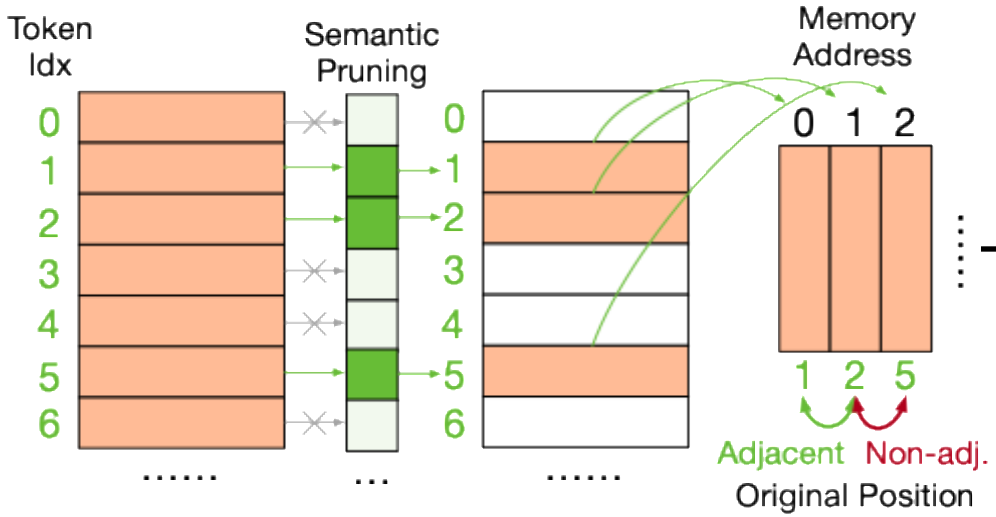
Workflow of Similarity Gather

- Detect similarity and concentrate output tiles



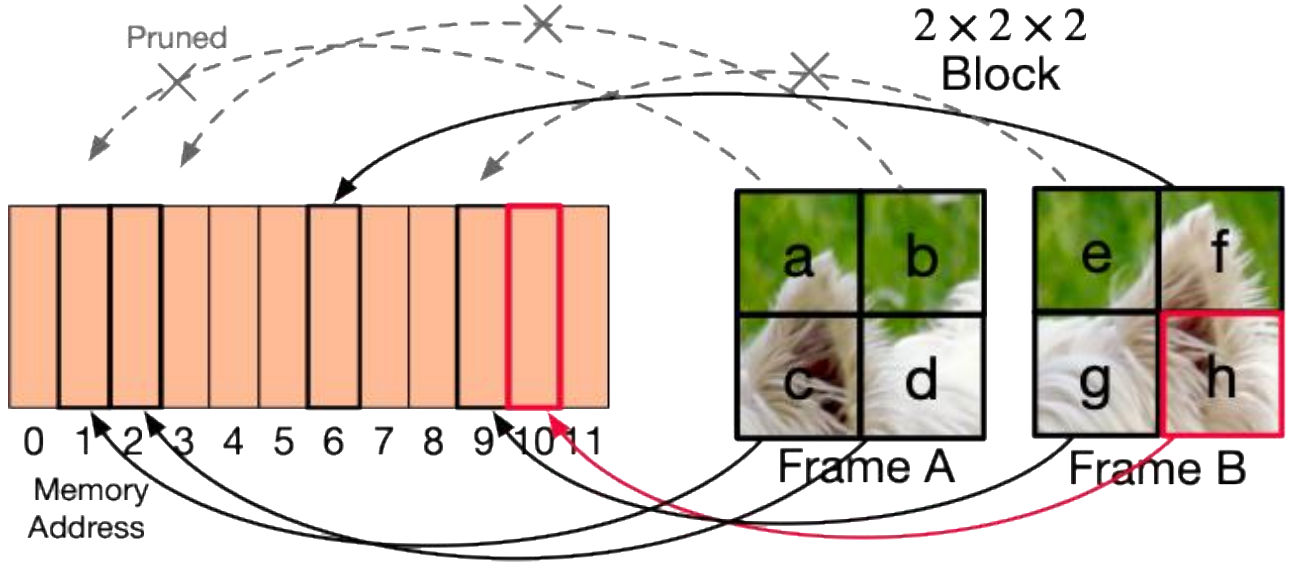
Challenges in Similarity Gather

- Require **original position** to match token in a block
  - **Pruning** breaks relation of position and memory address



Misalignment of address and position

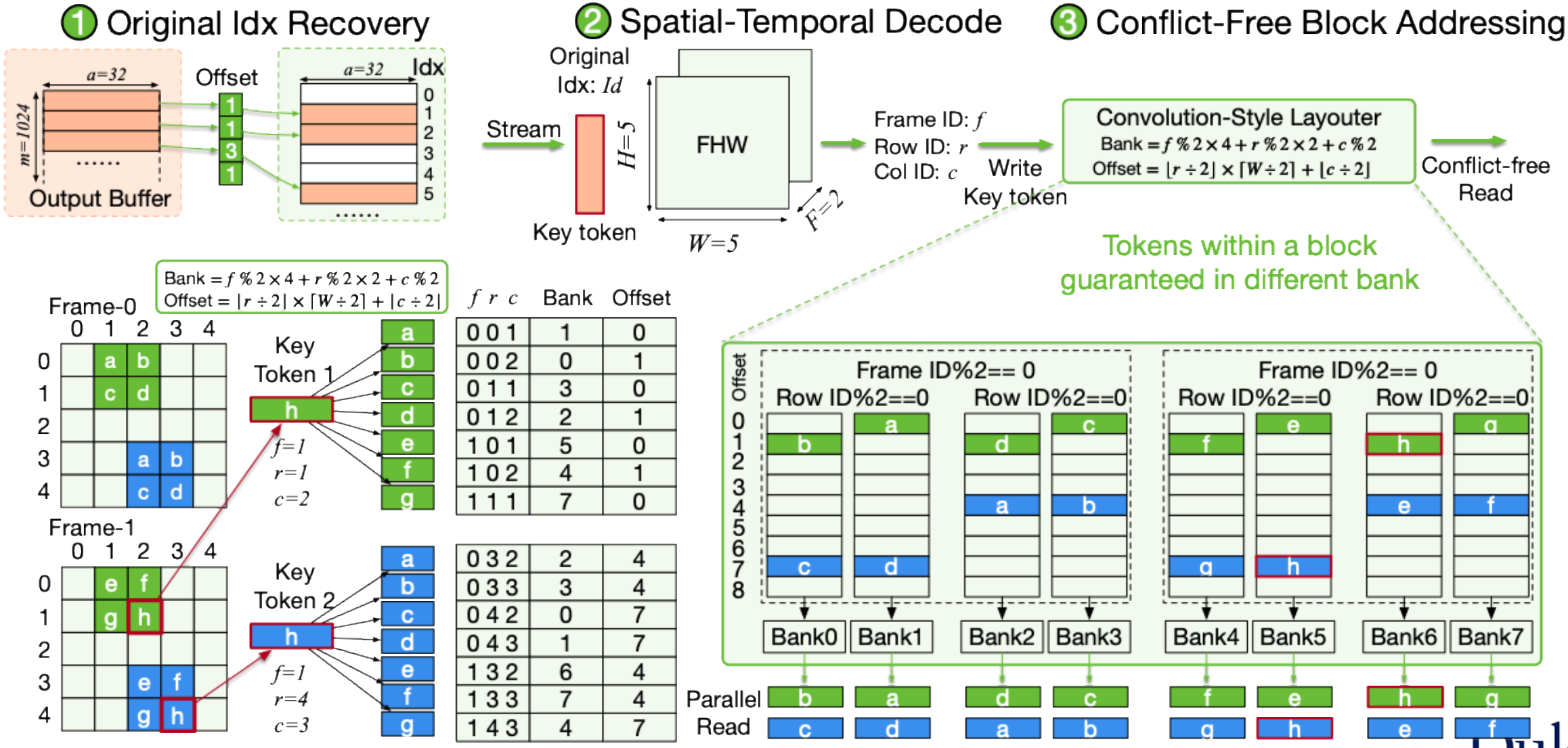
- Tokens in a block **not contiguous** in memory
  - **Memory conflict** in parallel matching



Irregular memory access within a block

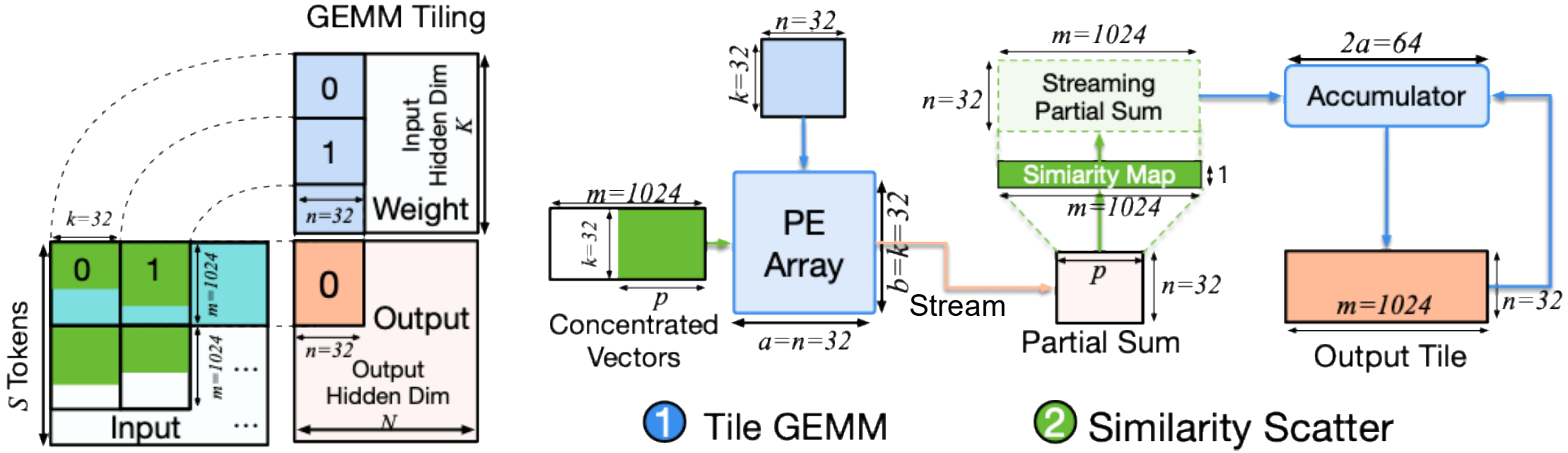
Duke

Convolution-style Layouter for Similarity Gather



Similarity Scatter

- **Restore** concentrated tile to original size
- **Accumulate** to get output tile



Contents

- Background
- Motivation
- Focus Architecture
- **Evaluation**

Duke

Evaluation Baselines and Tools

Methods	Type	Sparsity-level
Systolic Array	DNN accelerator	Dense
AdapTiV (MICRO'24)	ViT accelerator	Token-level
CMC (ASPLOS'24)	ViT accelerator	Token-level
NVIDIA Jetson Orin Nano	Edge GPU	Dense
FrameFusion (ICCV'25)	VLM pruning algorithm	Token-level
***Focus* (Ours)**	***VLM accelerator***	**Token-level & Vector-level**

Tools	Usage
Focus simulator	Overall simulation
ScaleSim-V2	Systolic array simulation
TSMC N28HPC+ Memory Compiler	Buffer evaluation
DRAMsim3	DRAM simulation
Synopsys Design Compiler	Logic synthesis

***Focus*: The First VLM architecture**

Full-stack open-source evaluation

Duke

Accuracy and Sparsity

Sparsity (%)

Model	Tasks	Original	Frame Fusion	Adaptiv	CMC	*Focus*
LLaVA-Video-7B	VMME	0.00	70.00	52.15	58.62	**82.82**
	MLVU	0.00	70.00	32.52	42.46	**78.26**
	MVB	0.00	70.00	41.07	53	**78.44**
LLaVa-OneVision-7B	VMME	0.00	70.00	36.8	47.95	**81.49**
	MLVU	0.00	70.00	39.55	35.48	**78.34**
	MVB	0.00	70.00	42.03	63.69	**85.49**
MiniCPM-V	VMME	0.00	70.00	49.27	57.2	**82.87**
	MLVU	0.00	70.00	41.88	35.23	**78.01**
	MVB	0.00	70.00	50.09	40.27	**75.99**

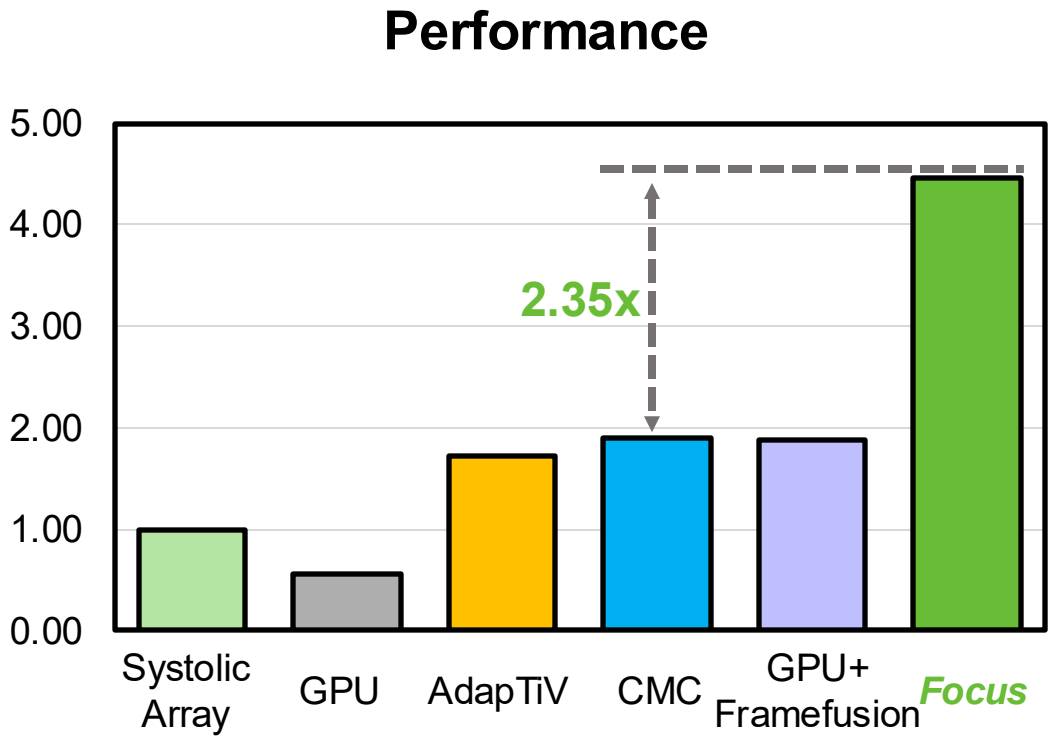
Accuracy (%)

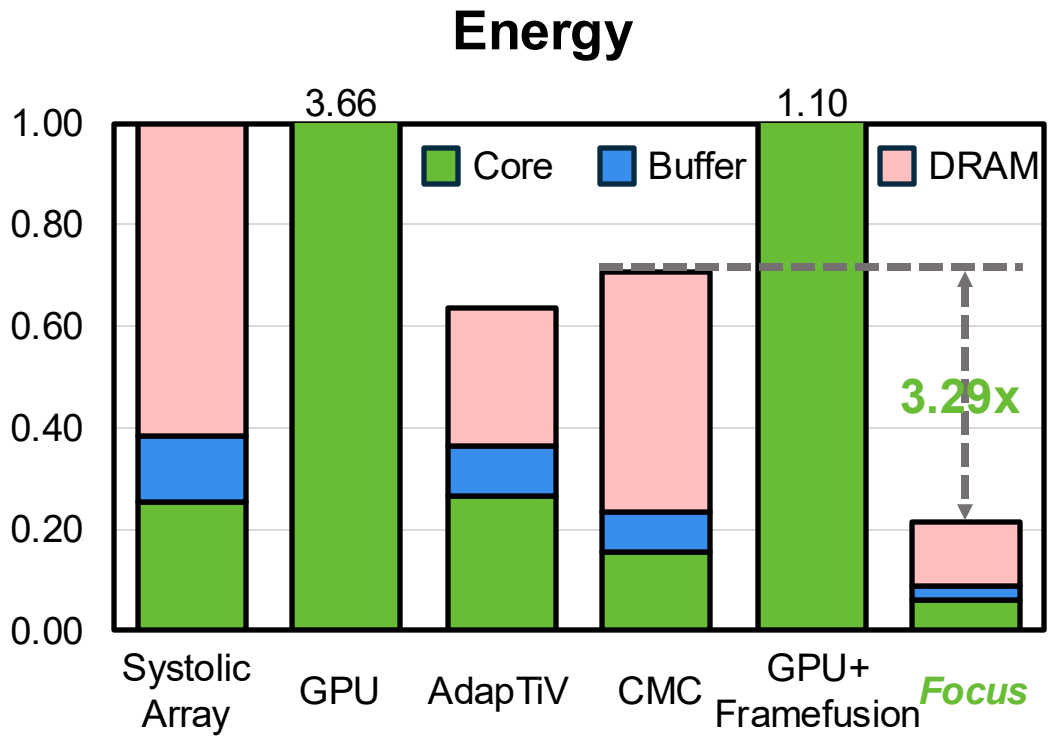
Original	Frame Fusion	Adaptiv	CMC	*Focus*
64.15	62.00	62.44	62.52	**62.74**
67.74	65.38	65.94	65.17	**65.99**
60.33	57.20	57.73	58.18	**58.20**
58.41	57.70	58.33	58.11	**58.70**
63.32	62.54	62.22	62.50	**62.52**
58.38	56.93	56.83	56.75	**56.78**
58.81	58.81	58.07	55.89	**58.30**
55.89	54.80	54.84	43.80	**53.59**
55.63	52.43	53.70	48.78	**54.30**

Focus Average sparsity: **80.19%** Average accuracy drop: **1.2%**

Duke

Performance and Energy





2.35x speedup over state-of-the-art

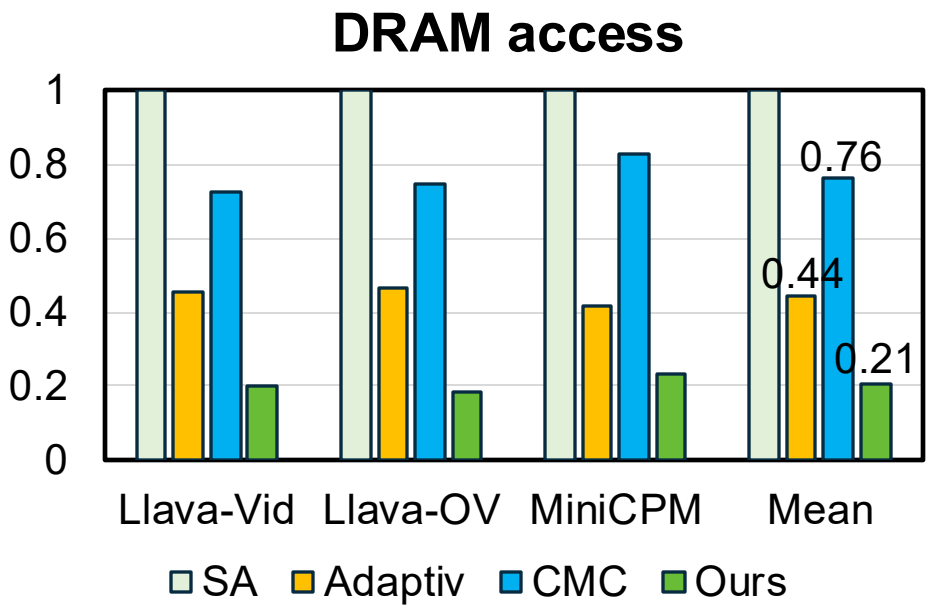
4.47x speedup over systolic array

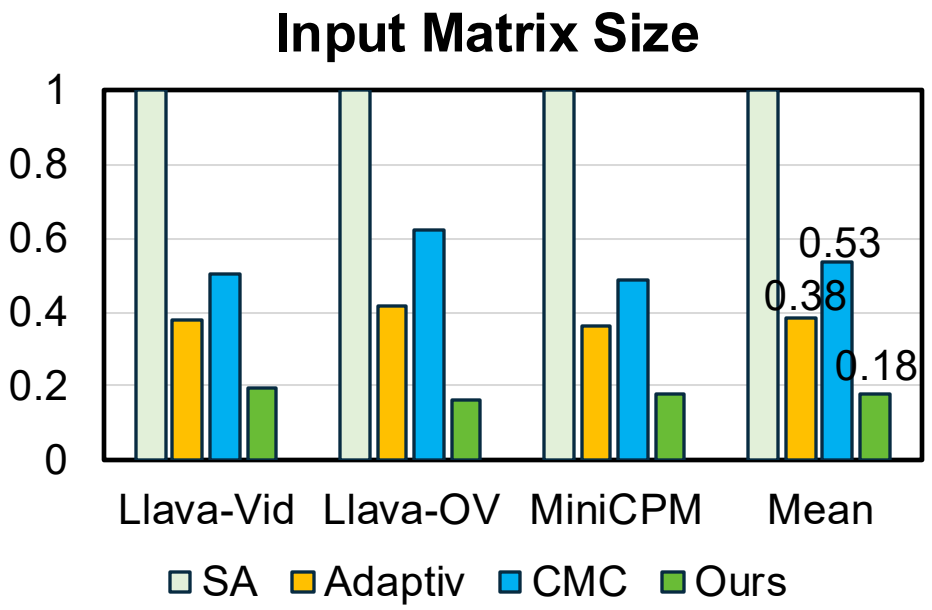
3.29x energy efficiency over state-of-the-art

4.67x energy efficiency over systolic array

Duke

Memory Access





3.7x DRAM reduction over state-of-the-art

4.9x DRAM reduction over systolic array

3.0x compression over state-of-the-art

5.6x compression over systolic array

Duke

Takeaways

- VLMs are the **foundation of next-generation AI**, enabling machines to understand the world through vision and language.
- But VLMs suffer from **massive fine-grained visual redundancy** across tokens, space, and time.
- ***Focus*** **co-designs algorithms and hardware** to remove this redundancy *in a streaming, **GEMM-friendly** way*.
- With **tiny hardware overhead**, ***Focus*** converts redundancy into major speed, energy, and memory efficiency gains.

Duke

Acknowledgements

Focus
HPCA 2026

Presenter: Bowen Duan

Chiyue Wei*, Cong Guo*, Junyao Zhang, Haoxuan Shan, Yifan Xu, Ziyue Zhang, Yudong Liu, Qinsi Wang, Changchun Zhou, Hai “Helen” Li, Yiran Chen

Thank you for listening.

Welcome to use our **full stack** open-source code at:

https://github.com/dubcyfor3/Focus

Duke
Center of Computational Evolutionary Intelligence (CEI)