CLAMP-ViT: Contrastive Data-Free Learning for Adaptive Post-Training Quantization of ViTs

Akshat Ramachandran[1], Souvik Kundu[2], Tushar Krishna[1]

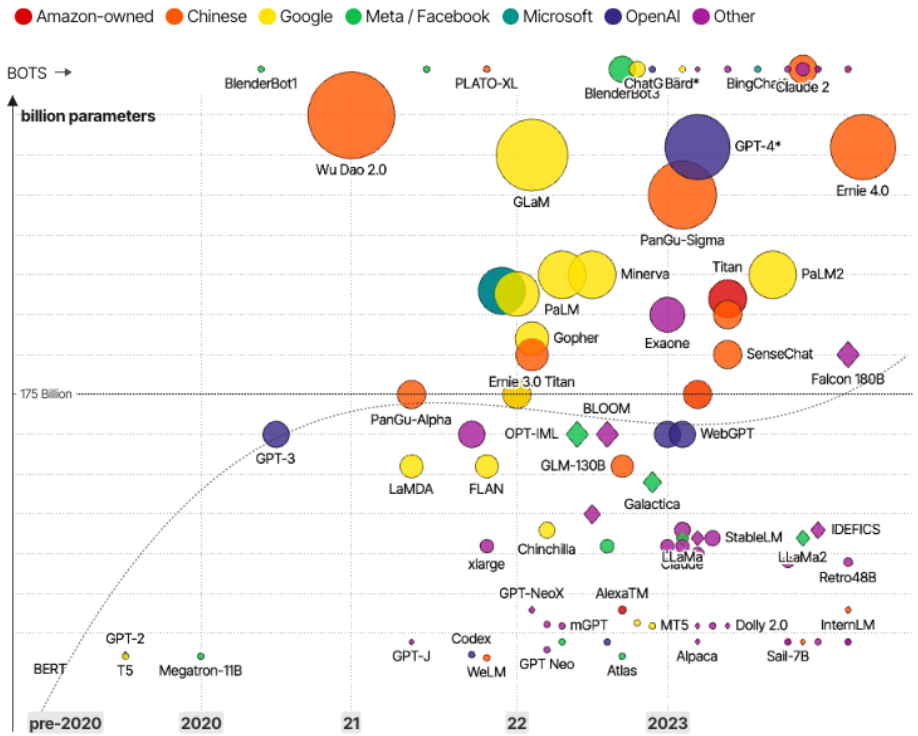
[1]Georgia Institute of Technology, [2]Intel Labs

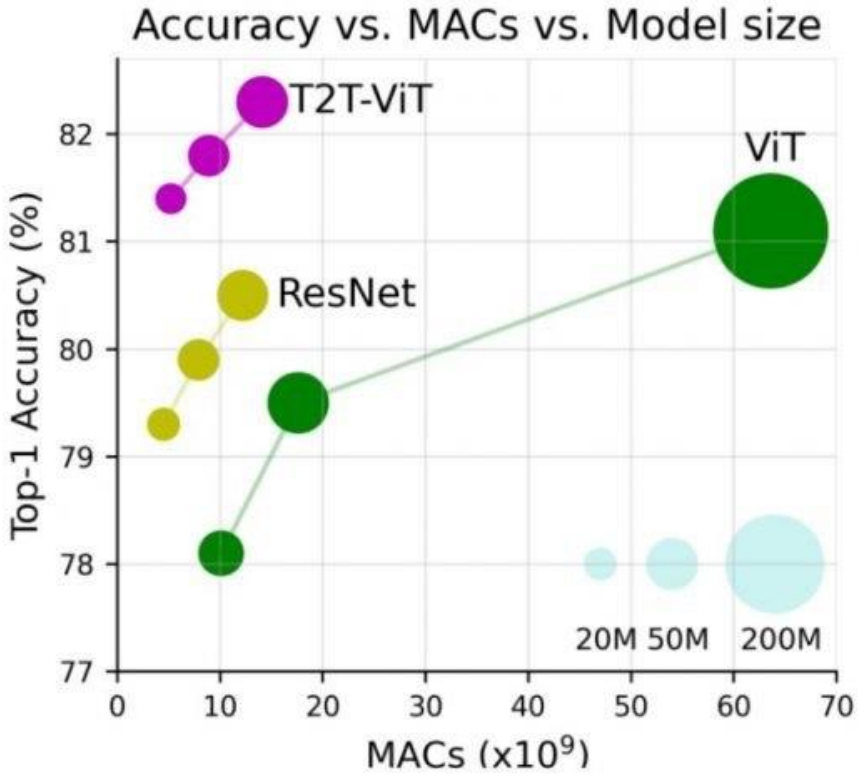
EUROPEAN CONFERENCE ON COMPUTER VISION

MILANO 2024

Contact: akshat.r@gatech.edu (Akshat Ramachandran)

From Bulky DNNs to Sleek Edge Deployment!

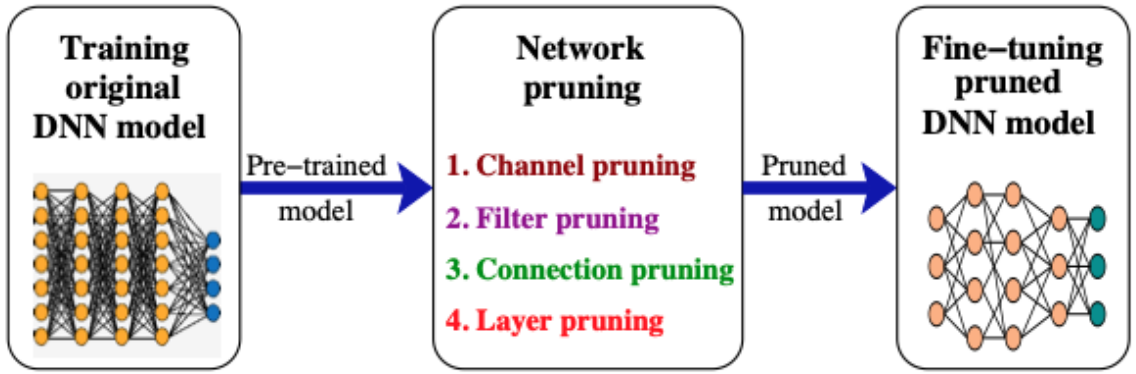




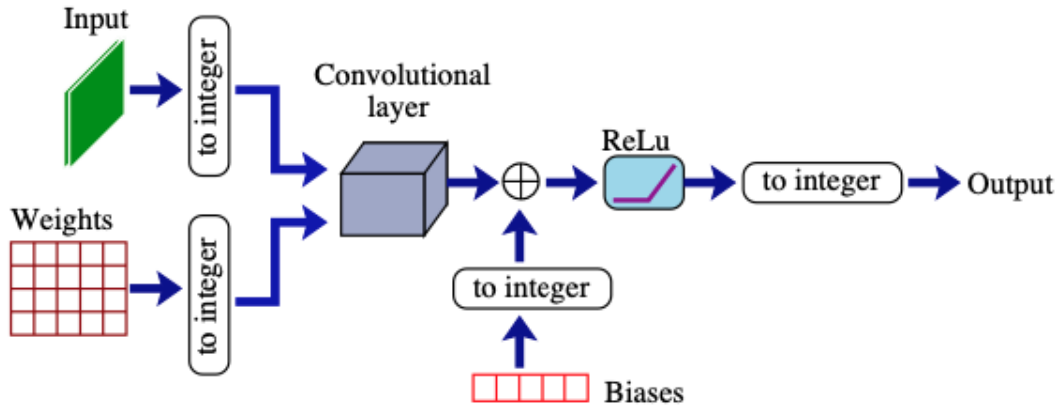
❖ YoY increase in DNN sizes leads to escalating computational and storage demands!
❖ Limited compute, storage resources and energy budget of edge devices (e.g., phones) makes deployment challenging!

From Bulky DNNs to Sleek Edge Deployment!

Model Compression techniques for efficient DNN deployment:



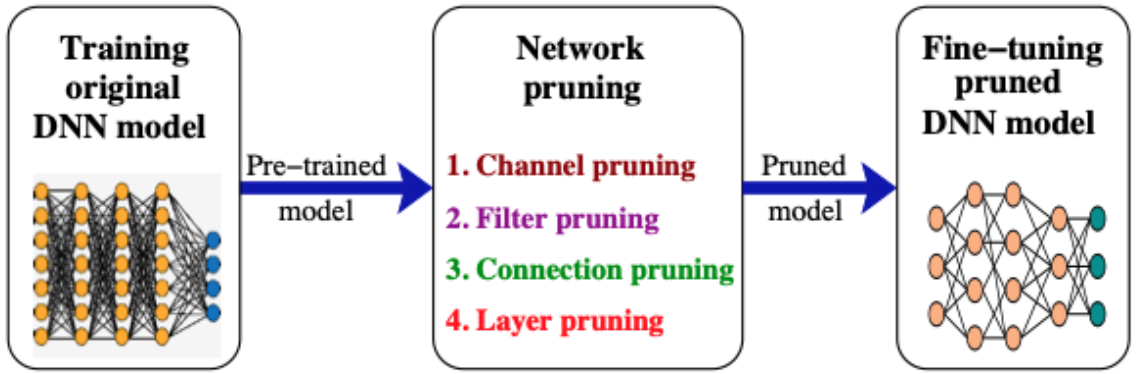
Pruning



Quantization

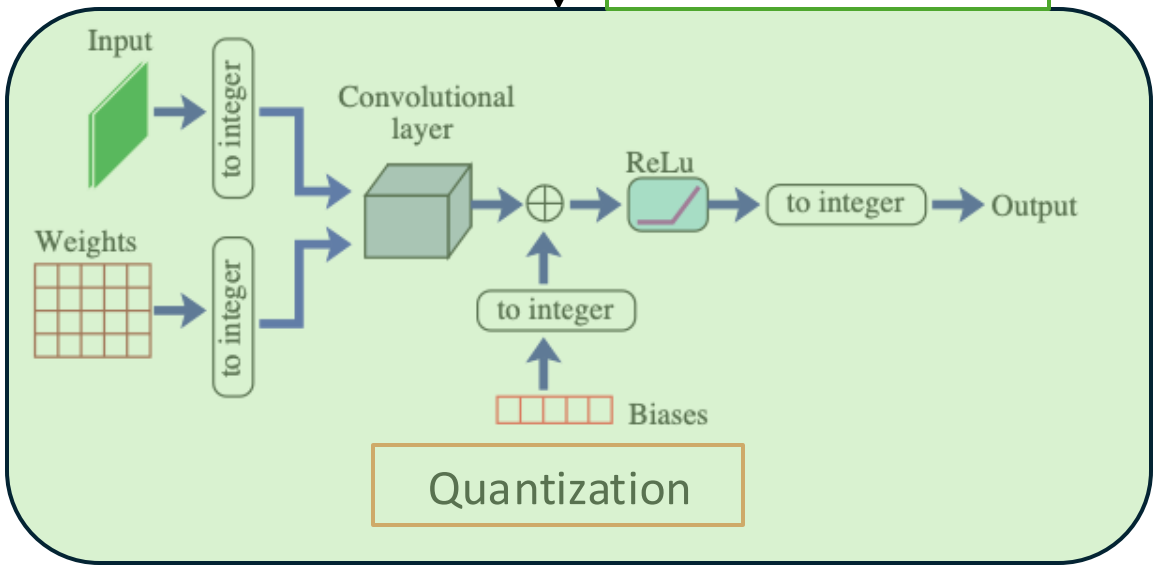
From Bulky DNNs to Sleek Edge Deployment!

Model Compression techniques for efficient DNN deployment:



Pruning

Focus of our work!



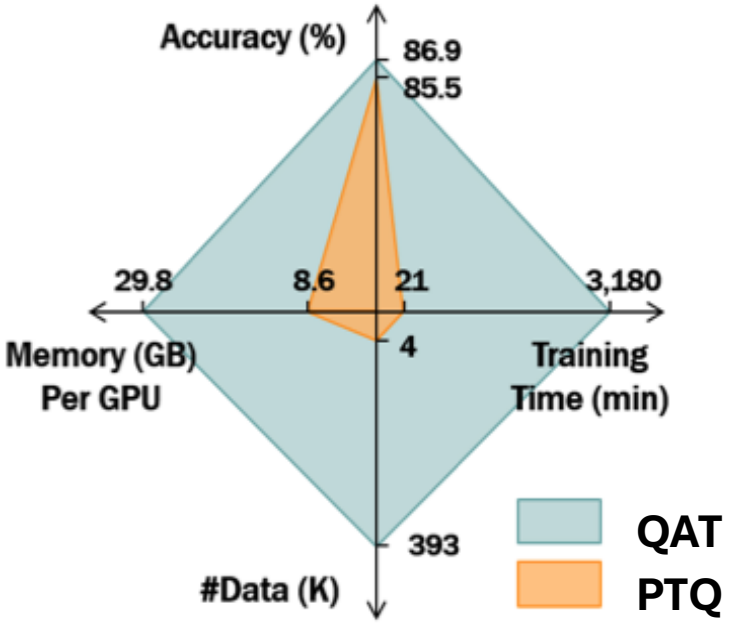
Quantization

Background: Types of Quantization Techniques

Quantization Aware Training (QAT):

- ✓ Higher Accuracy
- ✓ Improved Model Robustness

- ✗ Increased training complexity and time
- ✗ Large-scale data dependency



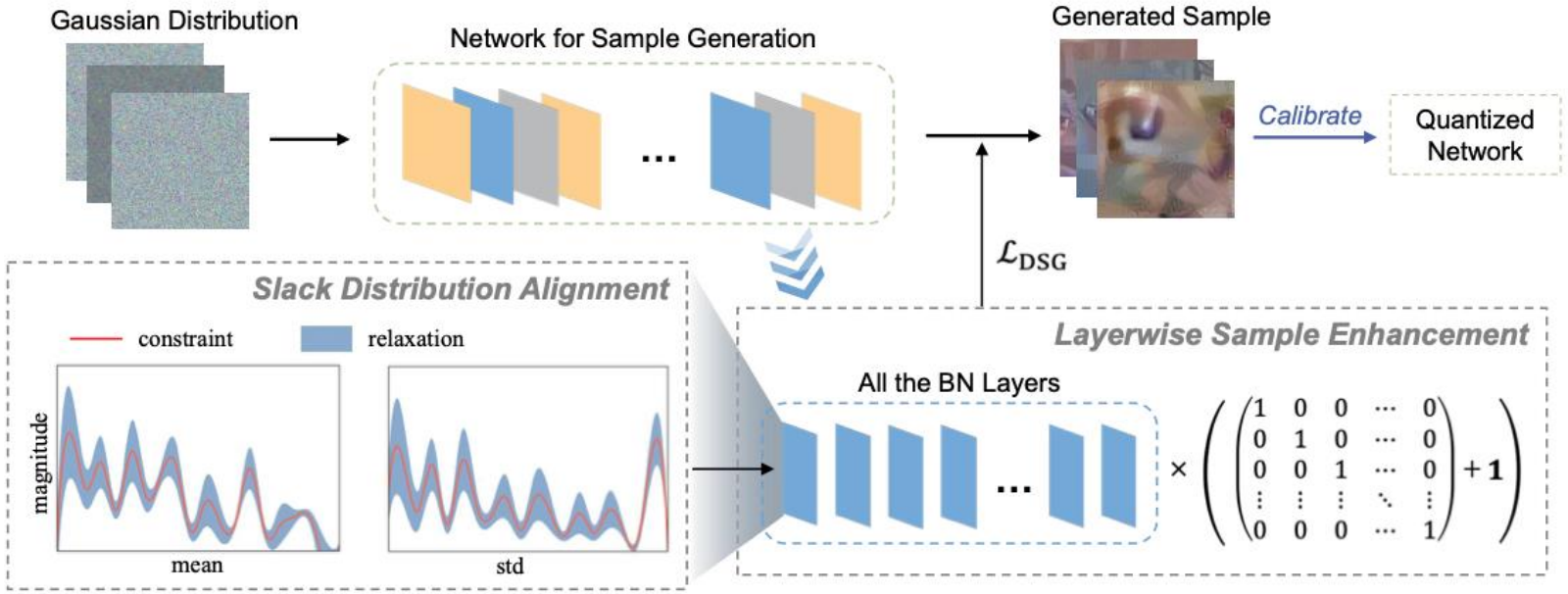
Focus of our work!

Post Training Quantization (PTQ):

- ✓ Speed and Simplicity
- ✓ Low data requirement (Can also be data-free)

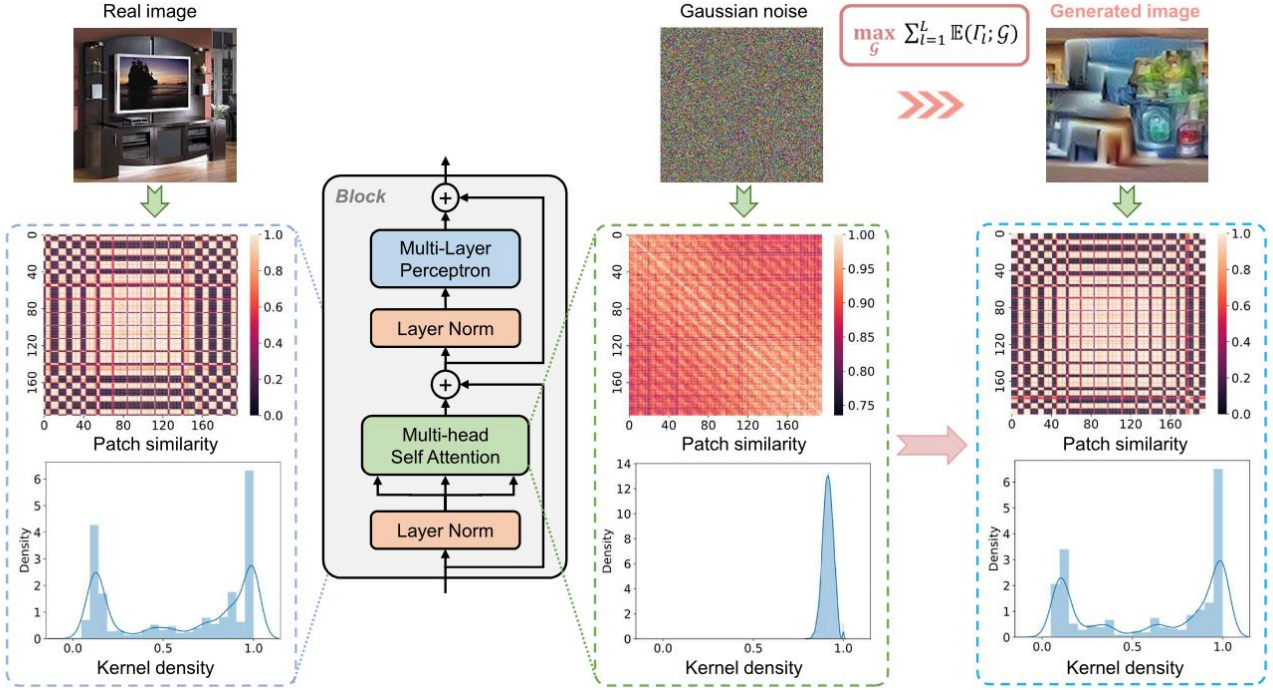
- ✗ Potential accuracy drop
- ✗ Sensitivity to calibration dataset

Data-Free Quantization before ViTs



- In CNNs, data-free quantization techniques synthesize data for calibrating the quantized model according to the batch normalization (BN) statistics of FP32.
- This is a reliable and efficient method for distilling data from the FP32 model for CNNs.

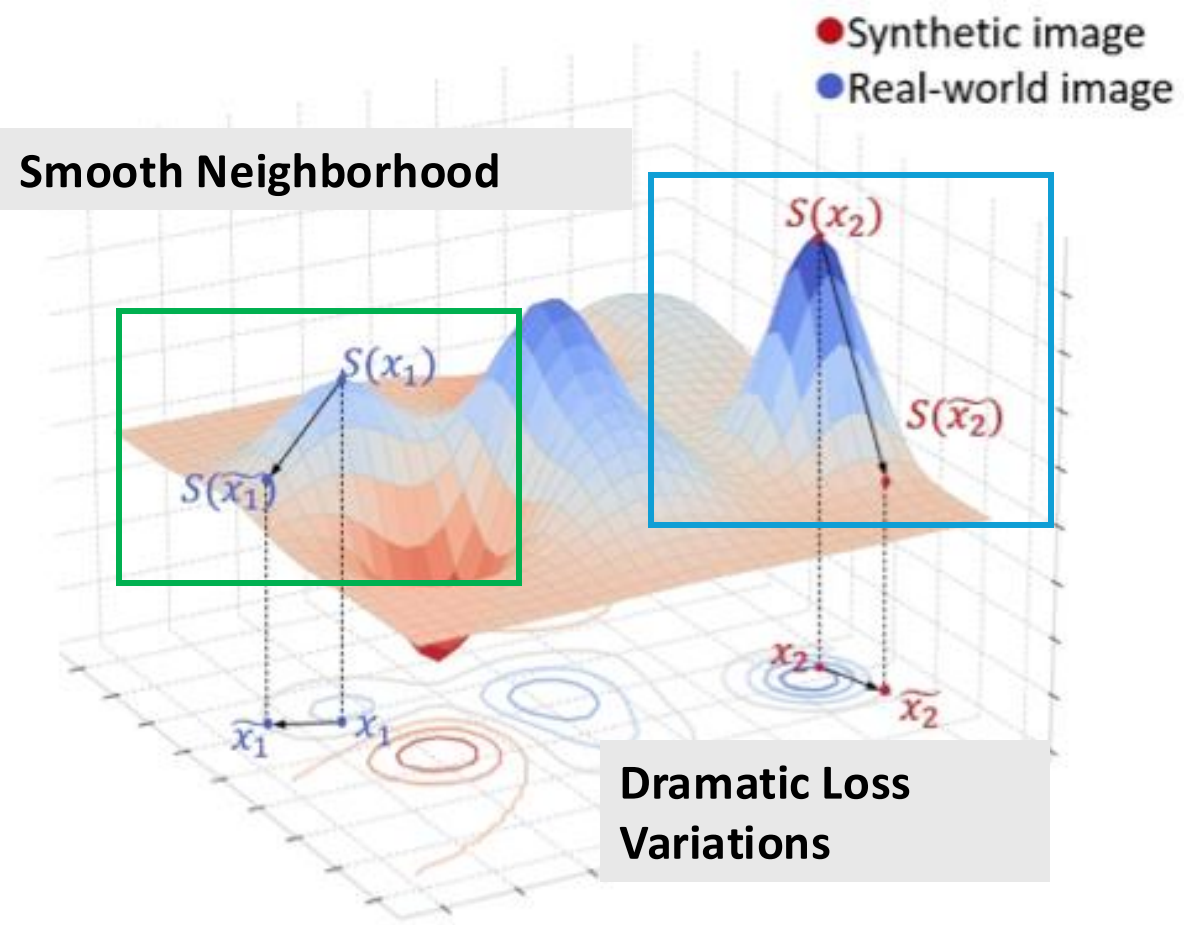
Data-Free Quantization: Transformer-Based models (ViTs)



Does not explore semantically meaningful interpatch relations!

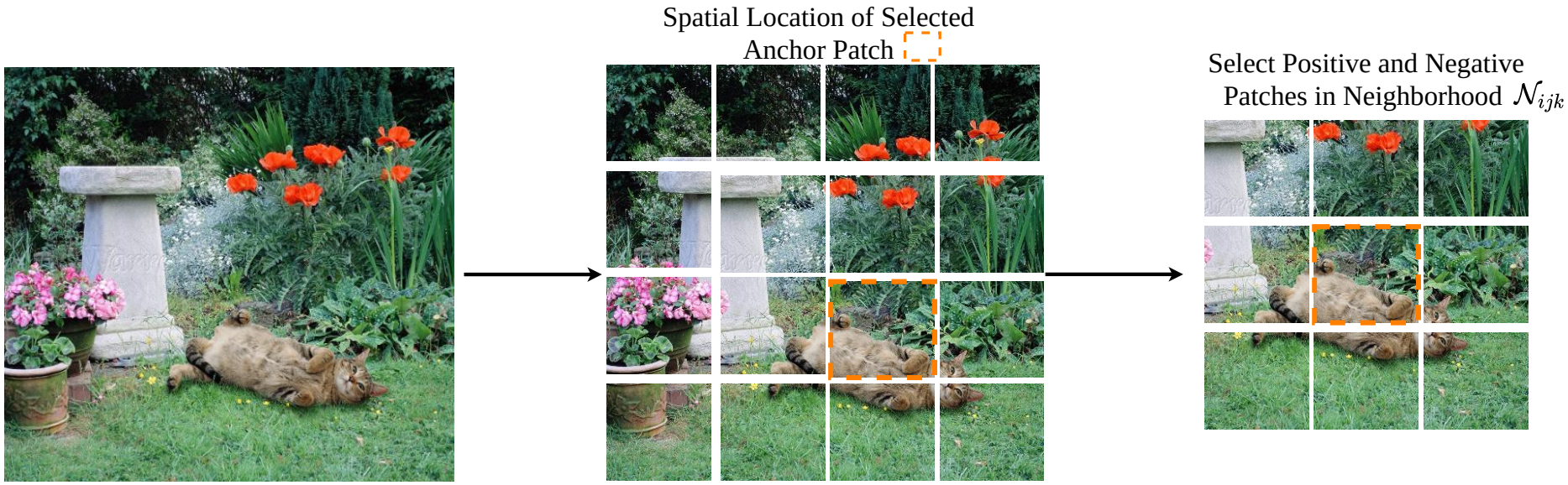
- ViTs or transformer-based models do not have a BN layer or any layer that holds statistics of the data it has been trained on.
- Previous methods rely on maximizing the entropy of the self-attention layer outputs.
- This ignores inter-patch relations and generates semantically non-informative and less realistic data.

Data-Free Quantization



- Using a global scheme without considering inter-patch relations does not result in realistic and robust images.
- The lower semantic content in generated synthetic images severely impacts quantized model generalizability.

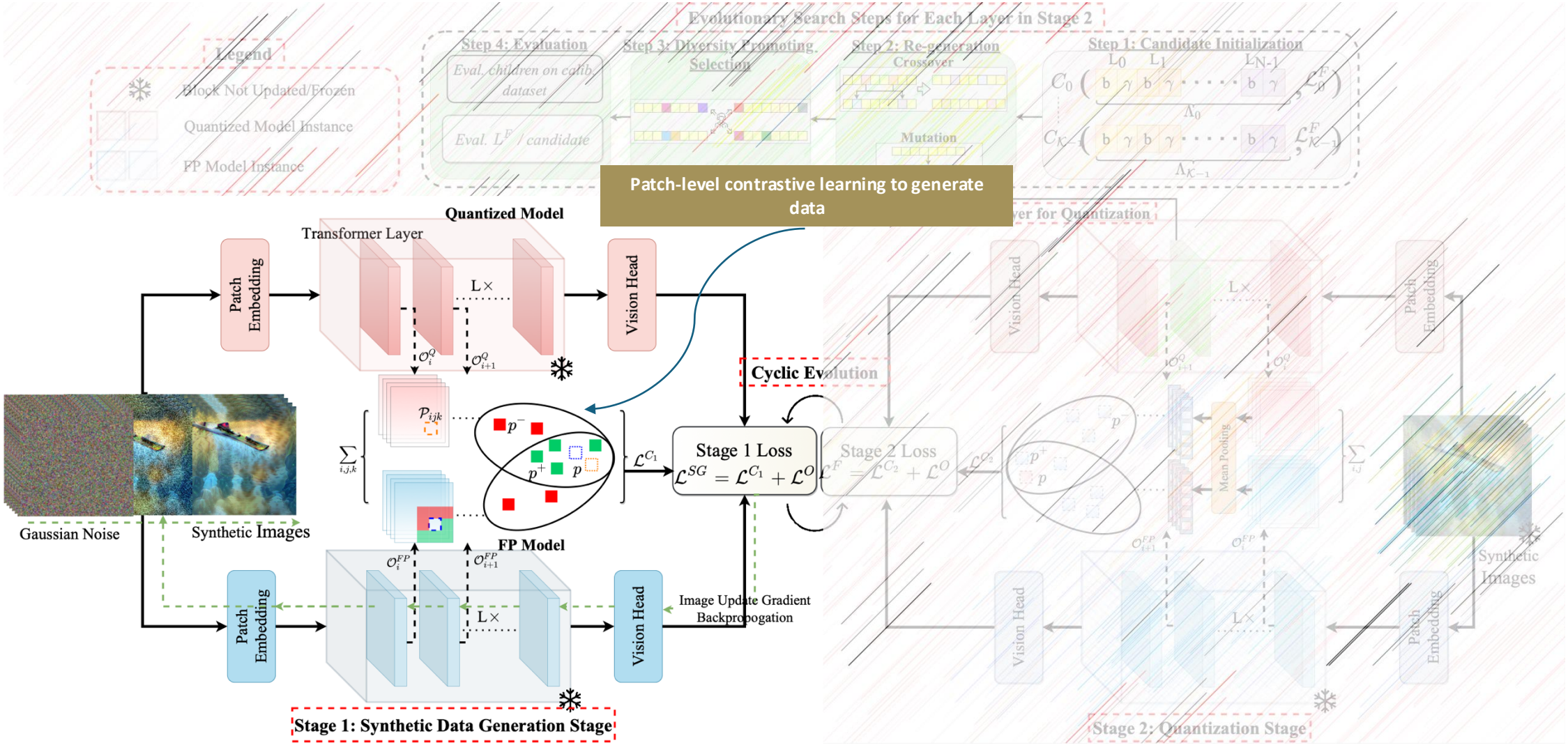
Data-Free Quantization: Proposed Approach



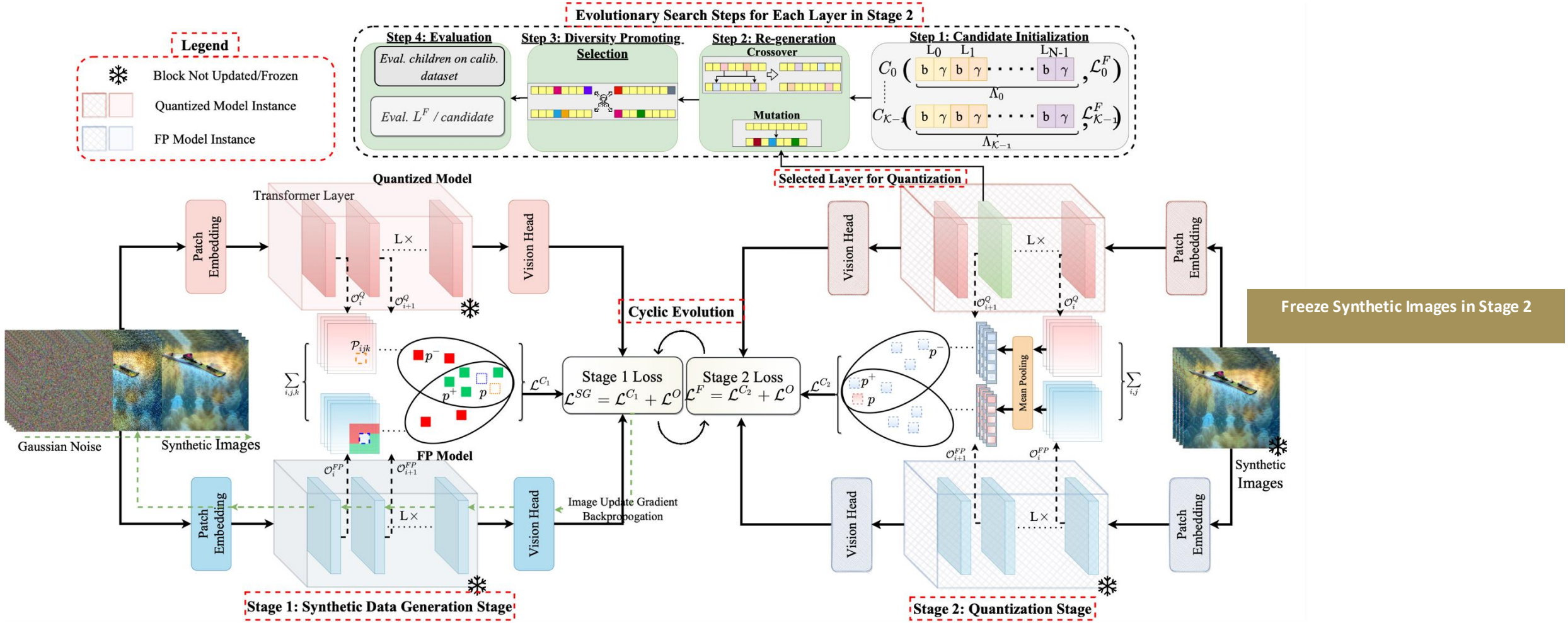
- We propose a data-generation scheme that leverages the architectural characteristics of ViTs i.e, patch-level attention and the inherent properties of real-images to generate semantically rich and meaningful data.

- We develop a contrastive learning scheme that treats semantically similar patches in a neighborhood as positive and others as negative.

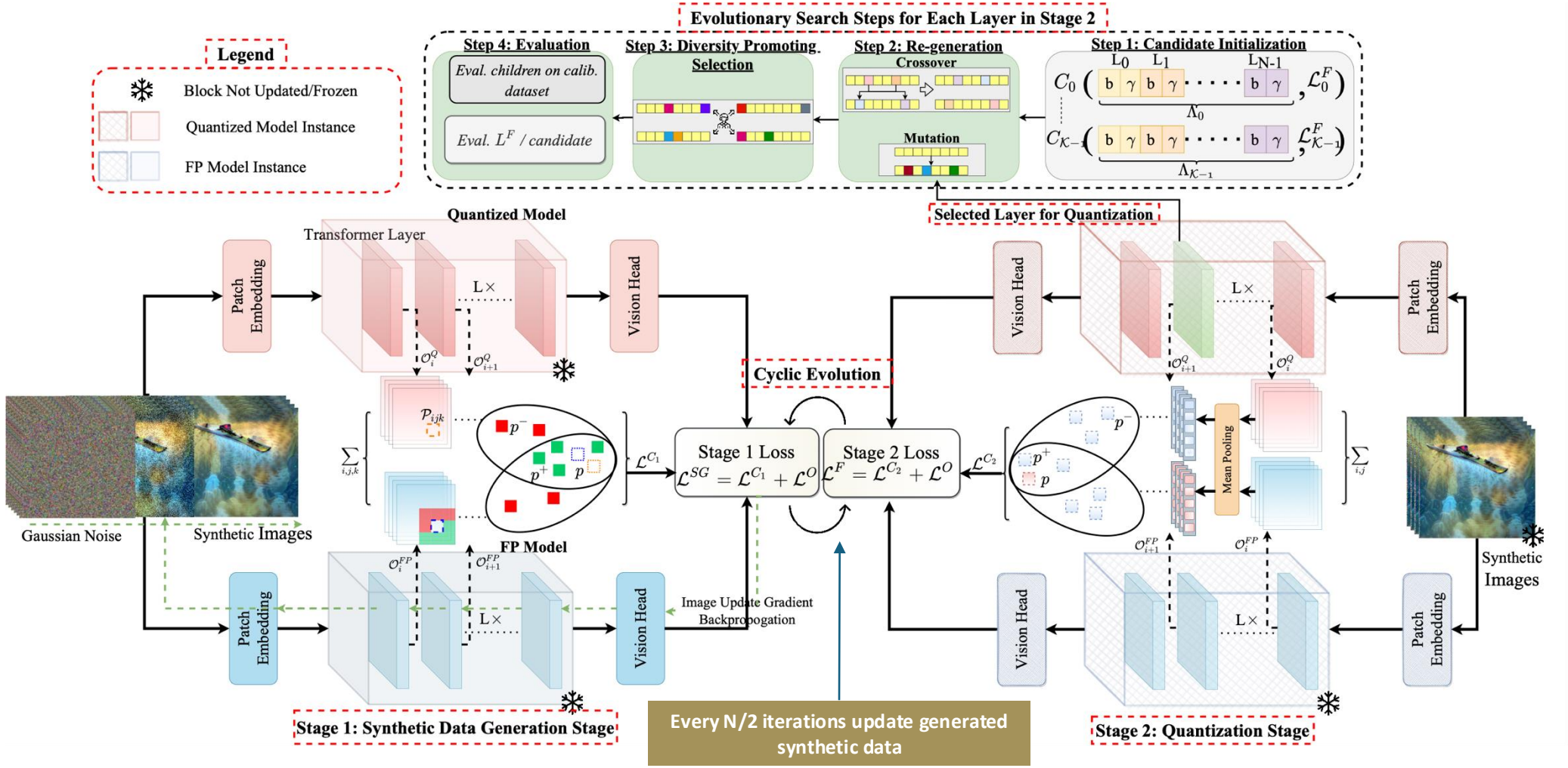
Data-Free Quantization: Complete Pipeline



Data-Free Quantization: Complete Pipeline



Data-Free Quantization: Complete Pipeline



Data-Free Quantization: Generated Samples

PSAQ-ViT v1

PSAQ-ViT v2

Ours

Notice the semantic meaningfulness!

Data-Free Quantization Performance

Model	Method	Data	#Images	W/A	Top-1	W/A	Top-1
ViT-B	Baseline	-	-	32/32	84.53	32/32	84.53
	PSAQ-ViT v1	S	32	8/8	37.36	4/8	25.34
	PTQ4ViT	R	32	8/8	**84.25**	4/8	67.16
	FQ-ViT	R	1000	8/8	83.31	4/8	**78.73**
	RepQ-ViT	R	32	8/8	81.45	4/8	76.29
	CLAMP-ViT (Ours)	S	32	8/8	[illegible]	4/8	**78.73**
DeiT-T	Baseline	-	-	32/32	72.21	32/32	72.21
	PSAQ-ViT v1	S	32	8/8	71.56	4/8	65.57
	PSAQ-ViT v2	S	32	8/8	72.17	4/8	68.61
	FQ-ViT	R	1000	8/8	71.61	4/8	66.91
	RepQ-ViT	R	32	8/8	72.05	4/8	68.75
	CLAMP-ViT (Ours)	S	32	8/8	**72.17**	4/8	**69.93**
DeiT-S	Baseline	-	-	32/32	[illegible]	32/32	79.85
	PSAQ-ViT v1	S	32	8/8	76.92	4/8	73.23
	PSAQ-ViT v2	S	32	8/8	**79.56**	4/8	76.36
	PTQ4ViT	R	32	8/8	79.47	4/8	-
	FQ-ViT	R	1000	8/8	79.17	4/8	76.93
	RepQ-ViT	R	32	8/8	79.55	4/8	76.75
	CLAMP-ViT (Ours)	S	32	8/8	79.55	4/8	**77.03**

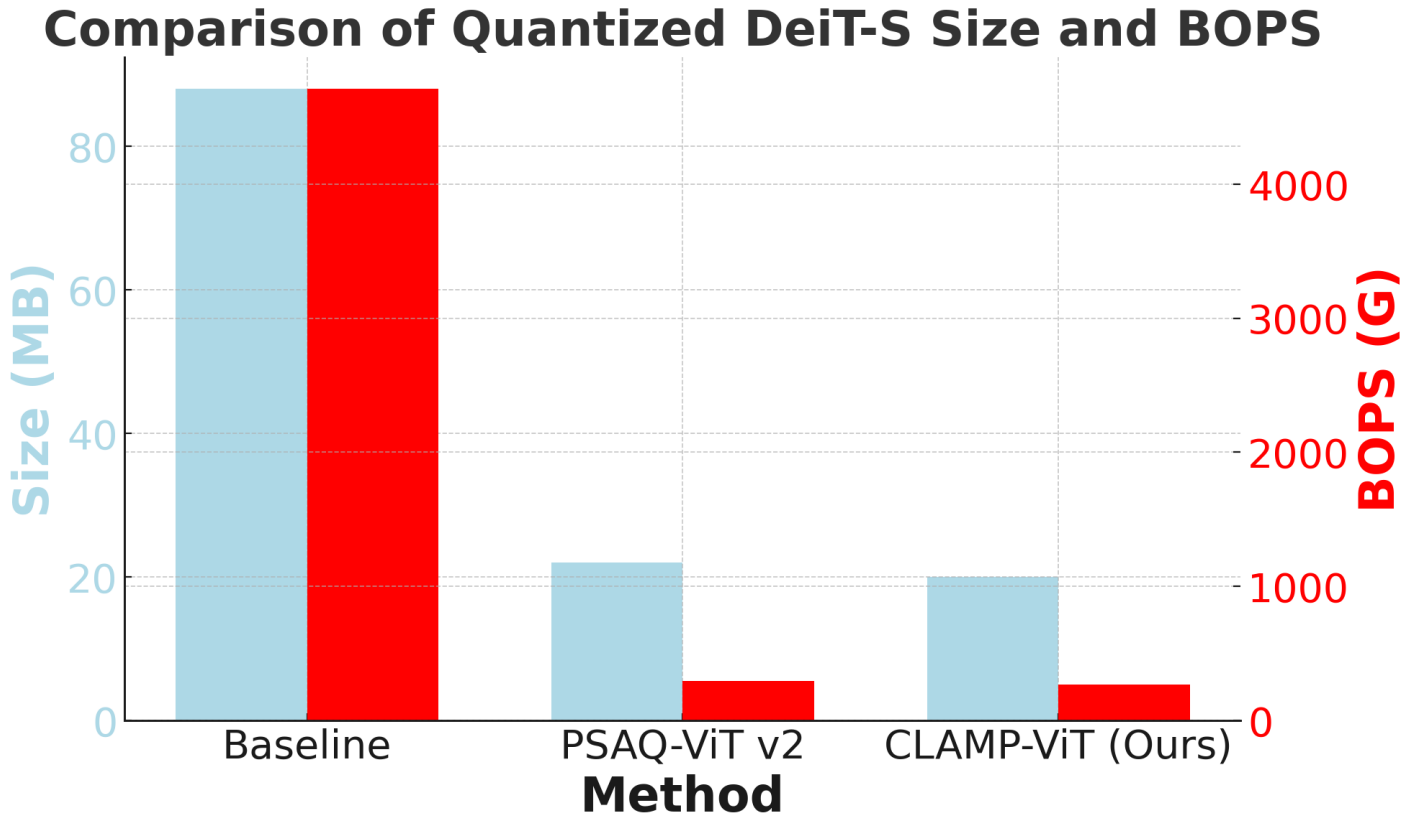
- **~2.2% Improvement over DFQ and ~1% Improvement over PTQ methods!**

Data-Free Quantization Performance

Model	Method	Data	#Images	W/A	Top-1	W/A	Top-1
ViT-B	Baseline	-	-	32/32	84.53	32/32	84.53
	PSAQ-ViT v1	S	32	8/8	37.36	4/8	25.34
	PTQ4ViT	R	32	8/8	**84.25**	4/8	67.16
	FQ-ViT	R	1000	8/8	83.31	4/8	**78.73**
	RepQ-ViT	R	32	8/8	81.45	4/8	76.29
	CLAMP-ViT (Ours)	S	32	8/8	84.19	4/8	**78.73**
DeiT-T	Baseline	-	-	32/32	72.21	32/32	72.21
	PSAQ-ViT v1	S	32	8/8	71.56	4/8	65.57
	PSAQ-ViT v2	S	32	8/8	72.17	4/8	68.61
	FQ-ViT	R	1000	8/8	71.61	4/8	66.91
	RepQ-ViT	R	32	8/8	72.05	4/8	68.75
	CLAMP-ViT (Ours)	S	32	8/8	**72.17**	4/8	**69.93**
DeiT-S	Baseline	-	-	32/32	79.85	32/32	79.85
	PSAQ-ViT v1	S	32	8/8	76.92	4/8	73.23
	PSAQ-ViT v2	S	32	8/8	**79.56**	4/8	76.36
	PTQ4ViT	R	32	8/8	79.47	4/8	-
	FQ-ViT	R	1000	8/8	79.17	4/8	76.93
	RepQ-ViT	R	32	8/8	79.55	4/8	76.75
	CLAMP-ViT (Ours)	S	32	8/8	79.55	4/8	**77.03**

- For W4/A8 our method shows significant performance boost over all the existing alternatives despite be

Quantized Model Stats



Thank you!

https://github.com/georgia-tech-synergy-lab/CLAMP-ViT

https://synergy.ece.gatech.edu/1443-2/