MimiQ: Low-Bit Data-Free Quantization of Vision Transformers with Encouraging Inter-Head Attention Similarity

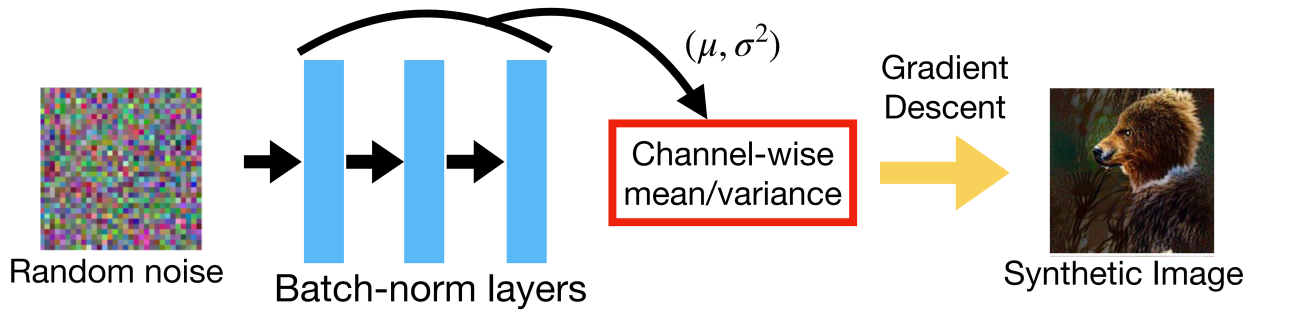
Kanghyun Choi[1] Hyeyoon Lee[1] Dain Kwon[1] SunJong Park[1] Kyuyeun Kim[2] Noseong Park[3] Jonghyun Choi[1] Jinho Lee[1]
[1]Seoul National University [2]Google [3]KAIST

AAAI-25 / IAAI-25 / EAAI-25
FEBRUARY 25 – MARCH 4, 2025 | PHILADELPHIA, USA

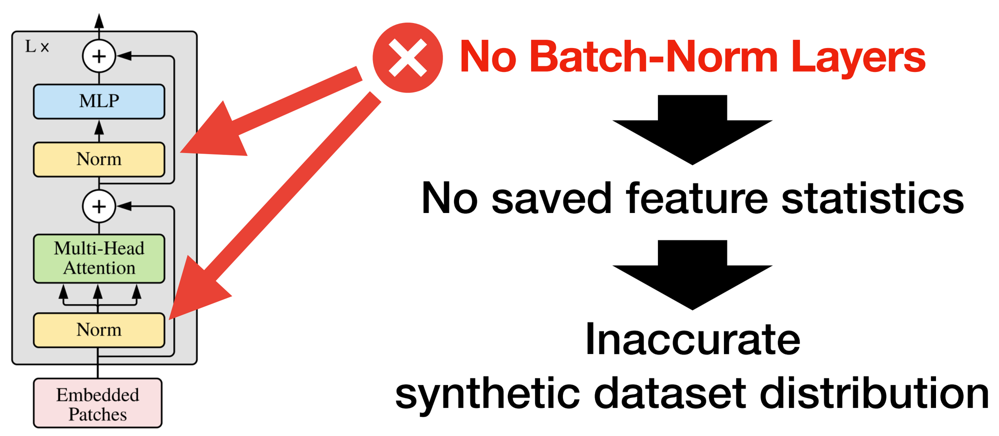
TL; DR: Enhancing head-wise attention similarity in Vision Transformers leads to better low-bit data-free quantization.

Backgrounds: Data-Free Quantization

- The original dataset is often **inaccessible** due to various reasons, such as **privacy, copyright, or protection.**
- **Data-free quantization (DFQ)** aims to quantize networks without the original dataset.
- Prior works for CNN utilize saved statistics in **batch-normalization layers** to create a synthetic dataset closer to the original.



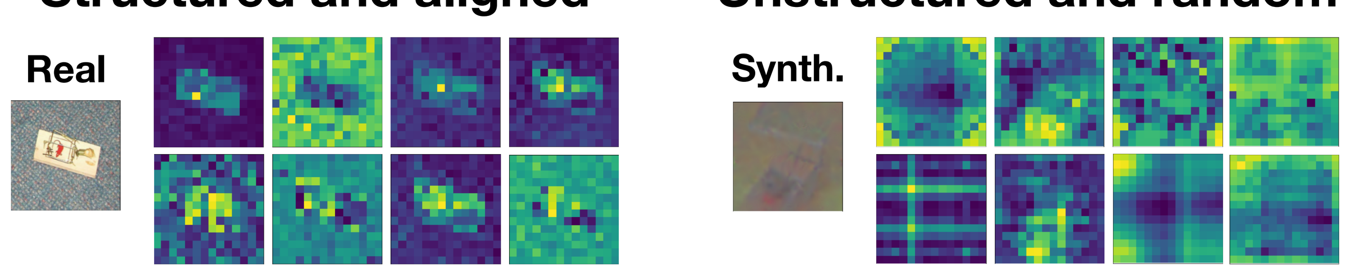
However, ViTs have no batch-norm layers



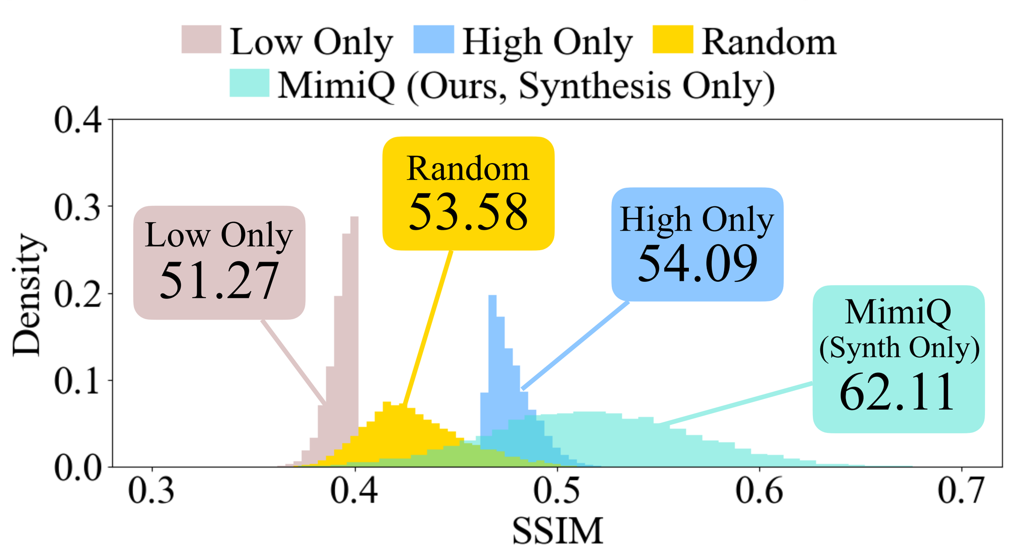
Motivation: Head-wise Attention Similarity

We examine **the head-wise attention map** of ViTs.

Structured and aligned **Unstructured and random**

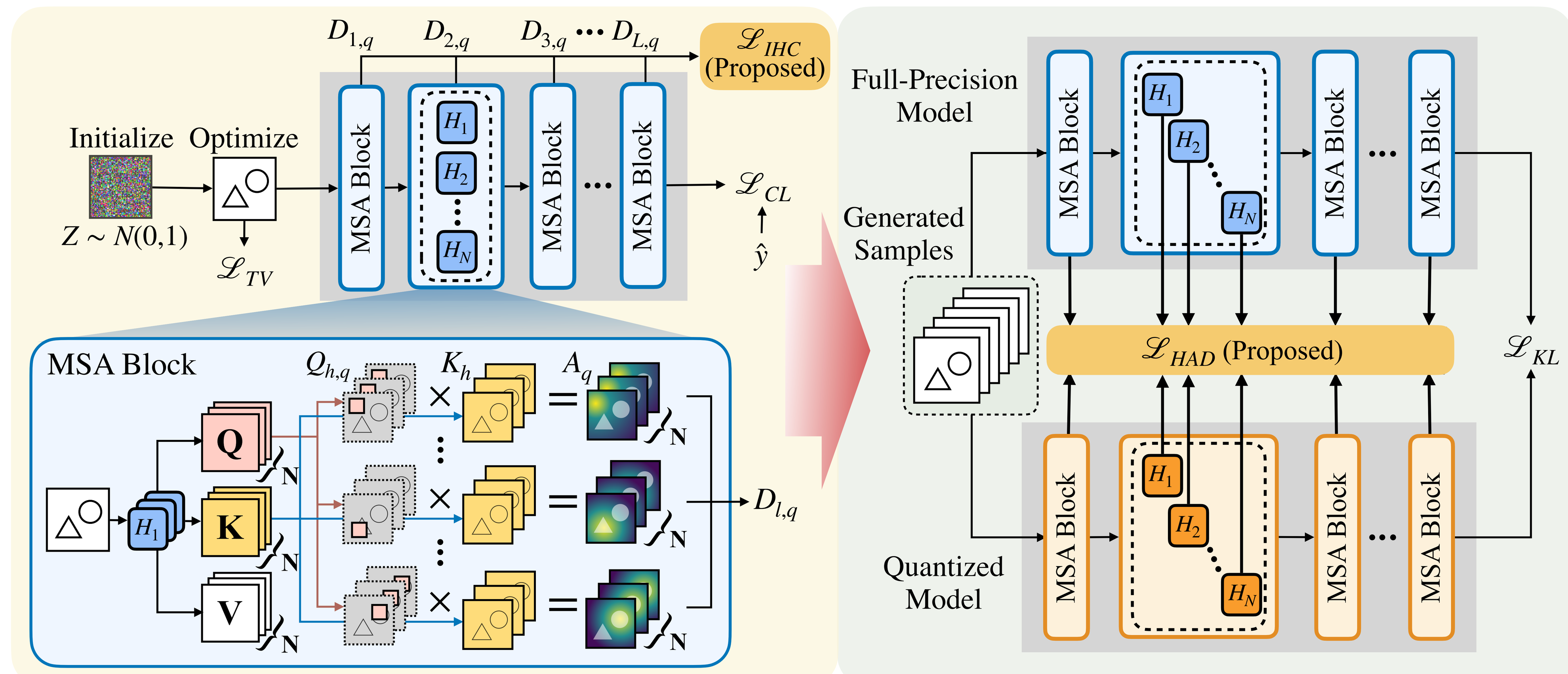


From the synthetic dataset, we sample low- and high-attention similarity data and test the quantization accuracy of ViTs.



Synthetic dataset with high attention similarity produces better quantization accuracy

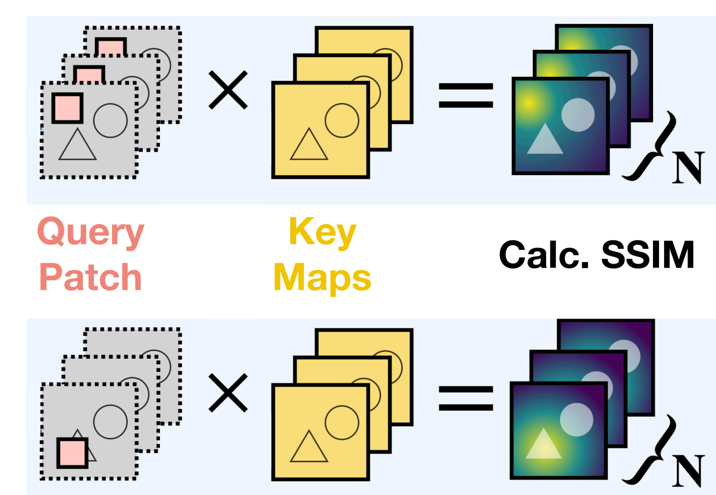
Overview of MimiQ Framework



Proposed Method 1: Sample Synthesis towards Inter-Head Similarity

Goal: Generate synthetic samples with high attention similarity

1. Collect attention map A_q of q-th query patch:
$$A_q = [Q_{1,q}K_1^T \quad Q_{2,q}K_2^T \quad \cdots \quad Q_{N,q}K_N^T]$$
2. Measure the distance D_q across N heads with distance metric f_{dist}. We choose structural similarity index measure (SSIM) for f_{dist}:
$$D_q = \frac{1}{N^2}\sum_i^N \sum_j^N f_{dist}(A_{q,i}, A_{q,j})$$
3. Compute Inter-head similarity loss: $\mathcal{L}_{IHC} = \frac{1}{LP}\sum_l^L \sum_q^P (1 - D_{l,q})$
4. Optimize synthetic samples with $\mathcal{L}_G = \mathcal{L}_{IHC} + \alpha\mathcal{L}_{CL} + \beta\mathcal{L}_{TV}$



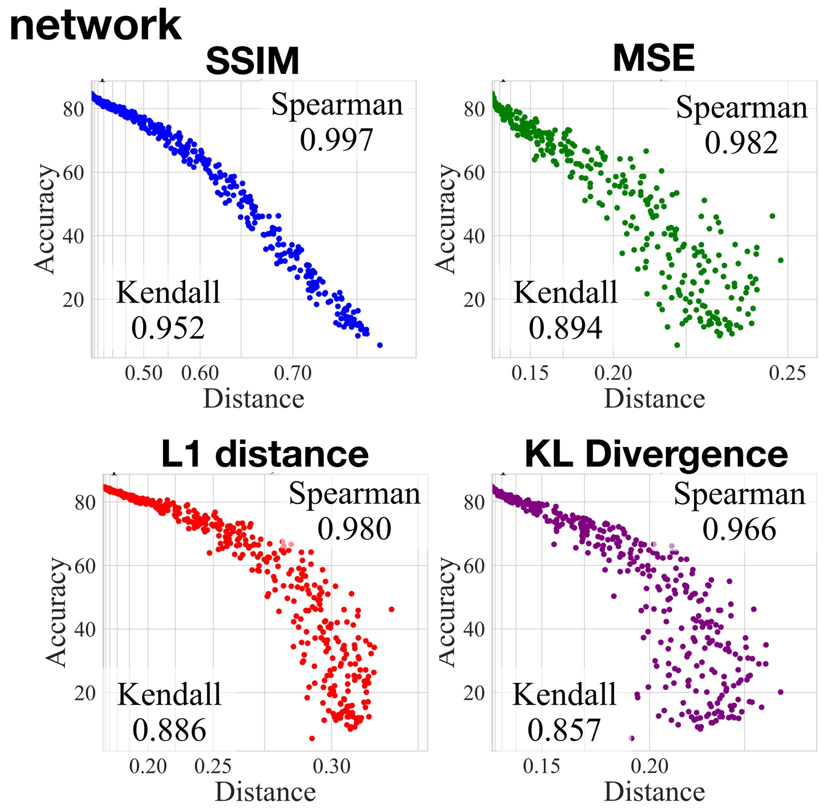
Proposed Method 2: Head-wise Structural Attention Distillation

Goal: Align attention maps of quantized network with full-precision network

1. Collect head-wise output H of teacher $\mathcal{T}$ and student $\mathcal{S}$.
2. Compute head-wise distance loss $\mathcal{L}_{HAD}$:
$$\mathcal{L}_{HAD} = \frac{1}{LN}\sum_l^L \sum_i^N g_{dist}(H_{l,i}^{\mathcal{T}}, H_{l,i}^{\mathcal{S}})$$
3. Train quantized network with $\mathcal{L}_T = \mathcal{L}_{KL}(f_{\mathcal{T}}(\hat{X}) || f_{\mathcal{S}}(\hat{X})) + \gamma\mathcal{L}_{HAD}$.

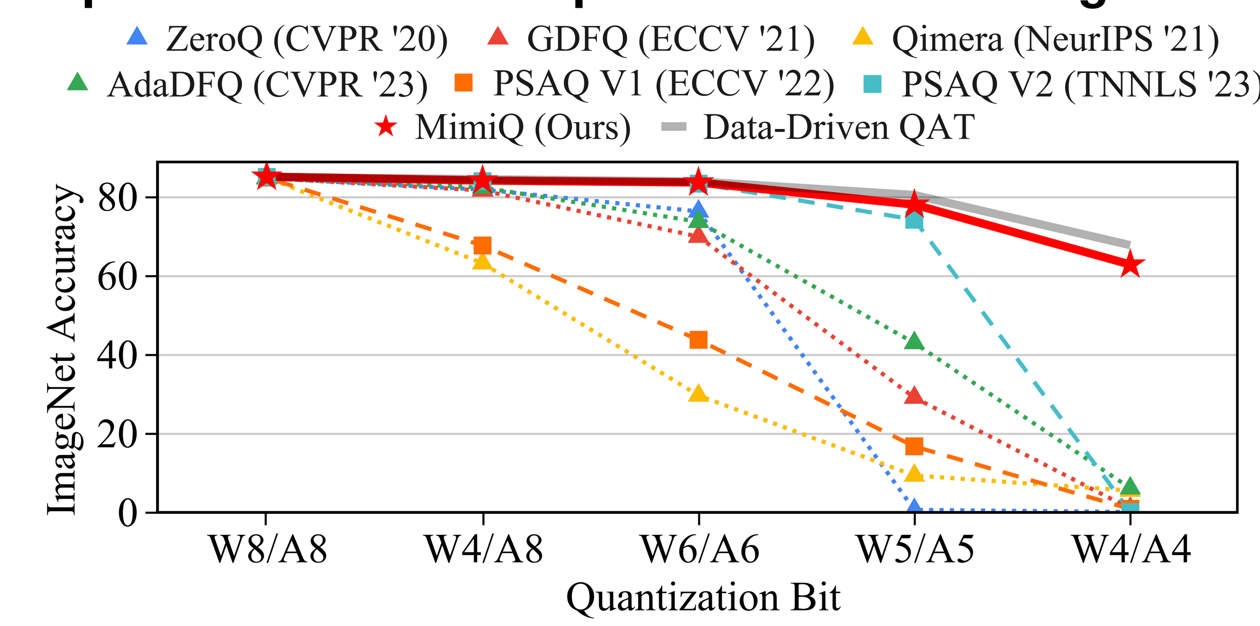
To choose g_{dist}, we randomly quantized a portion of attention heads and measured the accuracy and distance correlation of SSIM, mean-squared error, L1 distance, and KL divergence.

The results show that SSIM has the highest correlation.



Evaluation and Analyses

Comparison of different quantization bit settings

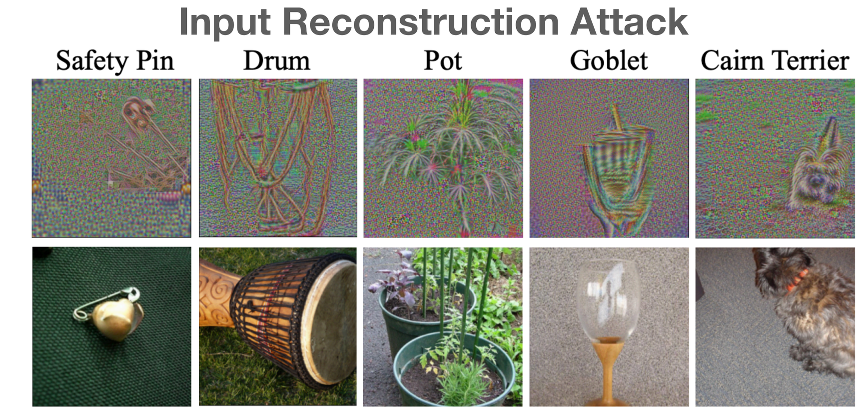


Comparison of quantization time and accuracy

Method	Type	Synth.	Quant.	Total	Acc.
GDFQ	QAT	-	10.70h	10.70h	11.73
AdaDFQ	QAT	-	8.44h	8.44h	6.21
PSAQ-V1	PTQ	0.11h	0.0002h	0.11h	0.94
PSAQ-V2	QAT	-	4.55h	4.55h	2.83
MimiQ-1k	QAT	1.98h	2.39h	4.37h	**59.32**
MimiQ-4k	QAT	7.92h	2.39h	10.31h	**62.59**
MimiQ-10k	QAT	19.79h	2.39h	22.18h	**62.91**

MimiQ preserves data privacy

Input Reconstruction Attack



Identity Attack and Model Inversion Attack

Measure	Train	Test
Synthetic/Real Distinguishability	99.97	99.99
Synthetic→Real Transferability	49.69	0.16

Input Reconstruction Attack: We measure LPIPS between MimiQ samples and the original dataset. The figure shows that MimiQ samples do not resemble specific images of the original data.

Model Inversion Attack: We show that our samples are clearly distinguishable from the real data and cannot be used to model inversion (stealing) attack.

GitHub Paper Author Page