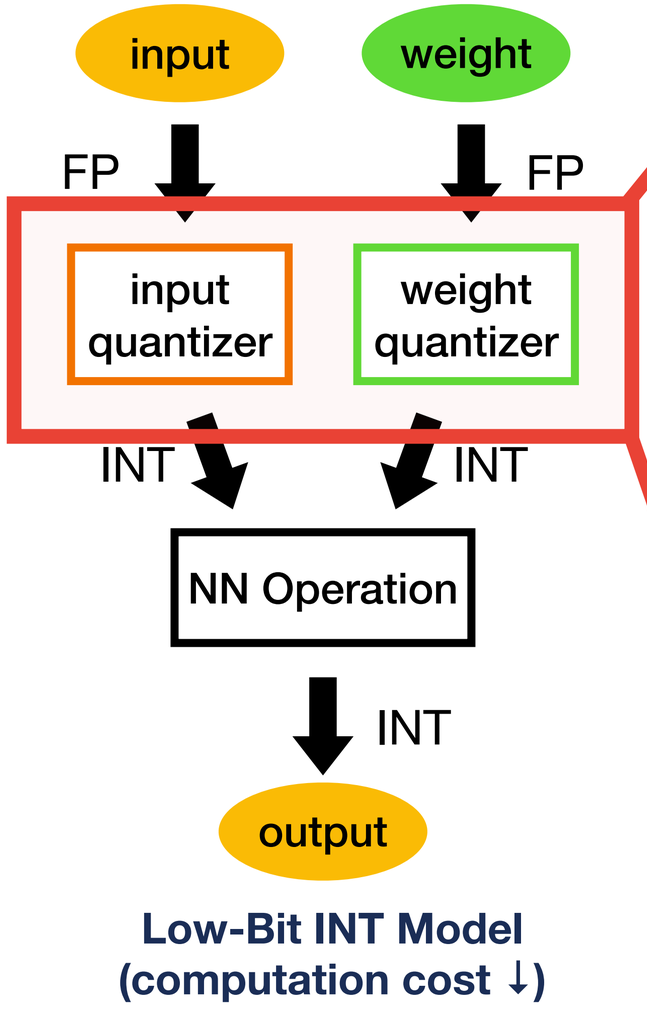
# MimiQ: Low-Bit Data-Free Quantization of Vision Transformers with Encouraging Inter-Head Attention Similarity

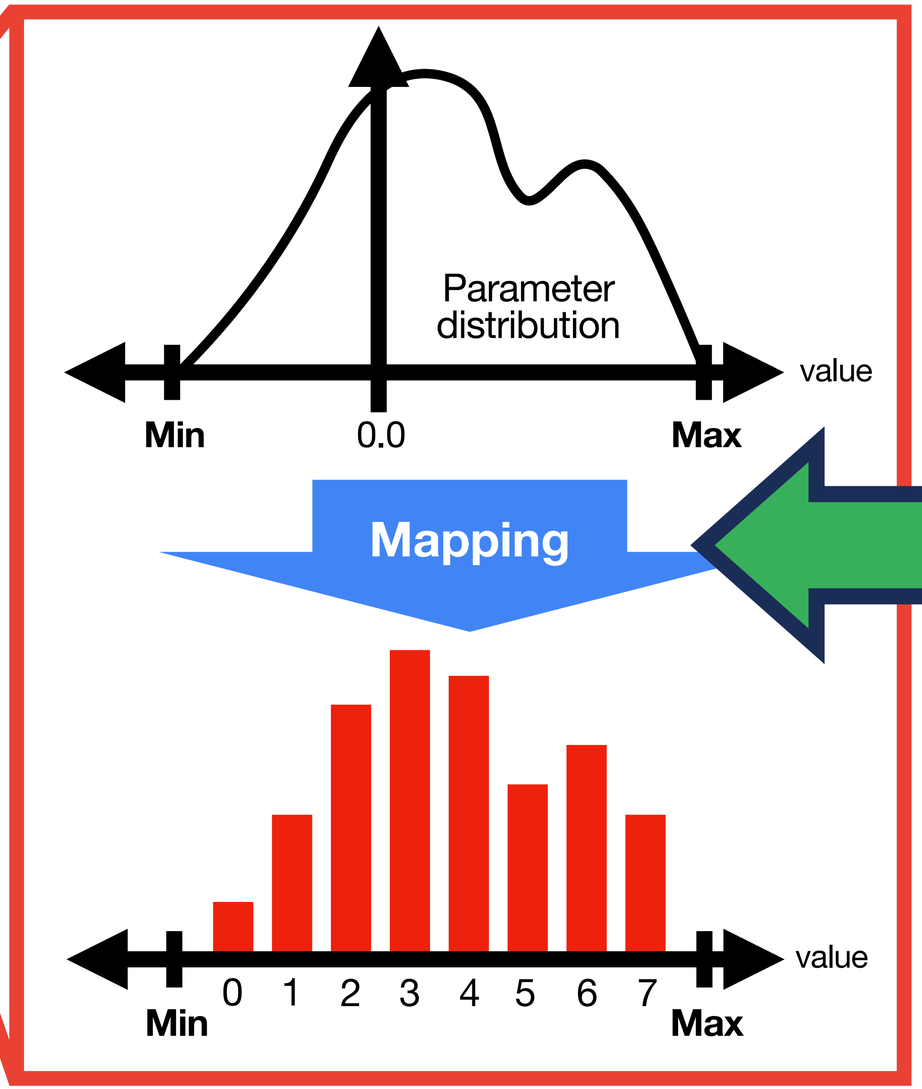
Kanghyun Choi[1] Hyeyoon Lee[1] Dain Kwon[1] SunJong Park[1]
Kyuyeun Kim[2] Noseong Park[3] Jonghyun Choi[1] Jinho Lee[1]

[1]Seoul National University
[2]Google
[3]KAIST

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

# Quantization Mapping




Require

**Calibration**

**Optimization**

**Training**

**Training Dataset**

# Calibration with the Original Datasets

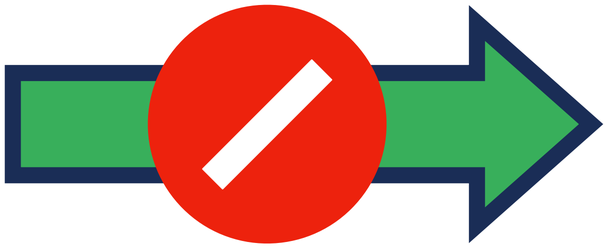
AAAI-25 / IAAI-25 / EAAI-25

**Require**

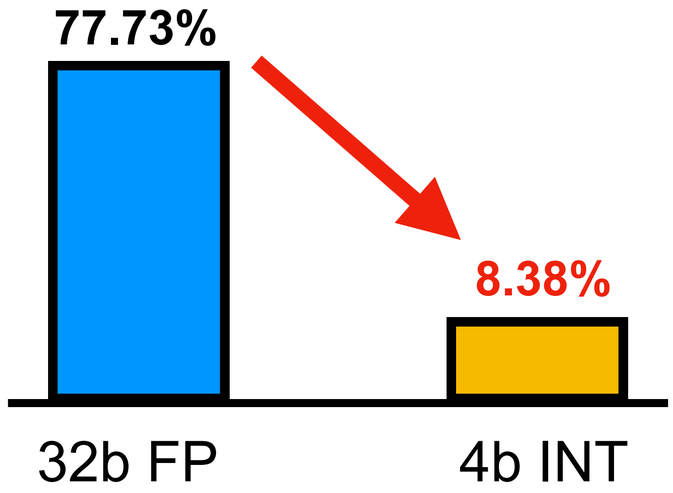
**Calibration**

**Optimization**

**Training**



Quantized network accuracy
w/o fine-tuning[1]




Necessary recalibration
with the original dataset

1 ResNet50, ImageNet-1k, 4bit quantization

# Calibration with the Original Datasets



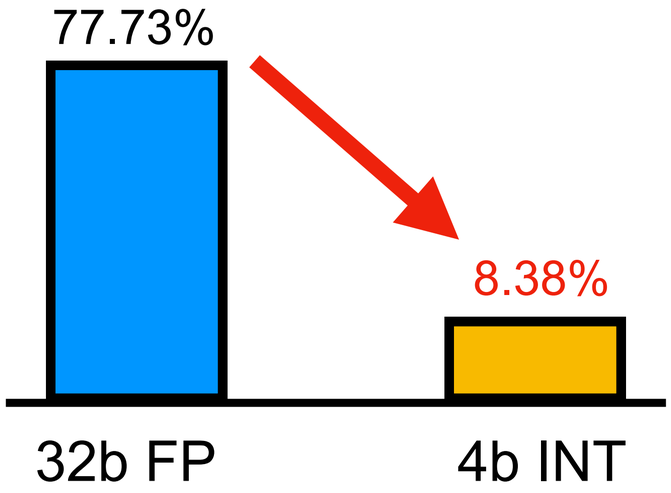
**The original dataset has issues**

- Copyright
- Privacy
- No public use
- Too large

# Data-Free Quantization



Quantized Network Accuracy
w/o fine-tuning[1]

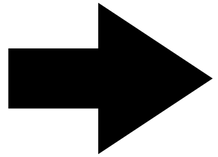



Necessary Recalibration
with Original Dataset

Inaccessible Dataset Problem

Privacy Protection Copyrights

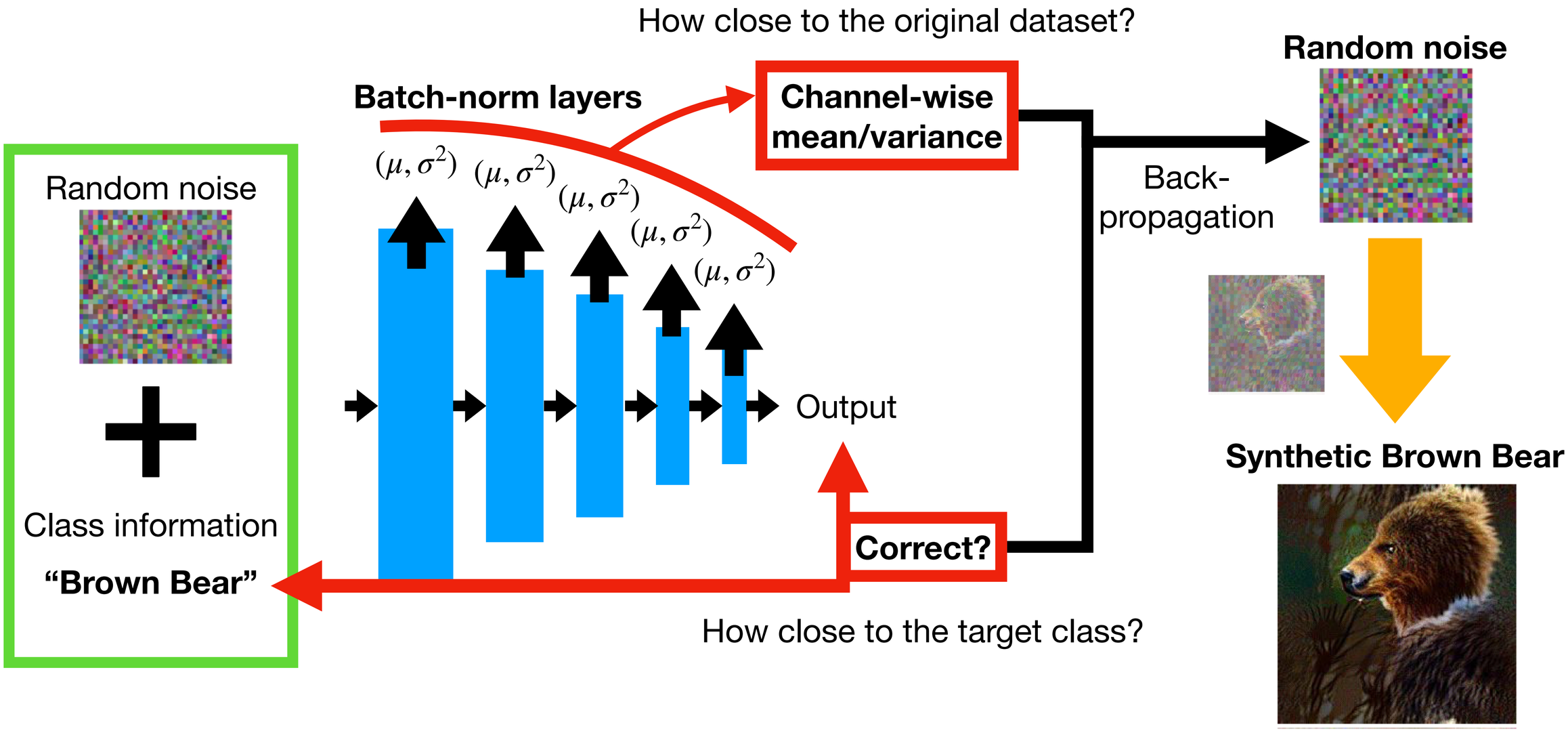
Fine-tuning
**without** the original dataset



**Data-Free**
**Quantization**

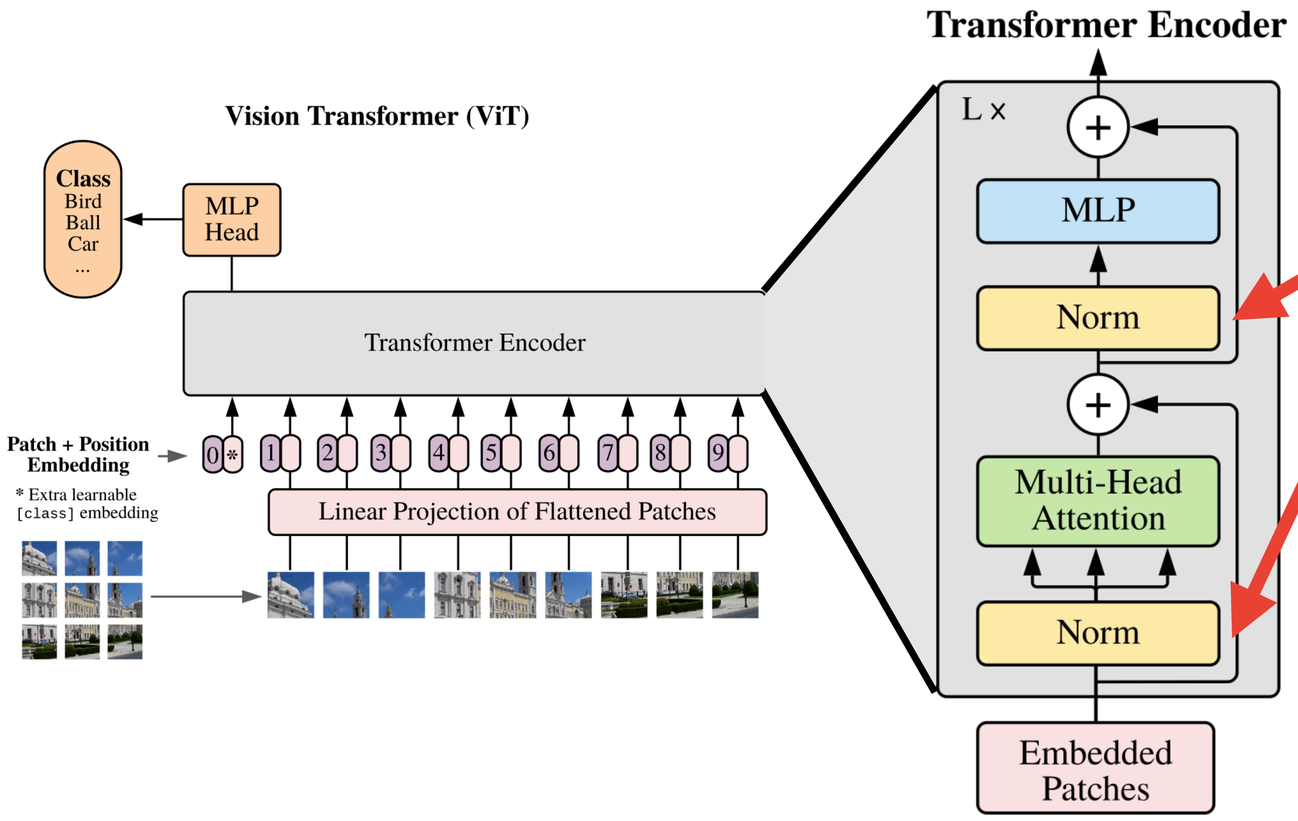
[1] ResNet50, ImageNet-1k, 4bit quantization

# Backgrounds: Data-Free CNN Quantization



# Data-Free ViT Quantization






**No Batch-Norm Layers**

**Using Instance-Norm Layer**

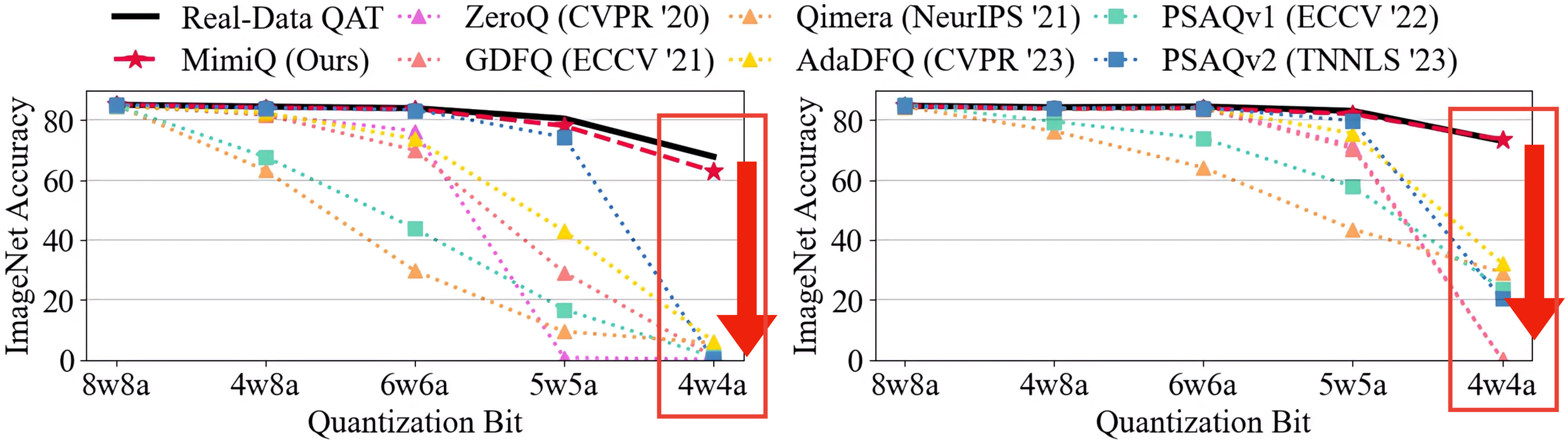
**No saved feature stats**

**Cannot create BN-based synthetic samples**

# Current Limitations: No Batch-Norm

AAAI-25 / IAAI-25 / EAAI-25

## Significant accuracy drop in low-bit quantization

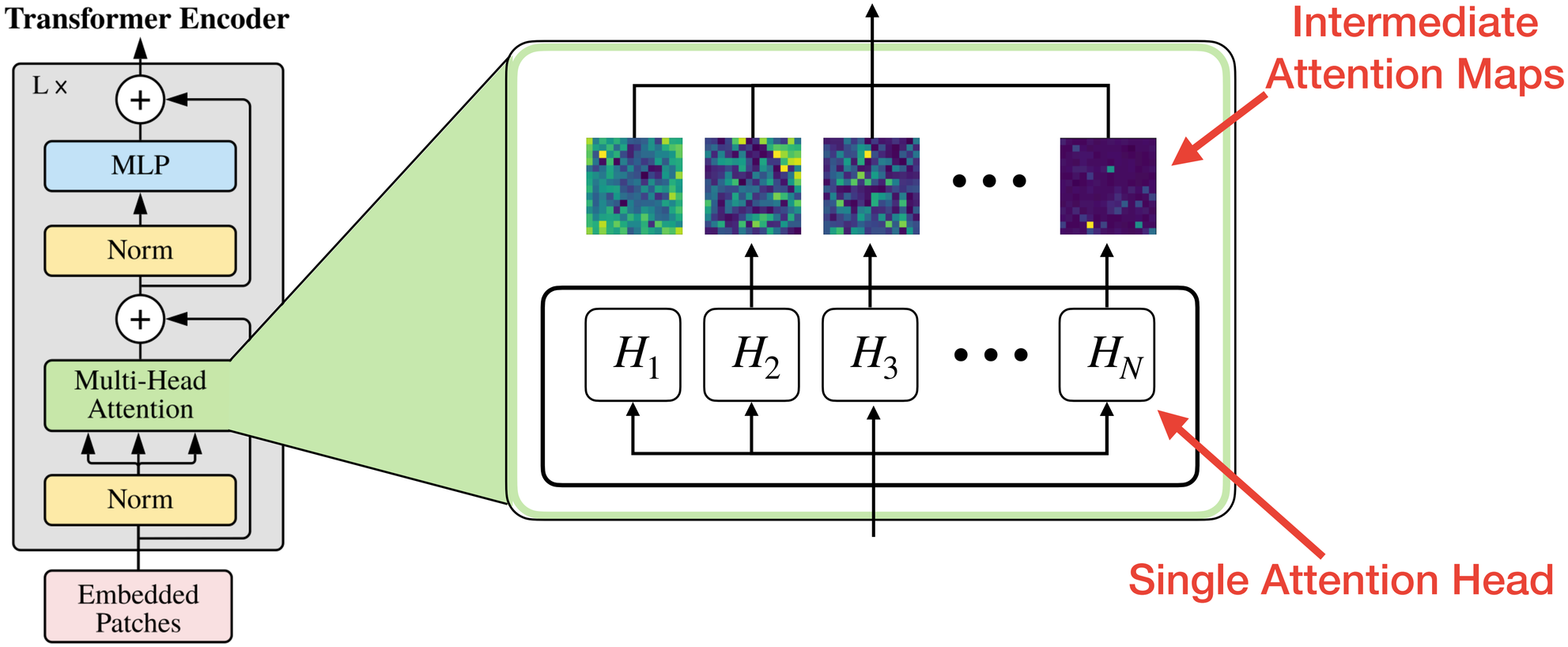



ViT-Base

Swin-Base

# Knowledge Extraction from Attention

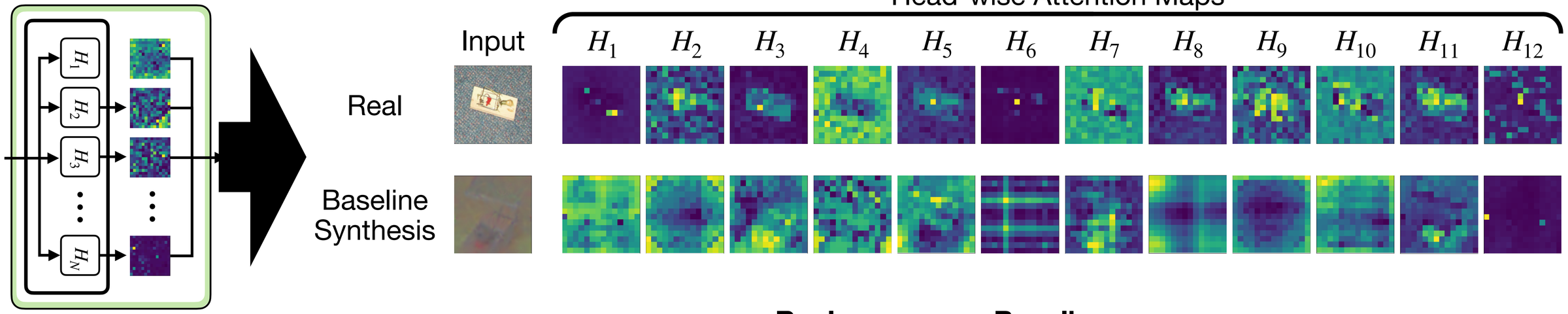





Extract prior knowledge from intermediate attention maps

# Knowledge Extraction from Attention

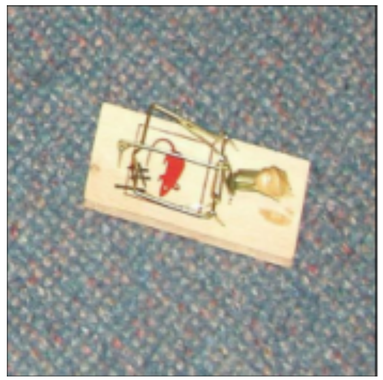
AAAI-25 / IAAI-25 / EAAI-25



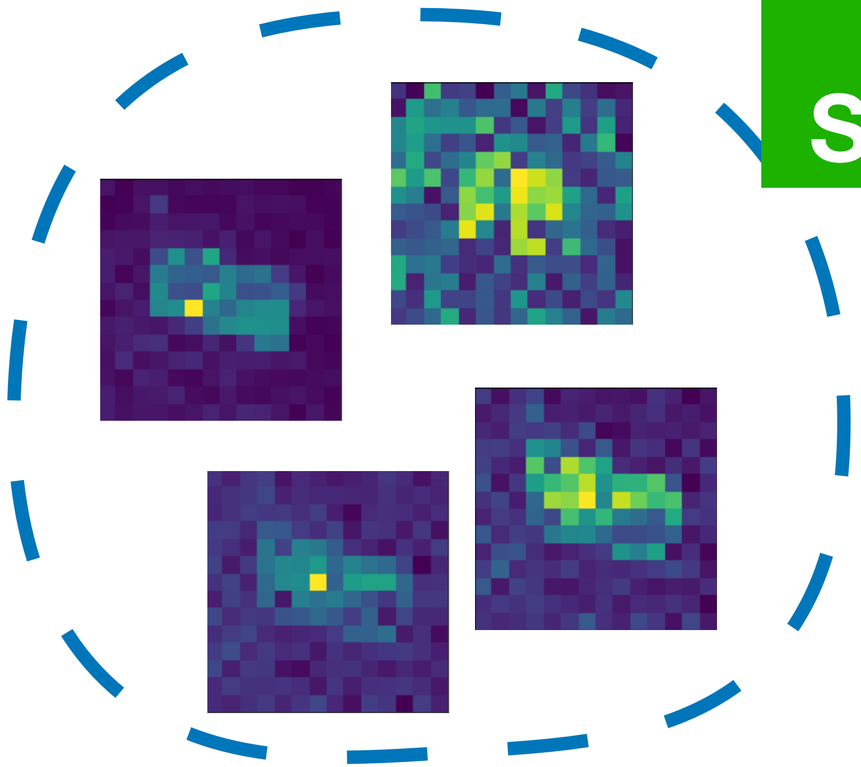

| | **Real** | **Baseline** |
|---|---|---|
| **Object Shape** | Structured | Random |
| **Object Position** | Correct | Unrecognizable |
| **Inter-head Similarity** | Aligned | Individualisitc |

# How to measure inter-head similarity?



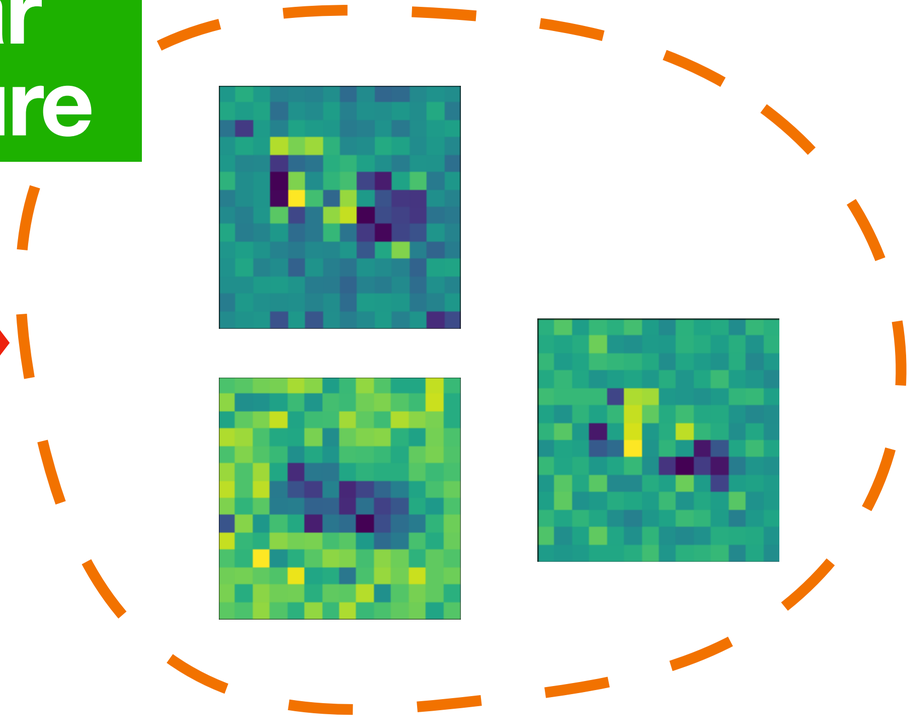


Input



Foreground-focusing attention heads

Similar Structure



Background-focusing attention heads

# Structure Similarity Index Measure



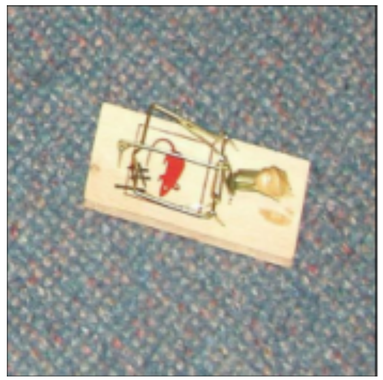
- SSIM: Image quality evaluation metric mimicking human perception
- Weighted product of luminance ($l$), contrast ($c$), structure ($s$)

$$SSIM(I_x, I_y) = l(I_x, I_y)^{\alpha} \cdot c(I_x, I_y)^{\beta} \cdot s(I_x, I_y)^{\gamma}$$
$$= \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}.$$

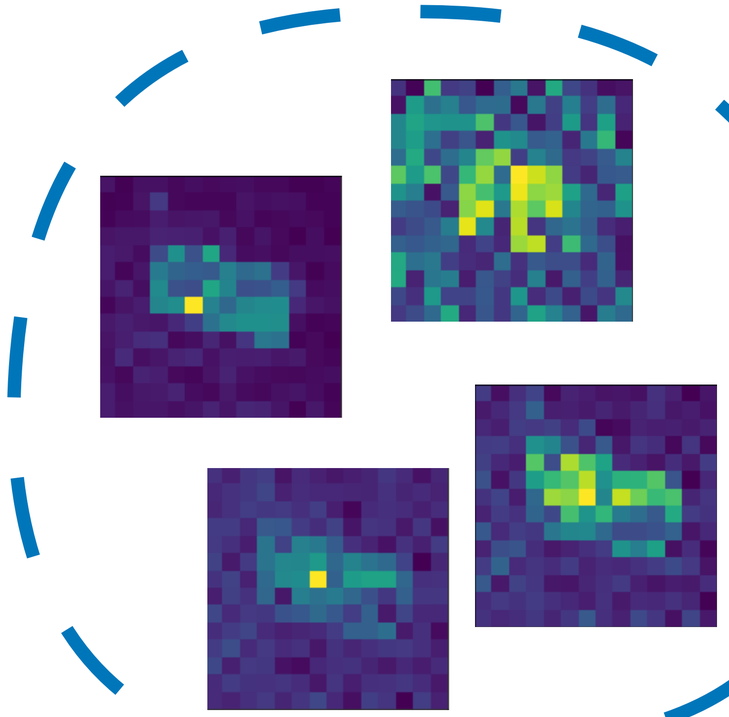
# How to measure inter-head similarity?

AAAI-25 / IAAI-25 / EAAI-25

## Structural Similarity!



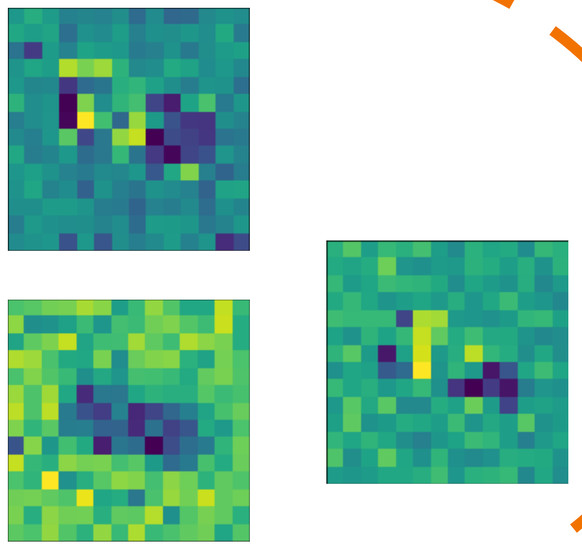
Input



Foreground-focusing attention heads

Similar Structure

Better SSIM!



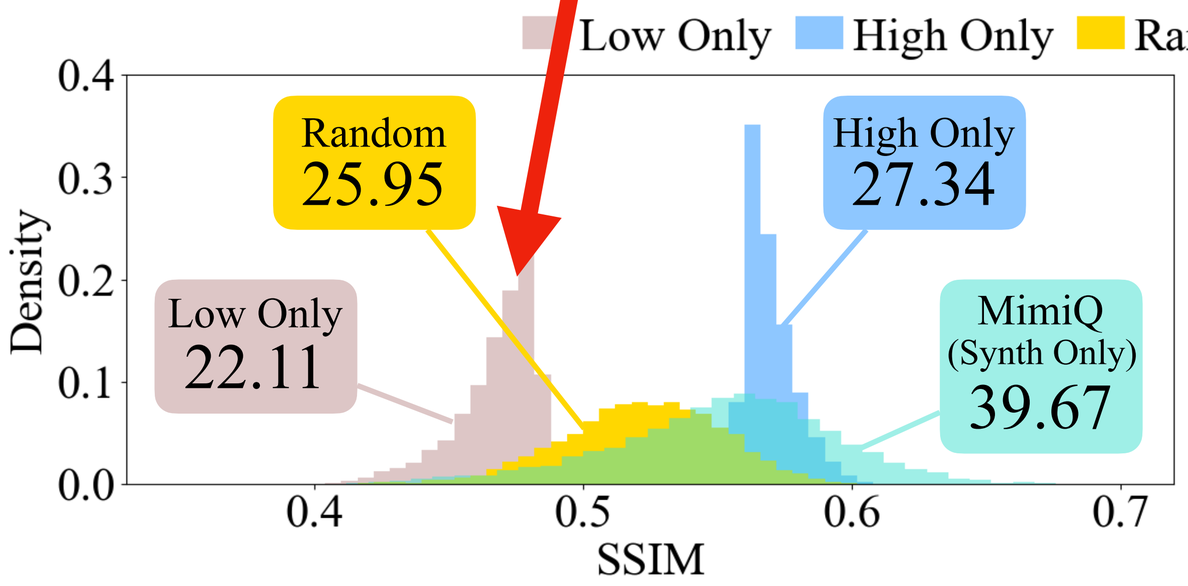
Background-focusing attention heads

# Motivational Study

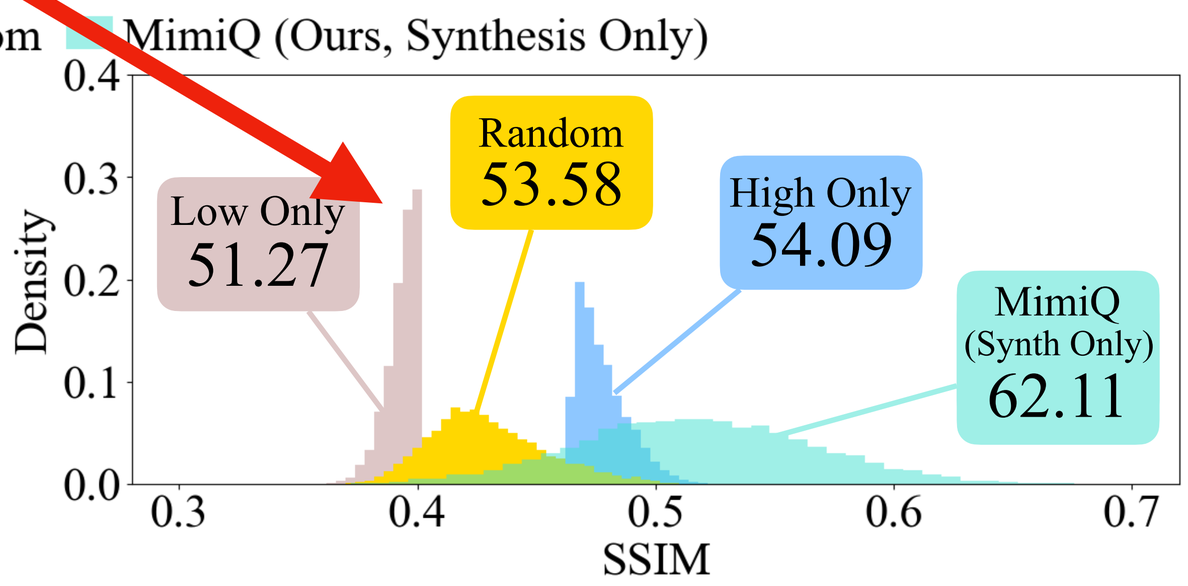
## Inter-head SSIM score of attention and quantization accuracy

AAAI-25 / IAAI-25 / EAAI-25

**Training only with low inter-head SSIM score** → **Low final accuracy**




**ViT-Base**



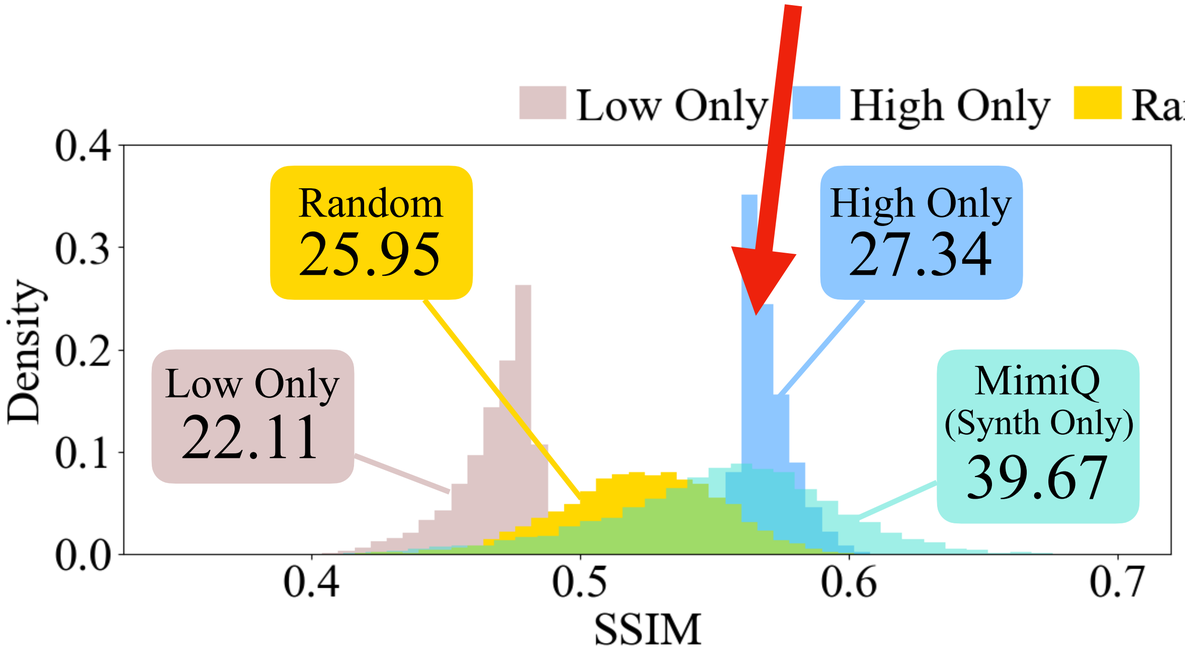

**Swin-Base**

# Motivational Study

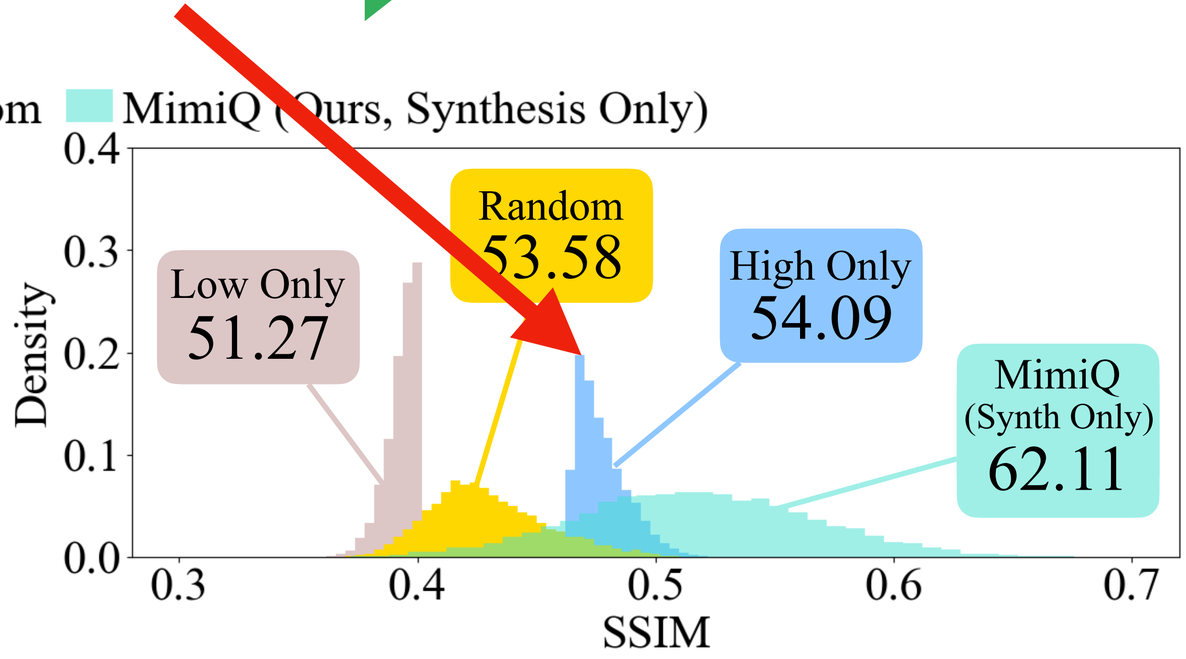
## Inter-head SSIM score of attention and quantization accuracy

AAAI-25 / IAAI-25 / EAAI-25

**Training only with high inter-head SSIM score** → **Higher final accuracy**



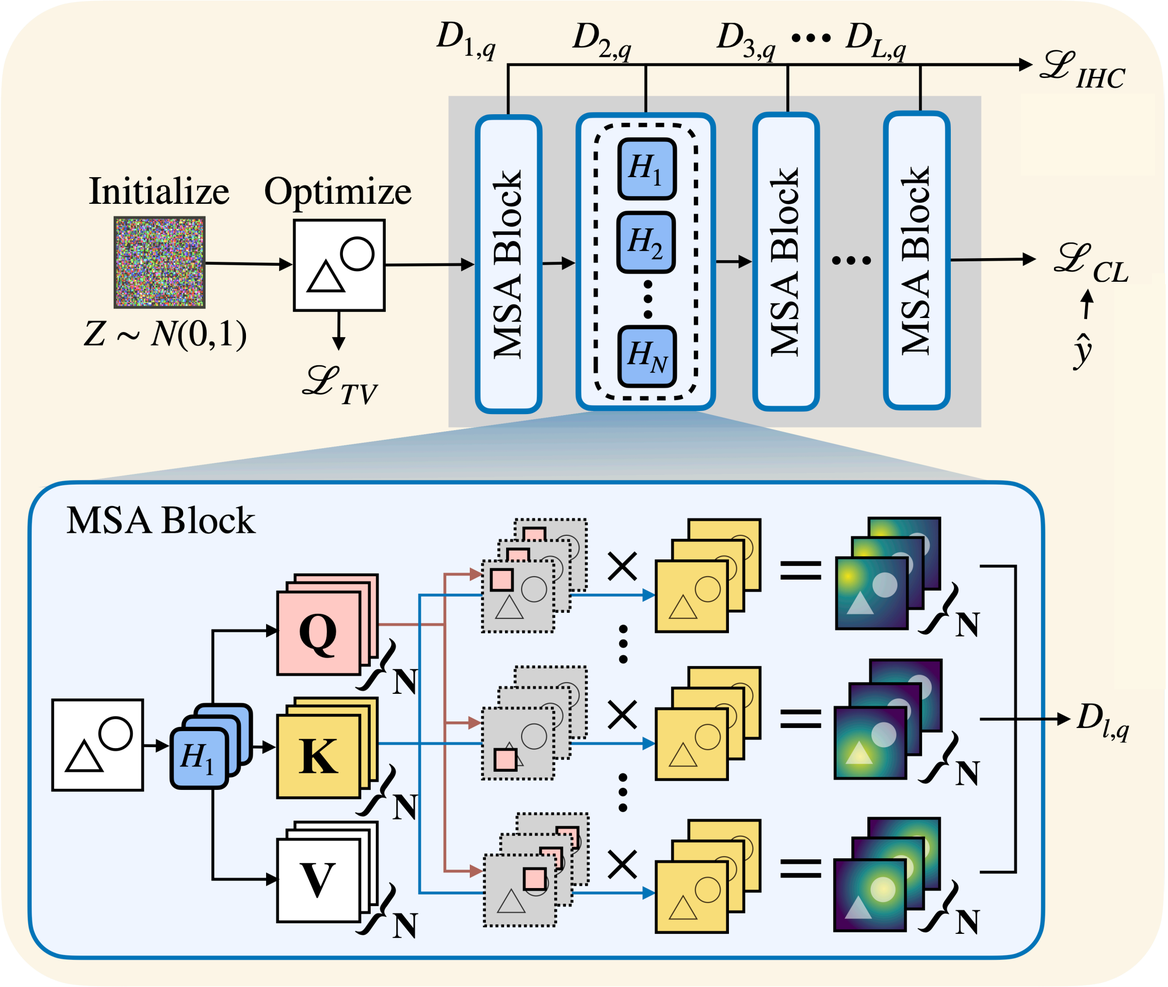

**ViT-Base**




**Swin-Base**

# Generating Synthetic Images

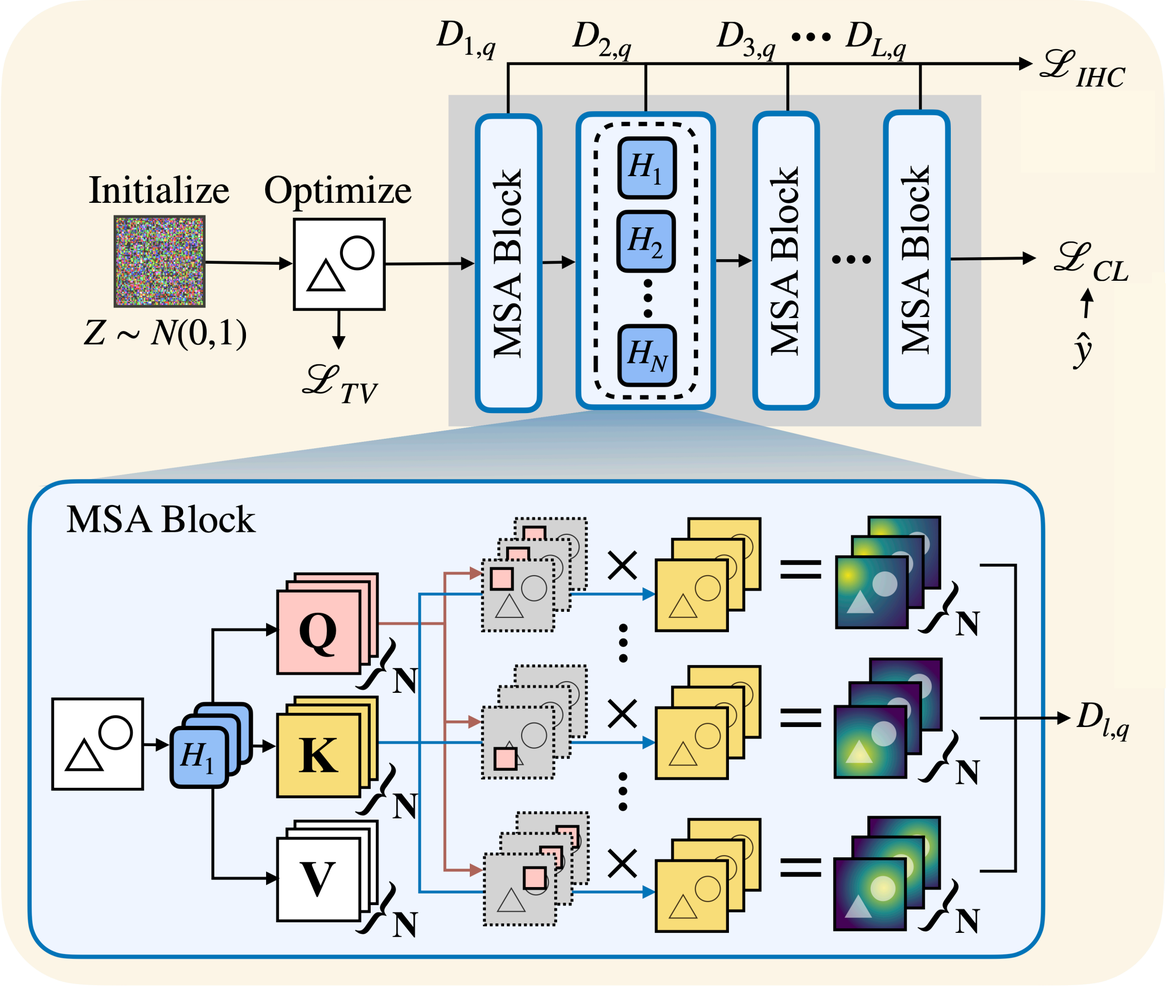





- Generate synthetic images towards high **absolute** SSIM value
  - SSIM > 0 : Positive correlation Foreground<->Foreground, vice versa
  - SSIM < 0 : Negative correlation Foreground<->Background, vice versa
- Both positive and negative correlation is recommended for better visual quality

# Generating Synthetic Images

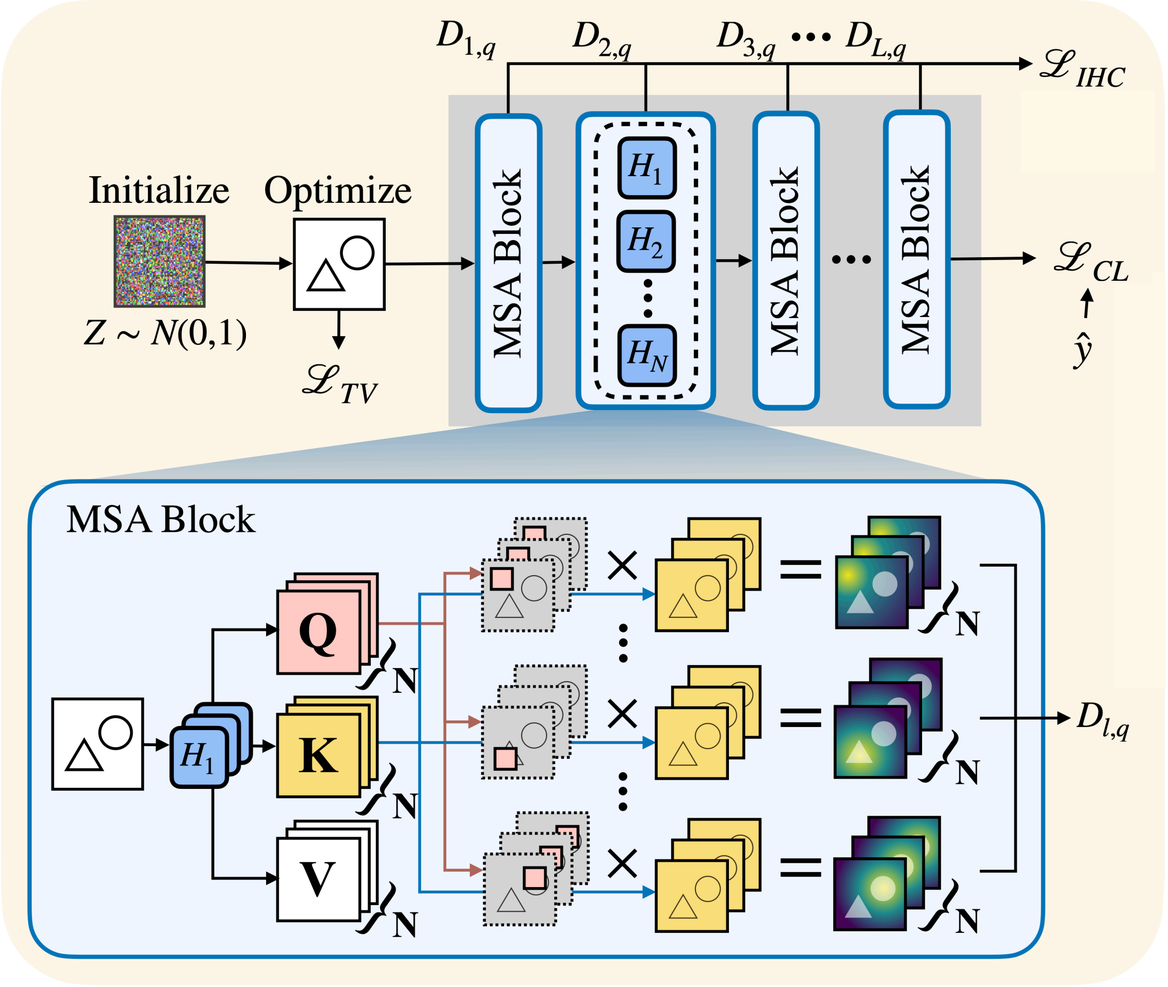


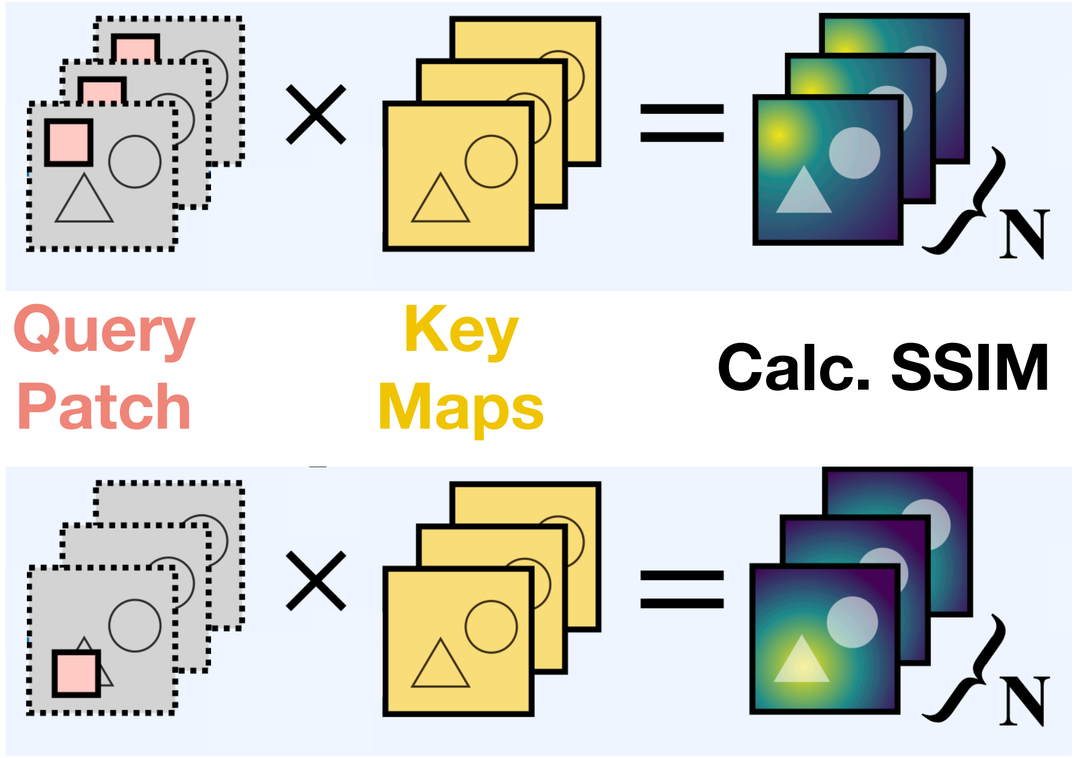



- $\mathscr{L}_{CL}$ : Cross-entropy loss
  - Optimize synthetic image to be classified as a pseudo label $\hat{y}$
- $\mathscr{L}_{TV}$ : Total variance loss
  - Reduces abrupt changes between nearby pixels
- $\mathscr{L}_{IHC}$: Maximize inter-head SSIM score

# Generating Synthetic Images

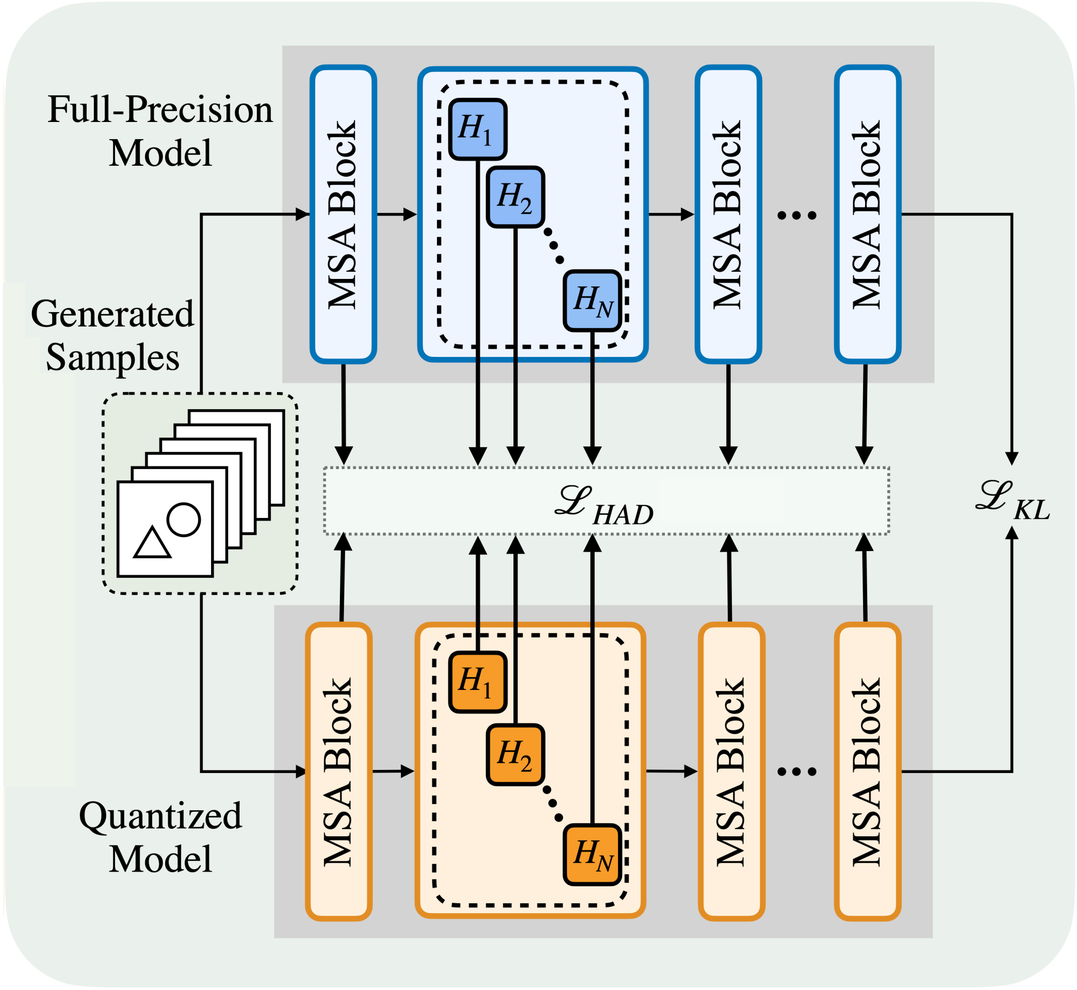




# SSIM-aware Attention Map Distillation



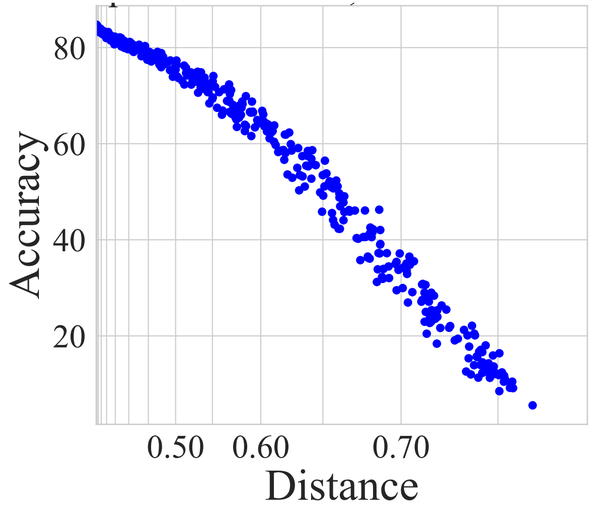



- **Head-wise** attention map distillation
- Reduce distance between attention maps
- How do we measure the distance between attention maps?

# SSIM-aware Attention Map Distillation

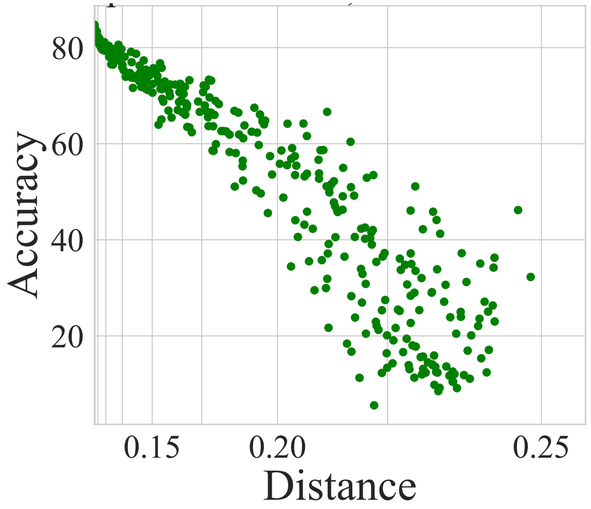
AAAI-25 / IAAI-25 / EAAI-25

## Randomly quantize each head of attention maps and plot distance metric and quantization accuracy



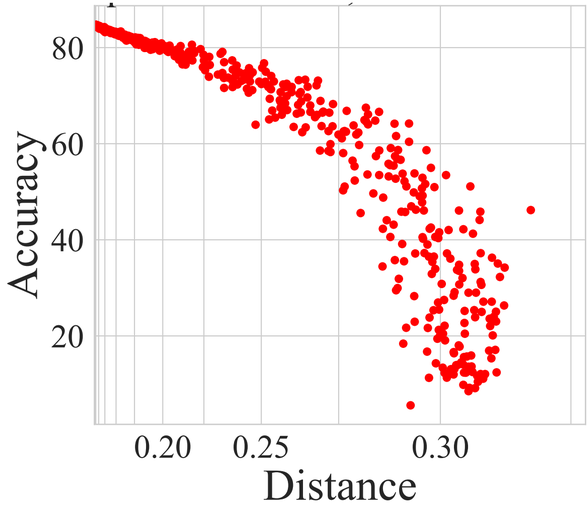

**SSIM**

**Spearman: 0.997**
**Kendall: 0.952**



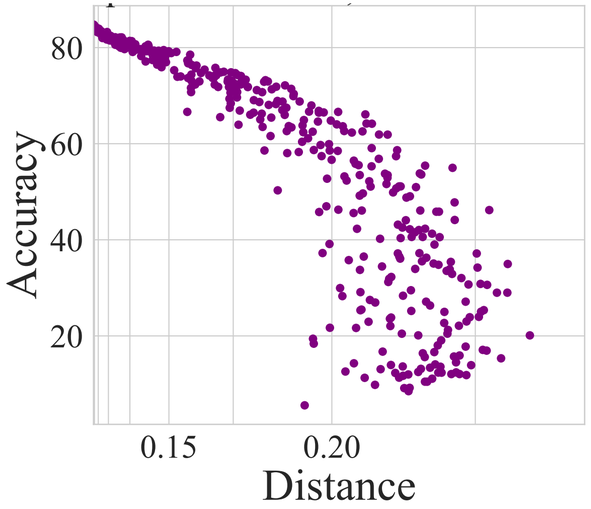

**MSE**

Spearman: 0.982
Kendall: 0.894




**L1-distance**

Spearman: 0.980
Kendall: 0.886

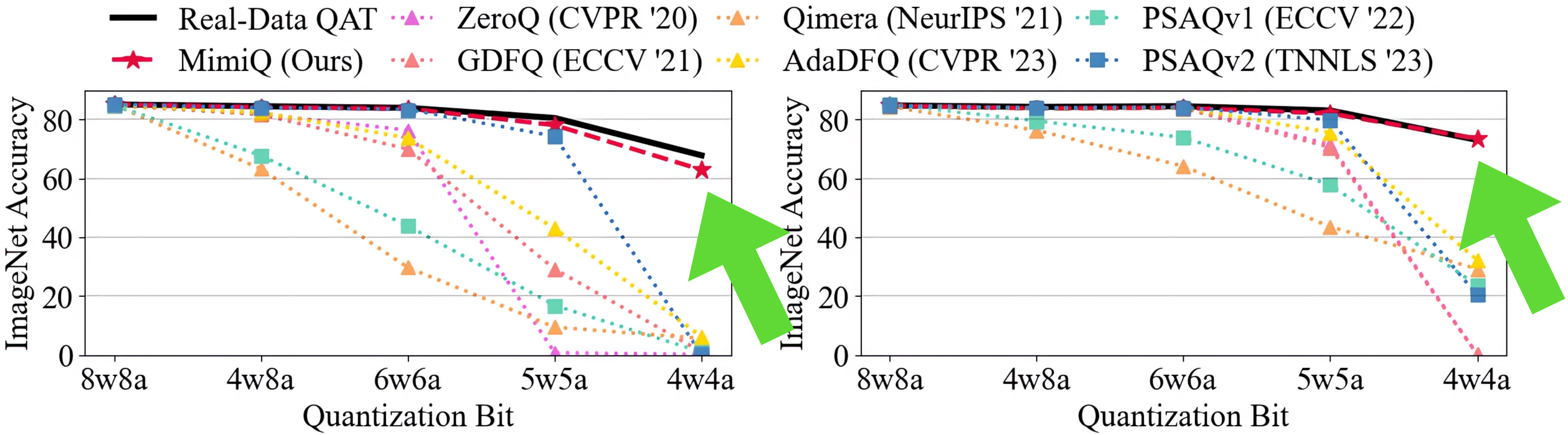



**KL-Divergence**

Spearman: 0.966
Kendall: 0.857

# Experimental Result: Accuracy






**ViT-Base** **Swin-Base**

**Preserve high accuracy in low-bit settings**

# Experimental Result: Accuracy

AAAI-25 / IAAI-25 / EAAI-25

| Bits | Methods | Target Arch. | Networks | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | ViT-T | ViT-S | ViT-B | DeiT-T | DeiT-S | DeiT-B | Swin-T | Swin-S | Swin-B |
| | Real-Data FT | - | 58.17 | 67.21 | 67.81 | 57.98 | 62.15 | 64.96 | 73.08 | 76.34 | 73.06 |
| W4/A4 | GDFQ | CNN | 2.95 | 4.62 | 11.73 | 25.96 | 22.12 | 30.04 | 42.08 | 41.93 | 36.04 |
| | Qimera | CNN | 0.57 | 7.02 | 5.61 | 15.18 | 11.37 | 32.49 | 47.98 | 39.64 | 29.27 |
| | AdaDFQ | CNN | 2.00 | 1.78 | 6.21 | 19.57 | 14.44 | 19.22 | 38.88 | 39.40 | 32.26 |
| | PSAQ-ViT | ViT | 0.67 | 0.15 | 0.94 | 19.61 | 5.90 | 8.74 | 22.71 | 9.26 | 23.69 |
| | PSAQ-ViT V2 | ViT | 1.54 | 4.14 | 2.83 | 22.82 | 32.57 | 45.81 | 50.42 | 39.10 | 39.26 |
| | MimiQ (Ours) | ViT | **42.99** | **55.69** | **62.91** | **52.03** | **62.72** | **74.10** | **69.33** | **70.46** | **73.49** |
| | Acc. Gain | | +40.04 | +48.68 | +51.18 | +26.07 | +30.15 | +28.28 | +18.91 | +28.53 | +34.23 |

**Accuracy gain up to 51.18%p**

# Analysis: Computational Costs

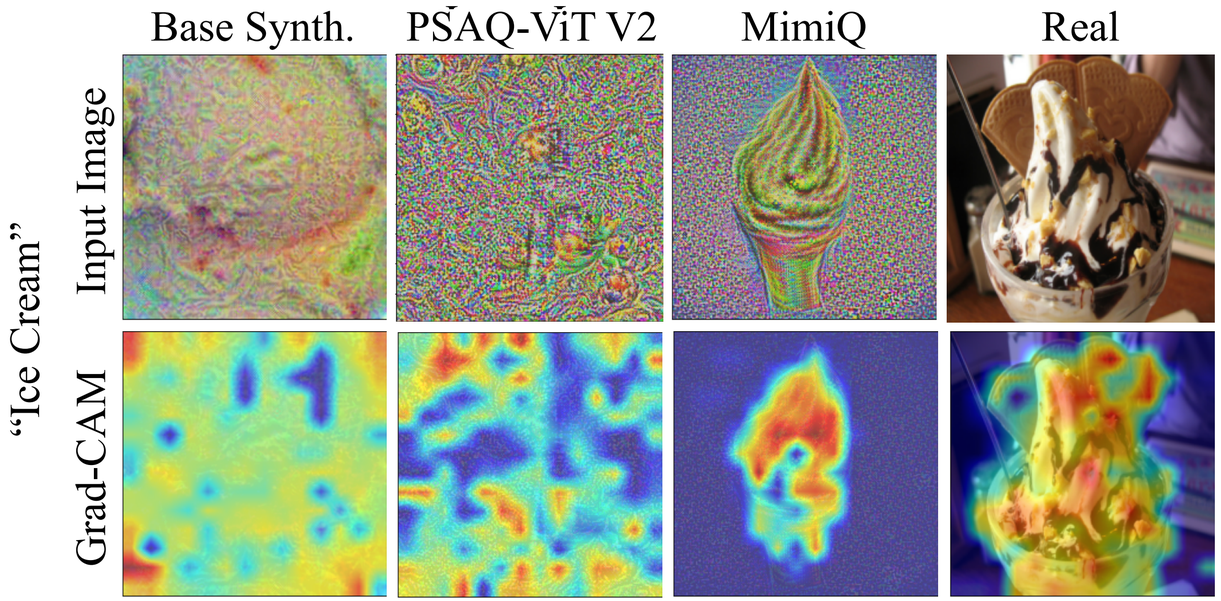


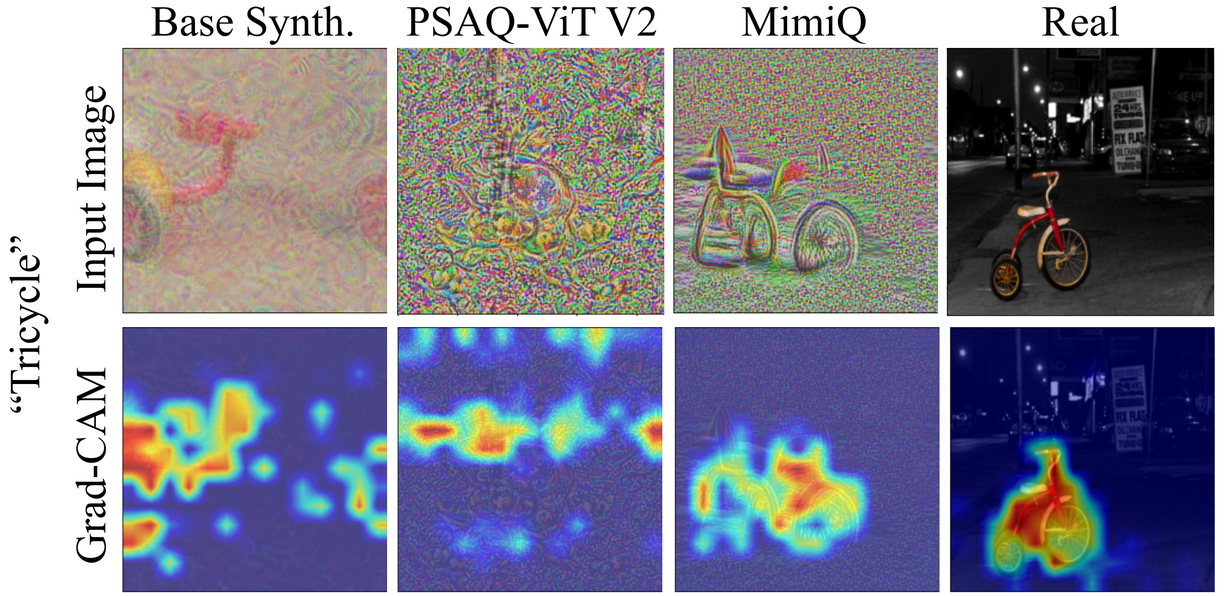
| Method | Type | Synth. | Quant. | Total | Acc. |
|---|---|---|---|---|---|
| GDFQ | QAT | - | 10.70h | 10.70h | 11.73 |
| AdaDFQ | QAT | - | 8.44h | 8.44h | 6.21 |
| PSAQ-V1 | PTQ | 0.11h | 0.0002h | 0.11h | 0.94 |
| PSAQ-V2 | QAT | - | 4.55h | 4.55h | 2.83 |
| MimiQ-1k | QAT | 1.98h | 2.39h | 4.37h | **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 requires synthetic data generation and quantization-aware training
- With similar costs, MimiQ shows superior accuracy
- More synthetic data shows further accuracy boost

# Analysis: GRAD-CAM





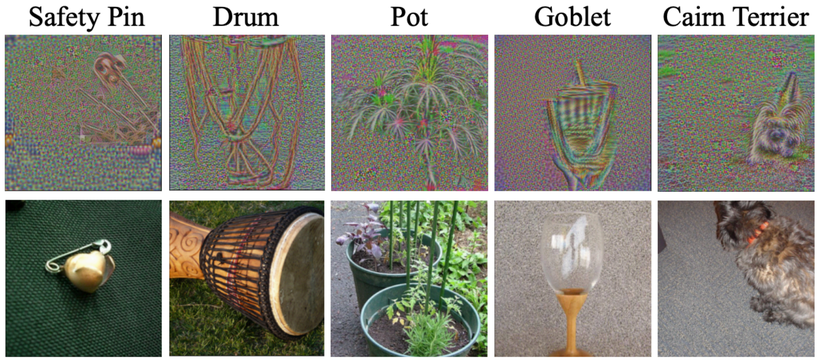




**Synthetic texture on the center**
**GRAD-CAM shows aligned attention on the desired texture information**

# Societal Concerns



## Input Reconstruction Attack




Most similar pairs of
synthetic and real images

Synthetic dataset
does not restore **exact** images
from the training set

## Model Inversion Attack

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

Adaptability of synthetic
data training to attacks

Synthetic and real samples
are **distinguishable** and has
**low transferability**

# Conclusion

AAAI-25 / IAAI-25 / EAAI-25

- Data-free quantization aims to tackle the scenario that the original dataset is inaccessible due to privacy concerns, the large dataset size, and copyright issues.
- We propose MimiQ, the first work to consider inter-head attention similarity with SSIM for synthetic sample generation and quantization-aware training.
- We show attention similarity of ViT models plays a crucial role in model training, affecting final accuracy