BEYOND A SINGLE DETERMINISTIC TEACHER: INVARIANT SYNTHESIS FOR DATA-FREE QUANTIZATION OF VISION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Data-Free Quantization (DFQ) of Vision Transformers (ViTs) enables post-training quantization without real calibration data by synthesizing surrogate calibration samples from a pre-trained full-precision teacher model. However, existing DFQ methods typically optimize synthetic samples against a single *deterministic* teacher, which overlooks their *sensitivity* to internal *structural variations*, such as attention-head responses and spatial token dependencies. As a result, synthesized samples may exhibit structural brittleness, leading to *unreliable calibration statistics* and degraded quantization performance. To address these issues, we revisit data-free synthesis for ViTs from the perspective of *invariant synthesis* over structurally perturbed teachers. We theoretically show that: 1) structural sensitivity of synthetic samples contributes to the discrepancy between real and synthetic calibration statistics; and 2) structurally perturbed teachers can expose response variations hidden by deterministic supervision. Motivated by these insights, we propose **DropIS**, a novel and effective **Drop**out-Induced **I**nvariant **S**ynthesis approach for DFQ of ViTs. Unlike conventional dropout regularization, DropIS employs structured dropout as a teacher probing mechanism during synthesis: *attention head dropout* suppresses excessive reliance on specific head responses, while *spatial token dropout* mitigates dependence on narrow local token regions. Furthermore, DropIS incorporates auxiliary attention-space regularizers to stabilize perturbed responses and synthesize semantically consistent calibration samples. Extensive experiments across diverse backbones and downstream tasks verify the advantages of DropIS over the state-of-the-art methods, achieving up to 8.46% Top-1 accuracy improvements on ImageNet under challenging low-bit regimes. *Our code is available in supplementary material package*.

1 INTRODUCTION

Vision Transformers (ViTs) Dosovitskiy (2020); Han et al. (2022a); Khan et al. (2022) have achieved remarkable success across diverse vision tasks Strudel et al. (2021); Zhang et al. (2021); Ma et al. (2023), owing to their ability to model long-range dependencies through self-attention Vaswani et al. (2017); Dosovitskiy (2020); Touvron et al. (2021); Liu et al. (2021b). Despite their effectiveness, deploying ViTs on resource-constrained devices remains challenging due to their substantial computational and memory overhead Lin et al. (2021). Post-training quantization (PTQ) Li et al.; Wei et al. (2022) provides a practical compression solution by reducing numerical precision without expensive retraining. However, PTQ generally relies on representative calibration data to estimate activation distributions and determine reliable quantization parameters Liu et al. (2021c); Hubara et al. (2018). In many practical scenarios (*e.g.*, social and medical domains), such calibration data may be inaccessible due to privacy or security restrictions Liu et al. (2021a).

As a compelling alternative, Data-Free Quantization (DFQ) Cai et al. (2020); Xu et al. (2020); Choi et al. (2021); Zhong et al. (2022); Qian et al. (2023b;a); Dung et al. (2024); Kim et al. (2025) is further proposed to synthesize surrogate calibration data from a pre-trained full-precision model P. However, quantizing ViTs remains particularly challenging compared with convolutional neural networks (CNNs), especially in low-bit regimes. Unlike CNNs, whose representations are primarily constructed through local convolutional operations, ViTs rely on *delicate interactions* among attention heads, token embeddings, and global token-to-token relations. As a result, quantization perturbations

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

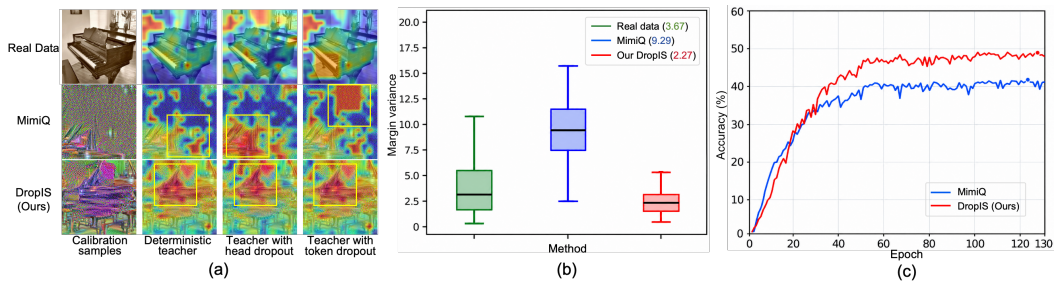


Figure 1: Motivation experiments beyond the deterministic teacher. (a) Synthetic samples optimized by a deterministic teacher may exhibit unstable attention responses under head or token perturbations. (b) Margin variance under perturbed teachers, where lower variance indicates stronger structural invariance (*Please refer to more details in Appendix.A.4*). (c) Our DropIS synthesizes semantically stable and transferable calibration samples compared to single-teacher synthesis methods (*e.g.*, MimiQ Choi et al. (2025)), yielding better quantization performance.

may disrupt the internal attention-token pathway and cause severe accuracy degradation Yuan et al. (2022); Li et al. (2023c); Wu et al. (2025). Recent efforts Li et al. (2022; 2023b); Ramachandran et al. (2024); Choi et al. (2025) for ViTs exploit different priors embedded in pre-trained transformers. PSAQ-ViT Li et al. (2022) and PSAQ-ViT V2 Li et al. (2023b) leverage patch similarity to separate foreground regions from background regions. CLAMP-ViT Ramachandran et al. (2024) further introduces patch-level contrastive learning to capture meaningful inter-patch relationships. More recently, MimiQ Choi et al. (2025) exploits inter-head attention similarity to improve global structural consistency. Although these methods have advanced DFQ for ViTs, they commonly synthesize samples by optimizing them against a *single deterministic* full-precision teacher, leaving the structural sensitivity of synthetic samples largely uncharacterized, hence suffering from large performance margin (see Fig.2(a) and Fig.1(c)).

This single-teacher synthesis paradigm implicitly assumes that a sample eliciting a confident prediction from P can serve as reliable calibration data for Q, however, is too restrictive for ViTs. Since ViT predictions are supported by *structured interactions* among multiple attention heads and spatial tokens, a high-confidence synthetic sample may still rely on a *narrow* and *brittle* internal pathway, such as a few dominant attention heads or a small local token region. Such calibration samples can be highly discriminative for the deterministic teacher, yet exhibit unstable internal responses, yielding brittle calibration statistics; see Fig.1(a). To understand this issue, we identify two primary limitations of existing DFQ methods for ViTs: 1) *structural brittleness*: synthesized samples may be highly confident on the deterministic teacher P, but their responses fluctuate significantly under moderate structural probes (*e.g.*, dropout), indicating sensitivity to internal structural variations; see Fig.1(b); 2) *limited transferability*: synthetic samples tailored only to P may yield unstable calibration statistics, increasing the real-synthetic calibration mismatch and reducing their reliability for calibrating Q; see Fig.1(a)(c).

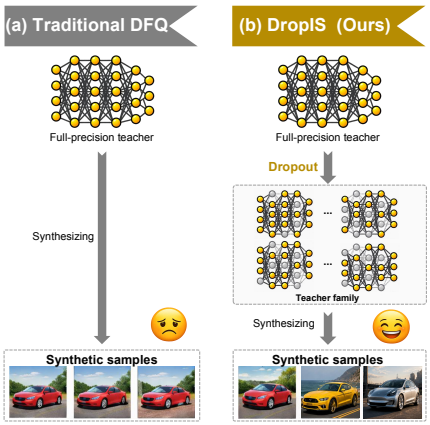


Figure 2: (a) Traditional DFQ arts against a deterministic teacher. (b) DropIS formulates invariant synthesis over structurally probed teacher family, yielding semantically stable and diverse synthetic samples.

To address the above issues, we revisit data-free synthesis for ViTs as an invariant optimization problem over a family of structurally probed teachers, and propose a novel and effective **Drop**out-Induced **I**nvariant **S**ynthesis approach for DFQ of ViTs, dubbed **DropIS**, as depicted in Fig.2(b) and Fig.3. Our motivation experiments and theoretical analysis *disclosing* that: 1) controlling the structural sensitivity of synthetic samples reduces one source of discrepancy between synthetic and real calibration statistics; 2) structural perturbations during synthesis enriches the supervision beyond the unperturbed teacher response, encouraging samples that are less tied to a particular internal computation path. Guided by these insights, beyond a single deterministic teacher, DropIS instantiates the general structural-probe family with two

structured dropout mechanisms: first, *attention head dropout* randomly suppresses entire attention heads to reduce dependence on specific head responses; second, *spatial token dropout* removes a contiguous block of patch tokens to discourage dependence on narrow local token regions. To stabilize the perturbed synthesis, we further endow DropIS with auxiliary regularizers on dropout-perturbed attention responses, which serve as tractable surrogates for structural stability, together with the task objective over structurally probed teachers, thereby encouraging *structurally stable* and *transferable* calibration samples.

We theoretically and empirically validate the effectiveness of DropIS, demonstrating substantial improvements over state-of-the-art methods across diverse ViT backbones and downstream tasks (*i.e.*, classification, detection and segmentation). For example, under challenging low-bit regimes on ImageNet, DropIS achieves up to 8.46% Top-1 accuracy gains over existing arts.

2 Invariant Synthesis for Data-Free Quantization of ViTs

2.1 Preliminaries: Data-Free Quantization of ViTs

Notation. For a Vision Transformer Dosovitskiy (2020); Khan et al. (2022), let $\mathbf{A}^{(l)} \in \mathbb{R}^{B\times H_l\times T\times T}$ denote the attention tensor of the l-th transformer layer, where B, H_l, and T are the batch size, number of heads, and token length, respectively. Let $\mathbf{X} \in \mathbb{R}^{B\times T\times D}$ be the token sequence after patch embedding, where D is the embedding dimension. For ViT-style architectures, the first token is the class token $\mathbf{x}_{\text{cls}}$, and the remaining patch tokens are denoted by $\mathbf{X}_p \in \mathbb{R}^{B\times N\times D}$ with $N = H \times W$.

Data-free quantization (DFQ) quantizes a pre-trained full-precision model P into a low-bit model Q without the original training data Xu et al. (2020); Choi et al. (2021); Qian et al. (2023a); Dung et al. (2024); Kim et al. (2025). A synthetic calibration set $\mathcal{D}_{\text{syn}}$ is therefore optimized from random noise under the supervision of P; below, we denote its induced synthesis distribution by $\mathcal{S}_\theta$. Given an input image x and pseudo-label y, P produces logits $f(x) \in \mathbb{R}^C$, where C is the number of classes. Standard data-free synthesis optimizes x with a classification loss and a total variation (TV) loss:

$$\mathcal{L}_{\text{cls}} = \text{CE}\big(f(x), y\big), \quad \mathcal{L}_{\text{tv}} = \|x_{:,:,1:,:} - x_{:,:,:-1,:}\|_2^2 + \|x_{:,:,:,1:} - x_{:,:,:,:-1}\|_2^2, \tag{1}$$

where $\text{CE}(\cdot,\cdot)$ denotes the cross-entropy loss and $\mathcal{L}_{\text{tv}}$ suppresses high-frequency artifacts. To stabilize synthesis under structured perturbations, we further adopt two attention-space regularizers as auxiliary constraints Li et al. (2022); Choi et al. (2025). Given the attention responses $\{\mathbf{A}^{(l)}\}_{l=1}^{L}$ of the current probed teacher, the SSIM-based structural regularizer Wang et al. (2004) encourages structurally coherent attention maps below:

$$\mathcal{L}_{\text{h}} = \frac{1}{L}\sum_{l=1}^{L}(1 - \text{SSIM}(\mathbf{A}^{(l)})^2). \tag{2}$$

In contrast, the similarity-entropy regularizer compute pairwise similarities between normalized attention vectors. Let $\mathbf{S}_l$ denote the flattened similarity vector at layer l and define the differential entropy surrogate as $\mathcal{H}_l = \frac{1}{2}\log(2\pi e\text{Var}(\mathbf{S}_l))$ Cover (1999); Shannon (1948), where $2\pi e$ is a constant, thus the similarity-entropy regularizer is formulated as

$$\mathcal{L}_{\text{s}} = -\frac{1}{L}\sum_{l=1}^{L}\mathcal{H}_l. \tag{3}$$

It is worth noting that $\mathcal{L}_{\text{h}}$ preserves the structural coherence of attention maps, while $\mathcal{L}_{\text{s}}$ promotes diverse attention similarity patterns and prevents token-relation collapse.

2.2 Theoretical Investigation: From Deterministic Supervision to Structural Stability

Prior studies have shown that ViT quantization is sensitive to the special distributions of activations and attention responses Yuan et al. (2022); Lin et al. (2021); Li et al. (2023c). We formalize how structural instability of synthetic samples contributes to calibration-statistic mismatch (Sec.2.2.1), and explain why supervision over structural probes captures response variations that are not characterized by a single deterministic teacher (Sec.2.2.2), serving as the theoretical bridge from the motivation in Sec.1 to the invariant synthesis (Sec.2.3).

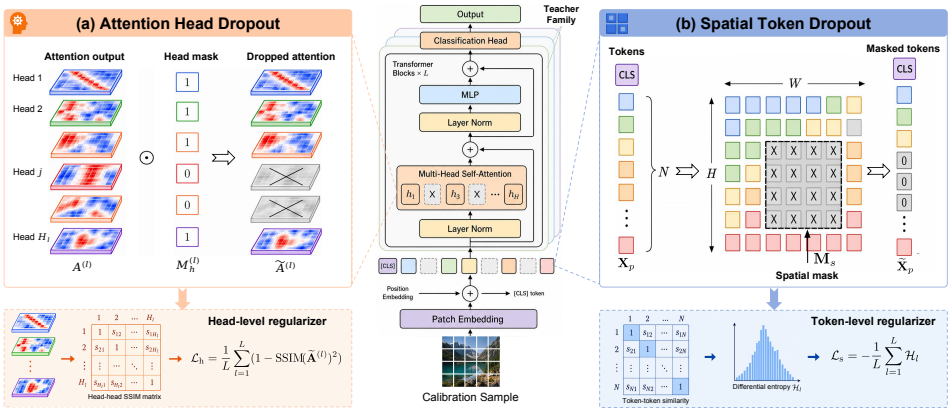


Figure 3: Illustration of our proposed DropIS. DropIS instantiates invariant synthesis over the general structural-probe family with complementary structured perturbations: (a) *attention head dropout* randomly suppresses entire attention heads, which prevents synthetic samples from over-relying on a small subset of dominant heads; (b) *spatial token dropout* removes a contiguous block of patch tokens, which prevents synthetic samples from depending excessively on narrow local regions.

2.2.1 Structural Stability as a Calibration Principle

Data-free quantization relies on synthetic samples to approximate real-data calibration statistics. However, deterministic teacher supervision does not explicitly characterize the sensitivity of synthetic samples to internal structural perturbations, which may lead to unstable calibration statistics. We therefore introduce a general family of structural probes to discuss structural stability. Formally, let $\xi \in \Xi$ denote a structural probe sampled from a prescribed distribution $q(\xi)$, and let ξ_0 be the identity probe such that $f(x;\xi_0) = f(x)$. The probe family is restricted to moderate internal interventions that are not intended to alter the task semantics of the input. For layer l, let $s_l(x;\xi)$ denote the sample-level calibration feature extracted from the teacher under probe ξ, such as activation moments or histogram/bin statistics used by the calibration rule. To characterize the impact of structural perturbations, we consider the expected feature $\bar{s}_l(x)$ of sample x under *structural probes* and the calibration statistic $T_{\mathcal{P}}^{(l)}$ collected from the *unperturbed* teacher over distribution $\mathcal{P}$, defined as

$$\bar{s}_l(x) := \mathbb{E}_{\xi\sim q}[s_l(x;\xi)], \qquad T_{\mathcal{P}}^{(l)} := \mathbb{E}_{x\sim\mathcal{P}}[s_l(x;\xi_0)], \tag{4}$$

where $\mathcal{P}$ denotes either the inaccessible real-data distribution $\mathcal{D}$ or the synthesized calibration distribution $\mathcal{S}_\theta$. Based on that, we quantify the structural sensitivity of calibration statistics by measuring the deviation between perturbed and unperturbed teacher responses:

$$I_{\mathcal{P}}^{(l)} := \mathbb{E}_{x\sim\mathcal{P}}\mathbb{E}_{\xi\sim q}[\|s_l(x;\xi) - s_l(x;\xi_0)\|_2], \tag{5}$$

where a smaller $I_{\mathcal{P}}^{(l)}$ indicates that the calibration statistic is more stable under structural probes. Furthermore, we define the remaining discrepancy between real and synthetic data after averaging over structural probes as

$$\Delta_{\text{str}}^{(l)}(\mathcal{D},\mathcal{S}_\theta) := \|\mathbb{E}_{\mathcal{D}}[\bar{s}_l(x)] - \mathbb{E}_{\mathcal{S}_\theta}[\bar{s}_l(x)]\|_2 . \tag{6}$$

The above allow us to decompose the real-synthetic calibration-statistic gap into three interpretable components: the intrinsic structural sensitivity of real data $I_{\mathcal{D}}^{(l)}$, the residual discrepancy between real and synthetic probe-averaged statistics $\Delta_{\text{str}}^{(l)}(\mathcal{D},\mathcal{S}_\theta)$, and the structural sensitivity of synthetic samples $I_{\mathcal{S}_\theta}^{(l)}$, leading to the following structural calibration-statistic bound.

Proposition 2.1 (Structural calibration-statistic decomposition)**.** *For every layer l,*

$$\|T_{\mathcal{D}}^{(l)} - T_{\mathcal{S}_\theta}^{(l)}\|_2 \le I_{\mathcal{D}}^{(l)} + \Delta_{\text{str}}^{(l)}(\mathcal{D},\mathcal{S}_\theta) + I_{\mathcal{S}_\theta}^{(l)}. \tag{7}$$

The proof is in Appendix A.2.1. Proposition 2.1 implies that the structural sensitivity of synthetic samples constitutes an explicit source of real-synthetic calibration mismatch. Therefore, when the probe family and other discrepancy terms remain comparable, *reducing $I_{\mathcal{S}_\theta}^{(l)}$ can tighten the upper bound on the calibration-statistic gap*, motivating invariant synthesis beyond deterministic teacher fitting. Under a mild stability assumption of the calibration procedure, the above analysis can be extended to characterize its impact on the calibrated quantized model (refer to Appendix A.2.2).

2.2.2 Structural Probes as Additional Supervision

We next investigate why supervision from structurally perturbed teachers provides additional information beyond a single deterministic teacher. Given teacher logits $f(x;\xi) \in \mathbb{R}^C$ under structural probe ξ, we define the smooth classification margin for a sample-label pair (x, y) as

$$\Gamma_\xi(x, y) := f_y(x;\xi) - \log \sum_{c \neq y} \exp(f_c(x;\xi)), \tag{8}$$

where a larger margin indicates stronger preference for the target class. A deterministic teacher only observes the identity response $\gamma(x, y) := \Gamma_{\xi_0}(x, y)$, yet it fails to reveal how the prediction changes when the internal structure of the teacher is perturbed. Therefore, we characterize such structural response variation by

$$V_{\mathrm{str}}(x, y) := \mathrm{Var}_{\xi\sim q}[\Gamma_\xi(x, y)], \tag{9}$$

where larger V_{str} indicates stronger sensitivity of the prediction margin to structural perturbations. Without additional assumptions, the deterministic margin $\gamma(x, y)$ cannot determine $V_{\mathrm{str}}(x, y)$, hence samples with similar confidence may exhibit different responses under structural perturbations. This motivates optimizing synthesis over a family of structurally perturbed teachers rather than relying only on the identity teacher. Specifically, we consider the structural-probe classification objective as

$$\mathcal{R}_{\mathrm{str}}(x, y) := \mathbb{E}_{\xi\sim q}[\ell(\Gamma_\xi(x, y))], \qquad \ell(u) := \log(1 + e^{-u}), \tag{10}$$

where $\ell(\cdot)$ is the classification loss induced by the margin. Let $\overline{\Gamma} := \mathbb{E}_{\xi\sim q}[\Gamma_\xi]$ denote the average margin over structural probes. Unlike deterministic supervision, $\mathcal{R}_{\mathrm{str}}$ depends on the full distribution of probed margins rather than only their mean. Therefore, The difference between $\mathcal{R}_{\mathrm{str}}$ and $\ell(\overline{\Gamma})$ captures the additional information introduced by structural response variations. Assuming the probed margins lie in a bounded interval $[a, b]$, we define

$$\mu_\ell := \min_{u\in[a,b]} \ell''(u) > 0, \qquad L_\ell := \max_{u\in[a,b]} \ell''(u). \tag{11}$$

Since the classification loss is smooth and convex, the above difference can be bounded by the variance of the probed margins, leading to the following proposition.

Proposition 2.2 (Variance-sensitive structural Jensen gap)**.** *For any structural-probe distribution q satisfying the bounded-margin condition, we obtain*

$$\frac{\mu_\ell}{2} V_{\mathrm{str}} \leq \mathcal{R}_{\mathrm{str}} - \ell(\overline{\Gamma}) \leq \frac{L_\ell}{2} V_{\mathrm{str}}. \tag{12}$$

The proof is in Appendix A.2.4. Proposition 2.2 implies that structural-probe supervision introduces a Jensen gap $\mathcal{R}_{\mathrm{str}} - \ell(\overline{\Gamma})$, bounded by the variance of prediction margins under structural perturbations, which allows the synthesis objective to account for prediction variations under internal perturbations. Accordingly, DropIS optimizes synthetic samples beyond a single deterministic response by enforcing consistency across structurally perturbed teachers.

2.3 DropIS: Dropout-Induced Invariant Calibration Sample Synthesis

Motivated by the above analysis, DropIS treats structural dropout as a teacher probing mechanism rather than conventional dropout regularization. Calibration samples are synthesized by optimizing the expected response over structurally perturbed teachers:

$$\min_x \ \mathcal{J}_{\mathrm{str}}(x, y), \qquad \mathcal{J}_{\mathrm{str}}(x, y) := \mathbb{E}_{\xi\sim q}[\mathcal{L}_{\mathrm{cls}}(x, y;\xi)], \tag{13}$$

where ξ denotes a sampled structural perturbation applied to the pre-trained teacher, and $\mathcal{L}_{\mathrm{cls}}$ represents the corresponding classification objective. For ViTs, DropIS instantiates the general structural probe along two complementary dimensions:

$$\xi = (\xi_h, \xi_s), \qquad q(\xi) = q_h(\xi_h) q_s(\xi_s), \tag{14}$$

where ξ_h and ξ_s correspond to attention-head and spatial-token perturbations, implemented by attention head dropout (2.3.1) and spatial-token dropout (Sec.2.3.2), respectively. The former reduces reliance on specific head responses, while the latter discourages reliance on localized token regions, providing complementary structural views of the same synthetic sample. In practice, the expectation

over structural probes in Eq.(13) is approximated using a Monte-Carlo estimator with K independently sampled structural realizations:

$$\widehat{\mathcal{J}}_{\mathrm{str}}(x,y) = \frac{1}{K}\sum_{k=1}^{K} \mathcal{L}_{\mathrm{cls}}^{(k)}(x,y;\xi^{(k)}), \qquad \xi^{(k)} = (\xi_h^{(k)}, \xi_s^{(k)}), \tag{15}$$

where each $\xi^{(k)}$ corresponds to a different realization of the structural probe distribution. The dropout masks are resampled during synthesis, allowing the same synthetic sample to be optimized under diverse structural perturbations rather than a single deterministic or fixed perturbed teacher. The two-factor variance decomposition of the structural probes and the corresponding Monte-Carlo estimation analysis are provided in Appendix A.2.5 and Appendix A.2.6.

2.3.1 Head-Wise Structural Probing via Attention Dropout

Inspired by structural variations in ViT attention Li et al. (2023a), we instantiate ξ_h by dropout on the head dimension. Unlike conventional dropout used for training regularization, attention-head dropout in DropIS probes whether synthetic samples rely excessively on individual attention heads. For the l-th transformer layer with H_l attention heads, we sample a head-wise mask $\mathbf{M}_h^{(l)} \in \mathbb{R}^{1\times H_l\times 1\times 1}$ as $\mathbf{M}_h^{(l)}(j) = \frac{z_j^{(l)}}{1-p_h^{(l)}+\epsilon}$, $z_j^{(l)} \sim \mathrm{Bernoulli}(1-p_h^{(l)})$, where $p_h^{(l)}$ denotes the head dropout rate and $\epsilon > 0$ is a small constant for numerical stability. The inverted scaling approximately preserves the expected magnitude of each head response at the perturbation site. Given the attention interaction tensor $\mathbf{A}^{(l)}$, the perturbed attention response is

$$\widetilde{\mathbf{A}}^{(l)} = \mathbf{A}^{(l)} \odot \mathbf{M}_h^{(l)}. \tag{16}$$

Samples that are sensitive to specific head responses experience larger variations under sampled masks, encouraging more distributed head-level evidence.

2.3.2 Spatial-Token Structural Probing via Token Dropout

Since ViTs can tolerate structured token removal Liu et al. (2023); Rao et al. (2021), we next instantiate ξ_s by dropout on the spatial token dimension. Different from these works that mainly use token removal for training regularization or inference efficiency, we adopt spatial token dropout as a structural probe to sample perturbed teachers during synthesis and reveal sensitivity to localized token regions. Instead of applying i.i.d. token dropout, we adopt block-wise spatial token dropout, since spatial token perturbations provide a direct way to evaluate and reduce sensitivity to localized token regions. Specifically, patch tokens are reshaped into a 2D grid and a binary spatial mask $\mathbf{M}_s^0 \in \{0,1\}^{B\times H\times W}$ is sampled by removing a square region Ω with side length r: $\mathbf{M}_s^0(u,v) = 0, (u,v) \in \Omega$. Given the dropout rate p_s, we rescale the mask as $\mathbf{M}_s = \frac{\mathbf{M}_s^0}{1-p_s+\epsilon}$. The perturbed patch tokens are then formulated as

$$\widetilde{\mathbf{X}}_p = \mathbf{X}_p \odot \mathrm{Flatten}(\mathbf{M}_s), \tag{17}$$

and the final token sequence is $\widetilde{\mathbf{X}} = [\mathbf{x}_{\mathrm{cls}}; \widetilde{\mathbf{X}}_p]$, where $\mathbf{x}_{\mathrm{cls}}$ is always preserved. Samples sensitive to localized token regions exhibit larger response variations under blockwise masking, which encourages more robust spatial representations. For Swin Transformer, the same perturbation is applied immediately after the patch embedding layer; when a square token layout cannot be reliably recovered, we fall back to random token dropout.

2.3.3 Invariant Synthesis Optimization

Based on the structural objective in Eq. (15), DropIS further introduces auxiliary attention-space regularization to preserve structural coherence during synthesis. Specifically, the perturbed attention responses are constrained by the head- and token-level regularizers defined in Eq. (2) and Eq. (3), together with the image priors in Eq. (1). The overall synthesis objective is formulated as:

$$\mathcal{L}_{\mathrm{total}} = \frac{1}{K}\sum_{k=1}^{K}\left(\mathcal{L}_{\mathrm{cls}}^{(k)} + \mathcal{L}_{\mathrm{tv}}^{(k)} + \lambda_{\mathrm{h}}\mathcal{L}_{\mathrm{h}}^{(k)} + \lambda_{\mathrm{s}}\mathcal{L}_{\mathrm{s}}^{(k)}\right). \tag{18}$$

where λ_{h} and λ_{s} balance the head- and token-level attention regularization terms. By optimizing synthetic samples over diverse structural realizations with attention-aware constraints, DropIS promotes calibration data that remain stable across internal teacher variations. The complete optimization procedure is summarized in Algorithm 1 of Appendix A.3.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETTINGS AND DETAILS

To evaluate our DropIS, we perform experiments on the typical image classification benchmark, *i.e.*, ImageNet (ILSVRC2012) Russakovsky et al. (2015), which comprises 1.2M training samples and 50k validation samples spanning 1,000 classes. *The experimental results on object detection and semantic segmentation tasks are deferred to Appendix A.8.* Under data-free scenarios, the evaluation of the quantized model Q relies solely on the validation set. As backbone architectures for the full-precision model P, we consider several representative Vision Transformer families, including ViT Dosovitskiy (2020), DeiT Touvron et al. (2021), and Swin Transformer Liu et al. (2021b), covering their tiny, small, and base variants.

Quantization. Following Choi et al. (2025), we adopt the uniform quantization to map floating-point parameters and activations θ_p of P into integer-valued representations θ_q in Q, formulated as

$$\theta_q = clip(\lfloor \theta_p \cdot s_f - z \rceil, T_{min}, T_{max}), \tag{19}$$

where s_f and z denote the scaling factor and zero-point, respectively. Here, $\lfloor \cdot \rceil$ represents the rounding operation, and $clip(\cdot, \cdot, \cdot)$ constrains the quantized values within the predefined range bounded by T_{min} and T_{max}.

Implementation Details. For *sample synthesis*, we optimize the synthetic inputs using Adam with an initial learning rate of 0.1, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a batch size of 32. A total of 10k synthetic samples are generated, where each mini-batch is updated for 2,000 iterations. For *network calibration*, we adopt SGD with Nesterov momentum of 0.9, a learning rate of 1×10^{-3}, and a batch size of 16. Q is calibrated for 200 epochs. During both the sample synthesis and calibration stages, we apply the same data augmentation strategy as SimCLR Chen et al. (2020). For the hyperparameters, we set λ_h and λ_s in Eq.(18) to 1; the dropout rates p_h and p_s to 0.3 and 0.2; the block size r in Sec.2.3.2 to 3; and K in Eq.(18) to 4. All experiments are conducted using the PyTorch framework Paszke et al. (2019) based on the code of MimiQ Choi et al. (2025), and run on an NVIDIA GeForce RTX 5090 GPU and an INTEL(R) XEON(R) GOLD 6530 CPU @ 2.10GHz. ***Due to page limitations, more discussions and experimental results are accessible in Appendix. A.5 - A.7 and A.9 - A.10*** .

3.2 COMPARISON WITH THE STATE-OF-THE-ARTS

To verify the superiority of our DropIS, we compare it with representative data-free quantization methods, including: 1) GDFQ Xu et al. (2020), Qimera Choi et al. (2021), and AdaDFQ Qian et al. (2023a) are originally proposed for CNNs; 2) for Vision Transformers, PSAQ-ViT Li et al. (2022), PSAQ-ViT V2 Li et al. (2023b), CLAMP-ViT Ramachandran et al. (2024), GenQ Li et al. (2024), MimiQ Choi et al. (2025) and MaskAQ Qian et al. (2026) improve synthetic data generation through patch-level similarity, sample diversity, contrastive learning, generative models, inter-head attention consistency and informative region coupling, which mainly optimize synthetic calibration samples under deterministic teacher supervision. For GDFQ, Qimera, AdaDFQ, and CLAMP-ViT, we adopt the ViT results reported in MimiQ. GenQ only reports the results under 4w4a with no code available. For MimiQ, we further reproduce the 3-bit results using its official implementation. Although 3-bit quantization typically causes a much larger accuracy degradation than higher-bit settings, we still include these results to highlight the advantages (*e.g.*, semantic stability and transferability to Q) of DropIS under more challenging low-bit regimes (*also refer to lower-bit results in Appendix.A.9*).

Tab.1 summarizes the following findings: *first*, DropIS consistently outperforms existing data-free quantization methods, with gains of up to 8.46% over the strongest competing approaches MaskAQ, supporting that the effectiveness of *invariant synthesis* over structurally perturbed teachers compared with optimization under a single *deterministic* teacher (Sec.2.2). In contrast, CNN-oriented methods, *e.g.*, GDFQ, Qimera and AdaDFQ, suffer severe performance degradation ($<$10%) or even convergence failure ($<$1%) when directly applied to ViTs, highlighting the *necessity* of transformer-specific calibration synthesis. Under matched settings, DropIS also surpasses CLAMP-ViT by at least 1.94% and slightly outperforms GenQ (76.53% *vs.* 76.10%) without relying on external generative priors. *Second*, DropIS generalizes consistently across diverse ViT architectures, achieving improvements of up to 5.95%, 8.46% and 6.78 % on ViT, DeiT and Swin Transformer, respectively, over MaskAQ and MimiQ. Such fact suggests that structurally invariant synthesis provides calibration samples that transfer more robustly to different quantized models (Sec.1). *Finally*, the advantage becomes more pronounced under ultra-low-bit quantization. Under *3*w*3*a, DropIS improves over

Table 1: Classification accuracy (%) comparison with state-of-the-art approaches on the ImageNet dataset. †: the results reproduced by author-provided code; -: no results are reported. FP: full precision; *m*w*n*a indicates the weights and activations are quantized to *m*-bit and *n*-bit. The best and second-best results are reported with **boldface** and underline, respectively.

Methods	Venue	Target Architecture	Bit Width	Backbone Networks						
				ViT-T	DeiT-T	DeiT-S	DeiT-B	Swin-T	Swin-S	Swin-B
Real-Data Fine-tuning	-	-	**FP**	72.01	72.21	79.85	81.85	81.35	83.23	85.27
MimiQ Choi et al. (2025)	AAAI	ViT	***3w3a***	8.64†	19.55†	27.39†	41.86†	42.90†	51.21†	46.56†
MaskAQ Qian et al. (2026)	ICML	ViT		11.50	22.65	30.41	43.28	44.98	-	-
DropIS (Ours)	-	ViT		**17.45**	**31.11**	**38.39**	**48.98**	**50.48**	**57.99**	**48.45**
Gains (↑)	-	ViT		**+5.95**	**+8.46**	**+7.98**	**+5.70**	**+5.50**	**+6.78**	**+1.89**
GDFQ Xu et al. (2020)	ECCV	CNN	***4w4a***	2.95	25.96	22.12	30.04	42.08	41.93	36.04
Qimera Choi et al. (2021)	NeurIPS	CNN		0.57	15.18	11.37	32.49	47.98	39.64	29.27
AdaDFQ Qian et al. (2023a)	CVPR	CNN		2.00	19.57	14.44	19.22	38.88	39.40	32.26
PSAQ-ViT Li et al. (2022)	ECCV	ViT		0.67	19.61	5.90	8.74	22.71	9.26	23.69
PSAQ-ViT V2 Li et al. (2023b)	T-NNLS	ViT		1.54	22.82	32.57	45.81	50.42	39.10	39.26
GenQ Li et al. (2024)	ECCV	ViT		-	-	-	76.10	-	-	70.08
CLAMP-ViT Ramachandran et al. (2024)	ECCV	ViT		-	0.15	0.13	0.11	-	-	-
MimiQ Choi et al. (2025)	AAAI	ViT		42.99	52.03	62.72	74.10	69.33	70.46	73.49
MaskAQ Qian et al. (2026)	ICML	ViT		44.90	53.10	63.81	74.41	70.40	-	-
DropIS (Ours)	-	ViT		**49.51**	**56.49**	**66.55**	**76.53**	**73.16**	**75.49**	**73.62**
Gains (↑)	-	ViT		**+4.61**	**+3.39**	**+2.74**	**+0.43**	**+2.76**	**+5.03**	**+0.13**
GDFQ Xu et al. (2020)	ECCV	CNN	***5w5a***	24.40	44.76	57.00	71.03	61.30	78.04	70.55
Qimera Choi et al. (2021)	NeurIPS	CNN		26.70	33.13	33.65	47.01	62.13	46.81	43.57
AdaDFQ Qian et al. (2023a)	CVPR	CNN		27.10	53.85	59.55	71.12	64.61	79.82	75.59
PSAQ-ViT Li et al. (2022)	ECCV	ViT		17.66	53.36	47.35	57.23	58.63	76.33	57.80
PSAQ-ViT V2 Li et al. (2023b)	T-NNLS	ViT		40.21	55.18	65.30	73.16	69.77	80.55	79.80
CLAMP-ViT Ramachandran et al. (2024)	ECCV	ViT		-	5.42	3.19	0.17	-	-	-
MimiQ Choi et al. (2025)	AAAI	ViT		62.40	63.40	72.59	78.20	76.39	80.75	82.05
MaskAQ Qian et al. (2026)	ICML	ViT		63.01	64.27	73.18	78.89	76.82	-	-
DropIS (Ours)	-	ViT		**65.26**	**65.78**	**74.24**	**79.07**	**77.63**	**80.86**	**82.28**
Gains (↑)	-	ViT		**+2.25**	**+1.51**	**+1.06**	**+0.18**	**+0.81**	**+0.11**	**+0.23**
GDFQ Xu et al. (2020)	ECCV	CNN	***4w8a***	62.65	65.82	76.49	80.03	78.90	81.47	83.63
Qimera Choi et al. (2021)	NeurIPS	CNN		61.80	61.90	70.10	72.38	73.93	72.22	76.35
AdaDFQ Qian et al. (2023a)	CVPR	CNN		64.67	67.71	76.92	80.49	79.70	82.07	83.78
PSAQ-ViT Li et al. (2022)	ECCV	ViT		59.59	66.16	76.56	80.05	79.06	81.89	79.51
PSAQ-ViT V2 Li et al. (2023b)	T-NNLS	ViT		66.78	68.23	78.27	81.15	79.98	82.04	83.90
CLAMP-ViT Ramachandran et al. (2024)	ECCV	ViT		-	65.71	76.44	79.76	-	-	-
MimiQ Choi et al. (2025)	AAAI	ViT		68.15	69.86	78.48	81.34	80.06	82.08	83.99
MaskAQ Qian et al. (2026)	ICML	ViT		68.36	69.95	79.44	81.40	80.19	-	-
DropIS (Ours)	-	ViT		**68.69**	**69.97**	**79.45**	**81.70**	**80.27**	**82.11**	**84.01**
Gains (↑)	-	ViT		**+0.33**	**+0.02**	**+0.01**	**+0.30**	**+0.08**	**+0.03**	**+0.02**

Table 2: (a) Ablation study of varied components under 3w3a setting. HD., TD. and Reg. denote Head Dropout, Token Dropout and auxiliary attention-space regularizers. (b) Parameter studies about dropout rate. (c) Different structural probe sampling strategies. *m*w*n*a indicates the weights and activations are quantized to *m*-bit and *n*-bit. The best results are reported with **boldface**.

(a)

HD.	TD.	Reg.	DeiT-T	Swin-T
✗	✗	✓	21.17	44.48
✓	✗	✓	30.28	48.86
✗	✓	✓	28.21	49.00
✓	✓	✗	29.42	48.70
✓	✓	✓	**31.11**	**50.48**

(b)

Dropout Rate (p_h, p_s)	Bit Width		
	3w3a	***4w4a***	***5w5a***
$(0.0, 0.0)$	44.48	69.79	76.25
$(0.15, 0.1)$	49.45	72.49	77.12
$(0.3, 0.2)$	**50.48**	**73.16**	**77.63**
$(0.6, 0.4)$	47.96	72.57	77.29

(c)

Probe Sampling Strategy		***3w3a***	***4w4a***
Fixed	Deterministic	31.32	64.24
	Global-Fixed	29.51	63.66
	Batch-Fixed	30.18	64.23
Dynamic	Iteration-Resampled	35.87	65.79
	DropIS (Ours)	**38.39**	**66.55**

MaskAQ and MimiQ by 6.04% on average, indicating greater robustness when aggressive quantization strongly perturbs internal attention-token representations. This is consistent with our motivation of synthesizing calibration samples that remain stable across diverse internal teacher responses (Sec.2.3).

3.3 Ablation Studies

Discussion on Each Component of DropIS. We validate the effectiveness of different DropIS components on DeiT-T and Swin-T over ImageNet, including variants without head dropout, token dropout, both dropouts, or the auxiliary regularizer. Tab.2(a) suggests the obvious accuracy advantages (at least 0.83% and 1.48%) of DropIS over other cases. Notably, both head and token dropout improve the performance, and their combination performs best (31.11% and 50.48%), indicating that such two perturbations are *complementary* and jointly improve synthesis invariance (Sec.2.3). Moreover, most of the improvement stems from the proposed structured teacher perturbations, whereas the attention-space regularizers provide additional stabilization during synthesis.

Parameter Studies. The dropout rates p_h and p_t determine the perturbation strength of attention head and spatial token dropout. To investigate their effect, we evaluate $(p_h, p_s) \in \{(0.0, 0.0); (0.15, 0.1); (0.3, 0.2); (0.6, 0.4)\}$ on Swin-T. Tab.2(b) reports that the optimal performance is achieved with $(p_h^*, p_s^*) = (0.3, 0.2)$. Both larger or smaller dropout rate lead to lower accuracy, especially under the challenging *3*w*3*a setting (50.48% to 47.96% or 49.45%). This suggests that overly weak perturbations provide insufficient structural exploration, whereas overly strong perturbations may discard useful semantic information. A moderate perturbation strength therefore offers a better balance between invariant exploration and semantic preservation (Sec.2.3).

3.4 Discussion on Different Structural Probe Sampling Strategies

We further investigate the effect of structural-probe sampling on DeiT-S by comparing: 1) *Deterministic Teacher*; 2) *Global-Fixed Probe* fixes one structural perturbation throughout synthesis; 3) *Batch-Fixed Probe* samples one perturbation for each batch and keeps it unchanged during optimization; 4) *Iteration-Resampled Probe* resamples one perturbation at each synthesis iteration; 5) *Multi-Probe Sampling (DropIS)*. Table 2(c) suggests that fixed probes provide limited gains, whereas dynamic resampling substantially improves performance. In particular, Iteration-Resampled improves the accuracy to 35.87% and 65.79% under 3w3a and 4w4a, indicating the benefit of exposing synthetic samples to diverse structural responses. By jointly considering multiple perturbations at each synthesis step, DropIS further improves the accuracy to 38.39% and 66.55%, outperforming Iteration-Resampled by 2.52% and 0.76%. The above demonstrate that broader structural-probe coverage yields more robust invariant supervision than deterministic or fixed-probe synthesis, with larger gains under aggressive quantization. (Sec.2.3).

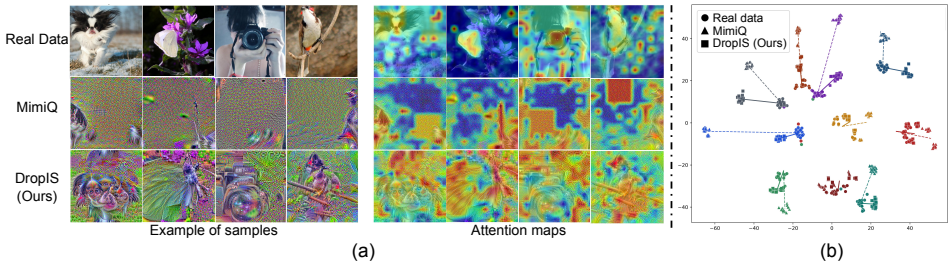


Figure 4: (a) Real data and synthetic samples, apart from the corresponding attention maps over perturbed teachers with head and token dropout. (b) t-SNE Van der Maaten & Hinton (2008) visualization of sample features across 10 categories, with each color representing one category.

3.5 Visual Analysis on Invariant Synthesis of DropIS

To further verify the merits of DropIS, we perform visual analysis on synthetic samples and attention maps with DeiT-S on ImageNet. Fig.4(a) compares real data with synthetic samples from MimiQ Choi et al. (2025) and our DropIS, together with their attention maps over perturbed teachers with head and token dropout; while Fig.4(b) shows the corresponding t-SNE Van der Maaten & Hinton (2008) feature distributions. As observed, the deterministic teacher synthesis tends to concentrate on narrow local evidence (row 2 in Fig.4(a)). In contrast, DropIS produces more spatially distributed and structurally coherent patterns (row 3 in Fig.4(a)), while their feature distribution is more similar to that of real data (■ and ● in Fig.4(b)). Moreover, DropIS maintains more stable attention responses under structural perturbations (row 3 in Fig.4(a)), suggesting reduced dependence on specific internal representation paths and supporting the motivation of *structural-invariant synthesis* in Sec.2.2.

4 Conclusion

In this paper, we formulate data-free calibration for ViTs as a structural-stability optimization problem beyond deterministic teachers, and reveal that deterministic supervision overlooks internal structural variations, producing synthetic samples with unstable responses under perturbations. Based on this insight, we propose **DropIS**, which introduces dropout-induced structural probes to generate stable and transferable calibration samples by exploring complementary variations in attention heads and spatial tokens. Extensive experiments across diverse architectures, tasks, and bit-width settings demonstrate the effectiveness of DropIS over existing data-free quantization methods.

AI USE STATEMENT

In this work, we used generative AI tools for language assistance, including grammar checking, improving writing clarity, and refining the presentation of the manuscript. We have not used generative AI tools for generating scientific ideas, designing the proposed method, conducting experiments, producing experimental results, creating datasets. These tasks were performed by the authors and are not applicable to the use of generative AI in this work.

All AI-assisted modifications were manually reviewed and verified by the authors. We ensured that the final manuscript accurately reflects the authors' original contributions and interpretations. The authors take full responsibility for the final content of this work, including all text, claims, experimental results, and conclusions.

ETHICS STATEMENT

This work does not involve human subjects, private data collection, or sensitive information. Our proposed data-free quantization framework operates solely on pre-trained models and synthetic data generated during optimization, without requiring access to original training datasets, thereby avoiding additional privacy risks.

We recognize that the proposed method inherits potential biases and limitations from the underlying pre-trained models and does not claim to address such issues. The goal of this research is to improve model efficiency and deployment accessibility, especially for resource-constrained applications. We encourage responsible evaluation and deployment of quantized models in real-world scenarios. This study follows standard research integrity principles, including transparent methodology description, fair experimental comparisons, and accurate reporting of results.

REPRODUCIBILITY STATEMENT

We make efforts to ensure the reproducibility of this work by providing comprehensive descriptions of the proposed DropIS framework and experimental protocols in the main paper and appendix. Detailed implementation settings, algorithm configurations, hyperparameters, and additional ablation studies are included in the appendix. The experiments are based on publicly available pre-trained Vision Transformer models and standard benchmark datasets, with the corresponding data processing and evaluation procedures documented accordingly. An anonymous implementation of the proposed method is provided as supplementary material to facilitate verification and future reproduction of our results.

REFERENCES

Yaohui Cai, Zhewei Yao, Zhen Dong, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Zeroq: A novel zero shot quantization framework. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13169–13178, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Kanghyun Choi, Deokki Hong, Noseong Park, Youngsok Kim, and Jinho Lee. Qimera: Data-free quantization with synthetic boundary supporting samples. *Advances in Neural Information Processing Systems*, 34, 2021.

Kanghyun Choi, Hyeyoon Lee, Dain Kwon, SunJong Park, Kyuyeun Kim, Noseong Park, Jonghyun Choi, and Jinho Lee. Mimiq: Low-bit data-free quantization of vision transformers with encouraging inter-head attention similarity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 16037–16045, 2025.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Hoang Anh Dung, Cuong Pham, Trung Le, Jianfei Cai, and Thanh-Toan Do. Sharpness-aware data generation for zero-shot quantization. In *Forty-first International Conference on Machine Learning*, 2024.

Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022a.

Tengda Han, Weidi Xie, and Andrew Zisserman. Turbo training with token dropout. *arXiv preprint arXiv:2210.04889*, 2022b.

Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Quantized neural networks: Training neural networks with low precision weights and activations. *journal of machine learning research*, 18(187):1–30, 2018.

Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

Minjun Kim, Jongjin Kim, and U Kang. Synq: Accurate zero-shot quantization by synthesis-aware fine-tuning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Bonan Li, Yinhan Hu, Xuecheng Nie, Congying Han, Xiangjian Jiang, Tiande Guo, and Luoqi Liu. Dropkey for vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22700–22709, 2023a.

Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. Brecq: Pushing the limit of post-training quantization by block reconstruction. In *International Conference on Learning Representations*.

Yuhang Li, Youngeun Kim, Donghyun Lee, Souvik Kundu, and Priyadarshini Panda. Genq: Quantization in low data regimes with generative synthetic data. In *European Conference on Computer Vision*, pp. 216–235. Springer, 2024.

Zhikai Li, Liping Ma, Mengjuan Chen, Junrui Xiao, and Qingyi Gu. Patch similarity aware data-free quantization for vision transformers. In *European conference on computer vision*, pp. 154–170. Springer, 2022.

Zhikai Li, Mengjuan Chen, Junrui Xiao, and Qingyi Gu. Psaq-vit v2: Toward accurate and general data-free quantization for vision transformers. *IEEE Transactions on Neural Networks and Learning Systems*, 2023b.

Zhikai Li, Junrui Xiao, Lianwei Yang, and Qingyi Gu. Repq-vit: Scale reparameterization for post-training quantization of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17227–17236, 2023c.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014.

Yang Lin, Tianyu Zhang, Peiqin Sun, Zheng Li, and Shuchang Zhou. Fq-vit: Post-training quantization for fully quantized vision transformer. *arXiv preprint arXiv:2111.13824*, 2021.

Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. When machine learning meets privacy: A survey and outlook. *ACM Computing Surveys (CSUR)*, 54(2):1–36, 2021a.

Yue Liu, Christos Matsoukas, Fredrik Strand, Hossein Azizpour, and Kevin Smith. Patchdropout: Economizing vision transformers using patch dropout. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3953–3962, 2023.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021b.

Zhenhua Liu, Yunhe Wang, Kai Han, Wei Zhang, Siwei Ma, and Wen Gao. Post-training quantization for vision transformer. *Advances in Neural Information Processing Systems*, 34:28092–28103, 2021c.

Yiwei Ma, Jiayi Ji, Xiaoshuai Sun, Yiyi Zhou, and Rongrong Ji. Towards local visual modeling for image captioning. *Pattern Recognition*, 138:109420, 2023.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Biao Qian, Yang Wang, Richang Hong, and Meng Wang. Adaptive data-free quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7960–7968, 2023a.

Biao Qian, Yang Wang, Richang Hong, and Meng Wang. Rethinking data-free quantization as a zero-sum game. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 9489–9497, 2023b.

Biao Qian, Yang Wang, Yong Wu, and Jungong Han. Selective coupling of decoupled informative regions: Masked attention alignment for data-free quantization of vision transformers. In *Forty-third International Conference on Machine Learning*, 2026. URL `https://openreview.net/forum?id=JeWtUdUk7l`.

Akshat Ramachandran, Souvik Kundu, and Tushar Krishna. Clamp-vit: Contrastive data-free learning for adaptive post-training quantization of vits. In *European Conference on Computer Vision*, pp. 307–325. Springer, 2024.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27 (3):379–423, 1948.

Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7262–7272, 2021.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Xiuying Wei, Ruihao Gong, Yuhang Li, Xianglong Liu, and Fengwei Yu. Qdrop: Randomly dropping quantization for extremely low-bit post-training quantization. *arXiv preprint arXiv:2203.05740*, 2022.

Zhuguanyu Wu, Jiayi Zhang, Jiaxin Chen, Jinyang Guo, Di Huang, and Yunhong Wang. Aphq-vit: Post-training quantization with average perturbation hessian based reconstruction for vision transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9686–9695, 2025.

Shoukai Xu, Haokun Li, Bohan Zhuang, Jing Liu, Jiezhang Cao, Chuangrun Liang, and Mingkui Tan. Generative low-bitwidth data free quantization. In *European Conference on Computer Vision*, pp. 1–17. Springer, 2020.

Zhihang Yuan, Chenhao Xue, Yiqi Chen, Qiang Wu, and Guangyu Sun. Ptq4vit: Post-training quantization for vision transformers with twin uniform quantization. In *European conference on computer vision*, pp. 191–207. Springer, 2022.

Zixiao Zhang, Xiaoqiang Lu, Guojin Cao, Yuting Yang, Licheng Jiao, and Fang Liu. Vit-yolo: Transformer-based yolo for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2799–2808, 2021.

Yunshan Zhong, Mingbao Lin, Gongrui Nan, Jianzhuang Liu, Baochang Zhang, Yonghong Tian, and Rongrong Ji. Intraq: Learning synthetic images with intra-class heterogeneity for zero-shot network quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12339–12348, 2022.

Yunshan Zhong, Yuyao Zhou, Yuxin Zhang, Wanchen Sui, Shen Li, Yong Li, Fei Chao, and Rongrong Ji. Semantic alignment and reinforcement for data-free quantization of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12479–12490, 2025.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5122–5130, 2017. doi: 10.1109/CVPR.2017.544.

A Appendix

A.1 Related Work

Data-Free Quantization for Vision Transformers. Data-free quantization (DFQ) quantizes pre-trained models without accessing the original training data by synthesizing surrogate calibration samples. Early methods, such as GDFQ Xu et al. (2020), Qimera Choi et al. (2021) and AdaDFQ Qian et al. (2023a), are mainly designed for convolutional networks and rely on statistic matching or generator-based synthesis. Recent ViT-oriented DFQ methods exploit transformer-specific priors: PSAQ-ViT Li et al. (2022) and PSAQ-ViT V2 Li et al. (2023b) introduces patch-similarity-aware synthesis; CLAMP-ViT Ramachandran et al. (2024) models inter-patch relationships via contrastive learning. Recently, MimiQ Choi et al. (2025) encourages inter-head attention similarity; while SARDFQ Zhong et al. (2025) improves the semantic alignment and richness of synthetic data. More recently, MaskAQ Qian et al. (2026) aims to preserve desirable mutual information with synthetic samples via informative region coupling. Despite these advances, existing methods predominantly synthesize calibration samples under a fixed deterministic teacher. In contrast, DropIS formulates ViT-DFQ as invariant synthesis over structurally perturbed teachers, encouraging synthetic samples to capture semantics shared across diverse internal teacher responses.

Dropout in Vision Transformers. Dropout-based stochastic regularization has been widely studied in Transformer architectures. DropKey Li et al. (2023a) regularizes self-attention by dropping keys, PatchDropout Liu et al. (2023) removes image patches for efficient ViT inference or training, and token dropout has been used to accelerate Transformer training Han et al. (2022b). These methods demonstrate the tolerance of Transformers to structured perturbations along attention or token dimensions. Unlike prior work that primarily employs dropout for regularization or efficiency, DropIS uses dropout as a structural probing mechanism during data-free synthesis. Specifically, head and spatial-token dropout expose synthetic samples to diverse internal teacher responses, discouraging dependence on individual attention heads or localized token regions and thereby promoting structural-invariant calibration synthesis.

A.2 Proofs and Auxiliary Results for Structural Stability Analysis

A.2.1 Proof of Proposition 2.1

Proposition 2.1 (Structural calibration-statistic decomposition). *For every layer l,*

$$\|T_{\mathcal{D}}^{(l)} - T_{\mathcal{S}_\theta}^{(l)}\|_2 \le I_{\mathcal{D}}^{(l)} + \Delta_{\mathrm{str}}^{(l)}(\mathcal{D}, \mathcal{S}_\theta) + I_{\mathcal{S}_\theta}^{(l)}. \tag{20}$$

Proof. Fix layer l and abbreviate $s_l^0(x) := s_l(x;\xi_0)$ and $\bar{s}_l(x) := \mathbb{E}_{\xi\sim q}[s_l(x;\xi)]$. Adding and subtracting the two probe-averaged means gives

$$T_{\mathcal{D}}^{(l)} - T_{\mathcal{S}_\theta}^{(l)} = \mathbb{E}_{\mathcal{D}}[s_l^0 - \bar{s}_l] + (\mathbb{E}_{\mathcal{D}}[\bar{s}_l] - \mathbb{E}_{\mathcal{S}_\theta}[\bar{s}_l]) + \mathbb{E}_{\mathcal{S}_\theta}[\bar{s}_l - s_l^0]. \tag{21}$$

By the triangle inequality,

$$\|T_{\mathcal{D}}^{(l)} - T_{\mathcal{S}_\theta}^{(l)}\|_2 \le \left\|\mathbb{E}_{\mathcal{D}}[s_l^0 - \bar{s}_l]\right\|_2 + \|\mathbb{E}_{\mathcal{D}}[\bar{s}_l] - \mathbb{E}_{\mathcal{S}_\theta}[\bar{s}_l]\|_2 \tag{22}$$

$$+ \left\|\mathbb{E}_{\mathcal{S}_\theta}[\bar{s}_l - s_l^0]\right\|_2. \tag{23}$$

For any distribution $\mathcal{P}$, Jensen's inequality first over x and then over ξ yields

$$\left\|\mathbb{E}_{\mathcal{P}}[s_l^0 - \bar{s}_l]\right\|_2 \le \mathbb{E}_{x\sim\mathcal{P}}\left[\|s_l(x;\xi_0) - \bar{s}_l(x)\|_2\right] \tag{24}$$

$$\le \mathbb{E}_{x\sim\mathcal{P}}\mathbb{E}_{\xi\sim q}\left[\|s_l(x;\xi_0) - s_l(x;\xi)\|_2\right] = I_{\mathcal{P}}^{(l)}. \tag{25}$$

The same argument bounds $\|\mathbb{E}_{\mathcal{P}}[\bar{s}_l - s_l^0]\|_2$ by $I_{\mathcal{P}}^{(l)}$. Substituting these inequalities and the definition of $\Delta_{\mathrm{str}}^{(l)}$ into Eq.(23) proves Eq.(7). □

A.2.2 Auxiliary Result: Conditional Local Calibration Consequence

Proposition 2.1 concerns calibration statistics and does not itself assume a smooth quantization map. A conditional consequence for the calibrated quantized model follows when the selected calibration procedure is locally stable. Let

$$\alpha_{\mathcal{P}} := \mathcal{A}(\{T_{\mathcal{P}}^{(l)}\}_{l=1}^{L}), \qquad \alpha_{\mathcal{S}} := \alpha_{\mathcal{S}_\theta}, \tag{26}$$

let Q_α denote the quantized model calibrated by parameters α, and define $\mathcal{R}_{\mathcal{D}}(Q) := \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathrm{CE}(f_Q(x), y)]$. Assume locally around the two compared calibration solutions that

$$\|\alpha_{\mathcal{S}} - \alpha_{\mathcal{D}}\|_2 \le L_{\mathcal{A}} \sum_{l=1}^{L} \|T_{\mathcal{S}_\theta}^{(l)} - T_{\mathcal{D}}^{(l)}\|_2, \tag{27}$$

$$|\mathcal{R}_{\mathcal{D}}(Q_\alpha) - \mathcal{R}_{\mathcal{D}}(Q_{\alpha'})| \le L_Q\|\alpha - \alpha'\|_2. \tag{28}$$

Then

$$|\mathcal{R}_{\mathcal{D}}(Q_{\alpha_{\mathcal{S}}}) - \mathcal{R}_{\mathcal{D}}(Q_{\alpha_{\mathcal{D}}})| \le L_Q L_{\mathcal{A}} \sum_{l=1}^{L} \left[I_{\mathcal{D}}^{(l)} + \Delta_{\mathrm{str}}^{(l)}(\mathcal{D}, \mathcal{S}_\theta) + I_{\mathcal{S}_\theta}^{(l)}\right]. \tag{29}$$

Indeed, the first local stability inequality bounds the calibration-parameter discrepancy by the statistic discrepancy, the second transfers this discrepancy to risk, and Proposition 2.1 is then applied layerwise. This is only a local conditional statement; it imposes no global smoothness requirement on rounding, clipping, or discrete calibration search.

A.2.3 Counterexample Supporting the Deterministic-Teacher Observation

The deterministic loss depends only on the identity response $\Gamma_{\xi_0}(x, y)$, whereas structural sensitivity depends on the complete response profile $\{\Gamma_\xi(x, y) : \xi \sim q\}$. For any non-degenerate probe family, choose a square-integrable scalar function $h : \Xi \to \mathbb{R}$ such that $h(\xi_0) = 0$ and $\mathrm{Var}_{\xi\sim q}[h(\xi)] > 0$. Consider

$$\Gamma_\xi^{(1)}(x, y) = M, \qquad \Gamma_\xi^{(2)}(x, y) = M + ah(\xi), \qquad a \ne 0. \tag{30}$$

Both profiles have the same deterministic margin M because $h(\xi_0) = 0$, but

$$V_{\mathrm{str}}^{(1)} = 0, \qquad V_{\mathrm{str}}^{(2)} = a^2\mathrm{Var}_{\xi\sim q}[h(\xi)] > 0. \tag{31}$$

Thus identical deterministic confidence can coexist with different structural sensitivities unless an additional model-specific relation between identity and off-identity responses is imposed.

A.2.4 Proof of Proposition 2.2

Proposition 2.2 (Variance-sensitive structural Jensen gap). *For any structural-probe distribution q satisfying the bounded-margin condition, we obtain*

$$\frac{\mu_\ell}{2} V_{\mathrm{str}} \le \mathcal{R}_{\mathrm{str}} - \ell(\overline{\Gamma}) \le \frac{L_\ell}{2} V_{\mathrm{str}}. \tag{32}$$

Proof. Fix (x, y) and write $Z := \Gamma_\xi(x, y)$ and $\bar{Z} := \mathbb{E}_{\xi\sim q}[Z]$. Since

$$\ell(u) = \log(1 + e^{-u}), \qquad \ell''(u) = \frac{e^{-u}}{(1 + e^{-u})^2} > 0, \tag{33}$$

ℓ is μ_ℓ-strongly convex and L_ℓ-smooth on $[a, b]$. Therefore, for any $u, v \in [a, b]$,

$$\ell(v) \ge \ell(u) + \ell'(u)(v - u) + \frac{\mu_\ell}{2}(v - u)^2, \tag{34}$$

$$\ell(v) \le \ell(u) + \ell'(u)(v - u) + \frac{L_\ell}{2}(v - u)^2. \tag{35}$$

Set $u = \bar{Z}$ and $v = Z$, and take expectation over ξ. Since $\mathbb{E}[Z - \bar{Z}] = 0$, the linear term vanishes, giving

$$\frac{\mu_\ell}{2}\mathrm{Var}_\xi[Z] \le \mathbb{E}_\xi[\ell(Z)] - \ell(\bar{Z}) \le \frac{L_\ell}{2}\mathrm{Var}_\xi[Z]. \tag{36}$$

Using the definitions of $\mathcal{R}_{\mathrm{str}}$ and V_{str} proves Eq.(12). □

A.2.5 Auxiliary Result: Two-Factor Structural Variance Decomposition

For the DropIS instantiation $\xi = (\xi_h, \xi_s)$ with the product distribution in Eq.(14), abbreviate $Z := \Gamma_{(\xi_h,\xi_s)}(x, y)$ and $\bar{Z} := \mathbb{E}[Z]$. Define

$$G_h(\xi_h) := \mathbb{E}_{\xi_s}[Z \mid \xi_h] - \bar{Z}, \tag{37}$$

$$G_s(\xi_s) := \mathbb{E}_{\xi_h}[Z \mid \xi_s] - \bar{Z}, \tag{38}$$

$$G_{hs}(\xi_h, \xi_s) := Z - \bar{Z} - G_h(\xi_h) - G_s(\xi_s). \tag{39}$$

Under the product distribution, these three components are zero-mean and pairwise orthogonal in $L_2(q)$. Hence

$$V_{\mathrm{str}} = V_h + V_s + V_{hs}, \qquad V_h := \mathbb{E}[G_h^2],\ V_s := \mathbb{E}[G_s^2],\ V_{hs} := \mathbb{E}[G_{hs}^2] \ge 0. \tag{40}$$

This decomposition only describes the chosen two-factor probe family; it does not claim that head and token perturbations exhaust every possible form of structural sensitivity.

A.2.6 Auxiliary Result: K-Probe Monte-Carlo Estimation

For the structural classification objective in Eq.(13), let $\xi^{(1)}, \ldots, \xi^{(K)} \overset{\text{i.i.d.}}{\sim} q$ denote sampled structural perturbations and define

$$\widehat{\mathcal{J}}_{\mathrm{str}}(x, y) := \frac{1}{K}\sum_{k=1}^{K} \mathcal{L}_{\mathrm{cls}}(x, y; \xi^{(k)}). \tag{41}$$

For fixed (x, y) and finite second moments,

$$\mathbb{E}[\widehat{\mathcal{J}}_{\mathrm{str}}] = \mathcal{J}_{\mathrm{str}}, \qquad \mathrm{Var}[\widehat{\mathcal{J}}_{\mathrm{str}}] = \frac{1}{K}\mathrm{Var}_q[\mathcal{L}]. \tag{42}$$

This follows immediately from independence of the K samples. If differentiation and expectation can be interchanged, the same argument gives an unbiased gradient estimator with

$$\mathrm{Cov}[\nabla_x \widehat{\mathcal{J}}_{\mathrm{str}}] = \frac{1}{K}\mathrm{Cov}_q[\nabla_x \mathcal{L}]. \tag{43}$$

The $1/K$ factor concerns Monte-Carlo sampling noise only and should not be interpreted as a $1/K$ improvement rate in invariance or quantized accuracy.

A.2.7 SCOPE OF THE THEORETICAL CLAIM

The main text establishes two core statements. First, for any prescribed family of moderate structural probes, the synthetic structural signature defect contributes to an upper bound on the real–synthetic calibration-statistic gap. Second, expected classification loss over the probe family contains a Jensen-gap term that is quantitatively sensitive to structural margin dispersion. The conditional calibration-risk consequence, two-factor variance decomposition, and K-probe Monte-Carlo result above are auxiliary results and are intentionally kept out of the main theoretical narrative. None of these statements requires the probes to approximate quantization noise.

The theory also deliberately stops short of claiming an exact implication from output-margin stability to invariance of every intermediate calibration statistic. DropIS therefore instantiates the general principle with attention-head and spatial-token probes and uses classification- and attention-space objectives as tractable structural surrogates. The quantities $I_{\mathcal{D}}^{(l)}$ and $\Delta_{\mathrm{str}}^{(l)}$ are analytical constructs and are never accessed by the data-free synthesis algorithm. The relevant empirical premise can be examined by comparing layerwise signature defects or calibration-statistic discrepancies across real data, baseline synthetic data, and DropIS samples, together with the resulting quantized accuracy.

A.3 MORE DETAILS OF OVERALL PIPELINE

We further summarize the synthesis procedure of DropIS in Algorithm 1. Different from conventional dropout-based regularization, DropIS treats structured dropout as a teacher probing mechanism during data synthesis. Specifically, at each synthesis iteration, DropIS samples a structural probe $\xi = (\xi_h, \xi_s)$ from the perturbation distribution $q(\xi) = q_h(\xi_h)q_s(\xi_s)$, where ξ_h and ξ_s correspond to attention-head and spatial-token perturbations, respectively.

The synthetic samples are then forwarded through both the original teacher and the structurally perturbed teacher to compute the invariant consistency objective. Meanwhile, the perturbed attention responses are collected to calculate auxiliary attention-space regularizers, which stabilize intermediate structural responses. The overall objective is optimized using Monte-Carlo estimation over multiple structural probes. By optimizing synthetic samples toward invariant responses across structurally perturbed teachers, DropIS generates calibration samples with reduced structural sensitivity and improved transferability to the quantized model.

A.4 MORE DETAILS OF MOTIVATION EXPERIMENTS IN FIG. 1 OF SEC. 1: MARGIN STABILITY UNDER STRUCTURAL PERTURBATIONS

To investigate the structural stability of synthetic samples generated by deterministic-teacher synthesis methods and DropIS, we evaluate their prediction margins under multiple structural perturbations (see Fig. 5 and Fig. 6). These experiments aim to examine whether the confidence obtained from the deterministic teacher can be preserved when the internal teacher structure is moderately perturbed. Our motivation experiments reveal the following observations:

- **Observation 1:** ***Single-teacher synthesis produces high-confidence samples with limited structural stability.***
  Synthetic samples optimized only on the identity teacher can achieve large deterministic margins. However, their prediction confidence may vary significantly when the teacher structure is perturbed by attention-head or spatial-token dropout. This indicates that such samples may rely on specific internal response patterns rather than representations that remain stable across different structural perturbations.
- **Observation 2:** ***Structurally stable synthetic samples preserve discriminative confidence under perturbations and provide more reliable calibration signals.***
  Synthetic samples with smaller margin variations and better agreement between deterministic and perturbed responses preserve their discriminative confidence more consistently across structural probes. Since quantization also introduces perturbations to internal representations, such structurally stable samples provide more transferable calibration information for the quantized model Q.

Specifically, following the margin definition in Sec. 2.2, we denote the deterministic margin under the identity teacher ξ_0 as

$$\gamma(x, y) = \Gamma_{\xi_0}(x, y). \tag{44}$$

Algorithm 1 DropIS: Dropout-Induced Invariant Synthesis

Require: Full-precision ViT teacher P; synthesis iterations T; number of synthesis batches N_B; number of structural probes K; probe distribution $q(\xi) = q_h(\xi_h)q_s(\xi_s)$
Ensure: Synthetic calibration set $\mathcal{D}_{\text{syn}}$

Initialize $\mathcal{D}_{\text{syn}} \leftarrow \emptyset$
for $n_B = 1$ to N_B **do**
  Initialize synthetic images $\mathbf{x}$ from random noise
  Sample pseudo labels $\mathbf{y}$
  for $t = 1$ to T **do**
    Initialize $\widehat{\mathcal{L}}_{\text{total},t} \leftarrow 0$
    for $k = 1$ to K **do**
      Sample structural probe:

$$\xi^{(k)} = (\xi_h^{(k)}, \xi_s^{(k)}) \sim q(\xi)$$

      Construct perturbed teacher $P^{\xi^{(k)}}$ by applying attention-head and spatial-token dropout
      Forward $\mathbf{x}$ through P and $P^{\xi^{(k)}}$ and collect perturbed attention responses
      Compute invariant synthesis loss:

$$\mathcal{L}_{\text{total},t}^{(k)} = \mathcal{L}_{\text{prior},t}^{(k)} + \lambda_h \mathcal{L}_{h,t}^{(k)} + \lambda_s \mathcal{L}_{s,t}^{(k)}$$

      Accumulate Monte-Carlo objective:

$$\widehat{\mathcal{L}}_{\text{total},t} \leftarrow \widehat{\mathcal{L}}_{\text{total},t} + \frac{1}{K}\mathcal{L}_{\text{total},t}^{(k)}$$

    end for
    Update synthetic images:

$$\mathbf{x} \leftarrow \mathbf{x} - \eta_t \nabla_{\mathbf{x}} \widehat{\mathcal{L}}_{\text{total},t}$$

    Clamp $\mathbf{x}$ into valid input range
  end for
  Add optimized samples:

$$(\mathbf{x}, \mathbf{y}) \rightarrow \mathcal{D}_{\text{syn}}$$

end for
return $\mathcal{D}_{\text{syn}}$

We randomly sample K structural perturbations $\{\xi^{(k)}\}_{k=1}^{K}$, where each perturbation consists of attention-head and spatial-token perturbations:

$$\xi^{(k)} = (\xi_h^{(k)}, \xi_s^{(k)}). \tag{45}$$

The corresponding perturbed margin is computed as

$$\Gamma_k(x, y) = f_y(x; \xi^{(k)}) - \log \sum_{c \neq y} \exp(f_c(x; \xi^{(k)})), \tag{46}$$

where $f(x; \xi^{(k)})$ denotes the teacher response under the k-th sampled structural perturbation. Based on these perturbed margins, we compute the mean perturbed margin:

$$\bar{\Gamma}(x, y) = \frac{1}{K} \sum_{k=1}^{K} \Gamma_k(x, y), \tag{47}$$

the margin degradation:

$$\Delta_\Gamma(x, y) = \gamma(x, y) - \bar{\Gamma}(x, y), \tag{48}$$

and the margin variance:

$$V_\Gamma(x, y) = \frac{1}{K} \sum_{k=1}^{K} \left(\Gamma_k(x, y) - \bar{\Gamma}(x, y)\right)^2, \tag{49}$$

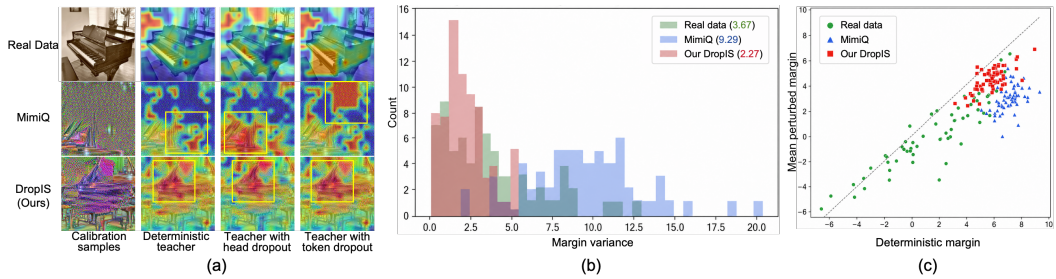


Figure 5: Motivation experiments beyond the deterministic teacher (copied from Fig. 1). (a) Attention visualization under deterministic and structurally perturbed teacher responses with head or token dropout. DropIS preserves more stable attention responses than single-teacher synthesis methods. (b) Margin variance distribution, where lower variance indicates stronger stability across structural perturbations. (c) Deterministic margin versus mean perturbed margin, where points closer to the diagonal indicate better confidence preservation under structural perturbations.

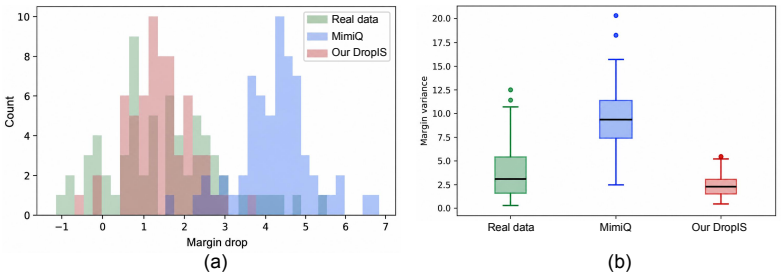


Figure 6: Additional margin stability analysis under structural perturbations. (a) Distribution of margin degradation $\Delta_\Gamma(x, y)$, which measures the average confidence reduction caused by structural perturbations. (b) Distribution of margin variance $V_\Gamma(x, y)$, where lower variance indicates stronger stability across structural probe realizations.

where $\Delta_\Gamma(x, y)$ measures the average confidence degradation caused by structural perturbations, while $V_\Gamma(x, y)$ characterizes the sensitivity of prediction margins to different structural probe realizations.

Fig. 5(a) illustrates examples of real data and synthetic samples generated by teacher-specific DFQ methods (*e.g.*, MimiQ Choi et al. (2025)) and our DropIS, together with their attention responses under deterministic and structurally perturbed teacher realizations with head or token dropout. We observe that samples generated by single-teacher synthesis methods may achieve strong confidence under the identity teacher, while their attention responses become less consistent after structural perturbations. In contrast, DropIS preserves more stable attention responses, indicating improved robustness to internal structural variations.

Fig. 5(c) illustrates the relationship between the deterministic margin $\gamma(x, y)$ and the mean perturbed margin $\bar{\Gamma}(x, y)$. The diagonal line represents perfect confidence preservation:

$$\bar{\Gamma}(x, y) = \gamma(x, y).$$

Therefore, this visualization reflects whether the confidence obtained from the deterministic teacher can be maintained under structural perturbations. Synthetic samples from single-teacher synthesis methods (*e.g.*, MimiQ Choi et al. (2025)) often achieve large deterministic margins but exhibit substantially lower perturbed margins, causing many points to deviate from the diagonal. This indicates that their confidence is sensitive to the original teacher structure. In contrast, DropIS samples are distributed closer to the diagonal, suggesting more consistent prediction margins across structural probe realizations.

Fig. 5(b) and Fig. 6(b) show the distribution of margin variance $V_\Gamma(x, y)$ across structural perturbations. The margin variance measures the sensitivity of each synthetic sample to different structural probe realizations. Single-teacher synthesis methods (*e.g.*, MimiQ Choi et al. (2025)) exhibit broader

and more long-tailed variance distributions, indicating larger prediction variations under structural perturbations. In comparison, DropIS produces a more concentrated distribution with lower variance values, demonstrating improved structural stability of the synthesized samples.

Fig. 6(a) visualizes the distribution of margin degradation $\Delta_\Gamma(x, y)$, which measures the confidence reduction when moving from the identity teacher to the structurally perturbed teacher family. Single-teacher synthesis methods exhibit larger margin degradation, indicating that their synthesized samples are more sensitive to internal structural changes. In contrast, DropIS achieves smaller margin degradation, suggesting that the synthesized samples preserve their discriminative confidence more effectively under head and token perturbations.

These observations are consistent with our theoretical analysis in Sec. 2.2: synthetic samples with higher structural sensitivity exhibit larger prediction variations under internal teacher perturbations, whereas structurally stable samples maintain consistent margins across different structural probe realizations. Therefore, optimizing synthetic samples over structural probes, rather than a single deterministic teacher response, provides more reliable calibration samples for quantized models.

A.5 Further Insights into Importance-aware DropIS

As aforementioned, random dropout may occasionally suppress semantically important attention heads or discriminative spatial tokens, introducing unnecessary perturbations during synthesis. To mitigate this issue, we introduce an importance-aware variant, denoted as **DropIS-I**, which preserves a subset of important structures before applying dropout, as illustrated in Fig. 7. Specifically, head importance is estimated by the mean magnitude of attention responses, while token importance is measured by feature norms. The top-ranked heads or tokens are protected, and dropout is applied only to the remaining structures. The results in Tab. 3 demonstrate that DropIS-I consistently improves over DropIS across all evaluated settings. Under the challenging 3w3a setting, DropIS-I improves the accuracy of DeiT-S, DeiT-B, and Swin-S by 0.52%, 0.67%, and 0.31%, respectively. Similar improvements are observed under 4w4a quantization, with gains of 0.39%, 0.42%, and 0.29% on the three backbones. These results indicate that preserving informative attention heads and spatial tokens can effectively reduce harmful structural perturbations while maintaining sufficient diversity during synthesis. Therefore, DropIS-I provides a better balance between *semantic preservation* and *perturbation diversity*. By incorporating importance-aware structural probing, DropIS further improves the reliability and transferability of synthesized calibration samples, leading to more effective calibration of the quantized model Q.

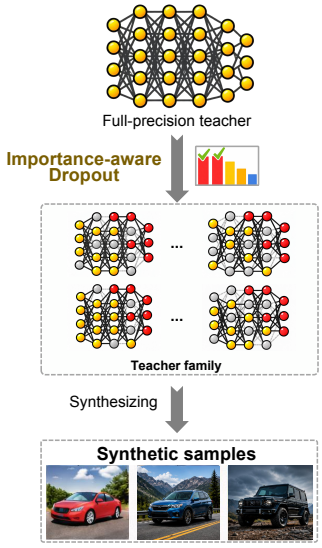


Figure 7: Illustration of importance-aware DropIS.

Table 3: Discussion on importance-aware DropIS. *mwna* indicates the weights and activations are quantized to *m*-bit and *n*-bit. The best results are reported with **boldface**.

Bit Width	Method	Backbone Networks		
		DeiT-S	DeiT-B	Swin-S
3w3a	DropIS	38.39	48.98	57.99
	DropIS-I	**38.91**	**49.65**	**58.30**
4w4a	DropIS	66.55	76.53	75.49
	DropIS-I	**66.94**	**76.95**	**75.78**

A.6 Parameter Studies

We further study the effect of several auxiliary hyper-parameters, including the number of structural probes K, the block size r in spatial token dropout, the balance parameters (λ_h and λ_s) of the attention-space regularizers and the number of synthetic samples.

A.6.1 NUMBER OF SAMPLED STRUCTURAL PROBES

We study the effect of the number of structural probes (K) sampled at each synthesis iteration. Specifically, we perform the experiments on DeiT-T by varying $K \in \{1, 2, 4, 6, 8, 16\}$, where $K = 1$ corresponds to evaluating each synthetic batch under a single structural perturbation, while larger K values approximate the structural-probe expectation using multiple independently sampled attention-head and spatial-token perturbations, which are resampled at every synthesis iteration.

Table 4 suggests that increasing K consistently improves the accuracy of quantized model, especially under lower-bit settings. For example, under 3w3a quantization, increasing K from 1 to 4 improves the accuracy from 25.69% to 31.11%, demonstrating that considering multiple structural responses provides more effective synthesis supervision. Further increasing K yields marginal improvements, while the synthesis cost grows approximately linearly. Although $K = 16$ achieves slightly higher accuracy under some settings, it introduces substantially larger computational overhead (12.53×). Therefore, we adopt $K = 4$ as the default setting, which provides a favorable trade-off between calibration performance and synthesis efficiency. These results verify that *the effectiveness of DropIS comes from approximating the structural-probe distribution with sufficient response diversity rather than simply increasing the number of perturbed forward passes*.

Table 4: Ablation study about the number of sampled structural probes K. The best results are reported with **boldface**.

Number of Structural Probes K		**1**	**2**	**4**	**6**	**8**	**16**
Speed		1×	1.77×	3.31×	4.84×	6.38×	12.53×
Bit Width	***3w3a***	25.69	29.52	31.11	31.63	30.11	**31.73**
	4w4a	53.51	55.65	56.49	56.61	**56.72**	56.63
	5w5a	64.40	65.48	65.78	65.88	65.80	**65.94**

A.6.2 EFFECT OF BLOCK SIZE r IN SEC. 2.3.2

The block size r in Sec. 2.3.2 determines the spatial granularity of perturbations applied in spatial token dropout. Intuitively, when r is too small, the perturbation only suppresses isolated tokens or weakly correlated local responses, making it behave similarly to conventional random token dropout. Such fine-grained perturbations may be insufficient to disrupt local token shortcuts, allowing the synthesized calibration samples to still exploit narrow discriminative regions to activate the deterministic teacher. In contrast, an overly large block size removes a substantial portion of spatially contiguous tokens, which may discard critical semantic evidence and further weaken the discriminative supervision required for effective sample synthesis.

To investigate the influence of the block size, we conduct ablation experiments on DeiT-S under 3-bit setting by varying the block size over $r \in \{1, 3, 5, 7, 9, 11\}$. Fig.8(a) illustrates that the performance exhibits a clear increasing-then-decreasing trend as r grows. The best result is achieved at a moderate block size $r^* = 3$, indicating that properly structured spatial perturbation is more effective than either overly fine-grained or overly aggressive masking. This observation supports our analysis that moderate block-wise token dropout can effectively discourage the synthesized samples from over-relying on localized token evidence, while still retaining sufficient semantic cues for stable and discriminative invariant synthesis. Moreover, DropIS remains relatively robust to the precise choice of r within a reasonable range ($r \in (1, 5)$), suggesting that the gains of DropIS are not attributed to a carefully tuned masking configuration, but rather stem from the general principle of enforcing semantic consistency under structured local spatial perturbations.

A.6.3 SENSITIVITY ANALYSIS ON BALANCE PARAMETERS IN EQ.(18)

We further investigate the effectiveness of the balance parameters λ_h and λ_s from two auxiliary attention-space regularizers ($\mathcal{L}_h$ and $\mathcal{L}_s$) in Eq.(18), which complement each other to stabilize the dropout-induced invariant synthesis process. In the ablation experiments, we testify varied $\lambda_h \in \{0.2, 0.5, 0.8, 1, 2, 5, 10\}$ and $\lambda_s \in \{0.2, 0.5, 0.8, 1, 2, 5, 10\}$ with DeiT-S under 3-bit setting. Fig.8(b) and Fig.8(c) report that the best performance is achieved at $\lambda_h^* = 5$ and $\lambda_s^* = 1$, indicating that carefully adjusting the balance parameters may obtain extra performance improvements.

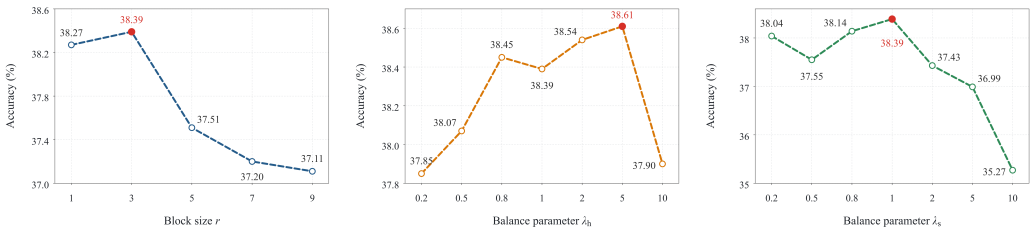


Figure 8: Parameter studies about (a) the block size r, (b) the balance parameter λ_h and (c) the balance parameter λ_s.

Meanwhile, DropIS remains stable within a broad range of balance parameters, since they serve as stabilizing auxiliary constraints rather than the primary source of improvement, which is consistent with *our proposal that the core effectiveness of DropIS comes from dropout-induced invariant synthesis, while $\mathcal{L}_h$ and $\mathcal{L}_t$ is exploited to further stabilize the perturbed attention responses (Sec.2.3.1 and 2.3.2).*

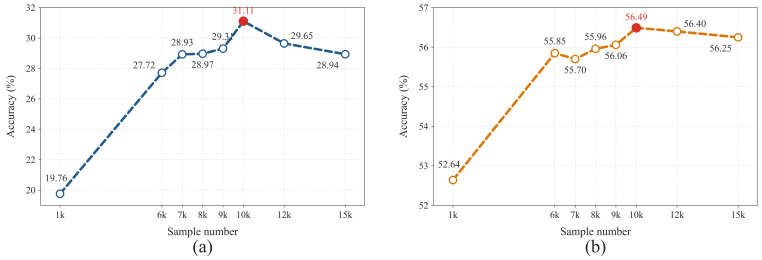


Figure 9: Parameter studies about the number of synthetic calibration samples under (a) 3-bit and (b) 4-bit settings.

A.6.4 Discussion on Number of Synthetic Calibration Samples

In data-free quantization, synthetic calibration samples are required to approximate the original calibration distribution, thereby providing representative activation statistics, attention responses, and token interactions for quantization calibration. We further investigate the effect of the synthetic sample budget by varying the number of synthetic samples as $\{1k, 6k, 7k, 8k, 9k, 10k, 12k, 15k\}$ on DeiT-T under both 3-bit and 4-bit settings. The results from Fig.9 show that the performance of Q is improved as more synthetic samples are used for calibration. The gain is more significant when the sample number is small, indicating that insufficient synthetic data limits the coverage of the teacher-induced semantic space. As the sample budget increases, the improvement gradually saturates, suggesting that the calibration set already covers most of the useful activation and attention statistics. Notably, even under a limited sample budget, DropIS can provide effective calibration samples, confirming that *DropIS produces calibration data with stronger structural stability and better transferability to Q by optimizing samples over dropout-induced perturbed teachers (Sec.2.3).*

A.7 Computational Overhead

To evaluate the computational cost of DropIS, we conduct an efficiency analysis with Swin-T under the *3w3a* setting on ImageNet. We report the synthesis time, calibration time, total wall-clock time, and relative time overhead in Table 5. Compared with MimiQ Choi et al. (2025), DropIS with $K = 1$ achieves a substantial accuracy improvement from 42.90% to 46.98% with only 1.009× total time overhead, confirming that *introducing a single structural probe incurs negligible additional cost*, since DropIS only requires lightweight head and token masking operations during synthesis and does not modify the subsequent calibration procedure. When increasing the number of sampled structural probes, DropIS obtains further accuracy improvements at the cost of additional synthesis time. Specifically, DropIS with $K = 4$ improves the accuracy to 50.48%, while the total time increases to

60.10 hours (3.34× overhead). The importance-aware variant DropIS-I further improves the accuracy to 50.88% with only marginal additional cost (3.35× overhead), indicating that the additional structural selection introduces limited computational burden. Therefore, the computational overhead of DropIS is mainly determined by the number of sampled structural probes K, which *provides a controllable trade-off between synthesis cost and calibration performance*. The default setting of $K = 4$ achieves significant accuracy gains while maintaining practical efficiency, demonstrating that DropIS is applicable for data-free quantization of ViTs.

Table 5: Efficiency analysis with Swin-T under *3*w*3*a setting on ImageNet. Synthesis Time reports the total wall-clock time for generating the same number of synthetic batches. Calibration Time denotes the time of subsequent calibration process with the synthetic samples. Relative Time Overhead is computed with respect to MimiQ. DropIS achieves the controllable accuracy-efficiency trade-off with different numbers of sampled structural probes.

Method	Accuracy (%)	Synthesis Time (Hour)	Calibration Time (Hour)	Total Time (Hour)	Relative Time Overhead
MimiQ Choi et al. (2025)	42.90	10.40h	7.6h	18.00h	1.00×
DropIS ($K = 1$)	46.98	10.56h	7.6h	18.16h	1.009×
DropIS ($K = 2$)	48.56	24.54h	7.6h	32.14h	1.78×
DropIS ($K = 4$)	50.48	52.50h	7.6h	60.10h	3.34×
DropIS-I ($K = 4$)	50.88	52.73h	7.6h	60.33h	3.35×

Table 6: Comparison with state-of-the-art approaches on the COCO dataset and ADE20K dataset. *m*w*n*a indicates the weights and activations are quantized to *m*-bit and *n*-bit. The best results are reported with **boldface**.

Methods	Venue	Bit width	COCO Dataset		ADE20K Dataset
			AP_{box}	AP_{mask}	mIoU
PSAQ-ViT Li et al. (2022)	ECCV' 22	***4w4a***	0.06	0.06	1.65
PSAQ-ViT V2 Li et al. (2023b)	T-NNLS' 23		4.52	5.03	3.83
MimiQ Choi et al. (2025)	AAAI' 25		26.41	26.63	29.92
DropIS (Ours)	-		**27.13**	**27.08**	**31.01**
PSAQ-ViT Li et al. (2022)	ECCV' 22	***5w5a***	0.41	0.46	20.26
PSAQ-ViT V2 Li et al. (2023b)	T-NNLS' 23		32.69	31.21	26.35
MimiQ Choi et al. (2025)	AAAI' 25		41.63	38.53	38.88
DropIS (Ours)	-		**42.35**	**39.05**	**40.23**

A.8 Additional Results on Object Detection and Semantic Segmentation

Beyond the image classification experiments presented in the main paper, we further examine the generalization ability of our DropIS on two dense prediction tasks, including object detection and semantic segmentation. For object detection, we employ Swin-T as the backbone and conduct experiments on the COCO dataset Lin et al. (2014), a challenging benchmark containing diverse visual scenes and fine-grained instance-level annotations. In addition, we evaluate semantic segmentation on the ADE20K dataset Zhou et al. (2017) with the same Swin-T backbone, where ADE20K provides dense pixel-level annotations covering a wide range of semantic categories and scene layouts. For COCO, we report both bounding-box average precision (AP_{box}) and instance-mask average precision (AP_{mask}). Following the standard COCO evaluation protocol, AP is computed by averaging over multiple IoU thresholds from 0.50 to 0.95 with a step size of 0.05, as well as over all object categories. For ADE20K, we adopt mean Intersection over Union (mIoU) as the evaluation metric. It is worth noting that the results of PSAQ-ViT Li et al. (2022) and PSAQ-ViT V2 Li et al. (2023b) are taken from those reported in MimiQ Choi et al. (2025). Tab.6 reports that DropIS consistently outperforms competing methods across downstream tasks. In particular, compared with the recent state-of-the-art method, *i.e.*, MimiQ, DropIS brings up to 0.72% and 1.35% performance gains on COCO and ADE20K, respectively. These results demonstrate that DropIS is not limited to image classification, but can also generalize well to dense prediction tasks, confirming its strong task-level compatibility and transferability.

Table 7: Classification accuracy (%) comparison under the extremely low-bit setting on ImageNet. †: the results reproduced by author-provided code. FP: full precision; *m*w*n*a indicates the weights and activations are quantized to *m*-bit and *n*-bit. The best results are reported with **boldface**.

Methods	Venue	Target Architecture	Bit width	Backbone Networks						
				ViT-T	DeiT-T	DeiT-S	DeiT-B	Swin-T	Swin-S	Swin-B
Real-Data Fine-tuning	-	-	**FP**	72.01	72.21	79.85	81.85	81.35	83.23	85.27
MimiQ Choi et al. (2025)	AAAI	ViT	***2w4a***	1.10†	1.21†	1.34†	1.75†	4.55†	9.37†	2.91†
DropIS (Ours)	-	ViT		**2.46**	**2.91**	**3.92**	**11.87**	**14.30**	**16.73**	**8.25**
Gains (↑)	-	ViT		**+1.36**	**+1.70**	**+2.58**	**+10.12**	**+9.75**	**+7.36**	**+5.34**

A.9 Additional Results under Lower-Bit Settings

To further examine the generality of DropIS in more challenging quantization regimes, we evaluate DropIS under the extremely low-bit *2*w*4*a setting. The results in Tab. 7 show that MimiQ suffers from severe accuracy degradation across all evaluated backbones, indicating that extremely low-bit quantization introduces substantial perturbations to internal feature transformations and makes calibration highly sensitive to the stability and transferability of synthetic samples. Despite this difficulty, DropIS consistently outperforms MimiQ on all evaluated architectures. The improvements are particularly significant on larger transformer backbones. For example, DropIS improves DeiT-B from 1.75% to 11.87%, achieving a substantial gain of $+10.12\%$. Similarly, on Swin-B, DropIS improves the accuracy from 2.91% to 8.25%, with a gain of $+5.34\%$. These results demonstrate that *structural-invariant synthesis becomes increasingly beneficial when quantization perturbations become more severe*, which is consistent with our theoretical analysis in Sec. 2.2. Under such an aggressive low-bit setting, synthetic samples optimized for a single deterministic teacher are more likely to overfit specific internal response patterns, resulting in limited transferability to the quantized model Q. In contrast, DropIS synthesizes calibration samples over dropout-induced structural perturbations, encouraging the samples to preserve discriminative responses under both attention-head and spatial-token variations. Consequently, *the generated samples exhibit stronger structural stability and improved calibration transferability*, leading to consistent performance gains even under extremely challenging low-bit quantization.

A.10 Additional Visual Results

We further provide additional visual results on synthetic calibration samples and attention maps with DeiT-S on ImageNet. Fig.10 illustrates the examples of real data and synthetic samples from MimiQ Choi et al. (2025) and our DropIS, apart from the corresponding attention maps over perturbed teachers with either head dropout or token dropout and both of them. It is observed that, the deterministic teacher synthesis tends to produce samples dominated by narrow local evidence. In contrast, synthetic samples from our DropIS exhibit more spatially distributed and structurally coherent patterns. Meanwhile, DropIS produces more stable attention responses under head and token perturbations, indicating that the synthesized semantics are less tied to a specific representation path, which is consistent with *our proposal of improving semantic stability and transferability to Q of synthetic samples via invariant synthesis beyond the deterministic teacher in Sec.2.2*.

A.11 Limitations and Future Work

Despite its effectiveness, DropIS has several limitations. First, it relies on iterative image synthesis and therefore introduces additional generation cost compared with approaches that avoid explicit sample optimization. Second, the current formulation focuses on attention-head and spatial-token perturbations, while other potentially informative structural dimensions, such as MLP channels and layer-wise adaptive perturbations, remain unexplored. Extending DropIS to more efficient synthesis schemes and broader structural probing mechanisms is left for future work.

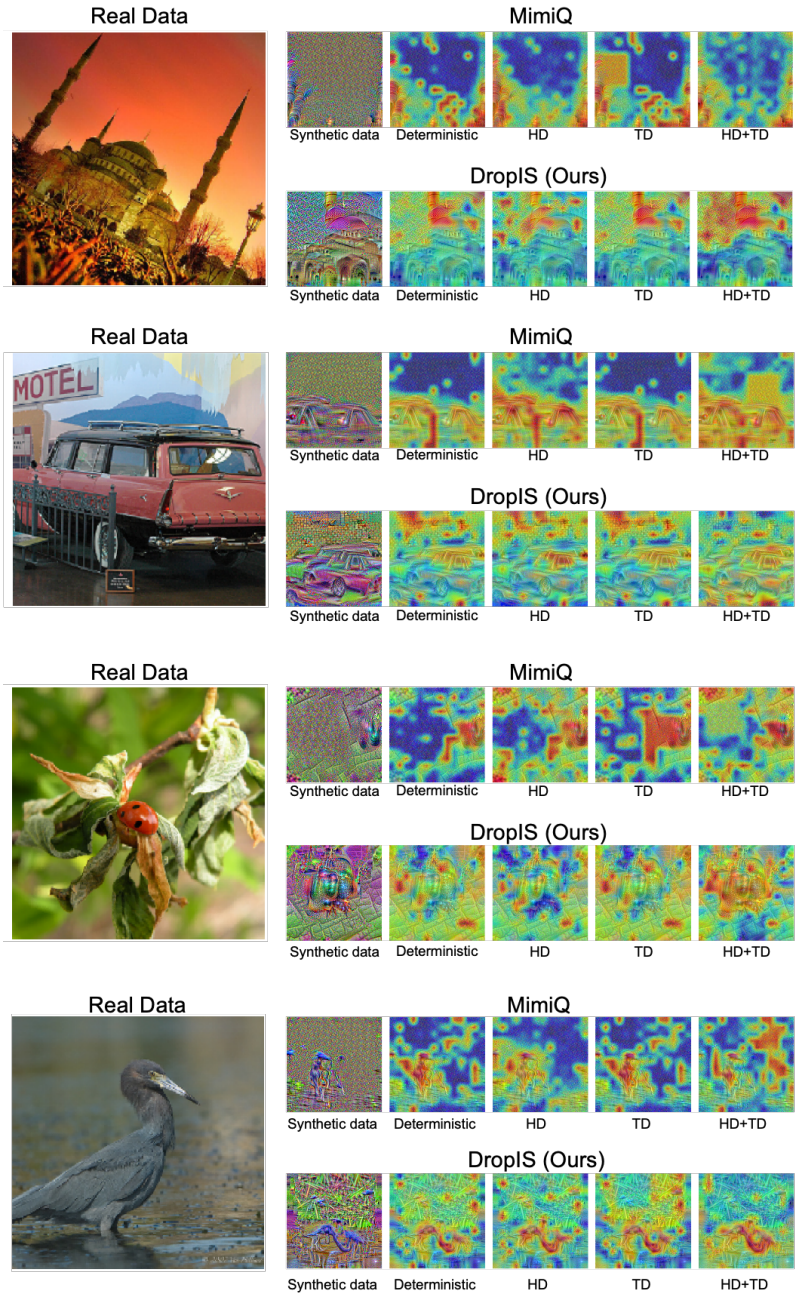


Figure 10: Examples of real data and synthetic samples, apart from the corresponding attention maps over deterministic teacher and perturbed teachers with head dropout (HD), token dropout (TD) and both of them (HD+TD). DropIS produces more stable attention responses under head and token perturbations.