VIGUARDQ: QUANTIZED VIDEO DEEPFAKE DETECTION FOR EDGE DEVICES

Anonymous authors
Paper under double-blind review

ABSTRACT

The growing use of edge devices, e.g. smartphones, for video conferencing and facial biometric authentication creates an increasing demand for efficient video deepfake detection. However, the substantial computational and memory requirements of existing detectors hinder their deployment on resource-constrained devices. Although quantization offers an effective means of model compression, conventional methods do not explicitly account for the subtle spatial artifacts and temporal inconsistencies essential to video deepfake detection, which can be distorted by low-precision representations. To address this challenge, we propose **VIGUARDQ**, a post-training quantization framework tailored to video deepfake detection on edge devices. First, *Dual-Evidence-Constrained (DEC) Rotation* mitigates forgery-evidence relevant quantization distortion by identifying a sensitive channel subspace using within-frame and cross-frame differential decision gradients, and optimizing orthogonal rotations to preserve more forgery-evidence sensitive information. Then *Evidence-guided Consistent Response (ECR) Calibration* refines the quantizer's consistency in response to evidence under aligned and opposite perturbations. They together improve the preservation of and responses to subtle forgery evidence and greatly improve quantization accuracy. Extensive experiments on six benchmark datasets and eleven deepfake detectors demonstrate that VIGUARDQ outperforms state-of-the-art quantization baselines. Specifically, we achieve **83.59%** AUC scores at a lower 13.4% compression rate, compared with **68.07%** achieved by the strongest baseline, PTQ4VM, at a 14.95% compression rate on the DFD dataset. **Real-world deployment on edge devices** further demonstrates our practical effectiveness. *Our code will be released upon acceptance.*

1 INTRODUCTION

Deepfake technologies are triggering a severe crisis in digital media authenticity, with manipulated facial biometric-related content posing profound threats to public trust and information dissemination (Goodfellow et al., 2020). Meanwhile, with the increasing reliance on edge devices for online meetings and payment authentication using facial biometrics, which may lead to serious financial and safety consequences, the demand for deepfake detection, especially video deepfake detection, on resource-limited devices such as laptops and smartphones has become urgent and critical (Mahendran et al., 2021). Compared to image deepfakes (Huang et al., 2025; Ba et al., 2024; Yan et al., 2024a; Tan et al., 2025), video deepfakes (Oorloff et al., 2024; Choi et al., 2024; Han et al., 2025) are more critical in edge scenarios: real-world applications such as video calls and continuous identity authentication inherently involve temporal streams rather than static images, requiring detection methods that can model cross-frame dynamics and temporal inconsistencies.

Nevertheless, despite the continuously improving accuracy of existing video deepfake detection approaches (Hashmi et al., 2025; Yermakov et al., 2026), most of them are computationally intensive and require high-end GPUs to perform the task. They are thus not suitable for deployment on resource-constrained edge devices for real-time applications. The huge gap between resource-intensive high detection accuracy and resource-constrained real-world deployment has become a critical bottleneck limiting the practical adoption of video deepfake forensics (Hinke-Navarro et al., 2025).

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.

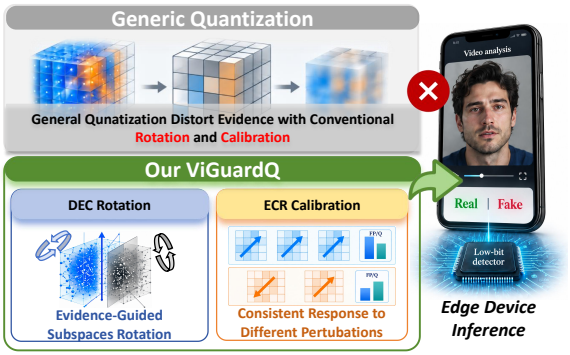

Figure 1: **ViGuardQ**, the first quantization of video deepfake detection for edge devices.

The Research community has recognized this real-world need. Relevant research on compression research were explored, such as Model Pruning (Yu & Xiang, 2023), Knowledge Distillation (Woo et al., 2022), Neural Architecture Search (Le & Woo, 2024), and Low-rank Approximation (Xu et al., 2024). But *compression methods remain impractical for resource-constrained edge devices* which all rely on full-precision weight representations with significant computational and memory overhead. More recent research on quantization then emerged to address such bottlenecks. However, *existing quantization methods remain limited for video deepfake detection* (Wu et al., 2025b; Li et al., 2022; Wu et al., 2025a; Chen et al., 2025; Tao et al., 2025; Zhang et al., 2026), as this task critically depends on fine-grained spatiotemporal forgery traces that are highly sensitive to quantization noise (Afchar et al., 2018). Existing quantization methods primarily focus on general representation fidelity with conventional calibration procedure and do not explicitly preserve the subtle forensic evidence and consistent decision responses required for reliable deepfake detection (Figure 1, top).

Hence, a dedicated quantization strategy for edge device deployable video deepfake detection should address two complementary requirements ***under a low-bit budget***: **R1). minimize** quantization distortion in discriminative spatial and temporal feature representations; and **R2). maintain** the robustness of the quantized detector through consistent responses under various perturbations.

In this paper, we present **ViGuardQ** (Figure 1), a quantization framework tailored for edge-device deployable video deepfake detection under a low-bit budget, integrating two complementary modules that address the challenges of forensic evidence distortion with **D**ual **E**vidence **C**onstrained (**DEC**) Rotation and loss of detector sensitivity to subtle forgery variations with **E**vidence-guided **C**onsistent **R**esponse (**ECR**) Calibration. Specifically, **1). DEC Rotation** identifies a decision-sensitive channel subspace from intra-frame and inter-frame differential decision gradients. It then optimizes orthogonal rotations within this subspace and its complement to obtain coordinates better suited to low-bit quantization. Since rounding and clipping errors depend on the chosen coordinates, these rotations can reduce quantization distortion while preserving the full-precision linear mapping through paired weight transformations. Spatially and temporally weighted reconstruction objectives prioritize errors along decision-sensitive directions, directing this reduction toward representations of subtle intra-frame forgery traces and inter-frame differences. **2). ECR Calibration**. For quantized video deepfake detection, maintaining response consistency with the FP model helps prevent low-bit errors from altering how subtle spatiotemporal forgery cues influence real/fake decisions. Feature reconstruction on original samples alone does not ensure this consistency. ECR Calibration addresses this gap through DEC-guided, temporally aligned and opposed perturbations that probe spatial evidence and cross-frame differences. It aligns the resulting signed changes in the FP and quantized detectors' real/fake margins, helping preserve consistent responses to forgery evidence.

Together, *DEC Rotation* and *ECR Calibration* address the challenge of using low-bit representations to faithfully capture the full-precision detector's spatiotemporal forensic behavior and preserve its consistent responses to subtle forgery variations. Extensive cross-evaluations on **6** widely used video deepfake benchmarks and **11** SoTA video deepfake detectors demonstrate our superiority. Specifically, we achieve **83.59%** AUC scores at a lower 13.4% compression rate, compared with **68.07%** achieved by the strongest baseline, PTQ4VM Zhu et al. (2026), at a 14.95% compression rate on the DFD dataset.

We summarize our contributions as follows:

- We propose ViGuardQ, the first quantization framework for video deepfake detection on edge devices, featuring novel *DEC Rotation* and *ECR Calibration* to maintain quantized detector's sensitivity to forgery-related evidence and link the calibration target to the detector's forensic behavior.

- We validate our approach on six benchmark datasets, demonstrating significant improvements in maintaining detection AUC scores compared to state-of-the-art quantization baselines, while achieving higher compression ratios and reduced peak memory usage.
- Deployment of VIGUARDQ in real-world scenarios on an Android-based smartphone system further demonstrates its practicality and real-time inference capability.

2 RELATED WORKS

Video Deepfake Detection Research on video deepfake detection mainly focuses on improving detection accuracy, from early methods based on handcrafted features (Radford et al., 2018; Mazaheri & Roy-Chowdhury, 2022; Liang et al., 2024), to mainstream learning-based methods, spanning spatiotemporal feature fusion (Krizhevsky et al., 2012; Simonyan & Zisserman, 2014; Rossler et al., 2019), 3D CNN (Panzino et al., 2024; Wang et al., 2021; Cui et al., 2025), and vision transformers (ViTs) (Diba et al., 2023; Usmani et al., 2025). Most of them are computationally expensive and parameter-heavy (Lanzino et al., 2024; Oorloff et al., 2024). Though (Luo et al., 2025) proposed a compressed 3D model via knowledge distillation, it suffered from significantly degraded detection accuracy. Notably, compared to image deepfake detection (Tao et al., 2025), video deepfake detection requires modeling temporal dependencies (Ciamarra et al., 2024; Yan et al., 2025) and spatial-temporal features, making it more complex and challenging for edge deployment, but existing works have rarely focused on dedicated model compression techniques for this task (Cui et al., 2025).

Model Compression and Quantization **Model compression** aims to address the computational and memory bottlenecks of deep learning models, such as Pruning (Yu & Xiang, 2023), Knowledge Distillation (Woo et al., 2022), Low Rank (Xu et al., 2024) and Neural Architecture Search (Le & Woo, 2024). But they inherently rely on full-precision weight representations, resulting in substantial memory overhead, and thus cannot be effectively deployed on edge devices. While **Quantization**, including Quantization-Aware Training (QAT) and Post-Training Quantization (PTQ), has emerged as an effective and lightweight approach for efficient edge deployment, by reducing the bit-width of model parameters and activations to significantly decrease memory usage and computational latency. QAT (Zhang, 2025; Jacob et al., 2018; Liu et al., 2021b) incorporates quantization into the training process, achieving high accuracy but requiring extensive retraining and large-scale datasets. While PTQ (Ma et al., 2024; Kuzmin et al., 2023; Zhang et al., 2025) performs quantization without retraining the entire model, relying only on a small amount of calibration data, making it more suitable for practical edge deployment scenarios. However, they cannot be directly applied to video deepfake detection, as the subtle forgery traces are highly vulnerable to quantization errors (Tao et al., 2025). The above quantization methods tend to smooth out subtle forgery traces during the compression process, leading to significant performance degradation (Lin et al., 2024). Some recent quantization methods target image deepfake detection (Tao et al., 2025), but they lack temporal feature preservation mechanisms for effective quantization of video deepfake detection.

3 METHOD

VIGUARDQ (Figure 2) addresses the significant challenge of preserving and responding to subtle forgery evidence under low-bit model quantization. At the representation level, low-bit quantization can distort subtle spatial and temporal forgery cues. *Dual-Evidence-Constrained* (DEC) Rotation addresses this by learning quantization-friendly coordinates within decision-sensitive subspaces and their complements, prioritizing the preservation of discriminative evidence. At the response level, feature preservation alone does not ensure that these cues influence predictions consistently after quantization. *Evidence-guided Consistent Response* (ECR) Calibration therefore aligns the FP and quantized detectors' decision changes under temporally aligned and opposed perturbations, helping retain consistent responses to variations in forgery evidence.

3.1 PROBLEM STATEMENT AND PRELIMINARIES

Problem Statement. We adopt post-training quantization (PTQ) to convert a pre-trained video deepfake detector $\mathcal{M}_{\mathrm{FP}}$ into a low-bit model $\mathcal{M}_Q$ using a small source calibration set, without

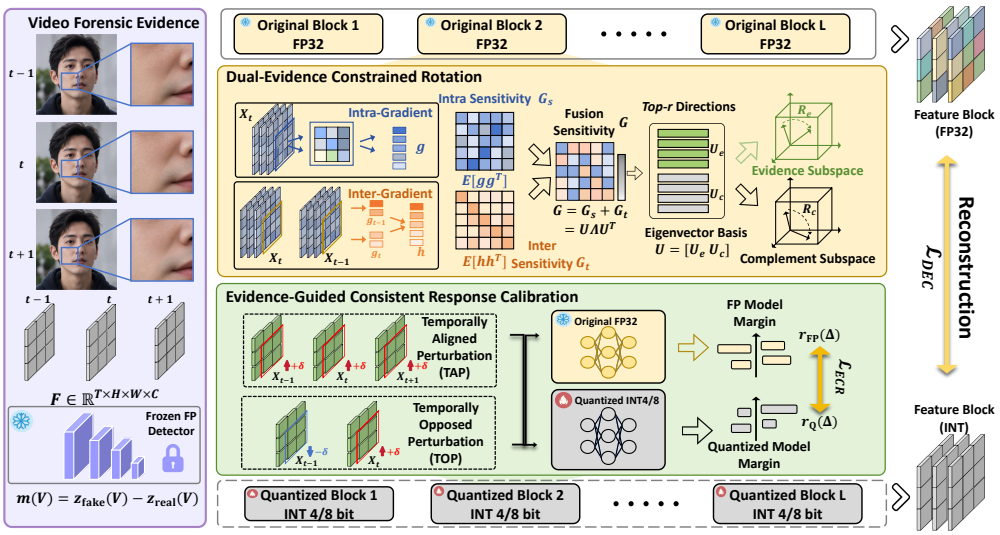

Figure 2: Overview of VIGUARDQ. **Dual-Evidence-Constrained (DEC) Rotation** aims to reduce quantization distortion with a decision-sensitive channel subspace. **Evidence-guided Consistent Response (ECR) Calibration** intends to maintain the quantized detector's consistent responses under various perturbations to subtle forgery traces.

retraining its original weights (Liu et al., 2021b). An input video is represented as $V = \{f_t\}_{t=1}^{T}$, where $f_t \in \mathbb{R}^{H\times W\times 3}$ denotes the t-th RGB frame. Following frame sampling, facial cropping, and spatial alignment, our goal is to reduce model storage and inference costs while preserving discriminative spatiotemporal forgery cues and the detector's sensitivity to these cues.

At a selected layer ℓ, the features are $F_\ell \in \mathbb{R}^{T_\ell\times P_\ell\times C_\ell}$, where the dimensions correspond to temporal positions, spatial locations, and channels, respectively. We use τ to index feature-level temporal positions, which may represent frames or tubelets. Reliable local correspondences p_τ between neighboring temporal positions are established during calibration.

Conventional Quantization. Following the PTQ formulation (Jacob et al., 2018), we define the integer code and its dequantized approximation for bit-width b and quantizer parameters $\theta = (s, \zeta)$ as follows:

$$q_{b,\theta}(x) = \mathrm{clip}\left(\lfloor x/s \rceil + \zeta,\, 0,\, 2^b - 1\right), \qquad \widetilde{\mathcal{Q}}_{b,\theta}(x) = s\big(q_{b,\theta}(x) - \zeta\big). \tag{1}$$

Here, $\lfloor\cdot\rceil$ denotes round-to-nearest. The operations apply elementwise, with quantizer parameters broadcast within predefined groups. During calibration, we refine quantization scales while keeping observer-initialized zero points and original detector weights fixed.

Rotation-based Quantization. Since rounding and clipping errors depend on the chosen coordinates, a suitable rotation can reduce quantization distortion (Ashkboos et al., 2024). The rotation acts only on channels and is not moved across LayerNorm, softmax, or positional operations. Specifically, F_ℓ is reshaped into $X \in \mathbb{R}^{N_\ell\times C_\ell}$, where $N_\ell = T_\ell P_\ell$. For a linear layer $Y = XW + b_0$, with $W \in \mathbb{R}^{C_\ell\times D_\ell}$ and $b_0 \in \mathbb{R}^{D_\ell}$, an orthogonal channel rotation $R \in \mathbb{R}^{C_\ell\times C_\ell}$ gives:

$$Z = XR, \qquad W_R = R^\top W, \qquad ZW_R + b_0 = XW + b_0. \tag{2}$$

Thus, the paired transformations preserve the full-precision linear mapping while changing the coordinates used for quantization. The quantized output is approximated by:

$$\widehat{Y} = \widetilde{\mathcal{Q}}_{b_a,\theta_a}(Z)\,\widetilde{\mathcal{Q}}_{b_w,\theta_w}(W_R) + b_0, \tag{3}$$

where b_a and b_w denote activation and weight bit-widths, and θ_a and θ_w are their respective quantizer parameters. The bias is broadcast across rows and left unquantized in this reference interface. However, **orthogonal rotation alone cannot ensure that quantization preserves subtle forgery cues or sensitivity to their changes**, motivating DEC Rotation and ECR Calibration.

3.2 DUAL-EVIDENCE-CONSTRAINED ROTATION

Low-bit quantization can distort subtle intra-frame forgery traces and inter-frame inconsistencies despite small overall reconstruction errors. DEC addresses this by jointly using the FP detector's sensitivities to intra-frame variations and inter-frame differences to identify a decision-sensitive channel subspace. It then optimizes orthogonal coordinates within this subspace and its complement, adapting representations to quantization while keeping sensitive directions separate.

Intra-frame and inter-frame differential sensitivity. Video deepfake detection relies on both subtle local forgery traces and cross-frame inconsistencies. Spatial sensitivity alone does not explicitly isolate sensitivity to temporal differences, whereas temporal differential sensitivity can overlook persistent spatial cues whose gradients largely cancel across neighboring frames. We therefore combine both statistics to identify decision-sensitive directions that capture complementary spatial and temporal evidence.

Specifically, decision margins provide the decision-level reference for sensitivity estimation and response calibration. The FP detector's video-level decision margin m_{FP} is defined as:

$$m_{\mathrm{FP}}(V) = z^{\mathrm{FP}}_{\mathrm{fake}}(V) - z^{\mathrm{FP}}_{\mathrm{real}}(V), \tag{4}$$

where the z terms denote class logits. For a single-logit classifier with the fake class as the positive class, the margin is the pre-sigmoid logit. Similarly, $m_Q(V)$ denotes the decision margin of the quantized detector.

Intra-frame differential sensitivity. Let $g_{\tau,p} = \nabla_{F_{\ell,\tau,p}} m_{\mathrm{FP}}(V)$ denote the column gradient restricted to the selected channel group. For a small local channel perturbation δ, the first-order margin change is approximately $g_{\tau,p}^{\top}\delta$. We summarize this intra-frame sensitivity through the gradient second-moment matrix:

$$G_s = \mathbb{E}\left[g_{\tau,p} g_{\tau,p}^{\top}\right]. \tag{5}$$

For any channel direction u, $u^{\top} G_s u$ measures the expected squared first-order response to a unit-amplitude perturbation along u.

Inter-frame differential sensitivity. To isolate sensitivity to inter-frame differences, consider matched locations $p_{\tau-1}$ and p_{τ}. Applying $-\delta/\sqrt{2}$ and $+\delta/\sqrt{2}$ at these locations changes their difference while preserving their mean. The resulting first-order margin change is approximately $h_{\tau,p_\tau}^{\top}\delta$, where:

$$h_{\tau,p_\tau} = \frac{g_{\tau,p_\tau} - g_{\tau-1,p_{\tau-1}}}{\sqrt{2}}, \qquad G_t = \mathbb{E}\left[h_{\tau,p_\tau} h_{\tau,p_\tau}^{\top}\right]. \tag{6}$$

Together, G_s captures spatial decision sensitivity, whereas G_t emphasizes sensitivity to temporal information. Note that both statistics are estimated directly from FP gradients at unperturbed source calibration features, without explicit perturbation evaluations. Real and fake clips receive equal total weight, and only valid temporal correspondences contribute to G_t. These statistics characterize the FP model's decision sensitivity rather than ground-truth forgery locations.

Normalization. To prevent either statistic from dominating solely because of its numerical scale, we normalize and combine the two symmetric positive semidefinite matrices:

$$\bar{G}_a = \frac{G_a}{\max(\mathrm{tr}(G_a), \eta_a)}, \quad a \in \{s, t\}, \qquad G = \alpha \bar{G}_s + (1-\alpha)\bar{G}_t, \tag{7}$$

where $\eta_a > 0$ is a fixed numerical floor and $\alpha \in [0, 1]$ balances the two sensitivity components. All settings are determined using source data only.

Remarks. For the eigen-decomposition $G = U\Lambda U^{\top}$, with eigenvalues in descending order, the leading r eigenvectors form the sensitive basis U_e, and the remaining eigenvectors form U_c. This selects the rank-r subspace with the largest combined sensitivity, with r fixed across basis comparisons. Since differential gradients are linear combinations of local gradients, their contribution is to reweight the selected directions toward temporal sensitivity, rather than necessarily expand the sampled gradient span.

Subspace-constrained block rotation. To adapt quantization coordinates without mixing decision-sensitive directions with their complement, DEC learns two separate orthogonal rotations within each channel group:

$$R = [U_e O_e, U_c O_c], \qquad O_e^\top O_e = I, \quad O_c^\top O_c = I. \tag{8}$$

This block structure ensures $\mathrm{span}(R_{:,1:r}) = \mathrm{span}(U_e)$, preserving the selected sensitive subspace while allowing its internal coordinates to adapt to quantization. Activation quantization groups do not cross the partition boundary, enabling separate scales for sensitive and complementary components without increasing their bit-width. The internal rotations are initialized to identity and optimized using an orthogonality-preserving Cayley parameterization (Wen & Yin, 2013). We apply rotations independently within channel groups to limit computational overhead. Both subspace partitions use the same bit-width, with separate quantization scales.

Evidence-weighted quantization reconstruction. The previous analysis identifies decision-sensitive subspaces and constrains their rotations. We further use such sensitivity information to prioritize reconstruction errors that affect forgery discrimination, guiding the optimization of rotations and quantization scales.

Let H_b and $\widehat{H}_b$ denote the FP and quantized block outputs in the same coordinates. Define:

$$e_{\tau,p} = \widehat{H}_{b,\tau,p} - H_{b,\tau,p}, \qquad d_{\tau,p_\tau} = e_{\tau,p_\tau} - e_{\tau-1,p_{\tau-1}}. \tag{9}$$

Here, d_{τ,p_τ} measures distortion of matched inter-frame differences, preserving temporal evidence rather than encouraging temporal smoothing. With sensitivity matrices estimated from FP model gradients at this output interface, we define:

$$\mathcal{L}_s = \mathbb{E}\left[e_{\tau,p} \bar{G}_s^{\mathrm{out}} e_{\tau,p}^\top\right], \qquad \mathcal{L}_t = \mathbb{E}\left[d_{\tau,p_\tau} \bar{G}_t^{\mathrm{out}} d_{\tau,p_\tau}^\top\right]. \tag{10}$$

These objectives penalize errors more strongly along locally and temporally sensitive directions, coupling the selected quantization coordinates with evidence-aware reconstruction.

We combine them with overall feature reconstruction and video-output alignment:

$$\mathcal{L}_{\mathrm{DEC}} = \widetilde{\mathcal{L}}_0 + \lambda_s \widetilde{\mathcal{L}}_s + \lambda_t \widetilde{\mathcal{L}}_t, \tag{11}$$

where $\mathcal{L}_0$ is the standard feature reconstruction loss and $\widetilde{\mathcal{L}}_k = \mathcal{L}_k / c_k$ uses fixed positive source reference scales. Only internal rotations and quantization scales are optimized; original weights, subspace bases, FP model statistics and targets, and zero points remain fixed. Rotations are then frozen for ECR.

3.3 Evidence-guided Consistent Response Calibration

Consistent responses to subtle forgery cues after quantization make the quantized detector more reliable and robust, a property that feature reconstruction on original calibration samples alone cannot ensure. We thus propose *Evidence-guided Consistent Response* (ECR) Calibration to reduce this discrepancy between the FP and quantized detectors. ECR introduces *Temporally Aligned Perturbation* (TAP) and *Temporally Opposed Perturbation* (TOP) to reveal response discrepancies under complementary temporal changes. It matches the resulting signed real/fake margin changes between FP and quantized detectors, promoting response consistency beyond the original samples.

To focus calibration on decision-relevant changes, we construct perturbation directions from DEC Rotation's selected subspace:

$$K_a = U_e^\top \bar{G}_a U_e, \qquad u_a = U_e v_a, \qquad a \in \{s, t\}, \tag{12}$$

where v_a is a unit eigenvector associated with a leading eigenvalue of K_a. The resulting unit directions u_s and u_t guide TAP and TOP, respectively, within the DEC Rotation-selected subspace.

Temporally Aligned Perturbation (TAP). To assess response consistency under same-direction changes across frames, TAP applies the same channel perturbation along a matched trajectory. At the input of layer ℓ, let $\mathcal{W}$ denote a short temporal window and p_τ the matched spatial location at temporal position τ. Define:

$$[P_s]_{\tau,p_\tau,:} = \frac{u_s^\top}{\sqrt{|\mathcal{W}|}}, \qquad \tau \in \mathcal{W}, \tag{13}$$

with zeros elsewhere. TAP applies identical changes at matched positions across the temporal window, preserving differences between adjacent matched features within the window, though differences across its boundaries may change.

Temporally Opposed Perturbation (TOP). To test sensitivity to cross-frame differences independently of the pair's shared component, TOP applies equal and opposite perturbations at neighboring matched locations:

$$[P_t]_{\tau,p_\tau,:} = \frac{u_t^\top}{\sqrt{2}}, \qquad [P_t]_{\tau-1,p_{\tau-1},:} = -\frac{u_t^\top}{\sqrt{2}}, \tag{14}$$

with zeros elsewhere. This changes the matched features' difference while preserving their mean, complementing TAP's temporally shared changes.

Both perturbation templates have unit Frobenius norm. Thus, $\Delta_a = \varepsilon_\ell P_a$, for $a \in \{s, t\}$, gives TAP and TOP equal perturbation energy at the same interface and amplitude.

Evidence-response Consistency. To preserve how forgery evidence influences the detector's decisions, ECR Calibration matches the signed decision changes induced by TAP and TOP, extending calibration beyond predictions on unperturbed videos. For $M \in \{\mathrm{FP}, Q\}$, let $F_\ell^M(V)$ denote the features produced by model M at interface ℓ, and let ϕ_ℓ^M map them to the video-level real/fake margin. Both models receive the same perturbation Δ in the original channel coordinates; for the quantized model, it is inserted before rotation and activation quantization.

We define each model's signed response relative to its own unperturbed prediction:

$$\boldsymbol{r}_{M,\ell}(V, \Delta) = \left[\phi_\ell^M(F + \sigma\Delta) - \phi_\ell^M(F)\right]_{\sigma=+1,-1}^\top \tag{15}$$

Using each model's own features includes the effects of preceding quantized layers. Rounding and clipping remain active in the quantized forward pass, so these finite responses capture sensitivity changes caused by quantization thresholds. Matching both perturbation polarities constrains the direction and magnitude of the response, rather than enforcing invariance to evidence changes.

For TAP ($a = s$) and TOP ($a = t$), the response-matching loss is

$$\mathcal{L}_{\mathrm{resp}}^a = \mathbb{E}_{V,\ell,\Delta_a}\left[\frac{\|\boldsymbol{r}_{Q,\ell}(V, \Delta_a) - \boldsymbol{r}_{\mathrm{FP},\ell}(V, \Delta_a)\|_2^2}{\sigma_{a,\ell}^2 + \epsilon_r}\right], \tag{16}$$

where $\sigma_{a,\ell}^2$ is the fixed mean squared FP response norm estimated from source reference clips and perturbations, and $\epsilon_r > 0$ ensures numerical stability.

To target sensitivity discrepancies that remain after DEC, we select the perturbation with the largest response mismatch from a finite candidate pool:

$$\Delta_a^\star = \arg\max_{\Delta \in \mathcal{P}_a(V,\ell)} \|\boldsymbol{r}_{Q,\ell}(V, \Delta) - \boldsymbol{r}_{\mathrm{FP},\ell}(V, \Delta)\|_2^2. \tag{17}$$

Calibration mixes these selected perturbations with uniformly sampled candidates, concentrating on poorly preserved responses while maintaining coverage. Selection is not differentiated, and all candidate evaluations are included in the calibration cost.

The ECR objective combines the two response-matching losses:

$$\mathcal{L}_{\mathrm{ECR}} = \gamma_s \mathcal{L}_{\mathrm{resp}}^s + \gamma_t \mathcal{L}_{\mathrm{resp}}^t, \tag{18}$$

3.4 Overall calibration and reconstruction.

Our calibration extends conventional feature reconstruction by preserving both evidence-sensitive representations and more robust consistent responses under different perturbations. **DEC Rotation** first identifies sensitive subspaces from FP spatial and temporal gradients, then jointly optimizes their internal rotations and quantization scales through evidence-weighted reconstruction. With rotations fixed, **ECR Calibration** further refines the scales by matching FP and quantized responses to TAP and TOP, emphasizing larger response mismatches. Note that the original detector weights remain frozen throughout, and the FP model and perturbation branches are removed after calibration, ensuring that ECR introduces no additional inference overhead.

Table 1: Cross-dataset evaluations on the DFD Zi et al. (2020) testing dataset, where all quantized detectors are trained on FF++ Rossler et al. (2019). The best results are highlighted in **bold**. More Results on dataset CDF-v1, CDF-v2, DFDC, DFDCP, DF40 are provided in Table 4 and Table 5 in the Appendix.

Detectors	Type	Storage Cost (MB)						Compression Rate (%)					Detection AUC (%)					
		Quantization Methods						Quantization Methods					Quantization Methods					
		BRECQ ICLR'21	Adalog ECCV'24	FIMA-Q CVPR'25	PTQ4VM ICLR'26	Ours	GT	BRECQ ICLR'21	Adalog ECCV'24	FIMA-Q CVPR'25	PTQ4VM ICLR'26	Ours	BRECQ ICLR'21	Adalog ECCV'24	FIMA-Q CVPR'25	PTQ4VM ICLR'26	Ours	GT
DeiT-Tiny	Backbones	13.8	10.7	9.9	7.9	**5.8**	23.7	58.2	45.1	41.8	33.3	**24.5**	55.4 (-20.8)	59.2 (-17.0)	58.1 (-18.1)	68.8 (-7.4)	**76.2**	84.2
DeiT-Small		28.5	16.9	15.1	13.2	**11.7**	84.1	33.9	20.1	18.0	15.7	**13.9**	53.1 (-26.0)	57.0 (-22.1)	60.5 (-18.6)	62.3 (-16.8)	**79.1**	85.0
DeiT-Base		83.4	40.8	34.7	34.3	**33.9**	330.2	25.3	12.4	10.5	10.4	**10.3**	62.5 (-21.0)	61.6 (-21.9)	66.8 (-16.7)	67.2 (-16.3)	**83.5**	88.8
ViT-Small		29.8	18.4	16.7	15.9	**11.8**	84.1	35.4	21.9	19.9	18.9	**14.0**	59.7 (-24.5)	62.9 (-21.3)	68.0 (-16.2)	69.0 (-15.2)	**84.2**	91.3
ViT-Base		87.9	37.2	34.9	35.8	**34.5**	330.2	26.6	11.3	10.6	10.8	**10.4**	60.8 (-24.1)	63.5 (-21.4)	67.7 (-17.2)	67.0 (-17.9)	**84.9**	93.7
Swin-Small		51.9	28.3	24.8	22.2	**21.9**	189.2	27.4	15.0	13.1	11.7	**11.6**	59.8 (-24.3)	58.5 (-25.6)	70.0 (-14.1)	73.6 (-10.5)	**84.1**	91.8
Swin-Base		82.1	41.9	35.2	**34.3**	35.6	334.8	24.5	12.5	10.5	**10.2**	10.6	55.2 (-30.3)	60.6 (-24.9)	67.4 (-18.1)	69.3 (-16.2)	**85.5**	93.2
LSDA	CVPR'24	66.5	36.1	31.2	**29.2**	30.8	253.2	26.3	14.3	12.3	**11.5**	12.2	56.6 (-30.2)	61.0 (-25.8)	64.7 (-22.1)	68.4 (-18.4)	**86.8**	94.3
NPR	CVPR'24	50.8	29.1	25.9	25.5	**23.9**	177.9	28.6	16.4	14.6	14.3	**13.4**	53.9 (-30.5)	59.9 (-24.5)	63.8 (-20.6)	71.0 (-13.4)	**84.4**	92.1
UCF	ICCV'23	31.4	19.8	17.6	14.7	**13.5**	91.6	34.3	21.6	19.2	16.0	**14.7**	47.1 (-38.0)	52.9 (-32.2)	56.7 (-28.4)	64.3 (-20.8)	**85.1**	91.2
Wavelet	WACV'25	143.7	69.2	**67.8**	69.9	71.8	615.0	23.4	11.3	**11.0**	11.4	11.7	47.8 (-37.9)	52.6 (-33.1)	57.9 (-27.8)	67.9 (-17.8)	**85.7**	92.0
Average	-	60.89	31.67	28.53	27.54	**26.84**	228.55	31.26	18.34	16.49	14.95	**13.40**	55.63 (-27.96)	59.06 (-24.53)	63.78 (-19.81)	68.07 (-15.52)	**83.59**	90.69

Table 2: Cross-dataset comparison with other compression methods on DFD, CDF-v1, and DFDC.

Methods	Venue&Year	Category	Edge Deployment	DFD			CDF-v1			DFDC		
				DeiT	ViT	Swin	DeiT	ViT	Swin	DeiT	ViT	Swin
GT	-	-	✗	88.8	93.7	93.2	87.1	89.6	91.7	75.6	77.8	77.4
ADD	AAAI'22	Knowledge Distillation	✗	51.8	51.5	51.0	50.0	50.2	51.0	39.2	39.8	40.0
X-Pruner	CVPR'23	Model Pruning	✗	62.1	63.5	65.0	69.9	71.0	71.5	53.1	54.0	54.1
GAC-FAS	CVPR'24	Neural Architecture Search	✗	52.8	53.1	53.3	47.0	51.9	53.1	42.1	46.0	43.5
One-bit	NIPS'24	Low-Rank	✗	55.6	57.2	55.5	52.1	51.5	51.7	44.5	45.2	45.7
DIVPrune	CVPR'25	Model Pruning	✗	69.3	71.2	70.5	73.0	73.5	75.0	58.2	55.7	60.1
Vit-linearizer	ICCV'25	Knowledge Distillation	✗	58.2	57.6	59.0	55.5	53.1	51.6	43.5	45.0	44.4
Ours	-	Quantization	✓	**83.5**	**84.9**	**85.5**	**80.4**	**83.0**	**83.1**	**66.6**	**69.8**	**68.5**

4 EXPERIMENT

6 Datasets: FaceForensics++ (Rossler et al., 2019), CelebDF (CDF) (Li et al., 2020), DeepfakeDetection (DFD) (Zi et al., 2020), and Deepfake Detection Challenge (DFDC) (Dolhansky et al., 2020), the preview version of DFDC (DFDCP) (Dolhansky et al., 2019), and DF40 (Yan et al., 2024b). We use a subset of the FF++ training partition as the calibration dataset to calibrate and fine-tune the quantized model. More details are provided in the Appendix A. **3 Evaluation Metrics** across three key dimensions: detection capability in metric Area Under the Curve (AUC), storage cost in MB, and compression ratio (%). More details are in Appendix C. **11 Video Deepfake Detectors:** 4 SOTA Video Deepfake Detection methods: LSDA (Yan et al., 2024a), NPR (Tan et al., 2024), UCF (Yan et al., 2023), and Wavelet (Baru et al., 2025), and 7 mainstream backbones: DeiT (Tiny, Small, Base) (Touvron et al., 2021), ViT (Small, Base) (Dosovitskiy, 2020), and Swin (Small, Base) (Liu et al., 2021a). **4 Quantization Baselines:** image-oriented quantization methods Adalog (Wu et al., 2024), and video-related quantization techniques, FIMA-Q (Wu et al., 2025b), BRECQ (Li et al., 2022), and PT4VM (Zhu et al., 2026). **Implementation Details** can be found in Appendix D.

4.1 RESULTS

Cross-dataset Evaluation of Quantization Performance Quantized models for each deepfake detector are trained on the FF++ dataset (Rossler et al., 2019) and evaluated on the CDF (v1, v2), DFD, DFDC, DFDCP, and DF40 datasets. Notice that the full-precision (FP) deepfake detection model, labeled "GT", serves as the upper bound for reference. In summary, VIGUARDQ achieves significantly higher detection performance with competitive storage costs and reduced compression rate. Table 1 presents a comprehensive cross-dataset evaluation of quantization methods on SoTA deepfake detectors on dataset DFD. For space reasons, experimental results on datasets CDF-v1, CDF-v2, DFDC, DFDCP, and DF40 are put in Appendix. Table 4 and Table 5. Visual analysis of attention maps and quantization ranges are presented in Fig 3.

Cross-Dataset Evaluation with Other Compression Methods Table 2 compares VIGUARDQ with other model compression techniques across multiple datasets and backbones. The results show

Table 3: Ablation studies of our module design measured on metric AUC.

Method	CDF-v1			CDF-v2			DFD			DFDC		
	DeiT-Base	ViT-Base	Swin-Base	DeiT-Base	ViT-Base	Swin-Base	DeiT-Base	ViT-Base	Swin-Base	DeiT-Base	ViT-Base	Swin-Base
✗ DEC, ✗ ECR	57.8	58.9	53.2	46.6	48.2	48.0	52.2	54.9	51.9	42.3	45.8	43.4
✓ DEC, ✗ ECR	72.2	68.5	73.0	66.5	69.2	71.3	73.6	71.9	75.6	60.9	62.7	64.1
✓ DEC, ✓ ECR (Ours)	**80.4**	**83.0**	**83.1**	**75.7**	**81.3**	**80.5**	**83.5**	**84.9**	**85.5**	**66.6**	**69.8**	**68.5**

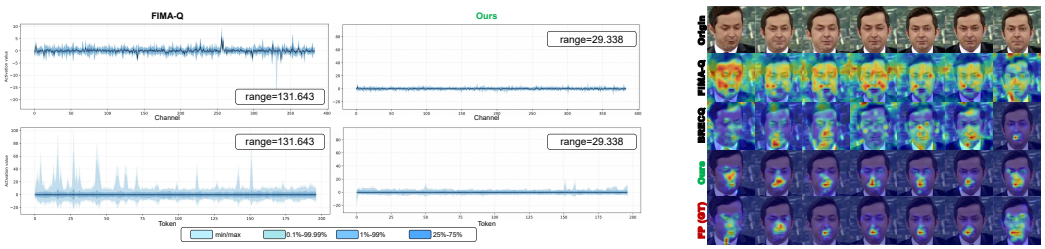

Figure 3: Left: Visualization of quantization ranges. Smaller color variations indicate narrower ranges and, at a fixed bit width, finer quantization resolution. Compared with conventional quantization, our DEC yields narrower channel ranges and tighter per-token min/max envelopes across tokens. Right: Visual comparison of salient forgery regions detected under quantization.

that our ViGuardQ achieves significantly higher detection accuracy than other compression methods. Moreover, as all other compression techniques require full-precision representations, only our approach enables effective deployment on resource-constrained edge devices.

Ablation Studies. We further perform ablation studies across all five datasets to assess the efficacy of each proposed module in ViGuardQ. As presented in Table 3, adding DEC substantially improves the baseline, while ECR provides further gains when applied after DEC, which validates the indispensable role of each component in our framework. Besides, additional ablation results on the DFDCP dataset (Dolhansky et al., 2020) are available in the Supplementary Materials.

More Discussions. Note that VIGUARDQ is designed to preserve the performance of the original detector as much as possible under low-bit quantization, rather than to improve its generalization or robustness. Consequently, the generalization capability and robustness to image degradations of the quantized detector are largely inherited from the original detector. Although improving these properties is important, it is beyond the scope of this paper. Additional results of robustness to image degradations and generalization to DF40 are provided in Appendix Table 7.

4.2 REAL-WORLD DEPLOYMENT ON EDGE DEVICES

We further deploy VIGUARDQ on Android mobile phones and conduct edge-device testing on Face Swap method (Zhu et al., 2021). During real-time inference, VIGUARDQ achieves an average latency of 90 ms per frame and consumes 92 mW of power while maintaining high detection accuracy. For broader real-world adoption, ViGuardQ can be further adapted to custom circuit boards (e.g., ASICs and FPGAs) for edge deployment scenarios. The proposed DEC can be aligned with the low-bit arithmetic capabilities of modern circuit designs. **Please refer to our supplemental demo video for real-world deployment on edge devices.**

5 CONCLUSION

We presented VIGUARDQ, a post-training quantization framework for video deepfake detection on edge devices. Specifically, DEC Rotation reduces distortion of discriminative within-frame cues and cross-frame differences through evidence-constrained rotations and weighted reconstruction. ECR Calibration complements representation preservation by aligning the FP and quantized detectors' decision responses to temporally aligned and opposed perturbations. Experiments across six benchmarks and eleven detectors, together with deployment on mobile devices, demonstrate the framework's effectiveness and practicality for resource-constrained video deepfake detection.

AI USE STATEMENT

In preparing this submission, the authors used generative AI tools to refine the language, improve readability, and enhance the visual presentation of teaser figures. All AI-assisted modifications were reviewed and verified by the authors, who take full responsibility for the final content, including the text, figures, claims, and associated artifacts.

REFERENCES

Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*, pp. 1–7. IEEE, 2018.

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L. Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. QuaRot: Outlier-free 4-bit inference in rotated LLMs. In *Advances in Neural Information Processing Systems*, 2024.

Zhongjie Ba, Qingyu Liu, Zhenguang Liu, Shuang Wu, Feng Lin, Li Lu, and Kui Ren. Exposing the deception: Uncovering more forgery clues for deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 719–728, 2024.

Lalith Bharadwaj Baru, Rohit Boddeda, Shilhora Akshay Patel, and Sai Mohan Gajapaka. Wavelet-driven generalizable framework for deepfake face forgery detection. In *Proceedings of the winter conference on applications of computer vision*, pp. 1661–1669, 2025.

Lei Chen, Yuan Meng, Chen Tang, Xinzhu Ma, Jingyan Jiang, Xin Wang, Zhi Wang, and Wenwu Zhu. Q-dit: Accurate post-training quantization for diffusion transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28306–28315, 2025.

Jongwook Choi, Taehoon Kim, Yonghyun Jeong, Seungryul Baek, and Jongwon Choi. Exploiting style latent flows for generalizing deepfake video detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1133–1143, 2024.

Andrea Ciamarra, Roberto Caldelli, and Alberto Del Bimbo. Temporal surface frame anomalies for deepfake video detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3837–3844, 2024.

Xinjie Cui, Yuezun Li, Ao Luo, Jiaran Zhou, and Junyu Dong. Forensics adapter: Adapting clip for generalizable face forgery detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19207–19217, 2025.

Ali Diba, Vivek Sharma, Mohammad Arzani, Luc Van Gool, et al. Spatio-temporal convolution-attention video network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 859–869, 2023.

Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019.

Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020.

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

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Yue-Hua Han, Tai-Ming Huang, Kai-Lung Hua, and Jun-Cheng Chen. Towards more general video-based deepfake detection through facial component guided adaptation for foundation model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22995–23005, 2025.

Ammarah Hashmi, Sahibzada Adil Shahzad, Chia-Wen Lin, Yu Tsao, and Hsin-Min Wang. Avtenet: A human-cognition-inspired audio-visual transformer-based ensemble network for video deepfake detection. *IEEE Transactions on Cognitive and Developmental Systems*, 2025.

Alejandro Hinke-Navarro, Mario Nieto-Hidalgo, Juan M Espín, and Juan E Tapia. Enhanced deep learning deepfake detection integrating handcrafted features. In *International Conference on Computer Analysis of Images and Patterns*, pp. 67–78. Springer, 2025.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. Sida: Social media image deepfake detection, localization and explanation with large multimodal model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28831–28841, 2025.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2704–2713, 2018.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Andrey Kuzmin, Markus Nagel, Mart Van Baalen, Arash Behboodi, and Tijmen Blankevoort. Pruning vs quantization: Which is better? *Advances in neural information processing systems*, 36: 62414–62427, 2023.

Romeo Lanzino, Federico Fontana, Anxhelo Diko, Marco Raoul Marini, and Luigi Cinque. Faster than lies: Real-time deepfake detection using binary neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3771–3780, 2024.

Binh M Le and Simon S Woo. Gradient alignment for cross-domain face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 188–199, 2024.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3207–3216, 2020.

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*, 2022.

Yachao Liang, Min Yu, Gang Li, Jianguo Jiang, Boquan Li, Feng Yu, Ning Zhang, Xiang Meng, and Weiqing Huang. Speechforensics: Audio-visual speech representation learning for face forgery detection. *Advances in Neural Information Processing Systems*, 37:86124–86144, 2024.

Li Lin, Xinan He, Yan Ju, Xin Wang, Feng Ding, and Shu Hu. Preserving fairness generalization in deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16815–16825, 2024.

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, 2021a.

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, 2021b.

Fei Luo, Anna Li, Jiguang He, Zitong Yu, Kaishun Wu, Bin Jiang, and Lu Wang. Improved multitask radar sensing via attention-based feature distillation and contrastive learning. *IEEE Transactions on Information Forensics and Security*, 2025.

Yuexiao Ma, Huixia Li, Xiawu Zheng, Feng Ling, Xuefeng Xiao, Rui Wang, Shilei Wen, Fei Chao, and Rongrong Ji. Outlier-aware slicing for post-training quantization in vision transformer. In *Forty-first International Conference on Machine Learning*, 2024.

Jagadish K Mahendran, Daniel T Barry, Anita K Nivedha, and Suchendra M Bhandarkar. Computer vision-based assistance system for the visually impaired using mobile edge artificial intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2418–2427, 2021.

Ghazal Mazaheri and Amit K Roy-Chowdhury. Detection and localization of facial expression manipulations. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1035–1045, 2022.

Trevine Oorloff, Surya Koppisetti, Nicolò Bonettini, Divyaraj Solanki, Ben Colman, Yaser Yacoob, Ali Shahriyari, and Gaurav Bharaj. Avff: Audio-visual feature fusion for video deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27102–27112, 2024.

Andrea Panzino, Simone Maurizio La Cava, Giulia Orrù, and Gian Luca Marcialis. Evaluating the integration of morph attack detection in automated face recognition systems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3827–3836, 2024.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018. URL `https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf`.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1–11, 2019.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28130–28139, 2024.

Chuangchuang Tan, Renshuai Tao, Huan Liu, Guanghua Gu, Baoyuan Wu, Yao Zhao, and Yunchao Wei. C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7184–7192, 2025.

Renshuai Tao, Ziheng Qin, Yifu Ding, Chuangchuang Tan, Jiakai Wang, and Wei Wang. Unlocking the potential of lightweight quantized models for deepfake detection. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 520–528, 2025.

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.

Shaheen Usmani, Sunil Kumar, and Debanjan Sadhya. Spatio-temporal knowledge distilled video vision transformer (stkd-vvit) for multimodal deepfake detection. *Neurocomputing*, 620:129256, 2025.

Zhibing Wang, Xin Li, Rongrong Ni, and Yao Zhao. Attention guided spatio-temporal artifacts extraction for deepfake detection. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 374–386. Springer, 2021.

Zaiwen Wen and Wotao Yin. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142:397–434, 2013. doi: 10.1007/s10107-012-0584-1.

Simon Woo et al. Add: Frequency attention and multi-view based knowledge distillation to detect low-quality compressed deepfake images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 122–130, 2022.

Junyi Wu, Zhiteng Li, Zheng Hui, Yulun Zhang, Linghe Kong, and Xiaokang Yang. Quantcache: Adaptive importance-guided quantization with hierarchical latent and layer caching for video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15035–15044, 2025a.

Zhuguanyu Wu, Jiaxin Chen, Hanwen Zhong, Di Huang, and Yunhong Wang. Adalog: Post-training quantization for vision transformers with adaptive logarithm quantizer. In *European Conference on Computer Vision*, pp. 411–427. Springer, 2024.

Zhuguanyu Wu, Shihe Wang, Jiayi Zhang, Jiaxin Chen, and Yunhong Wang. Fima-q: Post-training quantization for vision transformers by fisher information matrix approximation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14891–14900, 2025b.

Yuzhuang Xu, Xu Han, Zonghan Yang, Shuo Wang, Qingfu Zhu, Zhiyuan Liu, Weidong Liu, and Wanxiang Che. Onebit: Towards extremely low-bit large language models. *Advances in Neural Information Processing Systems*, 37:66357–66382, 2024.

Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. Ucf: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 22412–22423, 2023.

Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8984–8994, 2024a.

Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie Wang, Shouhong Ding, Yunsheng Wu, et al. Df40: Toward next-generation deepfake detection. *Advances in Neural Information Processing Systems*, 37:29387–29434, 2024b.

Zhiyuan Yan, Yandan Zhao, Shen Chen, Mingyi Guo, Xinghe Fu, Taiping Yao, Shouhong Ding, Yunsheng Wu, and Li Yuan. Generalizing deepfake video detection with plug-and-play: Video-level blending and spatiotemporal adapter tuning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12615–12625, 2025.

Andrii Yermakov, Jan Cech, Jiri Matas, and Mario Fritz. Deepfake detection that generalizes across benchmarks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 773–783, 2026.

Lu Yu and Wei Xiang. X-pruner: explainable pruning for vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24355–24363, 2023.

Jiaqi Zhang. Survey of quantization-aware training (qat) applications in deep learning quantization. In *Proceedings of the 2025 International Symposium on Artificial Intelligence and Computational Social Sciences*, pp. 431–442, 2025.

Jing Zhang, Zhikai Li, Chengzhi Hu, Xuewen Liu, and Qingyi Gu. Saq-sam: Semantically-aligned quantization for segment anything model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12547–12555, 2026.

Zhenwei Zhang, Fanhua Shang, Hongying Liu, Liang Wan, Wei Feng, and Yanming Hui. Qbasicvsr: Temporal awareness adaptation quantization for video super-resolution. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Tianrui Zhu, Houyuan Chen, Ruihao Gong, Michele Magno, Haotong Qin, and Kai Zhang. Post-training quantization for video matting. In *International Conference on Learning Representations*, volume 2026, pp. 62425–62442, 2026.

Yuhao Zhu, Qi Li, Jian Wang, Cheng-Zhong Xu, and Zhenan Sun. One shot face swapping on megapixels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4834–4844, 2021.

Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 2382–2390, 2020.

APPENDIX

A CALIBRATION

Post-training quantization (PTQ) uses a small representative calibration set to initialize and refine quantization parameters. For video deepfake detection, calibration must account for both quantization-induced distortion of discriminative features and changes in the detector's decision responses. Our **VIGUARDQ** addresses these objectives through DEC Rotation followed by ECR Calibration. The original detector weights remain frozen throughout both stages.

Calibration data and model preparation. We use a subset of the FaceForensics++ training partition as the source calibration set $\mathcal{D}_{\mathrm{cal}}$, without using target-domain evaluation data. Each video is processed through frame sampling, facial cropping, and spatial normalization to obtain a clip $V = \{f_t\}_{t=1}^{T}$. Reliable correspondences between neighboring feature-level temporal positions are established for the temporal objectives. Real and fake clips receive equal total weight when estimating sensitivity statistics.

We retain the original FP detector M_{FP} as a fixed reference and initialize M_Q from its weights. Rotation interfaces and fake-quantization modules are inserted at selected linear layers. Fake quantization includes rounding, clipping, and dequantization during calibration; low-bit integer representations are exported after optimization. FP parameter gradients are disabled, while gradients with respect to selected intermediate features remain available for sensitivity estimation.

DEC subspace construction. For each selected layer and channel group, we compute gradients of the FP model's video-level real/fake margin with respect to its input features. The local gradient second moments form G_s, while second moments of matched neighboring gradient differences form G_t. Both statistics are computed at the original calibration features, without explicit perturbed forward passes.

After trace normalization, the two matrices are combined using the weight α. The leading r eigenvectors define the sensitive basis U_e, and the remaining eigenvectors define its complement U_c. We initialize the internal rotations to identity and construct:

$$R_g = [U_{e,g}O_{e,g},\, U_{c,g}O_{c,g}], \qquad O_{e,g} = I, \quad O_{c,g} = I, \tag{19}$$

for channel group g. The groupwise matrices are assembled into a layer-level block-diagonal rotation R_ℓ. This construction permits coordinate adaptation within each subspace while preventing mixing between the sensitive and complementary partitions.

Quantizer initialization and DEC optimization. With the initial rotations inserted, each selected linear layer uses the paired transformations $Z = XR_\ell$ and $W_R = R_\ell^\top W$. We collect activation statistics with fake quantization bypassed and initialize activation and weight quantizers from the distributions of Z and W_R, respectively. Activation quantization groups do not cross the sensitive/complement boundary, allowing separate scales while retaining the same bit-width for both partitions. Observers are then disabled, and their initialized integer zero points remain fixed.

DEC jointly optimizes the internal rotations and positive quantization scales using evidence-weighted reconstruction. The local term prioritizes errors along decision-sensitive directions, while the temporal term penalizes distortion of matched cross-frame differences. The required sensitivity matrices are estimated separately at the corresponding block-output interfaces. Conventional feature reconstruction and video-output KL distillation provide additional constraints on overall fidelity.

Internal rotations are optimized through a Cayley parameterization, and straight-through surrogate gradients are used through fake quantization. The original weights, selected subspace bases, FP statistics and targets, and normalization constants remain fixed. Transformed weights W_R are recomputed whenever the rotations change; thus, the paired transformations preserve the FP linear mapping before quantization.

ECR perturbation construction. After DEC, we freeze all learned rotations. ECR constructs perturbation directions within the selected sensitive subspaces by projecting the FP sensitivity matrices

onto U_e and lifting their leading eigendirections back to the original channel coordinates. A fixed fraction of random unit directions within the same subspaces provides broader directional coverage.

Temporally Aligned Perturbation (TAP) applies identical channel changes along matched trajectories within short temporal windows. Temporally Opposed Perturbation (TOP) applies equal and opposite changes to neighboring matched features, preserving their mean. Both templates have unit Frobenius norm and are scaled by the same layer-specific amplitude for matched-energy comparisons. Amplitudes are source-validated multiples of a robust FP feature scale. Candidate perturbations, correspondences, and FP response normalization statistics remain fixed during ECR.

ECR response calibration. For each candidate perturbation, the FP and quantized models are evaluated at their respective original features and under both positive and negative perturbations. The same perturbation is applied in the original channel coordinates; in M_Q, it is inserted before the selected rotation and activation quantizer. Subtracting each model's original margin yields its signed response vector.

ECR refines only quantization scales by matching the FP and quantized response vectors for both TAP and TOP. Within a finite candidate pool, perturbations with larger response mismatches are mixed with uniformly sampled candidates to balance targeted correction and coverage. Candidate selection is not differentiated, and all candidate evaluations are included in calibration cost. The response objectives are combined with reconstruction on original calibration samples and video-output KL distillation. Throughout this stage, rotations, original weights, zero points, and FP reference quantities remain fixed.

Low-bit export and deployment. After calibration, the final transformed weights are quantized and stored as integer codes with their quantizer metadata. The deployed model retains the learned channel rotations and calibrated activation quantizers. The FP reference model, observers, gradient computations, and perturbation-evaluation branches are removed. Consequently, ECR introduces no additional inference operations, while DEC's rotation cost is included in deployment measurements. The complete procedure is summarized in Algorithm 1.

B Mathematical Details

This section explains the sensitivity-based subspace selection in DEC, the role of orthogonal rotations in quantization, and the finite-response constraints used by ECR.

B.1 Sensitivity-Based Subspace Selection

DEC combines within-frame and cross-frame differential sensitivity to select channel directions relevant to the FP detector's decision. Recall that

$$G_s = \mathbb{E}[gg^\top], \qquad G_t = \mathbb{E}[hh^\top], \tag{20}$$

where g is the within-frame decision gradient and h is the normalized difference between gradients at matched neighboring positions. For any channel direction v,

$$v^\top G_s v = \mathbb{E}[(g^\top v)^2] \geq 0, \qquad v^\top G_t v = \mathbb{E}[(h^\top v)^2] \geq 0. \tag{21}$$

Thus, both matrices are symmetric positive semidefinite. Their quadratic forms measure the expected squared first-order decision response to the corresponding perturbations.

Positive normalization and nonnegative mixing preserve this property:

$$\bar{G}_a = \frac{G_a}{\max(\mathrm{tr}(G_a), \eta_a)}, \qquad G = \alpha \bar{G}_s + (1-\alpha)\bar{G}_t, \tag{22}$$

where $a \in \{s, t\}$, $\eta_a > 0$, and $\alpha \in [0, 1]$. Let $\lambda_1 \geq \cdots \geq \lambda_c \geq 0$ be the eigenvalues of G for a channel group of width c. The leading r eigenvectors form a basis U_e satisfying

$$U_e \in \arg\max_{U^\top U = I_r} \mathrm{tr}(U^\top G U), \qquad \max_{U^\top U = I_r} \mathrm{tr}(U^\top G U) = \sum_{i=1}^{r} \lambda_i. \tag{23}$$

Algorithm 1 VIGUARDQ: DEC and ECR Calibration for Video Deepfake Detection

Input: Pre-trained FP detector M_{FP}; calibration set $\mathcal{D}_{\mathrm{cal}}$; clip length T; selected layers $\mathcal{S}$; bit-widths b_a, b_w; channel-group width c_g; subspace rank r; iteration budgets $N_{\mathrm{DEC}}, N_{\mathrm{ECR}}$
Output: Low-bit detector M_Q for edge deployment
Variables: Sensitivity matrices G_s, G_t; subspace bases U_e, U_c; internal rotations O_e, O_c; layer rotations R_ℓ; quantization scales s; zero points ζ; TAP/TOP perturbations Δ_s, Δ_t

Initialize M_Q from M_{FP} with rotation interfaces and fake-quantization modules
Freeze original detector weights; retain FP feature gradients for sensitivity estimation
Preprocess calibration clips and establish valid temporal correspondences
DEC: Construct evidence-guided coordinates
for each selected layer $\ell \in \mathcal{S}$ **do**
  for each channel group g **do**
    Compute FP margin gradients and matched temporal gradient differences
    Estimate G_s, G_t with balanced real/fake weights; normalize and combine them into G
    Obtain sensitive basis U_e and complement U_c from the eigendecomposition of G
    Initialize $O_e = I$, $O_c = I$ and set $R_g = [U_e O_e, U_c O_c]$
  end for
  Assemble R_ℓ from groupwise rotations and set $W_R = R_\ell^\top W$
end for
Collect rotated activation and weight statistics with fake quantization bypassed
Initialize scales and zero points; disable observers, fix zero points, and enable fake quantization

Estimate FP block-output sensitivities and fixed reconstruction reference scales
DEC: Optimize rotations and quantizers
for $i = 1, \ldots, N_{\mathrm{DEC}}$ **do**
  Sample calibration clips and compute quantized outputs
  Compute spatially and temporally weighted reconstruction, base reconstruction, and output KL
  Update internal rotations and positive quantization scales using straight-through surrogate gradients
  Recompute transformed weights from updated rotations
end for
Freeze all learned rotations
ECR: Calibrate decision responses
Construct fixed TAP/TOP candidate pools within the selected sensitive subspaces
Set matched-energy amplitudes and estimate fixed FP response-energy normalization statistics
for $i = 1, \ldots, N_{\mathrm{ECR}}$ **do**
  Sample clips, a selected layer, and TAP/TOP candidates
  Evaluate FP and quantized margins under $0, +\Delta, -\Delta$ at each model's own features
  Form signed responses relative to each model's original margin
  Mix maximum-mismatch candidates with uniformly sampled candidates
  Compute TAP/TOP response losses, base reconstruction, and output KL
  Update positive quantization scales only; keep rotations, original weights, and zero points fixed
end for
Export the calibrated model
Store quantized transformed weights as integer codes with quantizer metadata
Retain learned rotations and calibrated activation quantizers; remove calibration-only branches
return M_Q

Consequently, DEC selects the rank-r subspace that captures the largest aggregate sensitivity under the chosen mixture. This criterion concerns the FP model's decision behavior, rather than activation energy or ground-truth forgery localization.

The temporal statistic contributes through a different weighting of channel directions. Writing g_+ and g_- for gradients at matched neighboring positions gives

$$G_t = \frac{1}{2}\mathbb{E}\left[g_+ g_+^\top + g_- g_-^\top - g_+ g_-^\top - g_- g_+^\top\right]. \tag{24}$$

The cross terms distinguish directions with similar neighboring responses from those with contrasting responses. Although each differential gradient lies within the span of its endpoint gradients, these terms can change the leading eigenspace. Holding r fixed therefore allows the temporal ablation to evaluate changes in subspace selection without increasing its dimensionality. Whether this selection improves detection accuracy is assessed empirically.

B.2 Orthogonal Coordinates and Quantization Error

Within each channel group, DEC constructs:

$$R_g = [U_e O_e, U_c O_c], \qquad O_j = (I - A_j)(I + A_j)^{-1}, \qquad A_j^\top = -A_j, \tag{25}$$

where $j \in \{e, c\}$ and $[U_e, U_c]$ is an orthogonal basis.

For real skew-symmetric A_j, the matrix $I + A_j$ is invertible. $(I + A_j)x = 0$ implies:

$$0 = x^\top (I + A_j)x = \|x\|_2^2, \tag{26}$$

and hence $x = 0$. Since $I - A_j$ and $I + A_j$ commute, the Cayley transform satisfies $O_j^\top O_j = I$. Therefore, both R_g and the layer-level block-diagonal rotation R are orthogonal.

Moreover, internal rotations thus adapt the coordinates within the selected subspace while preserving its separation from the complement:

$$\mathrm{span}(U_e O_e) = \mathrm{span}(U_e). \tag{27}$$

Let X denote the feature matrix obtained by flattening the temporal and spatial dimensions, and consider $Y = XW + b_0$. Before quantization:

$$(XR)(R^\top W) + b_0 = XW + b_0. \tag{28}$$

The paired transformations preserve the original linear mapping. To characterize the error introduced by quantization, define:

$$E_a = \widetilde{\mathcal{Q}}_{b_a,\theta_a}(XR) - XR, \qquad E_w = \widetilde{\mathcal{Q}}_{b_w,\theta_w}(R^\top W) - R^\top W. \tag{29}$$

With the bias unchanged, the output error becomes:

$$\widehat{Y} - Y = XRE_w + E_a R^\top W + E_a E_w. \tag{30}$$

Using orthogonal invariance and submultiplicativity yields:

$$\|\widehat{Y} - Y\|_F \le \|X\|_2 \|E_w\|_F + \|E_a\|_F \|W\|_2 + \|E_a\|_F \|E_w\|_2, \tag{31}$$

where $\|\cdot\|_2$ denotes the spectral norm for matrices.

Although rotation preserves the FP mapping, the quantization errors E_a and E_w depend on the chosen coordinates. DEC exploits this dependence by optimizing rotations through evidence-weighted reconstruction, prioritizing errors along decision-relevant directions. Orthogonality alone does not guarantee smaller quantization error or unchanged predictions.

For completeness, an independently reconstructed activation: $\widehat{X} = \widetilde{\mathcal{Q}}_{b_a,\theta_a}(XR)R^\top$ satisfies

$$\|\widehat{X} - X\|_F = \|E_a R^\top\|_F = \|E_a\|_F. \tag{32}$$

Thus, inverse rotation preserves the error norm rather than eliminating quantization error; it is not required at the paired linear interface above.

B.3 Finite-Response Consistency

For either TAP or TOP, ECR evaluates both perturbation polarities. Suppressing the video and layer indices, define:

$$a_M = m_M^+ - m_M^0, \qquad b_M = m_M^- - m_M^0, \qquad M \in \{\mathrm{FP}, Q\}. \tag{33}$$

The odd and even components of these finite responses are

$$o_M = \frac{a_M - b_M}{2}, \qquad c_M = \frac{a_M + b_M}{2}. \tag{34}$$

Their matching errors satisfy the exact identity

$$(a_Q - a_{\mathrm{FP}})^2 + (b_Q - b_{\mathrm{FP}})^2 = 2(o_Q - o_{\mathrm{FP}})^2 + 2(c_Q - c_{\mathrm{FP}})^2. \tag{35}$$

Matching both polarities therefore constrains both the component that reverses with the perturbation and the component shared by the two responses. These constraints apply separately to TAP and TOP, which specify different temporal perturbation structures.

To interpret these components, consider a sufficiently smooth FP margin function m and a unit-norm feature direction P. With $m_\pm = m(F \pm \varepsilon P)$ and $m_0 = m(F)$, Taylor expansion gives:

$$\frac{m_+ - m_-}{2\varepsilon} = \langle \nabla m(F), P \rangle + O(\varepsilon^2), \tag{36}$$

$$\frac{m_+ + m_- - 2m_0}{\varepsilon^2} = \langle P, H_m(F) P \rangle + o(1), \tag{37}$$

where $H_m(F)$ is the Hessian with respect to the vectorized features. The first remainder holds, for example, when third derivatives are locally bounded; the second requires twice continuous differentiability.

These relations connect the odd response to directional first-order behavior and the even response to directional curvature. ECR nevertheless matches finite responses directly, retaining rounding and clipping in the quantized forward pass. It requires no smoothness assumption for the quantized model and does not recover its full Jacobian or Hessian. Moreover, the unscaled response loss does not equally weight derivative orders: at small amplitudes, the odd and even components are generally of orders ε and ε^2, respectively.

Finally, define the paired margin discrepancies:

$$d^\sigma = m_Q^\sigma - m_{\mathrm{FP}}^\sigma, \qquad \sigma \in \{0, +, -\}. \tag{38}$$

The response mismatch can then be written as

$$\boldsymbol{r}_Q - \boldsymbol{r}_{\mathrm{FP}} = \begin{bmatrix} d^+ - d^0 \\ d^- - d^0 \end{bmatrix}. \tag{39}$$

ECR therefore constrains how the FP–quantized margin discrepancy changes under perturbation, while the base calibration objective anchors predictions on the original samples. When $d^0 = 0$, response matching reduces to margin matching at the perturbed samples.

Each model is evaluated around its own interface features, so this identity remains valid when upstream quantization changes those features. Interpreting it as a difference between derivatives at a common feature point additionally requires identical interface inputs.

C Evaluation Metrics

For detection performance, we adopt the Area Under the Curve (AUC) as the primary metric to assess the model's ability to distinguish real from fake content across all datasets. For storage, we report the storage cost (MB) of each model to demonstrate the edge deployment feasibility of quantized models. For compression ratio, we calculate the ratio (%) between the original full-precision model and its quantized counterpart to quantify the extent of model compression.

D Implementation Details

We employ 7 mainstream detectors as the original backbones for model quantization, encompassing DeiT (Tiny, Small, Base) Touvron et al. (2021), ViT (Small, Base) Dosovitskiy (2020), and Swin (Small, Base) Liu et al. (2021a). All pre-trained detector parameters are initialized with pre-trained weights from the timm library. Following reconstruction-based PTQ approaches, 256 videos are randomly sampled from the FF++ dataset to serve as the calibration set for quantizing deepfake detectors. Consistent with established practices, our implemented quantization strategy utilizes per-channel quantization for weights and per-tensor quantization for activations.

Table 4: Cross on more datasets evaluations on detection AUC of quantization approaches on mainstream deepfake detection backbones and state-of-the-arts deepfake detectors. The best results are highlighted in **bold**.

Detectors	Type	CDF-v1 Li et al. (2020) (AUC%)						CDF-v2 Li et al. (2020) (AUC%)						DFDC Dolhansky et al. (2019) (AUC%)					
		Quantization Methods						Quantization Methods						Quantization Methods					
		BRECQ ICLR'21	Adalog ECCV'24	FIMA-Q CVPR'25	PTQ4VM ICLR'26	Ours	GT	BRECQ ICLR'21	Adalog ECCV'24	FIMA-Q CVPR'25	PTQ4VM ICLR'26	Ours	GT	BRECQ ICLR'21	Adalog ECCV'24	FIMA-Q CVPR'25	PTQ4VM ICLR'26	Ours	GT
DeiT-Tiny	Backbones	54.1 (-21.0)	59.9 (-15.2)	59.8 (-15.3)	62.2 (-12.9)	**75.1**	82.1	49.5 (-21.7)	51.4 (-19.8)	55.0 (-16.2)	52.8 (-18.4)	**71.2**	79.2	38.8 (-21.4)	41.1 (-19.1)	42.5 (-17.7)	47.8 (-12.4)	**60.2**	69.3
DeiT-Small		52.8 (-25.0)	55.8 (-22.0)	60.8 (-17.0)	64.1 (-13.7)	**77.8**	83.5	46.4 (-28.5)	54.0 (-20.9)	55.9 (-19.0)	55.7 (-19.2)	**74.9**	80.7	37.8 (-24.8)	41.8 (-20.8)	44.7 (-17.9)	49.0 (-13.6)	**62.6**	70.6
DeiT-Base		64.9 (-15.5)	62.6 (-17.8)	67.1 (-13.3)	69.3 (-11.1)	**80.4**	87.1	50.8 (-24.9)	52.1 (-23.6)	55.4 (-20.3)	57.7 (-18.0)	**75.7**	83.7	44.4 (-22.2)	43.3 (-23.3)	49.8 (-16.8)	51.7 (-14.9)	**66.6**	75.6
ViT-Small		62.0 (-20.9)	63.6 (-19.3)	68.6 (-14.3)	71.5 (-11.4)	**82.9**	87.8	48.8 (-31.9)	50.5 (-30.2)	57.8 (-22.9)	70.3 (-10.4)	**80.7**	85.7	44.7 (-24.6)	44.5 (-24.8)	49.5 (-19.8)	55.8 (-13.5)	**69.3**	76.2
ViT-Base		62.8 (-20.2)	63.9 (-19.1)	67.9 (-15.1)	70.3 (-12.7)	**83.0**	89.6	52.1 (-29.2)	51.1 (-30.2)	61.8 (-19.5)	62.9 (-18.4)	**81.3**	87.8	46.9 (-22.9)	46.8 (-23.0)	54.6 (-15.2)	55.9 (-13.9)	**69.8**	77.8
Swin-Small		61.6 (-21.2)	59.2 (-23.6)	70.5 (-12.3)	69.6 (-13.2)	**82.8**	89.7	49.4 (-29.4)	47.8 (-31.0)	60.1 (-18.7)	60.9 (-17.9)	**78.8**	86.1	47.6 (-19.5)	45.0 (-22.1)	55.6 (-11.5)	53.7 (-13.4)	**67.1**	75.0
Swin-Base		57.1 (-26.0)	61.4 (-21.7)	67.6 (-15.5)	70.0 (-13.1)	**83.1**	91.7	51.9 (-28.6)	48.9 (-31.6)	55.7 (-24.8)	63.7 (-16.8)	**80.5**	88.8	43.2 (-25.3)	45.2 (-23.3)	53.1 (-15.4)	54.6 (-13.9)	**68.5**	77.4
LSDA	CVPR'24	57.9 (-27.3)	61.9 (-23.3)	65.5 (-19.7)	74.1 (-11.1)	**85.2**	92.6	50.5 (-31.2)	53.1 (-28.6)	57.6 (-24.1)	74.9 (-6.8)	**81.7**	88.9	44.5 (-23.9)	48.6 (-19.8)	52.0 (-16.4)	57.0 (-11.4)	**68.4**	77.8
NPR	CVPR'24	56.0 (-28.7)	60.8 (-23.9)	63.1 (-21.6)	72.9 (-11.8)	**84.7**	91.5	50.2 (-30.6)	52.6 (-28.2)	55.3 (-25.5)	71.0 (-9.8)	**80.8**	87.1	42.1 (-26.5)	44.3 (-24.3)	53.8 (-14.8)	58.3 (-10.3)	**68.6**	80.0
UCF	ICCV'23	49.2 (-27.8)	53.7 (-23.3)	57.2 (-19.8)	67.1 (-9.9)	**77.0**	83.8	50.0 (-25.7)	46.0 (-29.7)	53.5 (-22.2)	60.8 (-14.9)	**75.7**	81.2	41.4 (-28.4)	44.6 (-25.2)	57.6 (-12.2)	56.3 (-13.5)	**69.8**	78.2
Wavelet	WACV'25	49.0 (-27.1)	53.9 (-22.2)	58.0 (-18.1)	61.5 (-14.6)	**76.1**	81.5	52.4 (-25.0)	47.5 (-29.9)	56.1 (-21.3)	65.9 (-11.5)	**77.4**	81.8	48.3 (-23.1)	45.7 (-25.7)	55.4 (-16.0)	58.9 (-12.5)	**71.4**	79.5

Table 5: Cross-dataset evaluations on the DFDCP Dolhansky et al. (2019) and DF40 Yan et al. (2024b) testing datasets, where all quantized detectors are trained on FF++ Rossler et al. (2019). The best results among quantized methods are highlighted in bold. Values in parentheses indicate the AUC difference between each baseline and our method (baseline AUC minus ours, in percentage points). Our method remains comparable to existing quantization approaches in terms of Compression Rate (%) and Storage Cost (MB), while achieving state-of-the-art (SOTA) detection AUC (%).

Detectors	Type	DFDCP Dolhansky et al. (2019) (AUC%)						DF40 Yan et al. (2024b) (AUC%)					
		Quantization Methods						Quantization Methods					
		BRECQ ICLR'21	Adalog ECCV'24	FIMA-Q CVPR'25	PTQ4VM ICLR'26	Ours	GT	BRECQ ICLR'21	Adalog ECCV'24	FIMA-Q CVPR'25	PTQ4VM ICLR'26	Ours	GT
DeiT-Tiny	Backbones	47.1(-21.7)	48.2(-20.6)	59.0(-9.8)	60.3(-8.5)	**68.8**	77.7	33.1(-15.2)	35.8(-12.5)	37.9(-10.4)	38.0(-10.3)	**48.3**	52.5
DeiT-Small		48.3(-23.0)	56.8(-14.5)	60.4(-10.9)	62.1(-9.2)	**71.3**	78.7	34.1(-14.6)	35.6(-13.1)	36.9(-11.8)	37.1(-11.6)	**48.7**	55.3
DeiT-Base		49.8(-24.7)	59.4(-15.1)	62.4(-12.1)	64.4(-10.1)	**74.5**	82.5	34.5(-13.9)	36.2(-12.2)	38.7(-9.7)	38.5(-9.9)	**48.4**	53.7
ViT-Small		50.6(-26.0)	59.2(-17.4)	62.9(-13.7)	68.7(-7.9)	**76.6**	82.4	37.6(-16.1)	36.3(-17.4)	40.5(-13.2)	40.3(-13.4)	**53.7**	57.1
ViT-Base		52.0(-25.1)	62.0(-15.1)	65.7(-11.4)	68.3(-8.8)	**77.1**	86.0	36.8(-15.5)	37.5(-14.8)	41.0(-11.3)	40.2(-12.1)	**52.3**	57.5
Swin-Small		50.8(-24.9)	60.4(-15.3)	63.7(-12.0)	66.1(-9.6)	**75.7**	84.0	33.8(-14.2)	35.2(-12.8)	37.8(-10.2)	37.3(-10.7)	**48.0**	54.0
Swin-Base		52.5(-24.8)	62.6(-14.7)	66.6(-10.7)	69.0(-8.3)	**77.3**	86.8	34.3(-15.8)	36.5(-13.6)	38.5(-11.6)	35.9(-14.2)	**50.1**	56.3
LSDA	CVPR'24	52.3(-23.9)	62.5(-13.7)	65.6(-10.6)	65.8(-10.4)	**76.2**	86.8	36.0(-16.4)	36.8(-15.6)	40.0(-12.4)	39.6(-12.8)	**52.4**	56.9
NPR	CVPR'24	53.3(-23.8)	62.6(-14.5)	67.2(-9.9)	68.1(-9.0)	**77.1**	87.1	33.6(-14.9)	35.2(-13.3)	37.7(-10.8)	37.2(-11.3)	**48.5**	53.9
UCF	ICCV'23	50.9(-24.7)	60.4(-15.2)	63.4(-12.2)	67.9(-7.7)	**75.6**	84.3	32.0(-13.6)	34.1(-11.5)	36.0(-9.6)	36.4(-9.2)	**45.6**	51.8
Wavelet	WACV'25	45.3(-31.7)	54.7(-22.3)	58.7(-18.3)	66.4(-10.6)	**77.0**	85.8	34.6(-15.9)	36.1(-14.4)	37.6(-12.9)	36.8(-13.7)	**50.5**	55.5

E EXPERIMENTS ON OTHER DATASETS

Tables 5 and 4 demonstrate the consistent cross-dataset superiority of VIGUARDQ, which achieves the highest AUC scores among all evaluated quantization methods across 11 detectors and five testing datasets. Averaged over all detectors, VIGUARDQ achieves accuracies of 75.20%, 49.68%, 80.74%, 78.06%, and 67.48% on DFDCP, DF40, CDF-v1, CDF-v2, and DFDC, respectively, outperforming PTQ4VM by 9.10, 11.75, 12.32, 14.74, and 13.03 percentage points. These improvements hold across both general-purpose transformer backbones and specialized deepfake detectors, demonstrating broad architectural applicability. Particularly, VIGUARDQ improves ViT-Base AUC scores on CDF-v2 from 62.9% to 81.3%, substantially narrowing the gap to its full-precision counterpart at 87.8%. Although a full-precision performance gap remains in most settings, the consistent gains on unseen datasets suggest that VIGUARDQ better preserves transferable forgery-discriminative information under quantization.

F EFFECTIVENESS OF TAP AND TOP

Table 6 demonstrates the effectiveness of TAP and the additional benefit of TOP within VIGUARDQ's ECR module. Across three backbones and five cross-dataset benchmarks, introducing TAP alone consistently improves AUC scores over the variant without ECR, increasing the average AUC scores from 67.87% to 72.87%. Incorporating TOP further improves all 15 evaluations,

Table 6: Ablation study of the Temporally Aligned Perturbation (TAP) and Temporally Opposed Perturbation (TOP). All detection models are trained on FF++ Rossler et al. (2019) and evaluated on multiple cross-dataset benchmarks.

Method	**CDF-v1** Li et al. (2020)			**CDF-v2** Li et al. (2020)			**DFD** Zi et al. (2020)			**DFDC** Dolhansky et al. (2020)			**DFDCP** Dolhansky et al. (2019)		
	DeiT-B	**ViT-B**	**Swin-B**	**DeiT-B**	**ViT-B**	**Swin-B**	**DeiT-B**	**ViT-B**	**Swin-B**	**DeiT-B**	**ViT-B**	**Swin-B**	**DeiT-B**	**ViT-B**	**Swin-B**
✗ ECR	72.2	68.5	73.0	66.5	69.2	71.3	73.6	71.9	75.6	60.9	62.7	64.1	61.4	64.5	62.7
✓ ECR (✓ TAP ✗ TOP)	74.2	75.1	76.3	69.9	76.8	76.2	77.6	79.5	78.3	63.5	66.6	65.8	68.4	70.9	74.0
✓ ECR (✓ TAP ✓ TOP)	**80.4**	**83.0**	**83.1**	**75.7**	**81.3**	**80.5**	**83.5**	**84.9**	**85.5**	**66.6**	**69.8**	**68.5**	**74.5**	**77.1**	**77.3**

Table 7: Robustness comparison between full-precision and quantized deepfake detectors under different perturbations on the DF40 and CDF-v1 datasets.

Backbone	**Precision**	**DF40**						**CDF-v1**					
		Clean	**Blur**	**JPEG**	**Artifacts**	**Noise**	**All**	**Clean**	**Blur**	**JPEG**	**Artifacts**	**Noise**	**All**
DeiT-Base	FP	53.7	49.5	51.7	52.0	49.8	47.6	87.1	77.5	80.1	76.3	79.4	75.2
	Quantized	48.4	43.5	47.2	45.4	45.0	42.2	80.4	69.5	72.3	70.0	70.5	66.7
ViT-Base	FP	57.5	54.0	55.1	52.8	49.7	46.9	89.6	76.4	78.1	79.3	75.8	74.7
	Quantized	52.3	49.6	51.0	47.4	46.8	40.1	83.0	68.5	71.4	73.0	66.2	66.1
Swin-Base	FP	56.3	52.3	49.2	47.5	51.0	47.6	91.7	78.2	73.6	82.1	81.5	72.5
	Quantized	50.1	46.9	42.1	41.6	44.4	42.1	83.1	70.4	66.8	74.4	72.0	66.1

yielding an additional average gain of 5.24 percentage points and raising the overall AUC scores to 78.11%. The complete ECR module therefore outperforms the variant without ECR by 10.24 percentage points on average, with improvements ranging from 4.4 to 14.6 percentage points. Notably, ViT-B AUC scores on CDF-v1 increases from 68.5% to 75.1% with TAP and further to 83.0% with TOP. These consistent improvements support the complementary benefits of combining TAP and TOP for preserving cross-dataset detection performance after quantization.

G Robustness Discussion

As shown in Table 7, image perturbations consistently degrade detection AUC scores on both DF40 and CDF-v1. Importantly, this degradation is already evident in the full-precision models, indicating that the performance drop mainly originates from the limited robustness of the underlying detectors rather than from our proposed VIGUARDQ. Across different backbones and perturbation types, the quantized models largely follow the same performance trends as their full-precision counterparts and maintain comparable accuracy gaps, demonstrating that VIGUARDQ can effectively preserve the robustness characteristics of the original models after quantization. Therefore, improving robustness in real-world settings primarily requires stronger full-precision detectors. In future work, we will explore more robust detector architectures and training strategies, and then apply VIGUARDQ to efficiently deploy them on edge devices.