MASKED QUATERNION PHASE MODELING FOR AI-GENERATED IMAGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Since Oppenheim and Lim's classic experiments, the phase of the Fourier spectrum is known to carry the structural content of an image. The spectral representations currently used to detect AI-generated images nevertheless rely almost entirely on the amplitude spectrum. We build a detector on the phase itself. We hypothesize that a real image inherits phase consistency from its physical imaging pipeline, across frequency bands and color channels, and that a generated image fails to reproduce this consistency. To capture this consistency, we take the quaternion Fourier transform (QFT) of the RGB image. Unlike three separate transforms, the QFT places the three color channels in one shared phase frame, as a luminance and a chrominance component whose relative phase is well defined. We pretrain a vision transformer on real images only to restore the quaternion phase of masked spectral regions and freeze the model, with its decoder, as the backbone. On this backbone, we build two scores: a phase-deviation score that measures how far an image departs from real-image phase structure without any generated image, and the score of a lightweight attention specialist trained on a single generator. Their decisions are combined by a logical OR at thresholds set on real training images. In the GenImage benchmark, real and generated images differ in compression format and resolution, and these alone predict the label. Under a protocol that canonicalizes both properties for every method, the detector averages 88.2% accuracy over eight generators, 14.9 points above the best of six retrained baselines. QPP transfers zero-shot to the nine generators of Synthbuster and remains stable under JPEG recompression down to a quality factor of 50. Controlled ablations attribute the gain to supervising the luminance–chrominance relative phase, a quantity defined in the shared phase frame that per-channel objectives leave unsupervised.

1 INTRODUCTION

It is well known that the structural content of an image is carried by the phase of its Fourier spectrum (Oppenheim & Lim, 1981). In the classic phase-swap experiments, an image reconstructed from its own phase and the amplitude spectrum of another image retains the edges and layout of the former, while little of the amplitude donor remains visible.

Frequency-domain detectors for generated images have nevertheless read the amplitude spectrum, almost exclusively, from the grid artifacts of GAN upsampling (Odena et al., 2016; Zhang et al., 2019; Frank et al., 2020; Durall et al., 2020) to learned spectral representations (Tan et al., 2024b; Karageorgiou et al., 2025). Even masked spectral self-supervision (Xie et al., 2023) constrains the phase only indirectly, and the detectors built on such pretraining extract only the real part (Karageorgiou et al., 2025). Where phase does appear, in the neighboring task of face forgery detection (Liu et al., 2021a) and in detecting generated images after compression (Li et al., 2026), phase is an auxiliary input to a supervised classifier, not a learned representation.

Yet if phase carries the structure of an image, the physical imaging pipeline should leave traces in phase: one lens, one sensor, and one illuminant produce the three color channels of a real image together, and their regularities should be visible in phase. Our starting hypothesis is that a representation of real-image phase makes a generated image's departure from these regularities measurable.

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.

The question is: "Which phase regularity should a detector read?" Image features arise where the phase components align across frequency bands, a property known as phase congruency (Morrone & Owens, 1987; Kovesi, 1999). Structure therefore lies in the interrelationships of phases, and it is these that a detector should read. Congruency alone, however, does not separate real from generated images, since any image with edges has congruency. What distinguishes a photograph is how these relations correspond across its color channels (Figure 1). In a photograph, the three channels are one optical event of a single lens, sensor, and illuminant. The illuminant, for instance, shifts the relative phase between luminance and chrominance of every spectral bin in one fixed direction (Appendix D.1.4). An image generator emits three channels without such an optical event, and nothing in its pipeline enforces these constraints. Whether its outputs nevertheless satisfy them is the empirical question this paper tests (Figure 1, Section 6).

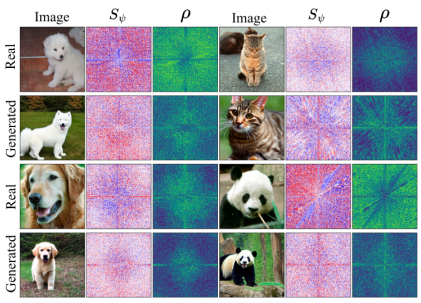


Figure 1: Color-coupling measure S_ψ and strength ρ of the quaternion Fourier transform (QFT; Section 3), which jointly characterize color coupling in the shared phase frame. In each block, the real image is above, and the generated one is below. In these examples, the generated images show weaker coupling (paler S_ψ, darker ρ) or, for the cat, radially streaked coupling. Figure 3 shows all quaternion quantities on more images.

The relative phase between luminance and chrominance exists only if both are measured against one reference axis; however, three separate Fourier transforms fix no such axis. We therefore propose Quaternion Phase Probing (QPP). QPP treats the RGB image as a single quaternion-valued signal and takes its QFT (Sangwine, 1996; Ell & Sangwine, 2007). The QFT places the three channels in one shared phase frame, as a luminance and a chrominance component whose relative phase is well defined. To capture the consistency across bands, we pretrain a vision transformer (ViT) (Dosovitskiy et al., 2021) on real images only to restore the phase of masked regions of the quaternion spectrum from the surrounding bands. This frozen ViT, with its decoder, is the backbone throughout the paper.

The detector is assembled on this backbone. Because the backbone learns restoration from real images only, its restoration reproduces a real-image phase structure. Where an input diverges from that structure, the restoration differs from the input, and the detector reads the size of that difference (Figure 2(b)). The *phase-deviation score* (Figure 2(c)) reads this difference as a distance from the restoration response of real training images and uses no generated images, and the *attention specialist* (Figure 2(d)) is a lightweight cross-attention head trained on the images of one generator. Their thresholds are declared on real training images, and their decisions are combined by a logical OR (Figure 2(e)). Because real images arrive with varied compression and resizing histories, every input first passes one fixed JPEG-and-resize step, the *imaging channel*, for every image to have the same compression and resizing stage, and this narrows the real-image reference (Section 4.2).

We evaluate on GenImage (Zhu et al., 2023b). Its real images are JPEG files, and its generated images are mostly PNG files at generator-specific resolutions, so format and resolution alone predict the label (Grommelt et al., 2024). We therefore re-encode and resize every image identically before it reaches any detector, and retrain six published detectors accordingly. Under this protocol, QPP averages 88.2% accuracy over the eight GenImage generators, against 73.3% for the best retrained baseline, and 74.8% even without the imaging channel. QPP transfers zero-shot to Synthbuster (Bammey, 2024) and is stable under JPEG recompression down to a quality factor of 50.

Our contributions are threefold.

- **Masked quaternion modeling (MQM).** A self-supervised objective that pretrains a ViT to restore the masked quaternion Fourier phase of real images, supervising the luminance–chrominance relative phase that per-channel objectives do not supervise.

- **A detector driven by quaternion phase.** A real-image-only phase-deviation score and a single-generator attention specialist on the frozen backbone, combined under thresholds declared on real training images.

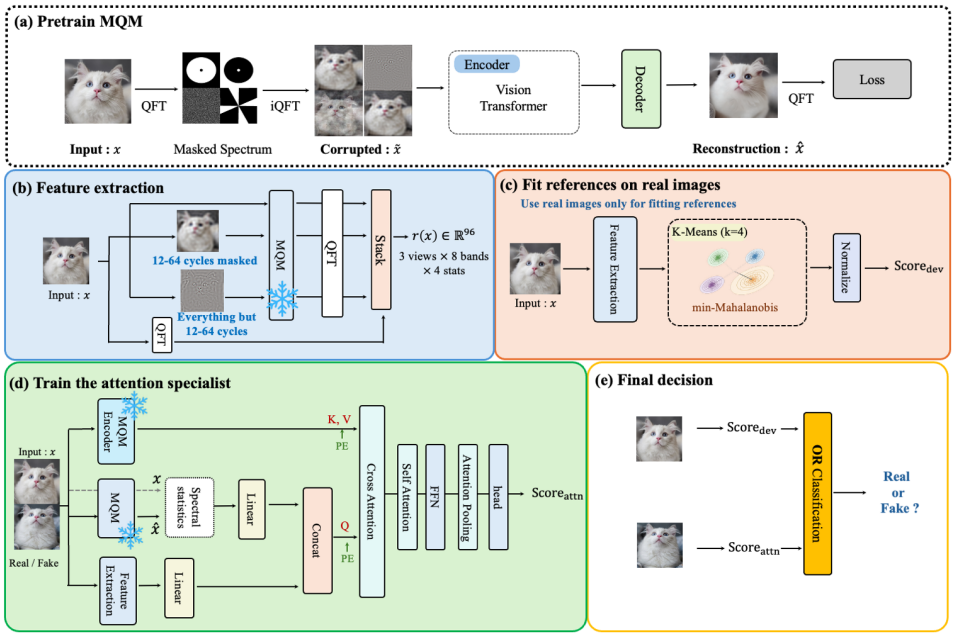


Figure 2: Overview of QPP. (a) Masked quaternion modeling (MQM): a real image's quaternion spectrum is corrupted, inverted to $\tilde{x}$, and restored by a ViT, with the loss equation 3 on the corrupted bins. This ViT and its decoder, frozen after pretraining (snowflakes), form the backbone. (b) Feature extraction: every input first passes the imaging channel described in Section 4.2. The frozen backbone then restores three views of the input, whose quaternion spectra are compared with the input's, yielding $r(x) \in \mathbb{R}^{96}$. (c) The phase-deviation score s_{dev}, a reference fitted on real training images alone equation 4. (d) The attention specialist is supervised on one generator. Its queries combine the probing tokens of (b) with band statistics of x and its restoration $\hat{x}$ and cross-attend to the patch tokens of the frozen encoder. (e) The decision equation 5 is an OR over the declared thresholds.

- **Evidence under a fair protocol.** With compression and resolution canonicalized for every method, 88.2% average accuracy, 14.9 points above the six retrained detectors, zero-shot transfer to Synthbuster, and robustness to JPEG recompression.

2 Related Work

Detecting generated images. Prior detectors can be placed on two axes: whether the decision is learned from generated examples or fitted on real images alone and, when the evidence is spectral, which quantity is read. On the first axis, nearly all detectors learn the decision from generated examples. Spatial classifiers train convolutional networks on real and generated images with aggressive augmentation (Wang et al., 2020) or isolate local traces of the upsampling path (Tan et al., 2024a), and foundation-model detectors classify in CLIP feature space (Ojha et al., 2023) or reshape that space with category prompts (Tan et al., 2025). On the second axis, spectral detectors read the spectrum almost entirely through its amplitude, from hand-analyzed artifacts to learned high-frequency representations (Tan et al., 2024b), spectral reconstruction similarity on a masked-spectrum backbone (Karageorgiou et al., 2025), and fingerprint analyses of diffusion models (Corvi et al., 2023; Ricker et al., 2024). Our phase-deviation score departs on both axes, being fitted on real images alone and reading the quaternion phase. The specialist, the only supervised component, sees a single generator.

Masked spectral self-supervision. Masked autoencoders (MAE; He et al., 2022) learn visual representations by restoring masked patches in pixel space, and masked frequency modeling moves the mask to the spectrum (Xie et al., 2023). In both, phase receives no term of its own and is supervised only implicitly, through reconstruction errors dominated by where the missing energy lands. Our task gives phase a term of its own, and the phase supervised is that of the quaternion spectrum. Thus, the task also supervises how the color channels relate, a quantity introduced next.

Quaternion signal processing. While the QFT analyzes a color image as one signal (Sangwine, 1996; Ell & Sangwine, 2007), the hypercomplex analytic signal framework extends phase to such signals (Bülow & Sommer, 2001). Together they give the three channels a single phase frame in which the luminance–chrominance relative phase of Section 3 is well defined. Quaternion spectral analysis has since provided color image processing with fixed formulas in edge detection (Sangwine, 1998), watermarking (Bas et al., 2003), and, closest to our use, saliency from the quaternion phase spectrum alone (Guo et al., 2008). Quaternion-valued neural networks (Parcollet et al., 2019) learn representations in the quaternion algebra but not of the quaternion Fourier phase. To our knowledge, our work is the first to make that phase the target of the masked pretraining described above.

3 Quaternion Fourier Transform

This section reviews the QFT and defines the coupling quantities applied throughout the paper.

A quaternion $q = q_0 + q_1\mathrm{i} + q_2\mathrm{j} + q_3\mathrm{k}$ extends the complex numbers to four dimensions (Hamilton, 1844). Following Sangwine (1996) and Ell & Sangwine (2007), we encode an RGB image as the pure quaternion field $q(m,n) = R(m,n)\,\mathrm{i} + G(m,n)\,\mathrm{j} + B(m,n)\,\mathrm{k}$ and take its left-sided discrete quaternion Fourier transform

$$F(u,v) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} e^{-\mu 2\pi \left(\frac{mu}{M} + \frac{nv}{N}\right)} q(m,n), \tag{1}$$

where $M \times N$ is the image size, (u,v) indexes frequencies, and the transform axis $\mu = (\mathrm{i}+\mathrm{j}+\mathrm{k})/\sqrt{3}$ is the achromatic gray line, as is standard for color images. Each coefficient separates into luminance and chrominance components, and the transform reduces to two complex Fourier transforms.

The quaternion algebra therefore adds no degrees of freedom beyond those two spectra (Ell & Sangwine, 2007). Its contribution is one shared phase frame, fixed once for all images (Appendix A.2), in which the luminance and chrominance components of every coefficient are measured against the same axis. Independent R, G, and B spectra, the per-channel alternative, provide no canonical frame (Sangwine, 1996). A relative phase between a luminance and a chrominance component is defined only after the color axes are rotated onto the gray line and the plane orthogonal to it, and per-channel objectives fix no such rotation.

We now define the two scalars read throughout the paper. In the shared phase frame, write the two components of a coefficient in polar form about the transform axis, $a_1 e^{\mu\beta_1}$ and $a_2 e^{\mu\beta_2}$, each a complex number with μ as the imaginary unit, so that their relative phase is $\beta_2 - \beta_1$. We summarize the color coupling of the coefficient by two scalars,

$$\rho = \frac{2\,a_1 a_2}{a_1^2 + a_2^2}, \qquad S_\psi = \rho\,\sin(\beta_2 - \beta_1). \tag{2}$$

The coupling strength ρ equals 1 when the two components have equal magnitude and 0 when either vanishes, so ρ measures how strongly they coexist. The coupling measure S_ψ[1] is the sine of the relative phase, weighted by ρ so that near-achromatic coefficients, where the relative phase is undefined, contribute little to any statistic derived from S_ψ. The pretraining objective supervises the relative phase itself through the unit cross-spectrum $\chi = e^{\mu(\beta_2-\beta_1)}$ (Section 4.1). Unlike S_ψ, χ retains the full angle rather than only its sine. Figure 1 visualizes S_ψ and ρ on real and generated images, and Appendix A derives the QFT in full.

4 Method

The proposed detector, QPP, is described in two parts: the pretraining task (Section 4.1, Figure 2(a)) and the detection model (Section 4.2, Figure 2(b–e)).

[1]The subscript refers to the quaternion polar form, in which $S_\psi = \sin 2\psi$ for the third phase angle ψ (Bülow & Sommer, 2001). Appendix A.5 shows that the two forms coincide.

4.1 MASKED QUATERNION MODELING

Masked quaternion modeling (MQM) pretrains a ViT encoder with a lightweight decoder, and the two, frozen together after pretraining, form the backbone (Figure 2(a)). The task is to restore the missing quaternion spectral content of real images. MQM builds on masked frequency modeling (MFM; Xie et al., 2023). MFM restores masked per-channel spectra, whereas MQM supervises the quaternion spectrum of Section 3 and, with it, the relative phase, a quantity per-channel objectives leave unsupervised. Given an ImageNet-1k photograph x (Russakovsky et al., 2015), used without labels, we delete a structured set $\mathcal{M}$ of bins from its quaternion spectrum F and invert the corrupted spectrum to an image $\tilde{x}$. The encoder and the decoder, a 1×1 convolution followed by pixel shuffle, restore an image $\hat{x}$ from $\tilde{x}$, and its quaternion spectrum $\hat{F}$ is compared with F on the deleted bins only. The mask is sampled from structured spectral regions that vary in both frequency and orientation (Appendix B.1, Figure 4).

Training objective. The loss has three terms,

$$\mathcal{L} \;=\; \mathcal{L}_{\text{spec}} \;+\; \lambda_{\text{cross}}\,\mathcal{L}_{\text{cross}} \;+\; \lambda_{\text{DC}}\,\mathcal{L}_{\text{DC}}. \tag{3}$$

The spectral term $\mathcal{L}_{\text{spec}}$ is the masked L_1 distance between $\hat{F}$ and F over the deleted bins, normalized by the clean amplitude, a masked spectral loss in the style of MFM applied to the quaternion spectrum. The cross term $\mathcal{L}_{\text{cross}}$, over the same bins, supervises the relative phase directly. The cross term compares the unit cross-spectrum $\hat{\chi}$ of the restoration with χ of Section 3, weighting each bin by its amplitude and its coupling strength ρ, so that weak or near-achromatic bins contribute little. The DC term $\mathcal{L}_{\text{DC}}$ anchors the global color, a component the masks never touch. Appendix B gives each term exactly, the mask construction, and the training configuration. That configuration keeps the architecture, data, and 300-epoch schedule of MFM. The encoder and decoder are then frozen and never fine-tuned on detection data. Fine-tuning on generated images would let the restoration itself learn generator cues, and the deviation score of Section 4.2 would no longer measure distance from real images alone.

4.2 DETECTION MODEL

The detection model (Figure 2(b–e)) comprises feature extraction without learned parameters (Figure 2(b)), the phase-deviation score fitted on real images alone (Figure 2(c)), the attention specialist trained on one generator (Figure 2(d)), and the decision (Figure 2(e)).

Imaging channel. QPP re-images every input, real or generated, through one fixed, label-free channel, JPEG compression at a quality factor of 75, followed by bicubic downscaling to $0.75\times$ and back to the original size. The channel is applied to every image the detector sees, the real images that fit the phase-deviation reference, the training images of the specialist, and every test image. The channel belongs to the detector, not to the representation, and MQM is pretrained once on unmodified photographs. The channel is intended to narrow the real-image reference. Each compression or resizing step in a real image's history moves the quaternion phase, and one shared channel removes much of that variation. The quality factor and the scale were fixed early and never tuned.

Feature extraction. Feature extraction (Figure 2(b)) reads the difference between an image and its restoration by the MQM backbone. Because the backbone learned restoration from real images only, that difference grows with the image's departure from real phase structure. Each image enters as five 224×224 crops, the four corners and the center (Section 5.2). The probing features $r(x)$ defined below are computed per crop, averaged over the crops by the phase-deviation score, and read on the center crop by the specialist. From the input image x we form three views, x itself and two corruptions from the pretraining family of Section 4.1, a deleted annulus of 12 to 64 cycles and its deleted complement, whose radii lie inside every annulus sampled during pretraining (Appendix B.1). The MQM backbone restores each view. We then compare the quaternion spectrum of each restoration with that of x over eight radial bands, octave-spaced up to 64 cycles (Appendix C.1), through four measures, the rotation of the quaternion phase, the two aspects of color coupling (S_ψ and χ), and the change in amplitude. Stacking 3 views $\times$ 8 bands $\times$ 4 measures yields the probing features $r(x) \in \mathbb{R}^{96}$ (Appendix C.1).

The phase-deviation score. The phase-deviation score (Figure 2(c)) measures how far an image departs from the real training images through a reference fitted on them in closed form without any generated image. Because real images do not follow a single Gaussian, we model the reference as four clusters of the probing features of the real training images, obtained by k-means. Each cluster j carries a mean η_j and a covariance Σ_j, and the score is the squared Mahalanobis distance to the *nearest* cluster,

$$s_{\mathrm{dev}}(x) \;=\; \min_{1\le j\le 4} \frac{\big(\bar{r}-\eta_j\big)^{\top}\Sigma_j^{-1}\big(\bar{r}-\eta_j\big) - m_j}{s_j}, \tag{4}$$

where $\bar{r}$ is $r(x)$ averaged over the five crops and (m_j, s_j) standardize each cluster's distance on its own members. An image scores high only when its restoration response is far from every real cluster, so real images score near zero and an unseen generator is detected to the extent that its images exhibit this departure (fit and choice of k: Appendix C.3, Table C.2).

The attention specialist. Because the phase-deviation score references no generated image, the specialist (Figure 2(d)) adds the supervised evidence available for the one generator seen at training time (Section 5.1). The specialist is the only component trained on generated images, with binary cross-entropy to distinguish the real training images from the training generator's images, and its score s_{attn} is standardized on the real training images. The head is a cross-attention layer whose query tokens carry the restoration signal in two forms. The first is the probing features of the center crop, kept as 24 tokens so that each view-and-band measurement remains a separate piece of evidence. The second is fixed spectral statistics. A fixed map $u(\cdot)$ collects 116 statistics of the quaternion spectrum of its argument, among them band statistics of the coupling pair (S_ψ, ρ) and band energies (Table C.1). The specialist reads these statistics of the residual image, $u(x-\hat{x})$, and their change under restoration, $u(x)-u(\hat{x})$, where $\hat{x}$ is the restoration of the intact view. The weight of such evidence depends on the image, since the coupling statistics of a nearly achromatic image carry no information. The query tokens therefore cross-attend to the patch tokens that the MQM encoder produces for the center crop, and a self-attention layer, a feed-forward block, and attention pooling produce the score (Appendix C.4).

Decision by declared thresholds. The two scores cover different generators. The deviation score responds to any departure from real-image phase, whereas the specialist responds to the one generator it was trained on, and on unseen generators its score is uninformative or inversely predictive (Table D.2). Score-level fusion would therefore give an uninformative specialist the same weight as an informative one. Instead, we threshold each score separately and combine the two decisions by a logical OR, so that each score adds detections only within its own false-positive budget. Thresholds tuned on test data would allow the test distribution into the decision, so we declare in advance the fraction of real images each score may misclassify: the one-sided tail probabilities $1-\Phi(1) = 15.87\%$ for the phase-deviation score and $1-\Phi(2) = 2.28\%$ for the specialist, with Φ the standard normal cumulative distribution function. The thresholds $\tau_{1\sigma}^{\mathrm{dev}}$ and $\tau_{2\sigma}^{\mathrm{attn}}$ are the corresponding empirical quantiles of the two scores on the real training images. The final decision is a logical OR over the two thresholded scores (Figure 2(e)):

$$\mathrm{fake}(x) \;=\; \big[\, s_{\mathrm{dev}}(x) \ge \tau_{1\sigma}^{\mathrm{dev}} \,\big] \;\vee\; \big[\, s_{\mathrm{attn}}(x) \ge \tau_{2\sigma}^{\mathrm{attn}} \,\big]. \tag{5}$$

The union bound caps the false-positive rate on the real training images at $15.87\% + 2.28\% = 18.15\%$. Appendix D.2.3 reports the measured rates there and on every other real set. The two levels differ by design. Generated images overlap only the upper tail of the real deviation distribution, so that the threshold must sit inside the tail, whereas the specialist separates the seen family sharply and a conservative level costs the specialist little.

5 Experiments

5.1 Datasets

GenImage (Zhu et al., 2023b) combines ImageNet photographs (Russakovsky et al., 2015) with images from eight generators, with matching class labels but not content. We calibrate the phase-deviation score and the declared thresholds on 40,000 real images from the training split, the real

Table 1: Accuracy (%) on the eight GenImage subsets with SD1.4 as the training generator, reported in the literature on unmodified images (top) and retrained under the canonicalized protocol of Section 5.2 with the imaging channel applied to every method (bottom). Bold: best per column in the bottom block.

Method	SD1.4	SD1.5	Wukong	MJ	GLIDE	VQDM	BigGAN	ADM	Avg.
Reported in the literature, unmodified images									
ResNet-50 (He et al., 2016)	99.9	99.7	98.2	54.9	61.9	56.6	52.0	53.5	72.1
Swin-T (Liu et al., 2021b)	99.9	99.8	99.1	62.1	67.6	62.3	57.6	49.8	74.8
GenDet (Zhu et al., 2023a)	96.1	96.1	92.8	89.6	78.4	66.5	75.0	58.0	81.6
PatchCraft (Zhong et al., 2023)	89.5	89.3	89.3	79.0	78.4	83.7	72.4	77.3	82.3
AIDE (Yan et al., 2025a)	99.7	99.8	98.7	79.4	91.8	80.3	66.9	78.5	86.9
NPR (Tan et al., 2024a)	98.2	97.9	96.9	81.0	89.8	84.1	84.2	76.9	88.6
FatFormer (Liu et al., 2024)	100.0	99.9	99.9	92.7	88.0	98.8	55.8	75.9	88.9
DRCT (Chen et al., 2024)	95.0	94.4	94.6	91.5	89.1	90.0	81.6	79.4	89.4
Effort (Yan et al., 2025b)	99.8	99.8	97.4	82.4	93.3	91.7	77.6	78.7	90.1
C2P-CLIP (Tan et al., 2025)	90.9	97.9	98.8	88.2	99.0	96.5	98.7	96.4	95.8
Retrained and re-evaluated under the canonicalized protocol									
NPR (Tan et al., 2024a)	90.8	90.5	90.6	66.2	54.1	53.0	46.8	48.4	67.6
NPR, official ProGAN weights (Tan et al., 2024a)	51.1	50.9	51.1	49.1	36.8	41.5	75.1	37.7	49.1
Swin-T (Liu et al., 2021b)	99.2	99.0	96.8	67.3	54.5	53.7	49.7	51.7	71.5
AIDE (Yan et al., 2025a)	97.4	97.5	97.0	65.0	61.9	51.3	51.0	51.0	71.5
LOTA (Wang et al., 2025)	97.0	97.0	97.0	63.8	49.3	50.7	49.1	49.7	69.2
SPAI (Karageorgiou et al., 2025)	87.5	87.3	87.2	61.1	52.1	50.4	50.3	49.0	65.6
C2P-CLIP (Tan et al., 2025)	**99.9**	**99.9**	**99.9**	**85.5**	50.3	50.7	49.9	50.1	73.3
QPP (ours)	90.6	91.2	89.9	75.1	**91.2**	**90.5**	**88.8**	**88.5**	**88.2**

training images of Section 4.2. The specialist is trained on 20,000 real images from the Stable Diffusion v1.4 (SD1.4) training subset and 20,000 SD1.4 outputs, SD1.4 being the sole seen generator. Of those, 5,000 are among the 40,000 (Appendix C.4). Evaluation uses the eight validation subsets, each with 6,000 generated and 6,000 real images.

Synthbuster (Bammey, 2024) provides raw, never resized or recompressed, native-resolution PNG outputs from nine generators, and the real class is all 123,287 unique MS-COCO photographs (Lin et al., 2014) (dataset details in Appendix E). No Synthbuster image is used for training or calibration, so Synthbuster measures zero-shot transfer. We report accuracy on the class-balanced GenImage subsets and balanced accuracy on Synthbuster, whose real pool is far larger than each generated set.

5.2 Experimental Setup

Canonicalized evaluation protocol. Because format and resolution alone predict the label on GenImage (Section 1), we remove both for every method. Every image, real or generated, for training or testing, is re-encoded at a JPEG quality factor of 96 and, if its shorter side exceeds 256 pixels, downscaled bicubically to that size, never upscaled. We call this canonicalization. The result is resized to 256×256 and read as five 224×224 crops: four corners and the center.

5.3 Results on GenImage

Against the reported results (top block of Table 1), QPP averages 88.2, comparable to NPR (88.6), FatFormer (88.9), and DRCT (89.4) but below Effort (90.1) and C2P-CLIP (95.8). QPP scores 88.5–91.2 on every subset except Midjourney (75.1), whereas supervised detectors fall far below outside their training family (FatFormer: 55.8 on BigGAN). Because the top block inherits the shortcuts of unmodified images (Grommelt et al., 2024), the bottom block is our main comparison.

Under the canonicalized protocol, the ranking changes, consistent with the share of reported accuracy that Grommelt et al. (2024) attributed to format and resolution. Every baseline in the bottom block is run from its public implementation, retrained where required on the same canonicalized images with SD1.4 as the seen generator, and given the same imaging channel (Section 4.2) before canonicalization; therefore all methods see identical inputs. Each baseline selects its checkpoint and threshold on SD1.4 validation, whereas QPP retains its declared thresholds. QPP's backbone is not retrained, and its reference, thresholds, and specialist are calibrated or trained on the same canonicalized images (Section 5.1). The leading detectors of the top block lose 15 to 23 points (NPR $88.6 \rightarrow 67.6$, AIDE $86.9 \rightarrow 71.5$, C2P-CLIP $95.8 \rightarrow 73.3$). QPP remains at 88.2. It leads on GLIDE, VQDM, BigGAN, and ADM, where retrained baselines often approach chance, and exceeds the best baseline average (73.3) by 14.9 points. Without the imaging channel, with every other component refitted (Tables 4 and D.10), the average falls to 74.8, still above the best baseline with the channel (73.3) and without it (72.2, Table D.10). Therefore, the lead does not depend on the channel alone.

Table 2: Balanced accuracy (%) on the nine Synthbuster generators. QPP: zero-shot from GenImage, against the full MS-COCO real pool. The baselines are published results of detectors trained on MS-COCO real images and SD outputs. The last row replaces the real class with 1,000 RAISE-1k photographs under a declared web-distribution standardization and is excluded from the bolding. Bold: best per column under the MS-COCO real class.

Method	DALL·E 2	DALL·E 3	Firefly	GLIDE	MJ v5	SD1.3	SD1.4	SD2	SDXL	Avg.
NPR (Tan et al., 2024a)	51.1	49.3	46.5	48.5	52.8	51.4	51.8	46.0	52.8	50.0
AIDE (Yan et al., 2025a)	34.9	33.7	24.8	65.0	57.5	74.1	73.7	53.2	68.4	53.9
FatFormer (Liu et al., 2024)	59.4	39.5	60.3	72.7	44.4	53.7	54.0	52.3	69.1	56.1
UnivFD (Ojha et al., 2023)	**83.5**	47.4	**89.9**	53.3	52.5	70.4	69.9	75.7	68.0	67.8
C2P-CLIP (Tan et al., 2025)	55.6	63.2	59.5	86.7	52.9	75.2	76.7	69.2	77.7	68.5
AlignedForensics (Rajan et al., 2025)	50.2	48.9	51.7	53.5	**98.7**	**98.8**	**98.8**	**98.6**	**97.3**	77.4
QPP (ours, zero-shot)	77.3	**74.8**	82.1	**91.5**	63.9	86.4	86.5	74.1	63.8	**77.8**
QPP (ours, RAISE-1k reals)	81.1	78.5	85.8	95.2	67.7	90.2	90.2	77.9	67.6	**81.6**

5.4 Zero-Shot Transfer to Synthbuster

Synthbuster is evaluated zero-shot with every component as trained or calibrated on GenImage (Section 5.1). With thresholds unchanged, QPP averages 77.8 balanced accuracy over the nine generators (Table 2), matching the best published baseline, AlignedForensics (77.4), a detector trained on the real domain and on four of the nine columns. The profiles are opposite: AlignedForensics scores 97–99 on the SD columns and Midjourney v5 and about 50 elsewhere, whereas QPP stays at or above 63.8 on every column and averages 77.9 on the five non-SD generators, against 60.6 for AlignedForensics and 65.3 for UnivFD, the best baseline there. Two controls in Appendix D.5 (Table D.12) bound the remaining compression-history asymmetry between JPEG MS-COCO and PNG outputs: matching the generated images' compression depth to MS-COCO lowers the average by 1.1 points, and replacing the real class with RAISE-1k photographs (Dang-Nguyen et al., 2015) gives 81.6.

6 Analysis

Pretraining objective of the backbone. This comparison changes only the backbone of Figure 2(a) and refits every trained or calibrated component on each backbone. We keep MQM's decoder architecture across encoders and compare MQM with four alternatives. Two are public. The MFM encoder (Xie et al., 2023) shares MQM's data, schedule, and architecture but uses a per-channel masked-spectrum objective. The MAE encoder (He et al., 2022) was pretrained on the same images for 1600 epochs with a pixel objective. Neither checkpoint ships with a decoder, so on each frozen encoder we fit MQM's decoder with the frequency loss of MFM (Appendix D.1.2). Two are controls, decoder included, pretrained end-to-end with MQM's data, masks, and schedule. MCFM (masked channel-wise frequency modeling) uses a per-channel FFT and drops $\mathcal{L}_{\text{cross}}$, and the no-cross control keeps the quaternion basis with $\lambda_{\text{cross}} = 0$.

In Table 3, the full system, (both scores under the OR decision) falls from 88.2 to 73.9 with the MFM encoder and to 78.1 with MAE (Appendix D.1.2). MCFM, MFM's per-channel spectrum under MQM's mask mixture, reaches 83.3 and the no-cross control 82.2, so most of the gap to the public encoders derives from the pretraining setup that MQM shares with its controls, the mask mixture, the form of the spectral loss, and a decoder trained jointly with the encoder, not from the quaternion representation. The quaternion basis alone adds nothing over MCFM, and supervising the relative phase through $\mathcal{L}_{\text{cross}}$ adds 6.0 points (Appendices B.4 and D.1.1). At readout, with the MQM representation fixed, the phase-rotation and coupling measures exceed the amplitude measure at equal dimensionality (70.8 and 71.6 against 65.8 alone, 73.7 against 70.3 as pairs), and dropping r^A raises the deviation score alone to 77.8 but lowers the full system to 87.5 (Table D.3, Appendix D.1.3), so amplitude helps only through the specialist.

Contribution of the two scores. Table 4 evaluates the two scores of equation 5 separately with all thresholds unchanged. The phase-deviation score alone averages 75.8, above every retrained baseline in Table 1, although that score uses no generated image, and the specialist alone averages 68.4. The OR of the two scores reaches 88.2 because each score detects a different set of generators. On SD1.4 and SD1.5, the family seen in training, the deviation score alone reaches 58–59 and the specialist supplies the rest. On GLIDE, VQDM, BigGAN, and ADM, the specialist (AUROC 0.25–0.40) rarely fires, so the deviation score carries the decision and the full system stays within 1.5

Table 3: Backbone comparison: full-system accuracy (%), GenImage average at declared thresholds, and illuminant-estimation error on Gehler-Shi (degrees, mean / median). All arms refit every trained or calibrated component. For the public encoders the decoder is fitted on the frozen encoder.

Backbone	Full system	Illuminant
MFM, public encoder, fitted decoder	73.9	3.66 / 2.89
MAE, public encoder, fitted decoder	78.1	3.08 / 2.57
MCFM, per-channel control (equation B.2)	83.3	3.19 / 2.51
MQM without $\mathcal{L}_{\text{cross}}$ ($\lambda_{\text{cross}} = 0$)	82.2	3.12 / 2.46
MQM (equation 3, QPP)	**88.2**	**2.81 / 2.22**

Table 4: Contribution of the two scores and of the imaging channel: accuracy (%), GenImage average, at the declared thresholds (Table D.10 for the channel).

Configuration	Avg.
Phase-deviation score only	75.8
Specialist only	68.4
Without imaging channel	74.8
Full system (Deviation $\vee$ Attention)	**88.2**

points of it (Table D.2). The declared point is the best of the seven settings across the two levels, and the tighter $(2\sigma, 2\sigma)$ point still averages 74.2. Fusing the scores instead of the OR loses 4.3–6.5 points at the same budget (Tables D.7 and D.8, Appendices D.2.1 and D.2.2).

A physical scene property in the representation. Does MQM's representation capture a physical property of the scene? As noted in Section 1, the scene illuminant, the color of the light source, shifts the luminance–chrominance relative phase in one fixed direction. This is exactly what the cross term $\mathcal{L}_{\text{cross}}$ supervises (Appendix D.1.4), so a representation trained to restore that phase should encode the illuminant, and we test whether it does. On Gehler-Shi (Gehler et al., 2008; Shi & Funt, 2010), a color-constancy benchmark with a measured illuminant per image, a linear probe on each frozen backbone predicts the illuminant and is scored by angular error (lower is better). MQM is the most accurate (2.81 degrees, Table 3), and the advantage comes from the cross term. Removing it raises the error by 0.31 degrees (95% bootstrap interval [0.16, 0.46]), while MAE and the two controls are indistinguishable from one another (Table D.6). The cross term thus makes a physical scene property more accessible to a linear probe. Removing it also lowers the GenImage average from 88.2 to 82.2 in those same two rows, a drop of 6.0 points. A single change to the objective moves both.

Role of the imaging channel. Table 4 tests the reading of Section 4.2 that the channel tightens the real-image reference of the deviation score. Withholding it costs QPP 13.4 points (88.2 to 74.8). Table D.10 (Appendix D.3) breaks the loss down by generator. It falls mostly on ADM (88.5 to 59.4), GLIDE, and VQDM, the generators the deviation score separates best, whereas the retrained baselines gain nothing comparable from the channel ($+1.7$ for C2P-CLIP, losses for Swin-T and NPR), having no real-image reference to tighten. The MCFM control reaches 83.3 with the same channel (Table 3), so the quaternion objective adds 4.9 points. The channel also absorbs upstream recompression: with thresholds unchanged, accuracy stays at 87.5–89.4 down to a JPEG quality factor of 50, whereas the published detectors lose 13 to 26 points by a quality factor of 70 (Table D.11, Appendix D.4).

7 Conclusion

We investigated whether the regularities that real images inherit from their imaging pipelines leave traces in phase and built a detector on a ViT pretrained to restore the masked quaternion phase. Under the canonicalized protocol, QPP averages 88.2% on GenImage against 73.3% for the best retrained baseline, stays ahead (74.8%) without its imaging channel, transfers zero-shot to Synthbuster, and is stable down to a JPEG quality factor of 50. Ablations trace the gain to supervising the luminance–chrominance relative phase of the shared phase frame, and a linear probe finds the scene illuminant more linearly decodable from MQM than from any of the four other pretrained backbones.

Two limitations follow. The phase-deviation score is a distance from a real-image reference, so the score assumes that an input may be re-imaged and retains the high frequencies the channel reads. On Chameleon (Yan et al., 2025a), whose real class has mixed platform histories, the score inverts (AUROC 0.382) and the full system falls to chance (0.512), while the specialist alone ranks at 0.744 (Appendix D.6). The specialist adds evidence only for its seen family, so Midjourney stays weakest (75.1, Table 1). Modeling imaging histories with phase is the next step.

AI USE STATEMENT

Claude was used to translate the main text into English and to assist with organizing the manuscript. The authors reviewed and edited all resulting text and take full responsibility for the content of the paper.

REPRODUCIBILITY STATEMENT

Upon acceptance, we will release the source code, the pretrained MQM backbone, the fitted phase-deviation reference, the trained specialist, and the declared thresholds, together with the scripts that reproduce every table. The pretraining objective, mask family, and training configuration are specified in Appendices B.1 to B.3, and the detector, its reference, and the specialist are specified in Appendix C. The evaluation protocol, including the canonicalization parameters and the imaging channel, is given in Sections 4.2 and 5.2, and the datasets, all publicly available, are described in Appendix E. All thresholds are quantiles of the real training images at the levels declared in Section 4.2, so no value in the paper depends on tuning against a test set.

REFERENCES

Adobe Inc. Adobe firefly. `https://firefly.adobe.com`, 2023.

Quentin Bammey. Synthbuster: Towards detection of diffusion model generated images. *IEEE Open Journal of Signal Processing*, 5:1–9, 2024.

Patrick Bas, Nicolas Le Bihan, and Jean-Marc Chassery. Color image watermarking using quaternion Fourier transform. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 521–524, 2003.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. Technical report, OpenAI, 2023.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations (ICLR)*, 2019.

Thomas Bülow and Gerald Sommer. Hypercomplex signals—a novel extension of the analytic signal to the multidimensional case. *IEEE Transactions on Signal Processing*, 49(11):2844–2852, 2001.

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *International Conference on Machine Learning (ICML)*, 2024.

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

Duc-Tien Dang-Nguyen, Cecilia Pasquini, Valentina Conotter, and Giulia Boato. RAISE: A raw images dataset for digital image forensics. In *Proceedings of the 6th ACM Multimedia Systems Conference (MMSys)*, pp. 219–224, 2015.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 8780–8794, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Todd A. Ell and Stephen J. Sangwine. Hypercomplex Fourier transforms of color images. *IEEE Transactions on Image Processing*, 16(1):22–35, 2007.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International Conference on Machine Learning (ICML)*, 2020.

Peter Vincent Gehler, Carsten Rother, Andrew Blake, Tom Minka, and Toby Sharp. Bayesian color constancy revisited. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

Patrick Grommelt, Louis Weiss, Franz-Josef Pfreundt, and Janis Keuper. Fake or JPEG? revealing common biases in generated image detection datasets. In *European Conference on Computer Vision (ECCV) Workshops*, 2024.

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10696–10706, 2022.

Chenlei Guo, Qi Ma, and Liming Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, 2008.

William Rowan Hamilton. On quaternions; or on a new system of imaginaries in algebra. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 25(163):10–13, 1844.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, 2022.

Huawei Noah's Ark Lab. Wukong text-to-image. `https://xihe.mindspore.cn/modelzoo/wukong`, 2022.

Ju Yeon Kang, Jaehong Park, Semin Kim, Ji Won Yoon, and Nam Soo Kim. SARE: Semantic-aware reconstruction error for generalizable diffusion-generated image detection. *arXiv preprint arXiv:2508.09487*, 2025.

Dimitrios Karageorgiou, Symeon Papadopoulos, Ioannis Kompatsiaris, and Efstratios Gavves. Any-resolution AI-generated image detection by spectral learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Peter Kovesi. Image features from phase congruency. *Videre: Journal of Computer Vision Research*, 1(3):1–26, 1999.

Kai Li, Wenqi Ren, Wei Wang, and Xiaochun Cao. Detecting compressed AI-generated images via phase spectrum robustness. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021a.

Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10770–10780, 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 *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, 2021b.

Midjourney, Inc. Midjourney. `https://www.midjourney.com`, 2022.

M. Concetta Morrone and Robyn A. Owens. Feature detection from local energy. *Pattern Recognition Letters*, 6(5):303–313, 1987.

Alexander Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning (ICML)*, pp. 16784–16804, 2022.

Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016. URL `http://distill.pub/2016/deconv-checkerboard`.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24480–24489, 2023.

A.V. Oppenheim and J.S. Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69(5): 529–541, 1981. doi: 10.1109/PROC.1981.12022.

Titouan Parcollet, Mirco Ravanelli, Mohamed Morchid, Georges Linarès, Chiheb Trabelsi, Renato De Mori, and Yoshua Bengio. Quaternion recurrent neural networks. In *International Conference on Learning Representations (ICLR)*, 2019.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations (ICLR)*, 2024.

Anirudh Sundara Rajan, Utkarsh Ojha, Jedidiah Schloesser, and Yong Jae Lee. Aligned datasets improve detection of latent diffusion-generated images. In *International Conference on Learning Representations (ICLR)*, 2025.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022.

Jonas Ricker, Simon Damm, Thorsten Holz, and Asja Fischer. Towards the detection of diffusion model deepfakes. In *International Conference on Computer Vision Theory and Applications (VISAPP)*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Stephen J. Sangwine. Fourier transforms of colour images using quaternion or hypercomplex numbers. *Electronics Letters*, 32(21):1979–1980, 1996.

Stephen J. Sangwine. Colour image edge detector based on quaternion convolution. *Electronics Letters*, 34(10):969–971, 1998.

Lilong Shi and Brian Funt. Re-processed version of the Gehler color constancy dataset of 568 images. `http://www.cs.sfu.ca/~colour/data/`, 2010.

Chuangchuang Tan, Huan Liu, 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28130–28139, 2024a.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *AAAI Conference on Artificial Intelligence*, 2024b.

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 *AAAI Conference on Artificial Intelligence*, 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 (ICML)*, 2021.

Hongsong Wang, Renxi Cheng, Yang Zhang, Chaolei Han, and Jie Gui. LOTA: Bit-planes guided AI-generated image detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17246–17255, 2025.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-generated images are surprisingly easy to spot... for now. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8695–8704, 2020.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for diffusion-generated image detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22445–22455, 2023. doi: 10.1109/ICCV51070.2023.02051.

Jiahao Xie, Wei Li, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Masked frequency modeling for self-supervised visual pre-training. In *International Conference on Learning Representations (ICLR)*, 2023.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for AI-generated image detection. In *International Conference on Learning Representations (ICLR)*, 2025a.

Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable AI-generated image detection. In *International Conference on Machine Learning (ICML)*, pp. 70268–70288, 2025b.

Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in GAN fake images. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2019.

Nan Zhong, Yiran Xu, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. PatchCraft: Exploring texture patch for efficient AI-generated image detection. *arXiv preprint arXiv:2311.12397*, 2023.

Mingjian Zhu, Hanting Chen, Mouxiao Huang, Wei Li, Hailin Hu, Jie Hu, and Yunhe Wang. GenDet: Towards good generalizations for AI-generated image detection. *arXiv preprint arXiv:2312.08880*, 2023a.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. GenImage: A million-scale benchmark for detecting AI-generated image. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023b.

A Details and Derivations of the Quaternion Fourier Transform (QFT)

A.1 Quaternion Algebra

Definition and multiplication. A quaternion q consists of a real part q_0 and three imaginary parts q_1, q_2, q_3,

$$q = q_0 + q_1\mathrm{i} + q_2\mathrm{j} + q_3\mathrm{k}, \qquad q_0, q_1, q_2, q_3 \in \mathbb{R}, \tag{A.1}$$

and the set of all quaternions is denoted by $\mathbb{H}$ (Hamilton, 1844). The imaginary units satisfy Hamilton's relations

$$\mathrm{i}^2 = \mathrm{j}^2 = \mathrm{k}^2 = \mathrm{ijk} = -1, \tag{A.2}$$

from which the products of distinct units follow by multiplying $\mathrm{ijk} = -1$ by a unit on the left or right (for instance, $\mathrm{ijk}\,\mathrm{k} = -\mathrm{k}$ gives $\mathrm{ij} = \mathrm{k}$):

$$\mathrm{ij} = \mathrm{k} = -\mathrm{ji}, \qquad \mathrm{jk} = \mathrm{i} = -\mathrm{kj}, \qquad \mathrm{ki} = \mathrm{j} = -\mathrm{ik}. \tag{A.3}$$

Distinct units therefore anticommute. Extending equation A.2 and equation A.3 bilinearly gives the Hamilton product of $p, q \in \mathbb{H}$,

$$\begin{aligned} pq = {} & (p_0q_0 - p_1q_1 - p_2q_2 - p_3q_3) \\ & + (p_0q_1 + p_1q_0 + p_2q_3 - p_3q_2)\,\mathrm{i} \\ & + (p_0q_2 - p_1q_3 + p_2q_0 + p_3q_1)\,\mathrm{j} \\ & + (p_0q_3 + p_1q_2 - p_2q_1 + p_3q_0)\,\mathrm{k}, \end{aligned} \tag{A.4}$$

which is associative and distributive but not commutative. Throughout the paper, juxtaposition denotes this product.

Conjugate, norm, and inner product. The conjugate negates the imaginary parts, and the norm is the Euclidean length in $\mathbb{R}^4$:

$$q^* = q_0 - q_1\mathrm{i} - q_2\mathrm{j} - q_3\mathrm{k}, \qquad |q|^2 = q\,q^* = q^*\,q = q_0^2 + q_1^2 + q_2^2 + q_3^2, \tag{A.5}$$

where the middle equalities follow from equation A.4 because all cross terms cancel. Conjugation reverses products, $(pq)^* = q^*p^*$, and the norm is multiplicative, $|pq| = |p|\,|q|$. Every nonzero quaternion has the inverse $q^{-1} = q^*/|q|^2$, so $\mathbb{H}$ is a division algebra. The Euclidean inner product of two quaternions is the real part of $p\,q^*$,

$$\langle p, q\rangle = \mathrm{Re}(p\,q^*) = p_0q_0 + p_1q_1 + p_2q_2 + p_3q_3, \tag{A.6}$$

as the real line of equation A.4 shows once q is replaced by q^*. Unit quaternions, $|q| = 1$, form the 3-sphere in $\mathbb{R}^4$. The unit phase p of a QFT coefficient (Appendix A.5) is one such unit quaternion, and for two unit quaternions $\langle \hat{p}, p\rangle \in [-1, 1]$ is the cosine of the angle between them, which is the quantity read by the phase-rotation measure r^p in equation C.1.

Pure quaternions and vectors. A quaternion with $q_0 = 0$ is called pure, and pure quaternions are identified with vectors of $\mathbb{R}^3$ through $q_1\mathrm{i} + q_2\mathrm{j} + q_3\mathrm{k} \leftrightarrow (q_1, q_2, q_3)$. For pure p and q, setting $p_0 = q_0 = 0$ in equation A.4

$$p\,q = -\langle p, q\rangle + p \times q, \tag{A.7}$$

leaves the negative dot product as the real part and the cross product as the vector part. Three consequences are used in this appendix. First, a unit pure quaternion μ satisfies $\mu^2 = -\langle \mu, \mu\rangle + \mu \times \mu = -1$: every unit vector is a square root of -1 and can function as the complex imaginary unit. Second, orthogonal pure quaternions anticommute, $p\,q = p \times q = -\,q \times p = -\,q\,p$, since their dot product vanishes. Third, a right-handed orthonormal triple $(\mu_1, \mu_2, \mu_3 = \mu_1 \times \mu_2)$ reproduces Hamilton's relations,

$$\mu_1^2 = \mu_2^2 = \mu_3^2 = -1, \qquad \mu_1\mu_2 = \mu_3, \quad \mu_2\mu_3 = \mu_1, \quad \mu_3\mu_1 = \mu_2, \tag{A.8}$$

by equation A.7 and the vector identity $a \times (b \times c) = b\,\langle a, c\rangle - c\,\langle a, b\rangle$. Any such triple is therefore a rotated copy of $(\mathrm{i}, \mathrm{j}, \mathrm{k})$, and all algebra done in one frame transfers verbatim to the other. This fact supports the frame described in Appendix A.2.

Exponential, complex subfields, and polar form. For a unit pure quaternion μ, the identity $\mu^2 = -1$ turns the power series of the exponential into Euler's formula,

$$e^{\mu\theta} = \sum_{n\geq 0} \frac{(\mu\theta)^n}{n!} = \cos\theta + \mu \sin\theta, \qquad \theta \in \mathbb{R}, \tag{A.9}$$

and the set $\mathbb{C}_\mu = \{a + b\,\mu : a, b \in \mathbb{R}\}$ is closed under the Hamilton product and isomorphic to the complex numbers, with μ in the role of the imaginary unit. It is this commutative subfield for μ, the transform axis, that the symplectic decomposition of Appendix A.2 and the reduction of Appendix A.3 exploit. Every nonzero quaternion has the polar form

$$q = |q|\, e^{\mu\theta}, \qquad \mu = \frac{q_1\mathrm{i} + q_2\mathrm{j} + q_3\mathrm{k}}{\sqrt{q_1^2 + q_2^2 + q_3^2}}, \quad \theta = \operatorname{atan2}\big(\sqrt{q_1^2 + q_2^2 + q_3^2},\, q_0\big) \in [0, \pi], \tag{A.10}$$

with the axis μ set by its own vector part. The QFT kernel $e^{-\mu\theta}$ of equation 1 is such an element with the axis fixed to the gray line for every bin. Exponentials about different axes do not commute, $e^{\mu\alpha}e^{\nu\beta} \neq e^{\nu\beta}e^{\mu\alpha}$ in general, because orthogonal axes anticommute by equation A.7. This is why the side on which the kernel multiplies the signal in equation 1 and the order of the three factors in the polar form equation A.17 must be specified. Finally, grouping the coordinates in pairs, $q = (q_0 + q_1\mathrm{i}) + (q_2 + q_3\mathrm{i})\,\mathrm{j}$, writes $\mathbb{H} = \mathbb{C} \oplus \mathbb{C}\,\mathrm{j}$, and the symplectic decomposition of Appendix A.2 is this Cayley–Dickson form with (i, j) replaced by the rotated pair (μ_1, μ_2).

A.2 Symplectic Decomposition along the QFT Axis

Let $\{\mu_1 = \mu,\ \mu_2,\ \mu_3 = \mu_1\mu_2\}$ be an orthonormal basis of the imaginary subspace, where $\mu = (\mathrm{i}+\mathrm{j}+\mathrm{k})/\sqrt{3}$ is the transform axis of equation 1, and μ_2 is a unit pure quaternion orthogonal to μ_1, fixed once for all images. Our implementation uses the opponent-color axis $\mu_2 = (\mathrm{j}-\mathrm{i})/\sqrt{2}$, giving $\mu_3 = \mu_1\mu_2 = (2\mathrm{k} - \mathrm{i} - \mathrm{j})/\sqrt{6}$. This basis defines the frame used in equation 2, and Appendix A.5 shows that the choice of μ_2 only sets the origin of the chrominance phase. Let (a, b, c, d) be the real coordinates of a quaternion q in the basis $\{1, \mu_1, \mu_2, \mu_3\}$. Factoring μ_2 out of the last two terms with $\mu_3 = \mu_1\mu_2$ decomposes q into simplex and perplex parts:

$$q \;=\; a + b\,\mu_1 + c\,\mu_2 + d\,\mu_1\mu_2 \;=\; \underbrace{(a + b\,\mu_1)}_{c_1} + \underbrace{(c + d\,\mu_1)}_{c_2}\,\mu_2, \tag{A.11}$$

where c_1 and c_2 live in the commutative subfield spanned by $\{1, \mu_1\}$ and are therefore isomorphic to complex numbers. Two multiplication rules govern how this subfield interacts with μ_2. Because μ_1 and μ_2 are orthogonal pure quaternions they anticommute, $\mu_1\mu_2 = -\mu_2\mu_1$, so for any real α,

$$\begin{aligned} e^{\mu_1\alpha}\,(a + b\,\mu_1) &= (a + b\,\mu_1)\,e^{\mu_1\alpha}, \\ e^{\mu_1\alpha}\,\mu_2 &= (\cos\alpha + \mu_1 \sin\alpha)\,\mu_2 = \mu_2\,(\cos\alpha - \mu_1 \sin\alpha) = \mu_2\,e^{-\mu_1\alpha}: \end{aligned} \tag{A.12}$$

The exponential commutes with the subfield and moving it across μ_2 conjugates its angle. Both rules are used repeatedly in Appendices A.3 and A.4. The choice of μ_2 is a convention: every unit pure quaternion orthogonal to μ_1 has the form $\cos\alpha\,\mu_2 + \sin\alpha\,\mu_3 = e^{\mu_1\alpha}\mu_2$ for some real α, and inserting $e^{-\mu_1\alpha}e^{\mu_1\alpha}$ into equation A.11 gives $c_2\,\mu_2 = \big(c_2\,e^{-\mu_1\alpha}\big)\big(e^{\mu_1\alpha}\mu_2\big)$, so switching to that axis multiplies c_2 by the unit complex number $e^{-\mu_1\alpha}$ and leaves c_1 and the moduli $|c_1|$, $|c_2|$ unchanged. Appendix A.5 reads this as a shift of the chrominance phase origin. For a spectral coefficient, c_1 carries the luminance content and c_2 the chrominance content of Section 3, and their moduli $a_1 = |c_1|$ and $a_2 = |c_2|$ are the quantities entering the coupling strength of equation 2.

Color coordinates in the frame. For the pure quaternion pixel $q(m, n) = R\,\mathrm{i} + G\,\mathrm{j} + B\,\mathrm{k}$ of Section 3, the coordinates in the basis $\{1, \mu_1, \mu_2, \mu_3\}$ are $a = 0$ and, by equation A.6,

$$b = \langle q, \mu_1\rangle = \frac{R + G + B}{\sqrt{3}}, \qquad c = \langle q, \mu_2\rangle = \frac{G - R}{\sqrt{2}}, \qquad d = \langle q, \mu_3\rangle = \frac{2B - R - G}{\sqrt{6}}, \tag{A.13}$$

thus, the simplex part $c_1 = b\,\mu_1$ is the luminance along the gray axis, and the perplex part $c_2 = c + d\,\mu_1$ captures the two opponent-color coordinates of the chromatic plane orthogonal to it. The terms luminance and chrominance for c_1 and c_2 pass to the spectrum through equation A.14: $F_1 = \mu_1\mathcal{F}(b)$ is the transform of the luminance signal, and $F_2 = \mathcal{F}(c + d\,\mu_1)$ that of the complex-valued chrominance signal, where $\mathcal{F}$ is the complex 2D DFT with μ_1 as the imaginary unit.

A.3 The Forward QFT and Its Reduction to Two Complex FFTs

The forward transform is the left-sided QFT of equation 1. Writing $\theta_{mn}(u,v) = 2\pi\left(\frac{mu}{M} + \frac{nv}{N}\right)$ and substituting the symplectic decomposition $q(m,n) = c_1(m,n) + c_2(m,n)\,\mu_2$ of equation A.11,

$$\begin{aligned} F(u,v) &= \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} e^{-\mu\,\theta_{mn}(u,v)}\,c_1(m,n) \;+\; \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} e^{-\mu\,\theta_{mn}(u,v)}\,c_2(m,n)\,\mu_2 \\ &= \underbrace{\sum_{m,n} c_1(m,n)\,e^{-\mu\,\theta_{mn}(u,v)}}_{F_1(u,v)} \;+\; \underbrace{\Big(\sum_{m,n} c_2(m,n)\,e^{-\mu\,\theta_{mn}(u,v)}\Big)}_{F_2(u,v)}\,\mu_2, \end{aligned} \tag{A.14}$$

where the second line moves each exponential across c_1 and c_2 by the first rule of equation A.12. Identifying $\mu_1 = \mu$ with the complex unit, F_1 and F_2 are standard complex 2D DFTs of c_1 and c_2. Therefore, the QFT costs exactly two complex FFTs (Ell & Sangwine, 2007), and $F = F_1 + F_2\,\mu_2$ is the symplectic decomposition of the spectrum in the same basis, and the moduli entering equation 2 are $a_1 = |F_1|$ and $a_2 = |F_2|$ per bin. By contrast, a right-sided transform places the exponential to the right of q, where the second rule of equation A.12 gives $c_2\,\mu_2\,e^{-\mu\theta} = c_2\,e^{+\mu\theta}\,\mu_2$: the perplex part would be transformed at the conjugate frequency, which is why the left-sided form is kept throughout.

A.4 The Inverse QFT

The inverse QFT replaces $-\mu$ by μ in equation 1 and normalizes by $1/MN$,

$$q(m,n) = \frac{1}{MN}\sum_{u=0}^{M-1}\sum_{v=0}^{N-1} e^{\mu 2\pi\left(\frac{mu}{M}+\frac{nv}{N}\right)}\,F(u,v), \tag{A.15}$$

where the exponential remains on the left because quaternion multiplication is non-commutative. Inversion follows by substituting equation 1 into equation A.15. The two exponentials share the axis μ and therefore combine, and the geometric sum $\sum_{u=0}^{M-1} e^{\mu 2\pi ku/M} = M\,\delta_{k,0}$ for $|k| < M$, which holds verbatim in the commutative subfield spanned by $\{1,\mu\}$, collapses the double sum:

$$\begin{aligned} \frac{1}{MN}\sum_{u,v} e^{\mu\,\theta_{mn}(u,v)}\,F(u,v) &= \sum_{m',n'}\Big(\frac{1}{MN}\sum_{u,v} e^{\mu 2\pi\left(\frac{(m-m')u}{M}+\frac{(n-n')v}{N}\right)}\Big)\,q(m',n') \\ &= \sum_{m',n'} \delta_{mm'}\,\delta_{nn'}\,q(m',n') \;=\; q(m,n). \end{aligned} \tag{A.16}$$

Applying the reduction of equation A.14 to equation A.15 similarly yields $q = \mathcal{F}^{-1}(F_1) + \mathcal{F}^{-1}(F_2)\,\mu_2$, where $\mathcal{F}^{-1}$ denotes the standard inverse complex 2D FFT, so the iQFT also costs exactly two complex FFTs. When a corrupted spectrum from Section 4.1 is inverted, its mask is real and symmetric under $(u,v) \mapsto (-u,-v)$. It therefore preserves the Hermitian symmetry of the spectrum of each real coordinate of q. Consequently, equation A.15 returns a field with zero real part and real-valued R, G, and B coordinates—that is, a valid image—which allows the masked spectrum to be passed to the encoder. Our implementation performs the basis change, the two FFTs, and the inverse basis change on GPU tensors.

A.5 Polar Form and Coupling Quantities of the QFT Phase

Each QFT coefficient factors into an amplitude and a unit-quaternion phase, $F(u,v) = A(u,v)\,p(u,v)$ with $A = |F|$ and $|p| = 1$. In the frame of Appendix A.2, p admits the polar form of Bülow & Sommer (2001),

$$p = e^{\mu_1\varphi}\,e^{\mu_3\psi}\,e^{\mu_2\theta}, \qquad \varphi \in [-\pi,\pi),\ \theta \in [-\tfrac{\pi}{2},\tfrac{\pi}{2}),\ \psi \in [-\tfrac{\pi}{4},\tfrac{\pi}{4}], \tag{A.17}$$

where φ and θ generalize the complex phase, and ψ is a third angle with no complex counterpart. The coupling measure of Section 3 takes its name from this angle. We define $S_\psi = \sin 2\psi$, derive the quadratic form that our implementation computes, and then show that it equals the form $\rho\sin(\beta_2 - \beta_1)$ of equation 2.

Derivation of $S_\psi = \sin 2\psi$. Expanding the last two factors of equation A.17 with $\mu_3\mu_2 = \mu_1\mu_2\mu_2 = -\mu_1$,

$$\begin{aligned} e^{\mu_3\psi}\, e^{\mu_2\theta} &= (\cos\psi + \mu_3 \sin\psi)(\cos\theta + \mu_2 \sin\theta) \\ &= \underbrace{\cos\psi\cos\theta}_{A} + \underbrace{(-\sin\psi\sin\theta)}_{B}\,\mu_1 + \underbrace{\cos\psi\sin\theta}_{C}\,\mu_2 + \underbrace{\sin\psi\cos\theta}_{D}\,\mu_3, \end{aligned} \tag{A.18}$$

and left-multiplying by $e^{\mu_1\varphi} = \cos\varphi + \mu_1 \sin\varphi$ with $\mu_1\mu_2 = \mu_3$ and $\mu_1\mu_3 = -\mu_2$ gives the coordinates of p,

$$\begin{aligned} p_w &= A\cos\varphi - B\sin\varphi, & p_x &= B\cos\varphi + A\sin\varphi, \\ p_y &= C\cos\varphi - D\sin\varphi, & p_z &= D\cos\varphi + C\sin\varphi. \end{aligned} \tag{A.19}$$

In the combination $p_w p_z - p_x p_y$, the cross terms in φ cancel and $\cos^2\varphi + \sin^2\varphi = 1$, leaving

$$2\,(p_w p_z - p_x p_y) \;=\; 2\,(AD - BC) \;=\; 2\sin\psi\cos\psi\,(\cos^2\theta + \sin^2\theta) \;=\; \sin 2\psi, \tag{A.20}$$

so that $S_\psi = \sin 2\psi = 2\,(p_w p_z - p_x p_y)$. As a smooth quadratic in the components of p, this form requires no angle extraction and avoids the $\psi = \pm\pi/4$ singularities of equation A.17. This is the form our implementation computes.

Derivation of $S_\psi = \rho\sin(\beta_2 - \beta_1)$. Write the symplectic components of the unit phase in polar form, $c_1(p) = p_w + p_x\mu_1 = |c_1|\, e^{\mu_1\beta_1}$ and $c_2(p) = p_y + p_z\mu_1 = |c_2|\, e^{\mu_1\beta_2}$, so that $p_w = |c_1|\cos\beta_1$, $p_x = |c_1|\sin\beta_1$, $p_y = |c_2|\cos\beta_2$, and $p_z = |c_2|\sin\beta_2$. Then

$$p_w p_z - p_x p_y \;=\; |c_1||c_2|\,(\cos\beta_1\sin\beta_2 - \sin\beta_1\cos\beta_2) \;=\; |c_1||c_2|\,\sin(\beta_2 - \beta_1). \tag{A.21}$$

For $F = A\,p$, the moduli of equation 2 are $a_1 = A\,|c_1|$ and $a_2 = A\,|c_2|$ with $|c_1|^2 + |c_2|^2 = 1$, so

$$\rho \;=\; \frac{2\,a_1 a_2}{a_1^2 + a_2^2} \;=\; 2\,|c_1||c_2|, \qquad \text{hence} \qquad S_\psi \;=\; \rho\,\sin(\beta_2 - \beta_1)\,: \tag{A.22}$$

which is equation 2, the coupling-strength-weighted sine of the luminance–chrominance relative phase $\beta_2 - \beta_1$. The unit cross-spectrum of Section 3 is $\chi = \overline{c_1}\,c_2/(a_1 a_2) = e^{\mu_1(\beta_2 - \beta_1)}$, formed from the luminance and chrominance components c_1 and c_2 of F with moduli a_1 and a_2, so that $\operatorname{Im}\chi = \sin(\beta_2 - \beta_1)$ and $S_\psi = \rho\operatorname{Im}\chi$. The bound $|S_\psi| \le \rho$ is immediate, and ρ is unchanged whether computed from the components of F or of p because the amplitude A cancels. The angle ψ alone conflates coupling strength with relative phase, which is why the main text reads coupling through the pair (S_ψ, ρ) rather than through ψ.

Frame dependence. By the change-of-axis identity of Appendix A.2, replacing μ_2 by another admissible axis $e^{\mu_1\alpha}\mu_2$ multiplies c_2, and hence χ, by $e^{-\mu_1\alpha}$, shifts β_2 to $\beta_2 - \alpha$, and leaves a_1, a_2, and ρ unchanged. The choice of μ_2 therefore only translates the origin of the relative phase, as the spatial origin translates the phase of a complex FFT. Every quantity in this paper is computed in the single fixed frame of Appendix A.2, so all losses, probing measures, and statistics read relative phases against one common origin.

B Pretraining Details for MQM

This appendix presents the mask family (Appendix B.1), the full loss (Appendix B.2), the training configuration (Appendix B.3), and the per-channel control MCFM (Appendix B.4) of the pretraining task in Section 4.1.

B.1 Mask Strategy

Section 4.1 draws each mask $\mathcal{M}$ from three shapes. With a probability of 0.5 it is an annulus $r_{\text{lo}} < \|(u, v)\| \le r_{\text{hi}}$, with r_{lo} and r_{hi} drawn uniformly from $[6, 12]$ and $[64, 96]$ cycles per image. With a probability of 0.25, two double-wedge sectors at a random orientation cover 30% of the plane. Otherwise, it is a conjugate-paired random scatter that covers 30% of the bins. With a probability of 0.5, the complement of the chosen region is corrupted instead. Therefore, a sample either fills

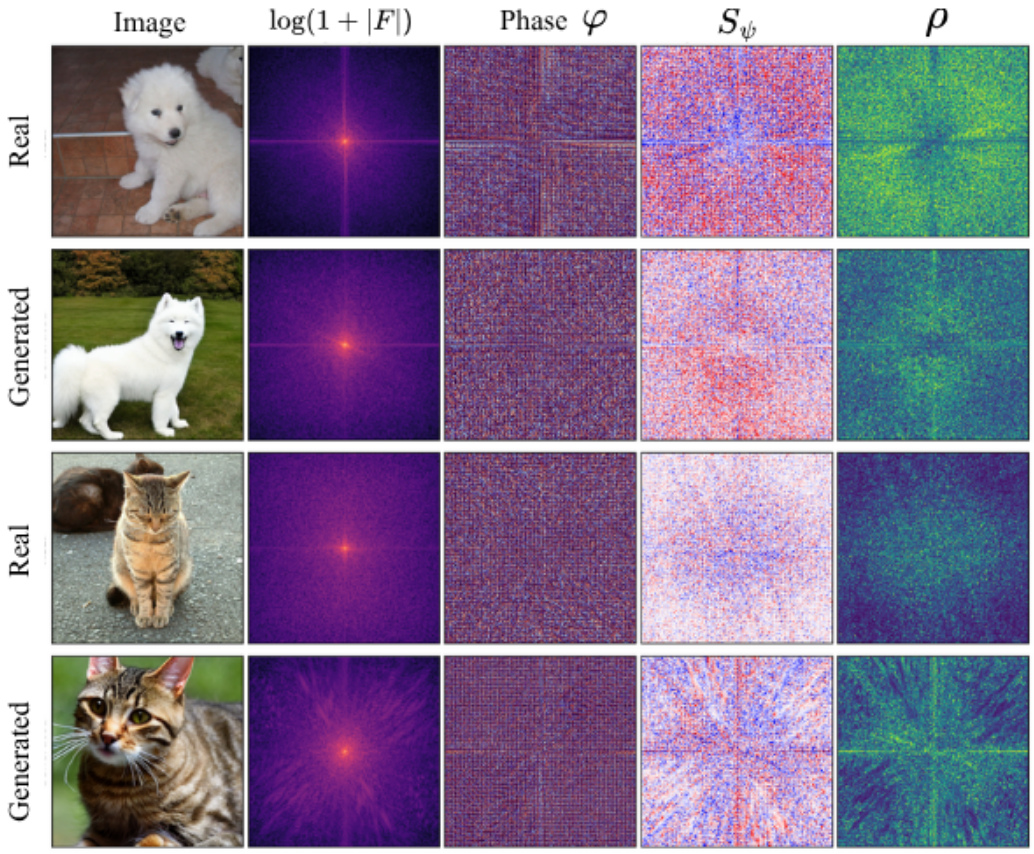


Figure 3: Quaternion Fourier quantities of real images and SD1.4 images (rows alternate real, generated). Columns show the input, log amplitude, luminance phase, S_ψ, and ρ.

in a band from its surroundings or reconstructs the surroundings from the band. Every mask is conjugate-symmetric and keeps the DC bin, so the corrupted spectrum inverts to a real image $\tilde{x}$ (Appendix A.4). The annulus is contiguous so that a masked bin cannot be interpolated from intact neighbors. The complement option supervises the low-frequency band in half of the samples. This setting performed best in the masking ablation of MFM (Xie et al., 2023). Figure 4 shows sixteen sampled masks with the corrupted inputs and the restorations. Keeping only the middle band gives an edge-like image, deleting it causes blur, and scatter produces ghosting.

B.2 Loss Function

Written out, the loss equation 3 is

$$\mathcal{L} = \underbrace{\frac{\sum_{\mathcal{M}} \left\|\hat{F} - F\right\|_1}{\sum_{\mathcal{M}} A}}_{\mathcal{L}_{\text{spec}}} + \lambda_{\text{cross}} \underbrace{\frac{\sum_{\mathcal{M}} \omega\, \ell(\hat{\chi}, \chi)}{\sum_{\mathcal{M}} \omega}}_{\mathcal{L}_{\text{cross}}} + \lambda_{\text{DC}}\, \mathcal{L}_{\text{DC}}, \qquad \omega = A\, \rho\, t, \tag{B.1}$$

where all sums run over $\mathcal{M}$, $A = |F|$ is the amplitude of the clean coefficient, ℓ is a smooth L_1 loss on the real and imaginary parts of the unit cross-spectra, ρ is the coupling strength of equation 2, and t is a taper that excludes the high-frequency band, which the surrounding context cannot predict reliably. Each term has its own normalizer, so each λ is the ratio of that term to $\mathcal{L}_{\text{spec}}$. In $\mathcal{L}_{\text{spec}}$, $\|\cdot\|_1$ sums the absolute values of the four quaternion components in the canonical basis $\{1, \mathrm{i}, \mathrm{j}, \mathrm{k}\}$, not in the frame of Appendix A.2. The L_1 norm changes with the frame, and in the canonical basis $\mathcal{L}_{\text{spec}}$ has the same form as the per-channel loss of MCFM equation B.2, a masked L_1 over the real components of each bin. The denominator is the clean amplitude, detached from the gradient, and only rescales the objective to order one. The gradient direction is that of the raw masked L_1, which

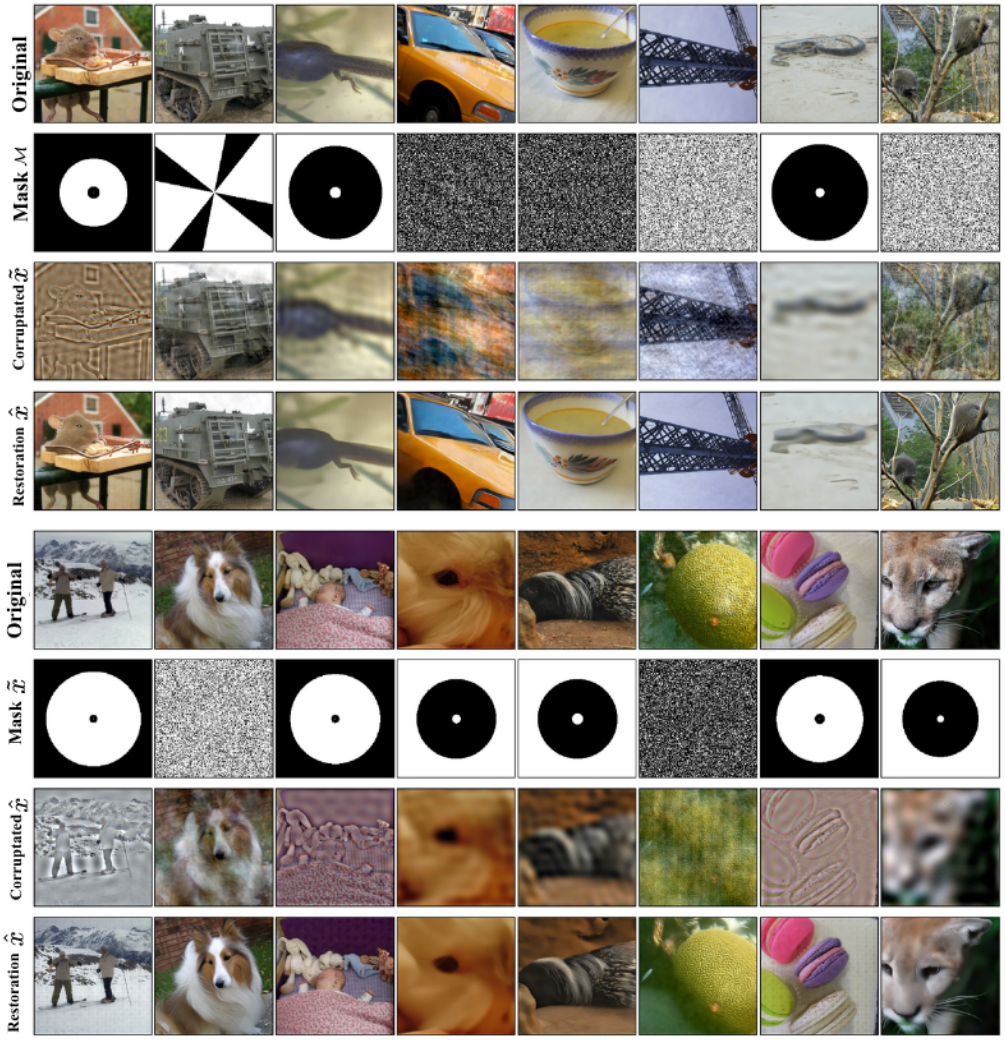


Figure 4: Pretraining corruptions of Section 4.1 on sixteen real validation images, in two blocks of four rows. Within each block, the rows show the image x, the sampled mask $\mathcal{M}$ (black bins corrupted, DC centered), the corrupted input $\tilde{x}$, and the restoration $\hat{x}$.

weights bins by their amplitude (Xie et al., 2023). The sums in $\mathcal{L}_{\text{spec}}$ and $\mathcal{L}_{\text{cross}}$ are pooled across the batch. In $\mathcal{L}_{\text{cross}}$, $\hat{\chi}$ is normalized by the moduli of the components of $\hat{F}$, clamped away from zero. Bins where either component of F vanishes have $\rho = 0$ and receive no cross supervision. The taper t decays from one to zero between 24 and 64 cycles. The DC coefficient is never masked, so the two masked terms never see it. The batch-pooled loss $\mathcal{L}_{\text{DC}} = \sum_b \|\hat{F}_b(0,0) - F_b(0,0)\|_1 / \sum_b |F_b(0,0)|$, with b the batch index, anchors it so that the global color cast must pass through the encoder.

B.3 Training Configuration

Table B.1 lists the pretraining configuration. The architecture, the data, and the 300-epoch schedule follow MFM (Xie et al., 2023). The public MFM checkpoint of Appendix D.1.2 therefore differs from MQM in the objective and in the mask family of Appendix B.1. For that, a decoder is fitted

there, whereas the MCFM control of Appendix B.4 shares MQM's mask family and differs only in the objective.

Table B.1: MQM pretraining configuration.

Encoder	ViT-B, patch size 16, input 224×224
Decoder	1×1 convolution + pixel shuffle
Data	ImageNet-1k, no labels
Epochs	300 (warmup 20)
Optimizer	AdamW, $\beta = (0.9, 0.95)$, weight decay 0.05
Learning rate	base 1.2×10^{-3}, cosine decay to 10^{-5}
Batch size	256 per GPU $\times$ 8 GPUs
Loss weights	$\lambda_{\text{cross}} = 0.5$, $\lambda_{\text{DC}} = 0.25$
Smooth-L_1 threshold of ℓ in equation B.1	0.2

B.4 The Per-Channel Control (MCFM)

To test the benefit of quaternion phase during pretraining, we construct a control backbone called masked channel-wise frequency modeling (MCFM), in which the three color channels are transformed independently using ordinary FFTs. Its loss is the per-channel counterpart of equation B.1:

$$\mathcal{L}_{\text{sc}} \;=\; \frac{\sum_{\mathcal{M}} \sum_{c} \left\| \hat{F}_c - F_c \right\|_1}{\sum_{\mathcal{M}} \sum_{c} |F_c|} \;+\; \lambda_{\text{DC}}\, \mathcal{L}_{\text{DC}}, \tag{B.2}$$

where F_c is the complex spectrum of channel c, and $\|\cdot\|_1$ sums the absolute real and imaginary components. The DC anchor is that of equation B.1, and $\lambda_{\text{DC}} = 0.25$ is the value of Table B.1. MCFM has no counterpart of $\mathcal{L}_{\text{cross}}$, since, like MFM, it supervises each channel's spectrum separately. All other training conditions follow Table B.1. MCFM is the per-channel control of Table 3.

C Detector Details and Calibration

This appendix specifies the parts of the detector of Section 4.2 that the main text summarizes: the probing measures (Appendix C.1), the statistics map (Appendix C.2), the real-image reference (Appendix C.3), and the specialist (Appendix C.4).

C.1 Restoration-Probing Measures

The probing features $r(x) \in \mathbb{R}^{96}$ of Section 4.2 compare three views of x with their restorations over eight bands and four measures. The views are x itself, x with the annulus between 12 and 64 cycles deleted from its quaternion spectrum, and x with the complement of that annulus deleted. The two corruptions are the two directions of the pretraining task, and their radii lie inside every pretraining annulus, whose inner radius is at most 12 and outer radius at least 64 cycles (Appendix B.1). The intact view measures the response to an image that needs no restoration. The network restores each view x_v to $\hat{x}_v$, and the quaternion spectrum $\hat{F}_v = \hat{A}_v\, \hat{p}_v$ of the restoration is compared with the spectrum $F = A\, p$ of x over eight radial bands $\mathcal{B}_b$ with boundaries $4, 8, 16, 32, 64, 96$, and 128 cycles:

$$r^{p}_{v,b} = \frac{\sum_{\mathcal{B}_b} A \left(1 - \langle \hat{p}_v, p \rangle\right)}{\sum_{\mathcal{B}_b} A}, \qquad r^{\psi}_{v,b} = \frac{\sum_{\mathcal{B}_b} A\rho \left| \hat{S}_{\psi,v} - S_{\psi} \right|}{\sum_{\mathcal{B}_b} A\rho},$$
$$r^{\chi}_{v,b} = \frac{\sum_{\mathcal{B}_b} A\rho \left\| \hat{\chi}_v - \chi \right\|_1}{\sum_{\mathcal{B}_b} A\rho}, \qquad r^{A}_{v,b} = \log \frac{\sum_{\mathcal{B}_b} \hat{A}_v}{\sum_{\mathcal{B}_b} A}. \tag{C.1}$$

Sums run over the bins of $\mathcal{B}_b$, $\langle \cdot, \cdot \rangle$ is the quaternion inner product of Appendix A.1, and $\|\cdot\|_1$ sums the absolute real and imaginary parts of the unit cross-spectrum χ of Section 3. The four measures read the rotation of the quaternion phase (r^{p}), the two coupling quantities of the shared phase frame (r^{ψ}, r^{χ}), and the energy lost or added (r^{A}). The coupling measures are weighted by $A\rho$ as in equation 3. Therefore, near-achromatic bins do not count. The band edges are octave-spaced up to

Table C.1: The 116 statistics of the map $u(\cdot)$, computed from the quaternion spectrum equation 1 with fixed formulas. The specialist reads them on $u(x - \hat{x})$ and $u(x) - u(\hat{x})$. Radial bands share the edges of the probing bands $\mathcal{B}_b$ above. Sectors split orientation into six and use three coarser high-frequency bands (32–64, 64–128, and above 128 cycles). Shells pair octaves $(f, 2f)$, and blocks use a 16×16 blockwise transform.

Group	Statistics	Count
Band moments	$A\rho$-weighted mean, $\lvert\cdot\rvert$-mean, and std of S_ψ; mean ρ; log band energy	$8 \times 5 = 40$
Sector moments	log energy, mean ρ, $\lvert S_\psi\rvert$-mean per sector and high band	$6 \times 3 \times 3 = 54$
Shell coherences	luminance phase-doubling coherence; $\lvert\Delta S_\psi\rvert$ across the octave	$4 \times 2 = 8$
Block covariation	local luminance vs. the two symplectic energies (slope, corr.); chroma ratio; energy corr.	6
Phase gradient	A-weighted quaternion phase step between neighboring bins, per band	8
Total		116

64 cycles, the range over which the shell coherences of Table C.1 pair octaves $(f, 2f)$, and then step by 32 cycles. On the 224×224 grid, the Nyquist frequency along an axis is 112 cycles, so the band from 96 to 128 cycles straddles it, and the last band holds only the corner bins, up to 158 cycles. These two bands lie outside every pretraining annulus but are corrupted by the complement masks of pretraining and by the complement view, so are read under corruptions the network was trained on. The deviation score reads $\bar{r}$, the features averaged over the five 224×224 crops of Section 5.2, and the specialist reads the center crop.

C.2 Spectral Statistics Read by the Specialist

This appendix lists the fixed statistics that the specialist of Section 4.2 reads in addition to the probing features. The map $u(\cdot)$ collects 116 statistics from the quaternion spectrum equation 1 of its argument, with fixed formulas and no learned parameters. Table C.1 lists the statistic groups. In addition to the amplitude A, the statistics read the coupling pair (S_ψ, ρ) of equation 2 through band and sector moments, octave coherences, blockwise covariation with luminance, and a phase-gradient coherence. The specialist evaluates $u(\cdot)$ on two images derived from the intact view x and its restoration $\hat{x}$: the residual image $x - \hat{x}$, giving $u(x - \hat{x})$, and the change of the statistics under restoration, $u(x) - u(\hat{x})$. Together, the two vectors record, band by band, where the restoration moved the image and in which quaternion quantity, without any learning.

C.3 Real-Image Reference and Choice of k

The number of clusters k in the reference of Section 4.2 is chosen on real images alone, and the choice is $k = 4$. More clusters describe the diversity of real images more finely, so we take the largest k whose fit still generalizes to held-out real images. No generated image and no detection metric enter the selection. Table C.2 reports the criterion for $k \in \{1, 2, 4, 8\}$. Twenty percent of the real training images are held out, and the reference is fitted on the remainder with the deployed procedure. The criterion is the mean standardized score of the held-out real images. It stays near zero as long as the fitted clusters describe the population, and it turns positive once the clusters begin to fit the particular split. The fit generalizes up to $k = 4$ and degrades at $k = 8$, which selects $k = 4$. The table also reports the unseen-generator AUROC of the deviation score refitted on all real training images. This column was measured after the selection and was not part of it, and it peaks at the selected value. Each cluster j carries its sample mean η_j and a covariance Σ_j obtained by shrinking the sample covariance by 0.1 toward a scaled identity. Because k-means is sensitive to initialization, the deployed fit uses the minimum-distortion run over ten seeds, a criterion that again consults no label.

Deployment of the score. The deployed score is equation 4 with these cluster parameters. Before thresholding, the minimum is additionally standardized on all real training images. This transformation is monotone and leaves the decision unchanged. All selections are fixed before any generator is consulted.

Table C.2: Choice of the number of clusters by the held-out fit criterion of the text. Values near zero indicate that the fitted clusters generalize. Unseen AUROC: AUROC of the deviation score averaged over the seven GenImage generators other than SD1.4, measured after the selection.

k	Held-out fit	Unseen AUROC
1	-0.009	0.839
2	-0.008	0.859
4 (selected)	-0.008	**0.867**
8	$+0.006$	0.839

C.4 Specialist Architecture

The specialist of Section 4.2 has 55 query tokens, a context of 49 patch tokens, and a small transformer between them. The query tokens are the 24 probing features of equation C.1, one four-dimensional token per view and band, and 31 statistics tokens, into which $u(x-\hat{x})$ and $u(x)-u(\hat{x})$ are grouped by the statistic groups of Table C.1: eight band tokens (the five band moments plus the phase-gradient statistic of that band), eighteen sector tokens, four shell tokens, and one block token, of dimension twelve, six, four, and twelve. Every query token is embedded into $D = 768$ dimensions with three learned embeddings, a radial-band identity, a group identity (probing or statistics), and a position. The context is the encoder's final-layer patch tokens of the center crop, average-pooled from 14×14 to 7×7 and linearly projected with its own position embeddings. All inputs are standardized on the real training images. The query tokens cross-attend to the context and pass through one self-attention layer and a feed-forward block, with 8 heads and pre-normalization throughout. Attention pooling with a learned query feeds a two-layer head that outputs s_{attn}. The specialist has about 10 million parameters, 12% of the backbone. It is trained for 40 epochs with binary cross-entropy on 20,000 real and 20,000 SD1.4 images, with the checkpoint chosen on held-out SD1.4 validation data. The real images are from the SD1.4 subset of the GenImage training split, and 5,000 of them are also among the 40,000 real training images of Section 5.1, which pool 5,000 from each of the eight training subsets.

D Additional Analysis

This appendix expands the analysis of Section 6: the backbone (Appendix D.1), the decision rule (Appendix D.2), the imaging channel and JPEG recompression (Appendices D.3 and D.4), and the two zero-shot benchmarks (Appendices D.5 and D.6).

D.1 Quaternion Representation and Pretraining Ablations

D.1.1 Detection Performance

This appendix gives the per-generator results behind the two controls of Table 3. The no-cross backbone retains the quaternion loss $\mathcal{L}_{\mathrm{spec}}$ and the DC anchor of equation 3, but sets $\lambda_{\mathrm{cross}} = 0$. The reference is refitted, and the specialist is retrained using the no-cross features, and every threshold is recomputed at the declared level from the corresponding real training images. Table D.1 reports per-generator results for MCFM, the no-cross backbone, and MQM.

Two results follow from Table D.1. First, the cross term $\mathcal{L}_{\mathrm{cross}}$ matters on every generator. Removing it reduces accuracy everywhere by 6.0 points on average. The largest loss falls on BigGAN, 19.7 points, and the other generators lose 0.1–9.0 points. The effect is not limited to the generators best separated by the phase-deviation score, since accuracy on SD1.4, SD1.5, and Wukong also decreases by 2.0–2.8 points. Second, the quaternion basis without the cross term is not superior to the per-channel control. Replacing the basis entirely with MCFM reduces the average by 4.9 points relative to MQM, with the largest losses on ADM and VQDM, but MCFM is not uniformly inferior. On BigGAN it exceeds QPP (91.6 versus 88.8), the one generator on which the full quaternion system does not lead. The two controls are close on both detection (83.3 and 82.2) and the illuminant probe (Appendix D.1.4), and MQM leads both.

Table D.1: Per-generator backbone comparison, full-system accuracy (%). Specialists and thresholds are refitted per backbone.

Generator	MCFM	MQM, no $\mathcal{L}_{\text{cross}}$	MQM (QPP)
SD1.4	87.2	88.6	**90.6**
SD1.5	87.6	88.4	**91.2**
Wukong	86.5	87.4	**89.9**
MJ	73.1	75.0	**75.1**
GLIDE	88.8	82.2	**91.2**
VQDM	82.3	86.7	**90.5**
BigGAN	**91.6**	69.1	88.8
ADM	69.6	80.5	**88.5**
Avg.	83.3	82.2	**88.2**

D.1.2 Public Encoders with a Linear Decoder

This appendix specifies the two public encoders in Table 3, how their decoders were obtained, and their per-generator results. Both keep the detector unchanged and replace only the backbone. The MFM checkpoint (Xie et al., 2023) is a ViT-B/16 pretrained on ImageNet-1k for 300 epochs with a per-channel masked-frequency loss, and the MAE checkpoint (He et al., 2022) is a ViT-B/16 pretrained on the same images for 1600 epochs with a masked-pixel loss. Both checkpoints contain the encoder only, so the decoder of Section 4.1 is fitted on each frozen encoder with the per-channel frequency loss of the MFM implementation (Xie et al., 2023), $\mathcal{L}_{\text{dec}} = \frac{1}{3|\mathcal{M}|}\sum_{\mathcal{M}}\sum_{c}|\hat{F}_c - F_c|$, the modulus of the complex difference averaged over the masked bins and the three channels, with no DC term. It differs from equation B.2 in the per-bin norm (modulus instead of the sum of absolute real and imaginary parts); the normalizer of equation B.2 only rescales it (Appendix B.2). The decoder is trained on ImageNet photographs for the same number of steps on every encoder. The reference, the specialist, and all thresholds are then refitted using the other backbones' procedure and from the same image lists.

Table D.2 reports per-generator results for the public encoders and MQM, for the full system, for the deviation score alone, and for the specialist score. The public encoders keep the seen family and lose the four generators that only the deviation score detects. Both retain the seen SD family and Wukong (82–90 for the full system, specialist AUROC 0.96–1.00) and fail on ADM, GLIDE, VQDM, and BigGAN. On these four generators, the specialist is inversely predictive for every backbone (AUROC 0.16–0.47). The decision, therefore, rests on the deviation score. With the public encoders, that score separates these generators less well than with MQM. Alone, it stays at 59.5–60.7 on ADM, and its AUROC over the eight generators is 0.785 and 0.795 against 0.848 for MQM.

Table D.2: Per-generator comparison of the public encoders with MQM. Full system: accuracy (%) at the declared thresholds. Deviation only: accuracy (%) at the declared 1σ threshold and AUROC of the deviation score. Specialist: AUROC of the specialist score. Every component is refitted per backbone.

	Full system			Deviation only, accuracy			Deviation only, AUROC			Specialist, AUROC		
Generator	MFM	MAE	MQM	MFM	MAE	MQM	MFM	MAE	MQM	MFM	MAE	MQM
SD1.4	82.4	89.9	**90.6**	64.5	64.9	58.3	0.759	0.774	0.720	0.963	0.996	0.992
SD1.5	82.2	90.3	**91.2**	64.0	64.9	58.7	0.757	0.772	0.721	0.960	0.995	0.992
Wukong	82.7	89.0	**89.9**	63.6	64.3	58.7	0.763	0.771	0.725	0.962	0.991	0.990
MJ	74.0	**83.0**	75.1	69.3	69.3	66.2	0.811	0.809	0.802	0.791	0.925	0.853
GLIDE	66.8	75.4	**91.2**	68.2	76.4	92.6	0.801	0.867	0.959	0.361	0.389	0.374
VQDM	73.1	76.5	**90.5**	74.3	77.4	91.8	0.838	0.869	0.958	0.328	0.474	0.401
BigGAN	71.5	60.4	**88.8**	72.7	61.6	90.1	0.830	0.733	0.961	0.244	0.160	0.255
ADM	58.3	59.9	**88.5**	59.5	60.7	89.9	0.718	0.762	0.941	0.314	0.459	0.284
Avg.	73.9	78.1	**88.2**	67.0	67.4	**75.8**	0.785	0.795	**0.848**	0.615	0.674	0.643

D.1.3 Phase and Amplitude Measures at Readout

Appendices D.1.1 and D.1.2 change the representation. This appendix keeps the MQM representation fixed and varies only the measures of equation C.1 that the deviation score reads. The question

is whether the deviation the score detects is carried by the phase-side measures. For each subset of the four measures, the probing features $\bar{r}$ are restricted to the corresponding columns, the real-image reference of Section 4.2 is refitted on the real training images with the deployed procedure, and the 1σ threshold is re-read from the same images. The specialist and its 2σ threshold are unchanged. Single measures (24 dimensions) and the two 48-dimensional pairs, the coupling pair (r^{ψ}, r^{χ}) defined in the shared phase frame against the non-coupling pair (r^{p}, r^{A}), are compared at equal dimensionality. Table D.3 reports the results.

At equal dimensionality, the phase-side measures exceed the amplitude measure. Alone, r^{χ} (71.6, AUROC 0.815) and r^{p} (70.8, 0.795) outperform r^{A} (65.8, 0.740), and the coupling pair outperforms the non-coupling pair (73.7 versus 70.3, AUROC 0.842 versus 0.786). The three phase measures without r^{A} (72 dimensions) exceed the deployed four measures on the deviation score alone (77.8 versus 75.8, AUROC 0.864 versus 0.848), so the amplitude measure adds nothing to the deviation score. In the full system, the deployed configuration is marginally higher (88.2 versus 87.5), so r^{A} contributes only through its complementarity with the specialist. Two caveats apply. The coupling measure r^{ψ} alone (64.8) is below the amplitude measure because the ρ weighting discards near-achromatic bins and leaves little information in isolation, so the coupling pair acts as a unit. Because the backbone is fixed, r^{A} measures the amplitude change of a restoration produced under a phase objective, not an amplitude representation. That is why the pretraining-level comparison of Appendix D.1.2 is needed.

Table D.3: Readout of subsets of the measures in equation C.1 from the fixed MQM representation, GenImage average. Deviation only: accuracy (%) at the re-read 1σ threshold and AUROC of the deviation score. Full system: accuracy (%) with the specialist deployed.

Measures	Dim.	Deviation only (acc)	Deviation only (AUROC)	Full system (acc)
r^{p}	24	70.8	0.795	84.3
r^{ψ}	24	64.8	0.750	76.3
r^{χ}	24	71.6	0.815	81.5
r^{A} (amplitude)	24	65.8	0.740	78.9
r^{ψ}, r^{χ} (coupling pair)	48	73.7	0.842	82.6
r^{p}, r^{A} (non-coupling pair)	48	70.3	0.786	82.9
$r^{p}, r^{\psi}, r^{\chi}$ (phase measures)	72	**77.8**	**0.864**	87.5
All four (deployed)	96	75.8	0.848	**88.2**

D.1.4 The Backbone outside Detection

This appendix evaluates the MQM backbone on two tasks beyond detection: ImageNet-1k classification, to check that the representation is not specific to detection, and illuminant estimation on Gehler-Shi, to test a prediction of the cross term. We use full fine-tuning for the former and a linear probe on the frozen backbone for the latter.

ImageNet-1k fine-tuning. Table D.4 reports end-to-end fine-tuning of the backbone for classification using the configuration in Table D.5. MQM reaches 82.1% top-1 accuracy, outperforming training from scratch and trailing the masked-spectrum and pixel-reconstruction objectives by 1.0 and 1.5 points. The quaternion-phase objective therefore costs at most 1.5 points of classification accuracy against the two public objectives; therefore, the backbone is not specific to detection.

Illuminant estimation probe. This probe predicts the color of the scene illuminant of a photograph from the frozen backbone's [CLS] feature with a linear ridge regression. We use Gehler-Shi (Gehler et al., 2008; Shi & Funt, 2010), 568 raw photographs with a measured illuminant, mask out the color checker, and report the angle between the predicted and measured illuminants over three camera-stratified folds. The ridge strength is selected using the training fold, and inputs are normalized with the ImageNet statistics of pretraining.

The illuminant is a per-channel gain $\mathrm{diag}(g_R, g_G, g_B)$. Split it into a mean gain $\bar{g}$ and a zero-mean cast $\Delta = \Delta_R\,\mathrm{i} + \Delta_G\,\mathrm{j} + \Delta_B\,\mathrm{k}$. In the coordinates equation A.13 and by linearity of the DFT,

Table D.4: ImageNet-1k top-1 (%) of ViT-B/16 after end-to-end fine-tuning (recipe in Table D.5). MAE and MFM are published numbers, and DeiT-B is supervised from scratch.

Backbone	Pretraining	PT epochs	Top-1
MAE (He et al., 2022)	pixel restoration	1600	83.6
MFM (Xie et al., 2023)	masked-spectrum restoration	300	83.1
DeiT-B (Touvron et al., 2021)	none (supervised, 300 epochs)	–	81.8
MCFM equation B.2	channel-wise restoration	300	81.9
MQM (ours)	equation 3	300	82.1

Table D.5: ImageNet-1k fine-tuning configuration.

Model	ViT-B/16, input 224×224, [CLS] head, LayerScale init 0.1
Epochs	100 (warmup 5)
Optimizer	AdamW, $\beta = (0.9, 0.999)$, weight decay 0.05
Learning rate	peak 8×10^{-3} at batch 2048, cosine decay to 10^{-6}
Layer-wise LR decay	0.8
Gradient clipping	none
Batch size	256 per GPU $\times$ 8 GPUs
Stochastic depth	0.1
Label smoothing	0.1
Augmentation	RandAugment (9, 0.5), Mixup 0.8, CutMix 1.0, random erasing 0.25
Evaluation	center crop 0.875, bicubic

the luminance and chrominance spectra become $F_1' \approx \bar{g}\, F_1$ and $F_2' \approx \bar{g}\, F_2 + \frac{\delta}{\sqrt{3}}\, \mathcal{F}(b)$, where $\delta = \langle \Delta, \mu_2 \rangle + \langle \Delta, \mu_3 \rangle\, \mu_1$ holds the opponent-color coordinates of the cast and the neglected terms couple Δ to the chromatic signal, which is considerably weaker than luminance in natural images. The illuminant thus adds a copy of the luminance spectrum to the chrominance spectrum in one fixed direction and shifts the relative phase χ of every bin. Because the cross term $\mathcal{L}_{\text{cross}}$ supervises χ, MQM should encode the illuminant, whereas a per-channel objective receives no such supervision.

MQM has the lowest mean and median error in Table D.6 (2.81/2.22 degrees). Removing the cross term raises it to 3.12/2.46, and the per-channel control MCFM gives 3.19/2.51. We compare backbones by a paired bootstrap: the 568 images are resampled with replacement 10,000 times, and on each resample we compute the difference in mean error between each backbone and MQM, $\Delta = \text{error}(X) - \text{error}(\text{MQM})$. MQM is lower than every other backbone ($\Delta = +0.31$ degrees [$+0.16$, $+0.46$] against the no-cross control, $+0.38$ [$+0.21$, $+0.56$] against MCFM, $+0.27$ [$+0.10$, $+0.45$] against MAE, and $+0.85$ [$+0.64$, $+1.06$] against MFM). The differences among MAE and the two controls are at most 0.11 degrees (0.04 between the no-cross control and MAE, 0.07 between the two controls, 0.11 between MCFM and MAE), each with an interval that includes zero. The intervals reflect variation over test images for fixed backbones and fitted probes; each backbone is pretrained once, so they do not include variation across pretraining seeds. These results are consistent with the cross term placing the illuminant, a physical scene property, in the representation. The coupling measures r^{ψ} and r^{χ} in equation C.1 read the same relative phase at detection time.

D.2 Analysis of the Decision Rule

D.2.1 Sensitivity of the Declared Thresholds

This appendix measures how much the result depends on the two declared levels of Section 4.2. Table D.7 varies them after all design choices have been fixed. The declared point $(1\sigma, 2\sigma)$, selected before this sweep, performs best among the seven measured settings. Sensitivity is asymmetric and concentrated in the deviation score. Moving its threshold by half a σ reduces accuracy by 4.0–5.3 points, and tightening it to 2σ reduces the average to 74.2. The specialist is less sensitive. Half a σ in either direction costs 1.1–1.3 points, and 3σ costs 4.2. The declared 2σ level is the measured optimum for the specialist. At the tighter $(2\sigma, 2\sigma)$ point, whose total declared budget is at most 4.6% of the real training images, the average is 74.2, still above the best retrained baseline (73.3) at its own operating point.

Table D.6: Illuminant estimation on Gehler-Shi with a frozen backbone and a linear probe. Angular error in degrees (lower is better). Gray-world and the RGB-statistics probe use no network. Δ is the mean error minus that of MQM, with a 95% paired-bootstrap interval over the 568 images.

Backbone	Mean	Median	Δ vs. MQM [95% CI]
Random initialization	5.16	4.13	–
Gray-world	5.09	3.86	–
RGB statistics	3.55	2.87	–
MFM (Xie et al., 2023)	3.66	2.89	$+0.85$ [$+0.64$, $+1.06$]
MAE (He et al., 2022)	3.08	2.57	$+0.27$ [$+0.10$, $+0.45$]
MCFM equation B.2	3.19	2.51	$+0.38$ [$+0.21$, $+0.56$]
MQM without $\mathcal{L}_{\text{cross}}$ ($\lambda_{\text{cross}} = 0$)	3.12	2.46	$+0.31$ [$+0.16$, $+0.46$]
MQM (ours)	**2.81**	**2.22**	–

Table D.7: Threshold sensitivity: accuracy (%) averaged over the eight GenImage subsets when the declared levels are moved. The declared point was fixed before this sweep.

Deviation	Specialist	Avg.
0.5σ	2σ	84.2
1σ (declared)	2σ (declared)	**88.2**
1.5σ	2σ	82.9
2σ	2σ	74.2
1σ	1.5σ	87.1
1σ	2.5σ	86.9
1σ	3σ	84.0

Table D.8: Decision-level OR versus score fusion, accuracy (%) averaged over the eight GenImage subsets. Every rule is thresholded at the same false-positive rate on the real training images (17.6%).

Rule	Avg.
OR over thresholded scores (equation 5)	**88.2**
Maximum of standardized scores	83.9
Sum of standardized scores	81.7

D.2.2 Decision-Level OR versus Score Fusion

Table D.8 compares the OR of equation 5 with fusing the two scores and thresholding once. Each score is standardized on the real training images, the fused score is their sum or their maximum, and its threshold is set so that it misclassifies the same fraction of real training images as the OR (17.6%). Every rule therefore spends the same false-positive budget. Both fusions lose accuracy. The sum loses 6.5 points, because the specialist score then enters every decision with the same weight whether or not it is informative, and it is inversely predictive on four generators (Table D.2). In the OR, the specialist counts only above its own threshold, so the real images it can misclassify are bounded by its declared 2.28%. The maximum loses 4.3 points, because a real image is misclassified whenever either score exceeds the single threshold, so the score with the wider real-image tail consumes most of the shared budget.

D.2.3 False-Positive Rate outside the Calibration Images

The declared levels of Section 4.2 fix the false-positive rate only on the real training images, where the thresholds are read. Table D.9 reports the measured rate of the OR decision, with thresholds unchanged, on every real class used in the paper. On the GenImage validation reals, drawn from the same distribution as the calibration images, the rate is 15.0% (14.6–15.4 across the eight subsets), slightly below the 17.6% of the calibration images. On the two Synthbuster real classes it is lower, and on Chameleon it exceeds the union bound, which is the failure discussed in Appendix D.6.

D.3 The Imaging Channel and Its Limitation

D.3.1 Effect on Detection Accuracy

This appendix measures what the imaging channel of Section 4.2 contributes by rebuilding the system without it. The channel, JPEG 75 followed by bicubic down-and-up scaling at $0.75\times$, is applied to both classes before the canonicalization of Section 5.2. We rebuild the complete system without this stage while retaining canonicalization. Every trained or calibrated component is refitted on

Table D.9: How often the OR decision equation 5 labels a real image as generated, on every real-image set used in the paper. The thresholds were set once, on the GenImage real training images, where the measured rate is 17.6% and the declared bound of Section 4.2 is 18.15%. Every row applies those thresholds to real images they never saw. The Synthbuster and Chameleon values are one minus the real-class acceptance in Tables D.12 and D.13.

Real-image set	Count	False-positive rate (%)
GenImage (Table 1)	48,000	15.0
Synthbuster, MS-COCO real class (Table 2)	123,287	12.8
Synthbuster, RAISE-1k real class (Table 2)	1,000	5.3
Chameleon	14,863	21.9

channel-free features. The reference is refitted on the real training images, the specialist is retrained, and all thresholds are recomputed at the declared levels. We repeat this procedure for the per-channel MCFM control of Appendix B.4 and for the three baselines that could be retrained within the same computational budget, namely NPR, Swin-T, and C2P-CLIP. Each baseline is retrained on canonicalized images without the channel.

Table D.10: Effect of the imaging channel, accuracy (%), GenImage. All components are refitted.

Method	Imaging channel	SD1.4	SD1.5	Wukong	MJ	GLIDE	VQDM	BigGAN	ADM	Avg.
QPP (MQM)	with	90.6	91.2	89.9	75.1	91.2	90.5	88.8	88.5	**88.2**
QPP (MQM)	without	87.0	86.0	84.2	62.8	73.3	73.0	72.9	59.4	74.8
MCFM	with	87.2	87.6	86.5	73.1	88.8	82.3	91.6	69.6	83.3
MCFM	without	85.0	83.9	82.3	62.1	62.9	73.8	68.8	58.1	72.1
C2P-CLIP	with	99.9	99.9	99.9	85.5	50.3	50.7	49.9	50.1	73.3
C2P-CLIP	without	99.8	99.8	98.8	73.8	51.1	49.9	49.9	49.9	71.6
Swin-T	with	99.2	99.0	96.8	67.3	54.5	53.7	49.7	51.7	71.5
Swin-T	without	99.3	99.1	96.4	66.3	60.1	50.4	52.7	51.6	72.0
NPR	with	90.8	90.5	90.6	66.2	54.1	53.0	46.8	48.4	67.6
NPR	without	94.5	93.9	90.6	63.2	71.5	54.5	57.7	52.0	72.2

The channel adds 13.4 points to the full system, and the gain goes to the deviation score rather than to the baselines (Table D.10). The gain is not uniform. Without the channel, accuracy on SD1.4, SD1.5, and Wukong decreases by 3.6–5.7 points, whereas ADM falls by 29 points to near chance while GLIDE, VQDM, BigGAN, and Midjourney lose 12–18 points. The MCFM control gains 11.2 points from the same stage, so the gain from the channel does not depend on the quaternion objective. The three retrained baselines show no comparable gain. C2P-CLIP improves by only 1.7 points, whereas Swin-T and NPR decline by 0.5 and 4.6 points, respectively. Thus, the channel does not improve detection uniformly but yields large gains for a deviation score fitted on real images. The contribution of the quaternion backbone, measured by the difference between QPP and the MCFM control, is 4.9 points with the channel and 2.7 points without it. The reduction is concentrated on ADM and VQDM, where the quaternion objective adds 18.9 and 8.2 points with the channel but 1.3 and -0.8 points without it.

D.3.2 Limitation: Inputs That Have Lost High-Frequency Content

The gain from the channel rests on the assumption stated in Section 7. The channel reads high-frequency content, so inputs that have already lost such content through stronger resizing or blurring fall outside the tested conditions. When the input must remain unmodified, the system averages 74.8 rather than 88.2, with the accuracy on ADM approaching chance. Nevertheless, it still outperforms every retrained baseline under the same canonicalization. The channel itself is fixed, label-free, and applied identically to every input. Appendix D.4 reports the same stage under upstream recompression.

D.4 Robustness to JPEG Recompression

This appendix tests QPP when every test image is JPEG-recompressed before it enters the imaging channel, that is, on the distributed file, as would occur during image delivery. The recompressed image then passes the imaging channel and the canonicalization of Section 5.2 unchanged. Table D.11

Table D.11: Accuracy (%), GenImage average, with test images JPEG-recompressed at quality q before the imaging channel and canonicalization. Baselines follow the re-evaluation of Kang et al. (2025) under their own preprocessing configurations. Each row is comparable only to its own clean value, which can differ marginally from Table 1. QPP keeps its declared thresholds.

Method	clean	$q = 90$	$q = 80$	$q = 70$	$q = 50$
AIDE (Yan et al., 2025a)	86.9	57.2	58.5	60.5	–
NPR (Tan et al., 2024a)	88.6	71.1	70.7	70.1	–
UnivFD (Ojha et al., 2023)	88.8	73.2	70.4	69.8	–
DRCT (Chen et al., 2024)	89.5	81.0	78.1	76.2	–
QPP (ours)	88.2	88.8	87.5	88.9	89.4

compares QPP, with thresholds unchanged, against published results when every test image is further JPEG-compressed at quality q. QPP stays between 87.5 and 89.4 down to $q = 50$, and at $q = 50$ it is 1.2 points above its clean value, whereas the published detectors lose 13 to 26 points by $q = 70$, although AIDE, UnivFD, and DRCT train with JPEG augmentation. The imaging channel of Section 4.2 re-images every fitting and test input through the same JPEG 75 stage, so a recompressed test image passes the same stage as the images on which the reference and thresholds were fitted. We attribute the stability, including the small rise, to this shared stage. After recompression at $q \leq 75$ a test image carries a compression history closer to that of the JPEG 75 real training images than a clean image does.

D.5 Synthbuster under Compression-Depth and Real-Domain Controls

This appendix bounds how much of the Synthbuster result in Table 2 could derive from the compression history of the real class. That class, MS-COCO, carries a JPEG compression history that the raw PNG generator outputs lack. The imaging channel of Section 4.2 narrows this asymmetry as a side effect but cannot equate a recompressed history with a once-compressed one. Two controls bound its effect, with every trained or calibrated component and threshold retained as for GenImage (Table D.12). The first control matches the compression depth of the two classes. Every generated image is pre-compressed at its native resolution with a JPEG quality sampled from the measured MS-COCO quality distribution. This lowers the balanced accuracy by 1.1 points (77.8 to 76.7) and the phase-deviation AUROC by 0.012 (0.848 to 0.836). The second control replaces the real class as an entire domain with 1,000 RAISE-1k photographs (Dang-Nguyen et al., 2015). They are developed from camera raw and standardized by a declared web-distribution stage, downscaling to a short side of 480 pixels followed by JPEG compression at quality 90, with both values taken from the measured distribution of the MS-COCO pool. This yields 81.6 with a phase-deviation AUROC of 0.860 (Table 2, last row, and Table D.12), above the 77.8 and 0.848 measured against MS-COCO. Because the generated images and thresholds are unchanged, that row differs from the MS-COCO row only in the real-side acceptance, the fraction of real images accepted as real (87.2% to 94.7%). The history asymmetry therefore accounts for at most 1.1 points, and the result holds with a different real domain.

Table D.12: Synthbuster controls on the compression history and the real domain, with every component and threshold as trained or calibrated on GenImage. Balanced accuracy (%) per generator, averaged over the nine generators, phase-deviation AUROC over the same nine, and acceptance (%) of the real class. As distributed: MS-COCO reals against raw PNG outputs (the QPP row of Table 2). Pre-compressed: every generated image JPEG-compressed at native resolution with a quality sampled from the measured MS-COCO distribution. RAISE-1k: the real class replaced by 1,000 RAISE-1k photographs under the declared web-distribution standardization (the last row of Table 2).

Condition	DALL·E 2	DALL·E 3	Firefly	GLIDE	MJ v5	SD1.3	SD1.4	SD2	SDXL	Avg.	AUROC	Real acc.
As distributed (MS-COCO reals, raw PNG)	77.3	74.8	82.1	91.5	63.9	86.4	86.5	74.1	63.8	77.8	0.848	87.2
Pre-compressed generated images	73.4	74.9	79.1	91.8	63.8	85.5	85.7	73.2	63.3	76.7	0.836	87.2
RAISE-1k reals, raw PNG	81.1	78.5	85.8	95.2	67.7	90.2	90.2	77.9	67.6	81.6	0.860	94.7

D.6 Stress Test under Real-Image Domain Shift: Chameleon

This appendix stress-tests QPP on a benchmark whose real class has heterogeneous imaging histories. Chameleon (Yan et al., 2025a) pairs generated images that human annotators could not distin-

guish from photographs with real images from the same online platforms and remains challenging for existing detectors. We evaluate the full benchmark zero-shot, using all 14,863 real and 11,170 generated images, with every component and threshold as trained or calibrated on GenImage and SD1.4 as the only seen generator.

Table D.13: AUROC on Chameleon. Baselines are literature values with their own training data. QPP is evaluated zero-shot, with every component as trained or calibrated on GenImage and SD1.4 as the only seen generator. The full system outputs a binary decision, so its AUROC equals its balanced accuracy at the declared thresholds.

Method	AUROC
CNNDetection (Wang et al., 2020)	0.560
UnivFD (Ojha et al., 2023)	0.627
NPR (Tan et al., 2024a)	0.690
SPAI (Karageorgiou et al., 2025)	0.789
DIRE (Wang et al., 2023)	0.814
QPP phase-deviation score (zero-shot)	0.382
QPP attention specialist (zero-shot)	0.744
QPP full system, OR decision (zero-shot)	0.512

The two scores of QPP differ on this benchmark. The specialist, which reads the quaternion features of the frozen backbone and saw only SD1.4, reaches an AUROC of 0.744, close to the published detectors (0.690–0.814), which were trained on their own data. The frozen features therefore still separate images that were selected to deceive human observers. By contrast, the phase-deviation score is inversely predictive, with an AUROC of 0.382. We attribute this inversion to the composition of the benchmark. The real class contains platform photographs with heterogeneous recompression and editing histories, whereas the generated class was selected to resemble clean photographs. Consequently, a score that measures deviation from photographic phase regularity ranks the real images as more anomalous. This is the limitation of Appendix D.3, with the real class itself carrying altered high-frequency content.

The two scores are therefore useful under different conditions. When the imaging pipeline leaves phase structure intact, the deviation score detects unseen generators, and the specialist contributes little outside the seen family (Section 6). For images whose high-frequency content has been altered during online distribution, however, the specialist remains discriminative while the one-class score inverts. The OR rule does not detect that the deviation score has inverted, and the declared budgets hold only on the calibration images. On Chameleon, the real-image false-positive rate rises to 21.9%, above the 18.15% bound (Table D.9), because the deviation score no longer ranks these real images as real (AUROC 0.382). The OR decision of the full system falls to chance level (AUROC 0.512), although the specialist alone still ranks at 0.744 (Table D.13).

E Dataset Details

GenImage. GenImage (Zhu et al., 2023b) pairs ImageNet photographs (Russakovsky et al., 2015) with images from eight generators: Stable Diffusion v1.4 and v1.5 (Rombach et al., 2022), ADM (Dhariwal & Nichol, 2021), GLIDE (Nichol et al., 2022), VQDM (Gu et al., 2022), Wukong (Huawei Noah's Ark Lab, 2022), Midjourney (Midjourney, Inc., 2022), and BigGAN (Brock et al., 2019). Text-to-image models are prompted with "photo of [class]" and ADM and BigGAN with the class label, so the real and generated classes share label semantics. Native resolutions range from 128^2 (BigGAN) through 256^2 (ADM, GLIDE, VQDM) and 512^2 (the Stable Diffusion family and Wukong) to 1024^2 (Midjourney). The real images are JPEG files, at quality factors ranging from 70 to 100, predominantly 96, and the generated images are typically lossless PNG files. These are the format and resolution shortcuts documented by Grommelt et al. (2024) and removed by the protocol of Section 5.2.

Synthbuster. Synthbuster (Bammey, 2024) takes the opposite approach. Its generated images are the raw, native-resolution PNG outputs of nine generators, never resized or recompressed: DALL·E 2 (Ramesh et al., 2022), DALL·E 3 (Betker et al., 2023), Adobe Firefly (Adobe Inc., 2023),

GLIDE, Midjourney v5, Stable Diffusion 1.3, 1.4, and 2, and SDXL (Podell et al., 2024), with 1,000 images each (999 for Midjourney v5) and native resolutions from 256^2 (GLIDE) to 2048^2 (Firefly). The prompts are derived one-to-one from the 1,000 uncompressed camera photographs of RAISE-1k (Dang-Nguyen et al., 2015); the content is controlled, and only the imaging source varies. As the real class, we use all 123,287 unique MS-COCO photographs (Lin et al., 2014) rather than the 1,000 RAISE-1k images. The larger pool estimates the real-image acceptance rate with much lower variance and is the real domain on which the baselines of Table 2 were trained, so the choice does not favor our system. No image from either class in this benchmark is used for training or calibration. The backbone, deviation score, specialist, and thresholds remain exactly as trained or calibrated on GenImage, so Synthbuster measures zero-shot transfer to both an unseen real-image domain and unseen generators.