Supplementary Information for

Entropy Change in Sampling Predicts Downstream Performance in Neural Networks

Xiu-Cheng Wang[2,3]†, Jun-Jie Zhang[1]*†, Nan Cheng[2,3], Long-Gang Pang[4], Taijiao Du[1], Deyu Meng[5,6]*

[1]Northwest Institute of Nuclear Technology, No. 28 Pingyu Road, Xi'an, 710024, Shaanxi, China.
[2]School of Telecommunications Engineering, Xidian University, No. 2 South Taibai Road, Xi'an, 710071, Shaanxi, China.
[3]State Key Laboratory of ISN, No. 2 South Taibai Road, Xi'an, 710071, Shaanxi, China.
[4]Key Laboratory of Quark and Lepton Physics (MOE) & Institute of Particle Physics, Central China Normal University, No. 152 Luoyu Road, Wuhan, 430079, Hubei, China.
[5]Ministry of Education Key Lab of Intelligent Networks and Network Security, Xi'an Jiaotong University, No. 28 Xianning West Road, Xi'an, 710049, Shaanxi, China.
[6]School of Mathematics and Statistics, Xi'an Jiaotong University, No. 28 Xianning West Road, Xi'an, 710049, Shaanxi, China.

*Corresponding author(s). E-mail(s): zjacob@mail.ustc.edu.cn; dymeng@mail.xjtu.edu.cn;
Contributing authors: xcwang_1@stu.xidian.edu.cn; dr.nan.cheng@ieee.org; lgpang@ccnu.edu.cn; dutaijiao@nint.ac.cn;
†Both authors contributed equally to this work.

This PDF includes

1. Supplementary Note 1: Theoretical Derivations and Proofs for Optimality of Random Image Masking
2. Supplementary Note 2: MIMO Over-the-Air Hardware Configuration
3. Supplementary Note 3: Ablation Studies on Husimi Parameters (`win`, σ, `hop`)

1 Supplementary Note 1: Theoretical Derivations and Proofs for Optimality of Random Image Masking

1.1 Physical Perspective: Symplectic Structure, Liouville's Theorem, and Entropy

In imaging and information theory, ideal optical transformations (such as Fourier transforms or lens propagation) can be viewed as canonical (symplectic) transformations in phase space. According to Liouville's theorem, such transformations strictly preserve phase-space volume and Gibbs-type entropy, ensuring that no information is lost or created. In contrast, *image masking is not a canonical transformation*: it multiplies the image in the spatial domain, which can disrupt the spectral distribution and potentially change the phase-space entropy.

The purpose of this work is to clarify how two common masking strategies affect phase-space entropy:

1. **Periodic masking (uniform subsampling).**
   This operation introduces structured replicas of the spectrum in frequency space. High-frequency energy is folded (aliased) into the low-frequency region, creating irreversible mixing in phase space. By strict concavity and Jensen's inequality, one can show that this aliasing always leads to entropy increase, and the increase is strict whenever overlapping replicas occur.
2. **Random masking (random subsampling).**
   Unlike periodic masking, random subsampling does not maintain fixed lattice coherence. In the limit of a large number of degrees of freedom, its effect in phase space is equivalent to adding a small, unbiased, isotropic perturbation in frequency. Since high frequencies are no longer systematically folded into the low-frequency baseband, the phase-space entropy is preserved in expectation, with only small $O(N^{-1/2})$ fluctuations.

Physical intuition.

Periodic masking acts like a *regular folding mechanism*: it systematically overlays high-frequency content onto the low-frequency region, causing irreversible information mixing and an increase in entropy. Random masking acts more like a *uniform jitter*: it breaks up coherent folding patterns, spreading energy in an unbiased way that does not accumulate into systematic aliasing. As the image size grows, the average phase-space entropy remains nearly unchanged.

In summary, the analysis shows a sharp contrast: **periodic subsampling necessarily increases entropy (irreversible loss of information), while random subsampling approximately preserves entropy in the statistical limit (information is nearly conserved).**

1.2 Stationarity in Natural Images

Natural images are often modeled as approximately stationary random processes: their local statistics change slowly across space. Formally, for an image function $I(x, y)$, we have

$$\mathbb{E}[I(x - a_x, y - a_y)] \approx \mathbb{E}[I(x, y)], \quad \forall (a_x, a_y) \in \mathbb{R}^2. \tag{S1}$$

Here, $\mathbb{E}[\cdot]$ denotes the ensemble average across samples; we assume weak ergodicity, so local spatial averaging approximates the ensemble expectation. This approximate stationarity motivates analyzing images in a joint phase-space representation, where spatial and spectral information coexist. Such a representation allows us to study how operations like masking affect both spatial detail and frequency content simultaneously.

1.3 Wigner Transform and Phase Space Density

To analyze the information entropy, we introduce the Wigner transform. For an image $I(\mathbf{x})$, where $\mathbf{x} = (x, y)$, the two-dimensional Wigner distribution is defined as:

$$W_I(\mathbf{x}; \mathbf{k}) = \int I\left(\mathbf{x} + \frac{\boldsymbol{\xi}}{2}\right) I^*\left(\mathbf{x} - \frac{\boldsymbol{\xi}}{2}\right) e^{-i\mathbf{k}\cdot\boldsymbol{\xi}} \mathrm{d}\boldsymbol{\xi}. \tag{S2}$$

The Wigner transform creates a phase-space representation of an image, combining spatial domain $\mathbf{x}$ and frequency domain $\mathbf{k} = (k_x, k_y)$. Natural images are not purely localized in space or frequency, so analyzing them in a joint domain avoids losing structural information. This makes it possible to evaluate how masking affects both spatial details and spectral content together.

We adopt the symmetric Fourier convention where the Wigner product–convolution carries a prefactor $(2\pi)^{-2}$; all delta masses and normalization constants are stated compatibly with that convention.

Since W_I can take negative values, we choose a *Cohen's class* kernel that is *known* to produce a non-negative distribution (e.g., the spectrogram/STFT or the Husimi/Q distribution, which correspond to convolving W_I with the Wigner of a window function such as a minimum-uncertainty Gaussian). Define

$$\rho_I(X) \;=\; (\Phi * W_I)(X), \qquad \Phi \geq 0, \quad X = (\mathbf{x}, \mathbf{k}), \tag{S3}$$

where $*$ denotes convolution in (x, y, k_x, k_y). For convenience, we write "$*$" for convolution in full phase space $(\mathbf{x}, \mathbf{k})$, and "$*_{\mathbb{T}^2}$" for convolution over the frequency torus $\mathbb{T}^2$ (in $\mathbf{k}$ only). Throughout this paper, we use periodic boundary conditions in the frequency domain ($\mathbb{T}^2$) for the analysis, ensuring consistency across all frequency space representations. This allows for proper application of symplectic transformations and the preservation of phase-space volume. *Not every* Gaussian convolution guarantees non-negativity; here we specifically use kernels (e.g., spectrogram/Husimi) that do.

We assume that the smoothing kernel Φ may vary across different domains to reflect the specific physical resolution of each measurement device. In some cases, Φ might represent the imaging resolution in MRI, or the antenna resolution in wireless MIMO systems.

1.4 Entropy and Liouville Dynamics

We quantify the spread of information in phase space using the (Gibbs-type) entropy of an *unnormalized* smoothed Wigner density $\rho(X)$:

$$S[\rho] \;:=\; -\int \rho(X)\, \ln \rho(X)\, \mathrm{d}X. \tag{S4}$$

When the image undergoes an ideal (canonical/symplectic) transformation, ρ evolves by a Liouville (continuity) equation

$$\partial_t \rho \;+\; \nabla_X \cdot (\rho V) \;=\; 0, \qquad \nabla_X \cdot V = 0, \tag{S5}$$

where V is the phase-space flow field. Under (S5) and suitable boundary conditions (torus or fast decay), phase-space volume is preserved and

$$\frac{\mathrm{d}}{\mathrm{d}t} S[\rho] \;=\; 0. \tag{S6}$$

In our analyses we often use a *band-normalized* density over the frequency torus $\mathcal{B} \subset \mathbb{T}^2$ at each spatial location:

$$m[\rho](\mathbf{x}) := \int_{\mathcal{B}} \rho(\mathbf{x}; \mathbf{k})\, \mathrm{d}\mathbf{k}, \qquad \mathcal{N}_{\mathcal{B}}[\rho]\,(\mathbf{x}; \mathbf{k}) := \frac{\rho(\mathbf{x}; \mathbf{k})}{m[\rho](\mathbf{x})}. \tag{S7}$$

The corresponding local band-entropy and a spatial aggregate are

$$s_{\mathcal{B}}[\mathcal{N}_{\mathcal{B}}[\rho]](\mathbf{x}) := -\int_{\mathcal{B}} \mathcal{N}_{\mathcal{B}}[\rho]\,(\mathbf{x}; \mathbf{k})\, \ln \mathcal{N}_{\mathcal{B}}[\rho]\,(\mathbf{x}; \mathbf{k})\, \mathrm{d}\mathbf{k}, \qquad S_{\mathcal{B}}[\mathcal{N}_{\mathcal{B}}[\rho]] := \int S_{\mathcal{B}}[\mathcal{N}_{\mathcal{B}}[\rho]](\mathbf{x})\, w(\mathbf{x})\, \mathrm{d}\mathbf{x}, \tag{S8}$$

where $w(\mathbf{x})$ is a specified weight (e.g., uniform or energy-weighted $w \propto m[\rho](\mathbf{x})$). Since $\mathcal{N}_{\mathcal{B}}[\rho]$ includes an $\mathbf{x}$-dependent normalization, exact Liouville conservation does not strictly apply to $S_{\mathcal{B}}[\mathcal{N}_{\mathcal{B}}[\rho]]$; it *serves as an ideal reference* when smoothing and normalization are fixed, and becomes exact for ρ under (S5).

1.5 Effect of Masking on Phase Space Entropy

A masking operation is defined as

$$J(\mathbf{x}) = M(\mathbf{x})\, I(\mathbf{x}), \tag{S9}$$

where $M(\mathbf{x})$ is the mask and $I(\mathbf{x})$ is the image. The bold vector $\mathbf{x} = (x, y)$. Applying the Wigner transform to the masked image $J = M \cdot I$ yields

$$W_J(\mathbf{x}, \mathbf{k}) = \int M\Big(\mathbf{x} + \tfrac{\boldsymbol{\xi}}{2}\Big)\, M^*\Big(\mathbf{x} - \tfrac{\boldsymbol{\xi}}{2}\Big)\; I\Big(\mathbf{x} + \tfrac{\boldsymbol{\xi}}{2}\Big)\, I^*\Big(\mathbf{x} - \tfrac{\boldsymbol{\xi}}{2}\Big)\; e^{-i\mathbf{k}\cdot\boldsymbol{\xi}}\, \mathrm{d}\boldsymbol{\xi}, \tag{S10}$$

where $\mathbf{k} = (k_x, k_y)$ and $\boldsymbol{\xi} = (\xi_x, \xi_y)$. We now proceed to express this in a more manageable form in 2D. Using the inverse-Fourier representation of the Wigner distribution, we have

$$I\Big(\mathbf{x} + \tfrac{\boldsymbol{\xi}}{2}\Big)\, I^*\Big(\mathbf{x} - \tfrac{\boldsymbol{\xi}}{2}\Big) = \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} W_I(\mathbf{x}, \mathbf{k}')\, e^{i\, \mathbf{k}' \cdot \boldsymbol{\xi}}\, \mathrm{d}\mathbf{k}', \tag{S11}$$

and similarly

$$M\Big(\mathbf{x} + \tfrac{\boldsymbol{\xi}}{2}\Big)\, M^*\Big(\mathbf{x} - \tfrac{\boldsymbol{\xi}}{2}\Big) = \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} W_M(\mathbf{x}, \mathbf{q})\, e^{i\, \mathbf{q} \cdot \boldsymbol{\xi}}\, \mathrm{d}\mathbf{q}. \tag{S12}$$

Substituting into the definition of W_J and integrating over $\boldsymbol{\xi}$ gives the product–convolution law in frequency:

$$W_J(\mathbf{x}, \mathbf{k}) = \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} W_M(\mathbf{x}, \mathbf{q})\, W_I\big(\mathbf{x}, \mathbf{k} - \mathbf{q}\big)\, \mathrm{d}\mathbf{q}. \tag{S13}$$

The numerical prefactor $(2\pi)^{-2}$ follows the symmetric Fourier normalization stated above, ensuring internal consistency across all Wigner and convolution relations.

Let the observed (nonnegative) phase-space density be $\rho = \Phi * W$ with a fixed nonnegative Cohen-class kernel Φ (convolution in both $\mathbf{x}$ and $\mathbf{k}$). From the product–convolution law, $\rho_J = \Phi * W_J = \frac{1}{(2\pi)^2}\, \Phi * \big(W_M *_{\mathbb{T}^2} W_I\big)$.

After smoothing both the mask and the image with the kernel Φ, we write

$$\rho_J = \Phi * W_J, \qquad \rho_I = \Phi * W_I, \qquad K_M := \Phi * W_M. \tag{S14}$$

Then, using the product–convolution law and the local-averaging approximation discussed above,

$$\rho_J(\boldsymbol{x}, \boldsymbol{k}) \;\approx\; \frac{1}{(2\pi)^2}\, K_M *_{\mathbb{T}^2} \rho_I(\boldsymbol{x}, \boldsymbol{k}). \tag{S15}$$

The kernel K_M is nonnegative and plays the role of an effective kernel that modulates both spatial and frequency components of the image. This convolution reflects how masking affects the phase-space distribution and hence its entropy.

Unlike the Liouville flow, which preserves phase space volume (i.e., no entropy change), the masking operation is not a canonical transformation and thus can break entropy conservation. The extent and direction of entropy change depend on the structure of the mask and on how the smoothing operation suppresses oscillatory cross-terms in W_M. In particular, periodic masks can introduce aliasing effects, leading to an increase in entropy. The smoothed phase-space density $\rho_J(\mathbf{x}, \mathbf{k})$ is a convex combination of shifted versions of $\rho_I(\mathbf{x}, \mathbf{k})$, and entropy is generally increased due to the mixing of spatial and frequency components in the distribution.

1.5.1 Periodic Masking Causes Entropy Increase (Uniform Subsampling)

Definition of a periodic (uniform subsampling) mask.

We model uniform subsampling by a 2D Dirac comb with spacings d_x, d_y:

$$M_{\text{per}}(x, y) \;=\; \sum_{m,n \in \mathbb{Z}} \delta(x - m d_x)\, \delta(y - n d_y). \tag{S16}$$

Within a large apodization window of area A (later $A \to \infty$), this corresponds to keeping only pixels on a regular lattice and discarding all others. By the 2D Poisson summation formula (in the sense of tempered distributions),

$$M_{\text{per}}(x, y) \;=\; \frac{1}{d_x d_y} \sum_{n \in \mathbb{Z}^2} \exp\Big(i\, \kappa_n \cdot (x, y)\Big), \qquad \kappa_n = \Big(\tfrac{2\pi n_x}{d_x},\, \tfrac{2\pi n_y}{d_y}\Big). \tag{S17}$$

Thus M_{per} admits a Fourier series with *equal* coefficients $c_n = \frac{1}{d_x d_y}$ for all $n \in \mathbb{Z}^2$.

Wigner distribution of the mask.

Substitute the Fourier series of M_{per} into the Wigner definition. Writing $M_{\mathrm{per}}(x,y) = \sum_n c_n e^{i\kappa_n\cdot(x,y)}$ with $c_n = \frac{1}{d_x d_y}$, the algebra is identical to the general periodic case, and yields

$$W_{M_{\mathrm{per}}}(x,y;k_x,k_y) = (2\pi)^2 \sum_{n,m\in\mathbb{Z}^2} c_n c_m^* \,\delta\big((k_x,k_y) - \tfrac{\kappa_n+\kappa_m}{2}\big)\; e^{i(\kappa_n-\kappa_m)\cdot(x,y)}. \tag{S18}$$

The *auto-terms* ($n = m$) are located at $k = \kappa_n$ and are independent of (x,y); the *cross-terms* ($n \neq m$) oscillate in (x,y) with spatial frequencies $\kappa_n - \kappa_m$ and sit at midpoints $\frac{\kappa_n+\kappa_m}{2}$ in frequency.

Smoothing and the effective kernel.

Let the observed (nonnegative) phase-space density be the Cohen-class smoothing of the Wigner distribution with a kernel Φ:

$$\rho_J \;=\; \Phi * W_J, \qquad K_{M_{\mathrm{per}}} \;:=\; \Phi * W_{M_{\mathrm{per}}}. \tag{S19}$$

Because Φ averages over a few unit cells in (x,y), the rapid oscillations $e^{i(\kappa_n-\kappa_m)\cdot(x,y)}$ for $n\neq m$ are suppressed by destructive averaging (Riemann–Lebesgue). Retaining the dominant auto-terms gives

$$K_{M_{\mathrm{per}}}(k_x,k_y) \;\approx\; (2\pi)^2 \sum_{n\in\mathbb{Z}^2} |c_n|^2 \,\big(\Phi *_{\mathbb{T}^2} \delta(\cdot - \kappa_n)\big)(k_x,k_y). \tag{S20}$$

Since $|c_n|^2 = \frac{1}{d_x^2 d_y^2}$ is constant, $K_{M_{\mathrm{per}}}$ is a *broadened frequency comb* centered at the reciprocal-lattice points $\{\kappa_n\}$ with nonnegative weights proportional to $|c_n|^2$.

Resulting phase-space density via the product–convolution law.

For a masked image $J = M_{\mathrm{per}} \cdot I$, the Wigner transform of the product obeys (in 2D, cf. the standard derivation in 1D and extending component-wise)

$$W_J(x,y;k_x,k_y) \;=\; \frac{1}{(2\pi)^2} \iint_{\mathbb{R}^2} W_{M_{\mathrm{per}}}(x,y;q_x,q_y)\; W_I\big(x,y;k_x - q_x,\; k_y - q_y\big)\,\mathrm{d}\mathbf{q}. \tag{S21}$$

Applying the linear smoothing Φ and using the same local-averaging approximation (under which Φ effectively commutes with the frequency convolution),

$$\rho_J \;=\; \Phi * W_J \;=\; \frac{1}{(2\pi)^2}\,(\Phi * W_{M_{\mathrm{per}}}) *_{\mathbb{T}^2} (\Phi * W_I) \;=\; \frac{1}{(2\pi)^2}\, K_{M_{\mathrm{per}}} *_{\mathbb{T}^2} \rho_I. \tag{S22}$$

On the frequency torus $\mathbb{T}^2$ (i.e., within one Nyquist cell modulo wrap-around), the Φ-averaged periodic kernel $K_{M_{\mathrm{per}}}$ is a broadened comb centered at the reciprocal-lattice points modulo the cell. Consequently,

$$\rho_J(x,y;\mathbf{k}) \;\approx\; \frac{1}{(2\pi)^2}\,\big(K_{M_{\mathrm{per}}} *_{\mathbb{T}^2} \rho_I\big)(x,y;\mathbf{k}). \tag{S23}$$

Using that $K_{M_{\mathrm{per}}}$ is (after spatial averaging) effectively independent of (x,y), we obtain the convex-mixture form

$$\mathcal{N}_{\mathcal{B}}[\rho_J]\,(\mathbf{x};\mathbf{k}) \;\approx\; \sum_{\kappa_n\in\mathcal{B}} w_n\, \mathcal{N}_{\mathcal{B}}[\rho_I]\,\big(\mathbf{x};\mathbf{k} - \kappa_n\big), \qquad w_n = \frac{\int_{\mathcal{B}} K_{M_{\mathrm{per}}}(\mathbf{k}-\kappa_n)\,\mathrm{d}\mathbf{k}}{\sum_{\kappa_m\in\mathcal{B}} \int_{\mathcal{B}} K_{M_{\mathrm{per}}}(\mathbf{k}-\kappa_m)\,\mathrm{d}\mathbf{k}}. \tag{S24}$$

Here $w_n \geq 0$ and $\sum_n w_n = 1$, and crucially w_n do not depend on $\mathbf{k}$ (and, after averaging, not on $\mathbf{x}$ either), so the right-hand side is a *pointwise* convex combination in $\mathbf{k}$ at each $\mathbf{x}$.

This formula shows that uniform subsampling replicates ρ_I on the reciprocal lattice and superposes the copies. Physically, frequencies outside the baseband are shifted by κ_n back into it—*high-frequency energy is folded (aliased) into low frequency*, producing irreversible mixing.

Entropy increase via Jensen's inequality.

Let $f(t) = -t \ln t$, which is strictly concave on $(0, 1]$. Since the normalized density after masking can be expressed as

$$\mathcal{N}_{\mathcal{B}}[\rho_J]\,(\mathbf{x};\mathbf{k}) = \sum_n w_n\, r_n(\mathbf{k}), \qquad r_n(\mathbf{k}) = \mathcal{N}_{\mathcal{B}}[\rho_I]\,(\mathbf{x};\mathbf{k} - \kappa_n), \tag{S25}$$

and $\sum_n w_n = 1$, Jensen's inequality yields

$$\int_{\mathcal{B}} f\Big(\mathcal{N}_{\mathcal{B}}[\rho_J]\,(\mathbf{x};\mathbf{k})\Big)\mathrm{d}\mathbf{k} \geq \sum_n w_n \int_{\mathcal{B}} f\big(r_n(\mathbf{k})\big)\mathrm{d}\mathbf{k}. \tag{S26}$$

That is,

$$s_{\mathcal{B}}[\mathcal{N}_{\mathcal{B}}[\rho_J]](\mathbf{x}) \geq s_{\mathcal{B}}[\mathcal{N}_{\mathcal{B}}[\rho_I]](\mathbf{x}), \tag{S27}$$

with equality only when all $r_n(\mathbf{k})$ (for $w_n > 0$) are identical almost everywhere on $\mathcal{B}$.

1.5.2 Random Masking Approximately Preserves Entropy (Random Subsampling)

Definition of a random subsampling mask.

Let the underlying fine grid have pixel pitches p_x, p_y and index set $\mathcal{G} = \{(mp_x, np_y) : m, n \in \mathbb{Z}\}$ within a large window of area A (later $A \to \infty$). Define an i.i.d. Bernoulli field $\{b_{m,n}\}$ with

$$b_{m,n} \in \{0, 1\}, \qquad \mathbb{P}(b_{m,n} = 1) = \rho \in (0, 1), \tag{S28}$$

and set the (energy-normalized) random subsampling mask

$$M_{\text{rand}}(x, y) \;=\; \frac{1}{\sqrt{\rho}} \sum_{(m,n)\in\mathbb{Z}^2} b_{m,n}\, \delta(x - mp_x)\, \delta(y - np_y). \tag{S29}$$

The global factor $1/\sqrt{\rho}$ keeps the average mask energy (Frobenius norm) independent of ρ, so we can compare entropies under a common normalization.[1]

Wigner distribution of the mask: mean–fluctuation decomposition.

Write $b_{m,n} = \rho + \epsilon_{m,n}$ with $\mathbb{E}[\epsilon_{m,n}] = 0$, $\mathrm{Var}(\epsilon_{m,n}) = \rho(1 - \rho)$ and independence across (m, n). Then

$$M_{\text{rand}} \;=\; \underbrace{\sqrt{\rho}\,\Gamma}_{\text{mean part}} \;+\; \underbrace{\frac{1}{\sqrt{\rho}} \sum \epsilon_{m,n}\, \delta(\cdot - mp_x)\delta(\cdot - np_y)}_{\text{zero-mean fluctuation}}, \tag{S30}$$

where $\Gamma(x, y) = \sum_{m,n} \delta(x - mp_x)\delta(y - np_y)$ is the full lattice comb. By bilinearity, the Wigner distribution splits into

$$W_{M_{\text{rand}}} \;=\; \rho\, W_\Gamma \;+\; W_\varepsilon \;+\; 2\,\mathrm{Re}\, W_{\Gamma,\varepsilon}, \tag{S31}$$

where W_ε is the Wigner of the fluctuation part and $W_{\Gamma,\varepsilon}$ the cross-Wigner. Taking expectations over the mask randomness yields

$$\mathbb{E}\big[W_{M_{\text{rand}}}\big] \;=\; \rho\, W_\Gamma \;+\; \mathbb{E}\big[W_\varepsilon\big], \quad \text{since} \quad \mathbb{E}\big[W_{\Gamma,\varepsilon}\big] = 0 \tag{S32}$$

(the cross term vanishes because one factor has zero mean).

We now analyze the two contributions:

(i) The mean term $\rho\, W_\Gamma$. As in the periodic case, W_Γ consists of sharp peaks at the reciprocal lattice $\kappa_{u,v} = (2\pi u/p_x,\, 2\pi v/p_y)$ plus oscillatory cross-terms in (x, y). After smoothing in (x, y) with kernel Φ over a few unit cells, the oscillatory cross-terms of W_Γ are suppressed (Riemann–Lebesgue averaging), leaving broadened peaks centered at all reciprocal lattice points $\kappa_{u,v}$. However, if one

---

[1] If one prefers unit amplitude on the kept pixels, i.e. $M = \sum b_{m,n}\delta(\cdot)$, one can perform unit-mass normalization on the final phase-space density ρ_J instead, with equivalent results; this paper chooses energy normalization for direct comparison with the "uniform mask" in terms of capacity and Fisher information.

restricts attention to the baseband detection region $|k_x| \leq \pi/p_x,\ |k_y| \leq \pi/p_y$, then only the central lobe at $\mathbf{k} = (0,0)$ remains inside the band. [2]

(ii) The fluctuation term $\mathbb{E}[W_\varepsilon]$. By independence and zero-mean, $\mathbb{E}[\epsilon_{m_1,n_1}\epsilon_{m_2,n_2}] = 0$ when $(m_1,n_1) \neq (m_2,n_2)$ and equals $\rho(1-\rho)$ on the diagonal. Tracing the definition of W shows that only the *diagonal* pairs force $\xi = 0$ and hence contribute a k-independent pedestal. Thus $\mathbb{E}[W_\varepsilon]$ produces a *flat (white) background in* k, localized in (x,y) on the lattice and becoming spatially uniform after smoothing by Φ over many unit cells.

Smoothing and the effective kernel.

Let $K_{M_{\text{rand}}} := \Phi * W_{M_{\text{rand}}}$. Under the i.i.d. Bernoulli model and Φ-averaging over many unit cells, we obtain

$$\mathbb{E}[K_{M_{\text{rand}}}](\mathbf{k}) = (2\pi)^2\,\delta(\mathbf{k}) + \alpha_A\,\mathbf{1}_{\mathcal{B}}(\mathbf{k}), \quad |\alpha_A| \leq \frac{C}{N}. \tag{S33}$$

Here $\delta(\mathbf{k})$ denotes the identity distribution on the frequency torus $\mathbb{T}^2$, satisfying $(g *_{\mathbb{T}^2} \delta)(\mathbf{k}) = g(\mathbf{k})$ for any test function g. With this convention, the convolution law $\rho_J = (2\pi)^{-2} K_{M_{\text{rand}}} *_{\mathbb{T}^2} \rho_I$ implies that

$$\mathbb{E}[\rho_J] = \rho_I + O\Big(\frac{1}{N}\Big). \tag{S34}$$

Moreover, since $K_{M_{\text{rand}}}$ is an average of $O(N)$ independent bounded contributions, standard concentration yields

$$\mathbb{E}\,\big\|K_{M_{\text{rand}}} - \mathbb{E}[K_{M_{\text{rand}}}]\big\|_{L^1(\mathcal{B})} = O\Big(\tfrac{1}{\sqrt{N}}\Big). \tag{S35}$$

Combining the mean bias $O(\frac{1}{N})$ with the fluctuations $O(\frac{1}{\sqrt{N}})$ yields

$$\mathbb{E}\,\big\|\rho_J - \rho_I\big\|_{L^1(\mathcal{B})} \;=\; O\Big(\tfrac{1}{\sqrt{N}}\Big). \tag{S36}$$

Band-normalization stability.

Let $\mathcal{N}_{\mathcal{B}}[\rho]$ denote the band-normalized density on $\mathcal{B}$, i.e.,

$$\mathcal{N}_{\mathcal{B}}[\rho]\,(\boldsymbol{x};\boldsymbol{k}) = \frac{\rho(\boldsymbol{x};\boldsymbol{k})}{m[\rho](\boldsymbol{x})}, \qquad m[\rho](\boldsymbol{x}) := \int_{\mathcal{B}} \rho(\boldsymbol{x};\tilde{\boldsymbol{k}})\,\mathrm{d}\tilde{\boldsymbol{k}}. \tag{S37}$$

Assume $m[\rho_I](\boldsymbol{x}), m[\rho_J](\boldsymbol{x}) \geq m_{\min} > 0$ almost everywhere. Then for each spatial location $\boldsymbol{x}$,

$$\big\|\mathcal{N}_{\mathcal{B}}[\rho_J]\,(\boldsymbol{x};\cdot) - \mathcal{N}_{\mathcal{B}}[\rho_I]\,(\boldsymbol{x};\cdot)\big\|_{L^1_{\mathbf{k}}(\mathcal{B})} \leq \frac{2}{m_{\min}}\,\big\|\rho_J(\boldsymbol{x};\cdot) - \rho_I(\boldsymbol{x};\cdot)\big\|_{L^1_{\mathbf{k}}(\mathcal{B})}. \tag{S38}$$

Integrating over $\mathbf{x}$ and using the $O(N^{-1/2})$ concentration bound above gives

$$\mathbb{E}\big\|\mathcal{N}_{\mathcal{B}}[\rho_J] - \mathcal{N}_{\mathcal{B}}[\rho_I]\,\big\|_{L^1_{\mathbf{x},\mathbf{k}}(\mathcal{B})} = O\Big(\tfrac{1}{\sqrt{N}}\Big). \tag{S39}$$

This shows that the normalization step is Lipschitz-stable with respect to the L^1 perturbation of ρ.

Physical explanation.

Periodic masking creates coherent spectral replicas on the reciprocal lattice, so high-frequency components are deterministically folded into the baseband, producing irreversible aliasing and entropy growth. Random masking, on the other hand, destroys lattice phase coherence: after spatial averaging by Φ, its spectrum contains only a central DC lobe plus a weak isotropic pedestal, which corresponds to unbiased white perturbations in frequency. Hence, high-frequency energy is not systematically folded into the low-frequency region, and the phase-space entropy remains statistically invariant up to $O(N^{-1/2})$ fluctuations. From an information-theoretic perspective, random masking results in a measurement operator with a near-isotropic singular-value spectrum and well-conditioned Fisher information, whereas periodic masking collapses certain singular values and creates spectral holes that reduce information capacity.

---

[2] Under a strictly periodic mask the Wigner spectrum forms a comb on the reciprocal lattice. With Bernoulli coefficients, however, lattice coherence is destroyed in expectation: after spatial averaging by Φ, nonzero harmonics lose phase coherence so that only a DC spike at $\mathbf{k} = \mathbf{0}$ remains coherent, while the residual appears as an approximately flat pedestal in k. The pedestal's contribution to the *smoothed kernel* decays like $O(1/N)$ due to averaging over N effectively independent spatial cells.

1.6 Assumptions, scope, and possible counterexamples

For clarity and to delimit the regime in which the arguments and asymptotic bounds below hold, we list here the explicit assumptions used in the derivations above, the finite–N conditions, and several noteworthy edge cases where the conclusions may fail.

A1. Fourier convention.

We adopt the forward/inverse Fourier transform convention (2D)

$$\widehat{f}(\mathbf{k}) = \int_{\mathbb{R}^2} f(\mathbf{x})\, e^{-i\mathbf{k}\cdot\mathbf{x}}\, \mathrm{d}\mathbf{x}, \qquad f(\mathbf{x}) = \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} \widehat{f}(\mathbf{k})\, e^{i\mathbf{k}\cdot\mathbf{x}}\, \mathrm{d}\mathbf{k}. \tag{S40}$$

All (2π) prefactors appearing in the Wigner product–convolution relations follow from this convention.

A2. Smooth, nonnegative Cohen kernel.

The smoothing kernel Φ belongs to the Schwartz class (or is a compactly supported, sufficiently smooth averaging kernel), satisfies $\Phi \geq 0$, and has an effective spatial averaging radius ℓ_Φ that covers many sampling unit cells (quantitatively: $\ell_\Phi \gg d$, where d denotes the lattice spacing of a periodic mask). This scale separation guarantees suppression of rapidly oscillatory cross-terms by the Riemann–Lebesgue lemma in the large-window limit.

A3. Mask models.

- *Periodic mask.* Modeled as an ideal Dirac comb $M_{\mathrm{per}}(x) = \sum_{m,n} \delta(x - md_x)\delta(y - nd_y)$ (or its apodized approximation on a large window of area A), with exact lattice coherence.
- *Random mask.* Modeled as i.i.d. Bernoulli samples $b_{m,n} \in \{0, 1\}$ with $\mathbb{P}(b_{m,n} = 1) = \rho \in (0, 1)$ and finite variance; the mask is energy–normalized as in the main text ($1/\sqrt{\rho}$ prefactor). Independence (or sufficiently fast decay of correlations) across lattice sites is required for the concentration bounds.

A4. Large-sample averaging and finite–N bounds.

Let N denote the effective number of independent spatial cells averaged by Φ (or present in the apodization window). The principal stochastic statements are asymptotic in $N \to \infty$; for finite N we obtain the explicit bounds stated in the text:

$$\mathbb{E}\|K_{M_{\mathrm{rand}}} - \mathbb{E}K_{M_{\mathrm{rand}}}\|_{L^1(\mathcal{B})} = O\Big(\tfrac{1}{\sqrt{N}}\Big), \qquad \mathbb{E}\|\rho_J - \rho_I\|_{L^1(\mathcal{B})} = O\Big(\tfrac{1}{\sqrt{N}}\Big), \tag{S41}$$

with the constants depending on the per-cell variance and the L^∞ bounds of the contributing terms; standard Hoeffding/Bernstein inequalities justify these rates under A3.

A5. Local energy (band) lower bound.

We assume the local band energy satisfies $m[\rho](\mathbf{x}) \geq m_{\min} > 0$ a.e., ensuring band–normalization is Lipschitz and avoiding division by numerically vanishing energy in the L^1 stability estimates.

A6. Image regularity and ergodicity.

The image ensemble has bounded energy and sufficiently regular Wigner distributions so that integrals and exchanges of limits used in the derivations are justified. We further assume weak ergodicity/stationarity at the scale of Φ so that local spatial averages represent ensemble expectations.

Remarks on scope. Under A1–A6 the periodic mask analysis yields a pointwise convex combination of band densities (hence Jensen applies) and the random mask analysis yields the stated expectation and concentration bounds. When any assumption above is violated (see next subsection), the stated conclusions can fail or require modification.

Possible counterexamples and limiting cases

The following non-exhaustive list highlights situations in which the above conclusions may not hold, or where the finite–N corrections may be large.

- **Lattice-invariant signals.** If the original band-normalized densities satisfy $r_n(\mathbf{k}) = \mathcal{N}_{\mathcal{B}}[\rho_I]\,(\mathbf{x}; \mathbf{k}-\kappa_n)$ that are identical (a.e.) for all κ_n with $w_n > 0$, then Jensen's inequality is tight and periodic subsampling does *not* increase the band entropy. Such signals are exceptional but conceptually important (they are "invariant under reciprocal lattice shifts").

- **Strongly correlated or structured masks.** If the random mask exhibits long-range correlations, clustering, or periodic substructure, the decomposition into a DC spike plus small flat pedestal may fail; the effective kernel K_M can keep coherent sidelobes and therefore produce deterministic aliasing-like effects.
- **Insufficient smoothing (ℓ_Φ too small).** If the smoothing kernel Φ averages over too few unit cells, oscillatory cross-terms in W_M may not be suppressed and the broadened-comb approximation for $K_{M_{\mathrm{per}}}$ breaks down.
- **Very small N.** For small numbers of independent cells, the $O(1/\sqrt{N})$ fluctuations can be non-negligible and may lead to observable entropy shifts.
- **Near-zero local energy ($m_{\min} \approx 0$).** When local band energy is vanishingly small the normalization step amplifies perturbations and the Lipschitz bound becomes uninformative.
- **Deterministic masks engineered to avoid aliasing.** Special deterministic sampling patterns (e.g., certain nonuniform sampling designs used in compressed sensing) can produce well-conditioned measurement operators without randomness; such constructions lie outside the Bernoulli model and must be analyzed case by case.

2 Supplementary Note 2: MIMO Over-the-Air Hardware Configuration

To validate the real-time performance and practical viability of the proposed ALWNN model, a hardware testbed was established for over-the-air (OTA) signal transmission and reception, simulating a realistic edge computing scenario. The experimental platform, comprises two National Instruments USRP-2901 Software-Defined Radios (SDRs) and a Raspberry Pi 4B for inference processing. One USRP-2901 unit, designated as the transmitter, was interfaced with a host PC running MATLAB R2024a for baseband signal generation and control. A 2ASK modulated waveforms, synthesized in software, emulating the signal characteristics specified in the RML2016.10a dataset, before being upconverted for OTA transmission. A second USRP-2901 served as the receiver, capturing the wireless signals and performing downconversion and digitization to produce complex in-phase and quadrature (I/Q) samples. The experiments were conducted in a Line-of-Sight (LOS) indoor environment with both SDRs tuned to a carrier frequency of 1.2 GHz and operating at a sampling rate of 5.6 MHz. The transmitter and receiver gains were set to 25 dB and 40 dB, respectively, to achieve an approximate signal-to-noise ratio (SNR) of 15 dB at the receiver input. Notably, the transmitter and receiver operated asynchronously, relying on their internal oscillators without a shared high-precision clock reference. This configuration intentionally introduces realistic channel impairments such as Carrier Frequency Offset (CFO) and Sampling Rate Offset (SRO).

3 Supplementary Note 3: Ablation Studies on Husimi Parameters

3.1 Visual Recognition

To evaluate the stability of entropy auditing with respect to the Husimi parameters, we systematically varied the window size (`win`), Gaussian width (σ), and hop size (`hop`) while computing the band-entropy change $\Delta S_{\mathcal{B}}$ on the ImageNet subsampling experiments. Across most configurations, the observed trends align well with theoretical expectations: as the sampling rate decreases (larger k), the magnitude of $\Delta S_{\mathcal{B}}$ increases, indicating a stronger perturbation of the local spectral structure. Furthermore, for any fixed sampling rate, the Uniform masking consistently yields larger $\Delta S_{\mathcal{B}}$ values than the Random masking. This confirms that periodic subsampling introduces greater spectral disorder, while random subsampling statistically preserves local entropy, in agreement with the findings of He *et al.* (2022) in Masked Autoencoders (MAE) [1]. Together, these results demonstrate that the Husimi-based entropy change serves as a robust *pre-training indicator* of downstream task difficulty in visual recognition.

However, certain parameter settings produce anomalous results at extremely sparse sampling (e.g., $k = 8$). Specifically, when `win`=64, each Husimi window covers only a 4×4 patch region, so under Uniform masking all pixels within the window may become constant, leading to degeneracy in the local spectral estimate. A similar artifact arises when `hop`=192, where the window stride is so large that many regions—particularly under Uniform masking—are insufficiently sampled. In both cases, the computed $\Delta S_{\mathcal{B}}$ becomes unreliable due to the breakdown of local spectral statistics. These observations highlight that the Husimi parameters should be chosen in a physically consistent manner, ensuring that the window and hop sizes reflect the effective spatial–frequency resolution of the underlying acquisition process.

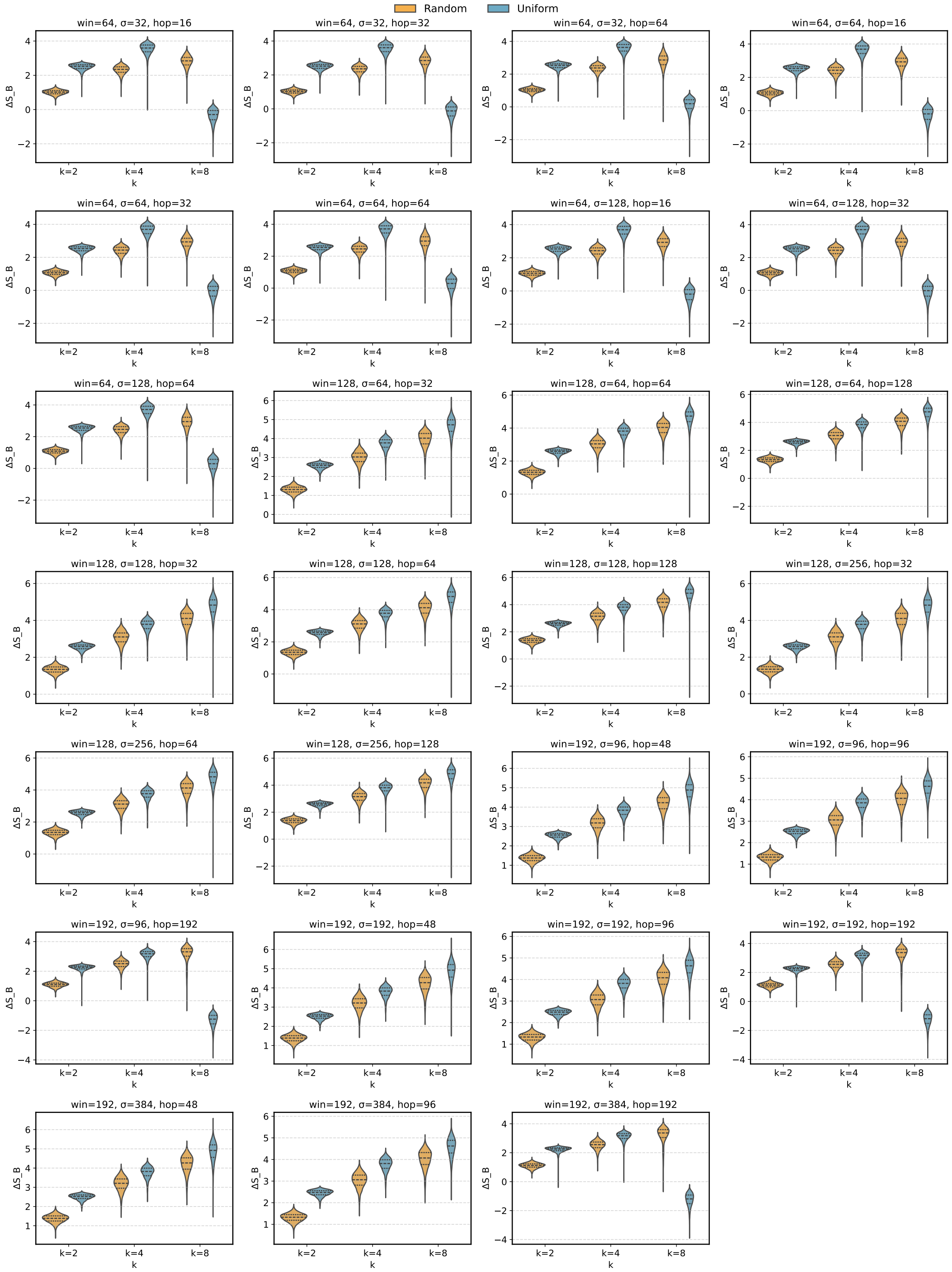

Fig. 1 **Ablation of Husimi parameters on ImageNet.** Each panel corresponds to one combination of (`win`, σ, `hop`). In nearly all cases, $\Delta S_{\mathcal{B}}$ increases monotonically with sampling interval k, and Uniform masking exhibits higher entropy change than Random masking. Deviations at extremely sparse sampling (notably at `win` = 64 and `hop` = 192) arise from insufficient spatial coverage within each Husimi window, demonstrating the importance of physically meaningful parameter choices.

3.2 MIMO

To ascertain the robustness of the proposed entropy metric with respect to its analysis parameters, we conducted an ablation study for the MIMO channel scenario. We varied the window size while holding the Gaussian width ($\sigma = 1$) and hop size (hop=1) constant. This parameterization is due to the fundamental structural differences between natural images and MIMO channel matrices. While images exhibit strong 2D spatial correlations, the pertinent correlations within a channel matrix exist primarily along one-dimensional axes, corresponding to the antenna arrays at the transmitter or receiver. Setting hop=1 ensures a high-resolution scan across the antenna elements without downsampling, thereby capturing the fine-grained local spectral variations essential for assessing channel quality. Concurrently, a small $\sigma = 1$ effectively localizes the Gaussian window's influence, concentrating the analysis on a single antenna's data stream (i.e., a row or column) at each step. This approach forces the 2D analysis framework to approximate a series of 1D evaluations, thereby minimizing cross-dimensional interference from physically uncorrelated antenna data and preserving the physical integrity of the spectral entropy measurement for the MIMO channel.

Across the majority of configurations, the observed trends aligned closely with theoretical expectations: as the masking interval d increases, the magnitude of the band-entropy change, $\Delta S_{\mathcal{B}}$, grows, indicating a more significant perturbation of the local spectral structure of the channel response. Furthermore, for any fixed sampling rate, the Uniform masking scheme consistently yields larger $\Delta S_{\mathcal{B}}$ values than the Random masking scheme. This finding suggests that periodic subsampling introduces greater spectral disorder into the channel estimate, while random subsampling statistically preserves local entropy. Collectively, these results demonstrate that the entropy-based analysis provides a robust characterization of spectral degradation under different subsampling strategies and is not overly sensitive to the selection of the window size.

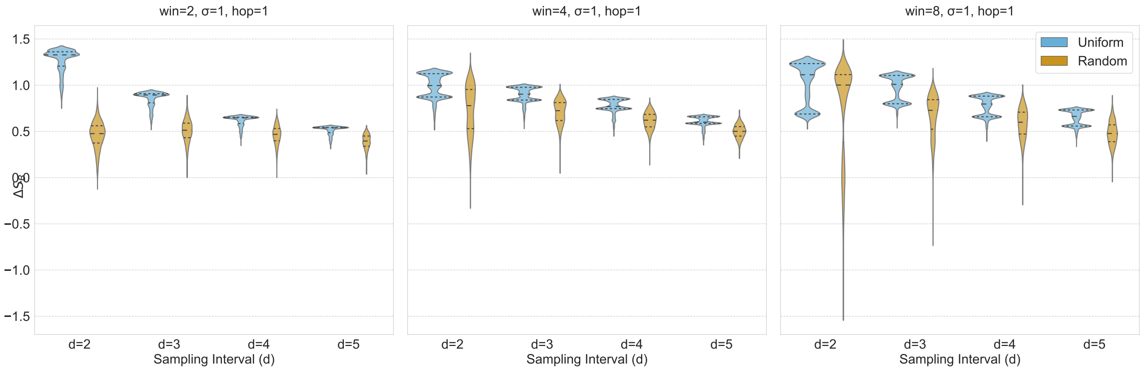

Fig. 2 **Ablation of Husimi parameters on MIMO.** Each panel corresponds to a specific parameter configuration of win, σ, and hop. Across all configurations, the $\Delta S_{\mathcal{B}}$, exhibits a monotonic decrease as the masking interval d increases. Furthermore, the Uniform masking scheme consistently yields a greater entropy perturbation than the Random scheme.

3.3 MRI: Ablation on Husimi Parameters

To evaluate the stability of the Husimi-based entropy auditing in magnetic resonance imaging (MRI), we conducted an extensive grid search across the parameter space of the Gaussian window (`win`), its standard deviation (σ), and hop size (`hop`). For each parameter combination, the band-entropy change $\Delta S_{\mathcal{B}}$ was computed from the same set of fully sampled k-space data under three canonical mask types—Random, Uniform, and Poisson—and multiple acceleration ratios. Each configuration was processed using the same pipeline described in Main; the analysis was performed entirely prior to training, and the results were visualized as a 6×6 grid of violin plots (Fig. 3).

We systematically varied $\texttt{win} \in \{12, 24, 48, 96\}$, $\sigma/\texttt{win} \in \{0.5, 1, 2\}$, and $\texttt{hop}/\texttt{win} \in \{0.25, 0.5, 1\}$, producing $4 \times 3 \times 3 = 36$ configurations. For each combination, $\Delta S_{\mathcal{B}}$ values across all slices and sampling ratios were aggregated per mask type to produce violin plots that summarize their statistical distributions. All subplots share a common color code (orange = Random, blue = Uniform, green = Poisson) and display the parameter triple ($\texttt{win}, \sigma, \texttt{hop}$) in their title; each panel's vertical scale is locally adaptive to emphasize within-condition trends.

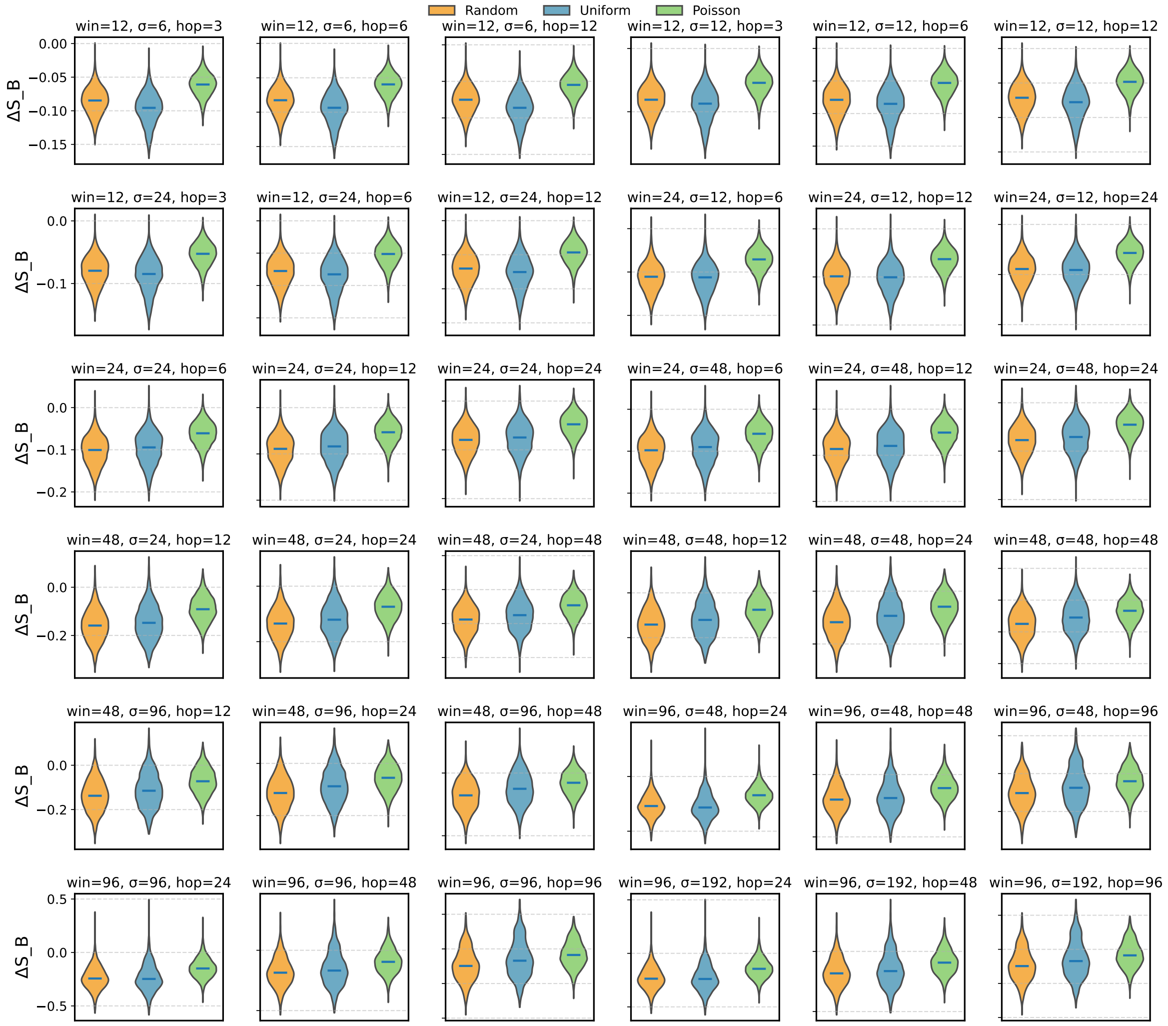

Fig. 3 Ablation of Husimi parameters on MRI. Each subplot corresponds to one combination of $(\texttt{win}, \sigma, \texttt{hop})$; violin plots summarize the distribution of $\Delta S_{\mathcal{B}}$ across all slices and acceleration ratios for three mask types (Random, Uniform, Poisson). The overall ranking Poisson>Uniform>Random remains stable across nearly all parameter settings, consistent with the empirical reconstruction results in the main text. The ordering partially reverses only at extremely small windows ($\texttt{win}$=12), where the Husimi window covers insufficient spatial–frequency support to distinguish the structured aliasing of Uniform from the stochastic loss in Random. This highlights the importance of selecting Husimi parameters that are physically and architecturally consistent with both the imaging resolution and the receptive-field scale of the downstream network.

Across the majority of settings, the ordering of mean $\Delta S_{\mathcal{B}}$ values is highly stable:

$$\mathsf{Poisson} \;>\; \mathsf{Uniform} \;>\; \mathsf{Random}.$$

This hierarchy exactly matches the empirical reconstruction ranking (Poisson → best PSNR/SSIM, Random → worst) reported in the main text, demonstrating that the Husimi-based entropy measure faithfully captures the relative information preservation of different undersampling patterns. The Poisson mask consistently produces the smallest $|\Delta S_{\mathcal{B}}|$, indicating that its variable-density sampling preserves the original spectral organization most effectively.

When the window size is very small ($\texttt{win}$=12), however, the ordering between Uniform and Random reverses. In this regime, the Husimi window spans too few pixels to include sufficient structural information, and the entropy estimate becomes dominated by local intensity variations rather than global aliasing structure. Physically, this corresponds to an under-resolved phase-space measurement: the local spectrogram fails to capture the coherent folding patterns induced by uniform subsampling. From a representational perspective, the network's convolutional receptive fields are relatively broad compared with such a small Husimi window, so the entropy measure loses sensitivity to the periodicity that the network can still exploit.

These findings emphasize that the choice of Husimi parameters should reflect not only the imaging physics (spatial–frequency resolution of the acquisition) but also the representational scale of the downstream model. In other words, the Husimi transform serves as a conceptual bridge between the *measurement domain* and the *learning domain*: its parameters determine at what scale the instrument-resolved structure is compared to the model's feature extraction capability. When chosen appropriately, the band-entropy change $\Delta S_{\mathcal{B}}$ provides a stable, physically interpretable predictor of reconstruction difficulty across undersampling strategies.

References

[1] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16000–16009 (2022)