Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Statistical Order in Representation Learning: Sufficiency, Architectural Blindness, and Generalization Bounds via Maximum Entropy

Yunguo Yu

yuyunguo@gmail.com

Zyter|TruCare https://orcid.org/0009-0006-2003-3813

Research Article

Keywords: representation learning, maximum entropy, higher-order statistics, sufficient statistics, distribution shift, generalization bounds

Posted Date: April 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9273811/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Statistical Order in Representation Learning: Sufficiency, Architectural Blindness, and Generalization Bounds via Maximum Entropy

Yunguo Yu yunguo.yu@gmail.com
AI Innovations & Prototyping, Zyter|TruCare
Vienna, VA, USA

Abstract

We establish three formal results about the role of statistical order in representation learning. **(i) Sufficiency:** under a maximum-entropy (Gibbs) data model, a representation that preserves all K-order multipoint statistics is a sufficient statistic for the model's natural parameters, and is equivalent to saturating the mutual information $I(Z;\Phi_K(X)) = H(\Phi_K(X))$. **(ii) Architectural blindness:** for any $r < K$, there exist distributions that agree on all r-order statistics yet differ at order K; consequently, no function with interaction order r can distinguish them. This provides a formal explanation for the texture bias of convolutional networks and their failure on shape-based tasks. **(iii) Generalization bounds:** out-of-distribution error for any predictor depending on K-order statistics is bounded by the K-order statistical discrepancy D_K between training and test distributions, an interpretable, computable quantity related to polynomial-kernel MMD.

We develop these results within a unified maximum-entropy framework that places data distributions as exponential families constrained by multipoint statistics, connecting statistical physics, biological vision, and machine learning. Primate visual cortex provides independent motivation: V2 — but not V1 — is selectively sensitive to higher-order multipoint correlations (Yu et al., 2015), suggesting the brain incrementally constructs sufficient statistics for increasing order K.

We validate all three predictions on controlled binary-texture stimuli that isolate statistical structure at each order while holding lower-order statistics fixed. At every target order $K \in \{2, 3, 4\}$, classifiers using order-$r < K$ features perform at chance while order-$r = K$ features yield perfect accuracy, providing an exact empirical realization of the architectural blindness theorem. D_K tracks the generalization gap of a diffusion model ($\rho = 0.94$), and transformer layer depth monotonically increases K-sufficiency, consistent with the information-theoretic characterization.

These results unify texture bias, adversarial vulnerability, and distribution shift under a single statistical framework, and suggest D_K as a principled pre-deployment diagnostic for generalization gaps.

Keywords: representation learning, maximum entropy, higher-order statistics, sufficient statistics, distribution shift, generalization bounds

1 Introduction

Deep neural networks have achieved state-of-the-art performance across vision, language, and multimodal tasks (LeCun et al., 1998; He et al., 2016; Dosovitskiy et al., 2021). Yet they fail systematically under distribution shift (Hendrycks and Dietterich, 2019), adversarial perturbations (Szegedy et al., 2014; Goodfellow et al., 2015), and texture versus shape manipulations (Geirhos et al., 2019). These failures are not explained by classical learning-theoretic bounds alone; they point to a representational deficiency.

A complementary perspective comes from the statistics of natural images. Julesz (1962) conjectured that human texture discrimination is governed by higher-order statistics, a hypothesis confirmed and operationalized in the parametric texture models of Portilla and Simoncelli (2000) and the neural style-transfer work of Gatys et al. (2015). Meaningful structure in images is encoded not

only in pairwise correlations but also in higher-order (multipoint) statistics. Natural scene statistics exhibit non-Gaussian structure captured precisely by these multipoint correlations (Simoncelli and Olshausen, 2001), and hierarchical visual cortex areas are tuned to progressively higher-order features (Olshausen and Field, 1996; Yu et al., 2015).

Despite this evidence, *contemporary representation-learning objectives do not explicitly enforce preservation of higher-order statistical structure.* Contrastive objectives (Chen et al., 2020) align features at the level of global embeddings; generative models typically constrain marginal or pairwise statistics; and empirical risk minimization says nothing about the order of the statistics that drive predictions. As a result, learned representations may match training statistics up to second order while being blind to structure that distinguishes classes, textures, or semantic compositions.

This work. We develop a theoretical framework grounded in the *maximum-entropy principle* (Jaynes, 1957) and establish the following contributions, each of which is new to the best of our knowledge:

1. **K-order sufficiency (Theorem 1):** We formalize learned representations as *K-order sufficient statistics* in the Fisher–Neyman sense: a representation Z is K-sufficient iff $\Phi_K(X)$ is a.s. recoverable from Z, or equivalently $I(Z;\Phi_K(X)) = H(\Phi_K(X))$. This is the first characterization of representation quality in terms of exponential-family sufficiency.
2. **Architectural blindness (Theorems 2 & 4):** We give an explicit parity construction proving that *no* function with interaction order $r < K$ can distinguish two distributions that agree on all r-order statistics. We show this implies that convolutional networks with bounded receptive field are formally incapable of learning features encoded only in higher-order structure.
3. **Order-aware generalization bound (Theorem 3):** We prove that prediction error under distribution shift is bounded by the K-order statistical discrepancy D_K, a computable moment-space analog of polynomial-kernel MMD (Gretton et al., 2012).
4. **Information-theoretic characterization (Theorem 5):** We show K-sufficiency is equivalent to saturating the mutual information $I(Z;\Phi_K) = H(\Phi_K)$, connecting the framework to the information bottleneck (Tishby et al., 1999).
5. **Unification:** We show that Gram-matrix style transfer (Gatys et al., 2015), Portilla–Simoncelli texture synthesis, and the biological V1–V4 hierarchy are all special cases of increasing K-order MaxEnt models in appropriate feature spaces.

Organization. Section 2 reviews related work. Section 3 introduces notation and preliminaries. Section 4 develops the multipoint statistical framework. Section 5 presents the main theorems with complete proofs. Section 6 establishes connections to existing methods. Section 7 sketches a higher-order learning objective as a proposed direction for future work. Section 8 provides illustrative examples. Section 9 presents empirical validation on controlled binary-texture stimuli. Section 10 discusses implications and limitations. Section 11 concludes.

Our results suggest that the choice of statistical order K plays a role analogous to model capacity, but along a fundamentally different axis: not complexity of functions, but complexity of interactions.

2 Related Work

Maximum entropy and statistical physics. The maximum-entropy principle was introduced by Jaynes (1957) as a foundation for statistical inference: the least-biased distribution consistent with observed constraints is the one maximizing entropy. In statistical physics, this yields the Gibbs/Boltzmann distribution, the foundation of equilibrium thermodynamics (Mézard and Montanari, 2009). Exponential families, the mathematical language of maximum-entropy models, are the subject of an extensive modern treatment (Wainwright and Jordan, 2008). We leverage this machinery

directly, showing that data distributions constrained by multipoint statistics are exponential families whose sufficient statistics are precisely the multipoint monomials.

Texture perception and higher-order image statistics. Julesz (1962) conjectured that textures with identical nth-order statistics are perceptually indistinguishable. Portilla and Simoncelli (2000) operationalized this by constructing a parametric texture model from joint statistics of complex wavelet coefficients — effectively second-order statistics in a rich feature space. Gatys et al. (2015, 2016) showed that Gram matrices of deep convolutional features (second-order statistics in feature space) capture enough texture information for convincing synthesis and style transfer. Our framework provides a unified account: these methods correspond to maximum-entropy distributions constrained by $K = 2$ statistics in an appropriate feature space, and their limitations reflect the absence of higher-order constraints.

Distribution shift and domain adaptation. Ben-David et al. (2010) established that transfer error is bounded by a divergence between source and target distributions; our Theorem 13 provides an analogous bound in terms of the multipoint statistical discrepancy D_K, giving a finer-grained, order-aware decomposition. Hendrycks and Dietterich (2019) benchmark neural-network robustness to common corruptions, and Geirhos et al. (2019) demonstrate that ImageNet-trained CNNs rely predominantly on texture (local, low-order) cues rather than shape (global, higher-order) cues. Our Theorem 16 formally explains this bias: architectures with limited receptive field or local inductive bias belong to function classes $\mathcal{F}_r$ for small r and are provably blind to higher-order structure.

Moment matching and kernel two-sample testing. Gretton et al. (2012) introduced Maximum Mean Discrepancy (MMD), the supremum of expected function differences over a reproducing kernel Hilbert space (RKHS). When instantiated with a degree-K polynomial kernel, MMD recovers a weighted variant of our D_K discrepancy (up to normalization constants), because the feature map of a polynomial kernel is precisely the vector of monomials up to degree K. The kernel mean embedding literature (Smola et al., 2007) further characterizes when moment matching to finite order is sufficient to identify a distribution — a precursor to our Theorem 10. Our contribution relative to this line is the explicit connection between D_K and *representation sufficiency*: we show that D_K bounds the error of any predictor that is K-sufficient, and that architectural limitations (Theorem 16) prevent certain models from computing functions necessary to reduce D_K. Additionally, D_K is defined directly in moment space and does not require kernel selection, making it interpretable as an order-indexed hierarchy of distributional discrepancies.

Adversarial examples. Szegedy et al. (2014); Goodfellow et al. (2015) showed that small imperceptible perturbations can fool classifiers. Several works have connected adversarial vulnerability to distributional discrepancies (Ben-David et al., 2010). Our framework adds a statistical-order dimension: adversarial inputs that preserve low-order statistics but disrupt higher-order ones are invisible to any K-constrained model.

Representation learning and information bottleneck. Bengio et al. (2013) survey representation learning objectives. Tishby et al. (1999); Tishby and Zaslavsky (2015) propose the information bottleneck (IB) as a principle for learning compact, task-relevant representations. Our Theorem 18 provides an information-theoretic characterization of K-sufficiency that connects directly to the IB: a K-sufficient representation saturates the mutual information between Z and $\Phi_K(X)$.

Biological vision. Sparse coding models of V1 (Olshausen and Field, 1996) explain simple-cell receptive fields as second-order statistical features of natural scenes. V2 neurons respond selectively to multipoint (higher-order) correlations (Yu et al., 2015), and V2-like texture sensitivity can be induced in artificial networks by changing training objectives (Yu, 2026). Freeman and Simoncelli (2011) showed that V2 model responses predict perceptual sensitivity to the same class of higher-order correlations used in our binary-texture experiments. Responses in higher visual areas (V4, IT) correspond to progressively more complex features (Simoncelli and Olshausen, 2001). Our framework

formalizes this hierarchy: each cortical stage performs K-order sufficient compression for increasing K.

3 Preliminaries

3.1 Random Fields and Multipoint Statistics

Let $\Omega = \{1, \dots, N\}$ be a finite index set (e.g., pixel or feature coordinates). Let $X = (X_1, \dots, X_N) \in \mathcal{X} \subseteq \mathbb{R}^N$ be a random vector (random field) with distribution p.

Definition 1 (Multipoint monomial) *For $S \subseteq \Omega$ with $|S| = k$, the* k-order monomial *is*

$$\phi_S(X) = \prod_{i \in S} X_i. \tag{1}$$

Definition 2 (K-order statistic vector) *The* K-order statistic vector *of X is*

$$\Phi_K(X) = \big(\phi_S(X)\big)_{S \subseteq \Omega,\, |S| \le K}, \tag{2}$$

a vector of dimension $M_K = \sum_{k=1}^{K} \binom{N}{k}$.

Remark 3 *$K = 1$: means. $K = 2$: means and covariances (after centering). $K \ge 3$: captures skewness, kurtosis, and multi-way correlations essential for non-Gaussian structure.*

3.2 Exponential Families and Maximum Entropy

Let $\boldsymbol{c} = (c_S)_{|S| \le K}$ be a vector of target moment constraints. The *maximum-entropy (MaxEnt) distribution* subject to $\mathbb{E}_p[\phi_S(X)] = c_S$ for all $|S| \le K$ is the solution to

$$\max_p H(p) \quad \text{subject to} \quad \mathbb{E}_p[\phi_S(X)] = c_S, \quad \forall S,\, |S| \le K. \tag{3}$$

Proposition 4 (Gibbs distribution; Jaynes 1957) *The unique solution is the Gibbs distribution*

$$p^*_{MaxEnt}(x) = \frac{1}{Z(\boldsymbol{\lambda})} \exp\Big(\sum_{|S| \le K} \lambda_S \phi_S(x) \Big). \tag{4}$$

where $\boldsymbol{\lambda} = (\lambda_S)$ are Lagrange multipliers (natural parameters) and $Z(\boldsymbol{\lambda}) = \int \exp(\cdots)\, \mathrm{d}x$ is the partition function. This is an exponential family with sufficient statistic $\Phi_K(X)$ and log-partition function $A(\boldsymbol{\lambda}) = \log Z(\boldsymbol{\lambda})$.

The natural parameters $\boldsymbol{\lambda}$ are in bijective correspondence with the moments $\boldsymbol{c}$ (by strict convexity of A), so the MaxEnt distribution is uniquely determined by either parameterization.

3.3 Sufficiency (Classical Definition)

Let $p_{\boldsymbol{\lambda}}$ be a parametric family indexed by $\boldsymbol{\lambda}$. A statistic $T(X)$ is *sufficient* for $\boldsymbol{\lambda}$ if $p(X \mid T(X), \boldsymbol{\lambda}) = p(X \mid T(X))$, i.e., X and $\boldsymbol{\lambda}$ are conditionally independent given $T(X)$. By the Fisher–Neyman factorization theorem, T is sufficient iff $p_{\boldsymbol{\lambda}}(X) = g_{\boldsymbol{\lambda}}(T(X)) \cdot h(X)$ for some functions g and h. For the exponential family p^*, the minimal sufficient statistic is $\Phi_K(X)$.

3.4 Information-Theoretic Notation

We write $I(A; B)$ for mutual information, $H(A)$ for Shannon entropy, and $D_{\mathrm{KL}}(p \,\|\, q)$ for KL divergence, all in nats. Equalities and inequalities hold almost surely unless stated otherwise.

4 Multipoint Statistical Framework

4.1 K-Order Sufficient Representations

Let $Z = f_\theta(X) \in \mathcal{Z}$ be a learned representation.

Definition 5 (K-order sufficiency) *Z is* K-order sufficient *(for distribution p^*) if there exists a measurable function $h : \mathcal{Z} \to \mathbb{R}^{M_K}$ such that*

$$\Phi_K(X) = h(Z) \quad p^*\text{-almost surely.} \tag{5}$$

Equivalently, $\Phi_K(X)$ is measurable with respect to the σ-algebra generated by Z. Intuitively, Z is K-sufficient if all K-order statistical information in X can be perfectly reconstructed from Z.

Remark 6 *This is the standard notion of deterministic sufficiency: Z retains all information in $\Phi_K(X)$ without loss. It does* not *require $Z = \Phi_K(X)$; any injective function of $\Phi_K(X)$ suffices. In particular, a neural network f_θ is K-sufficient if the mapping $X \mapsto (\Phi_K(X), f_\theta(X))$ has the same range as $X \mapsto f_\theta(X)$ up to a bijection on the Φ_K component.*

Remark 7 (Comparison with the literature) *Some works define sufficiency via the conditional expectation condition $\mathbb{E}[\phi_S(X) \mid Z] = \mathbb{E}[\phi_S(X)]$, which instead expresses* independence *of ϕ_S from Z — the opposite of what is needed. Definition 5 uses the standard Fisher–Neyman formulation and is consistent with classical statistics (Cover and Thomas, 2006).*

4.2 MaxEnt Distributions as Data Models

Under the MaxEnt hypothesis, the data-generating distribution is $p^* = p^*_{\boldsymbol{\lambda}}$ for some unknown $\boldsymbol{\lambda}$. Inference about $\boldsymbol{\lambda}$ is equivalent to inference about $\Phi_K(X)$ (by Proposition 4). A representation Z that is K-sufficient therefore retains all inferential content about the data-generating process.

This motivates the following design principle: *a representation should be at least K-sufficient for the highest order K at which the data exhibits structured correlations.*

5 Main Theoretical Results

5.1 Theorem 1: Sufficiency and Distribution Preservation

Theorem 8 (Sufficiency and natural parameters) *Let p^* be the MaxEnt distribution of Proposition 4. If Z is K-order sufficient (Definition 5), then Z is a sufficient statistic for the natural parameter $\boldsymbol{\lambda}$ of the exponential family p^*.*

Proof Since $\Phi_K(X)$ is minimal sufficient for $\boldsymbol{\lambda}$ in the exponential family p^*, and $\Phi_K(X)$ is a measurable function of Z by Definition 5, it follows that Z is also sufficient for $\boldsymbol{\lambda}$ by the data-processing property of sufficient statistics. ■

Corollary 9 *Any two K-sufficient representations Z_1 and Z_2 of X yield identical posterior distributions over $\boldsymbol{\lambda}$: $p(\boldsymbol{\lambda} \mid Z_1) = p(\boldsymbol{\lambda} \mid Z_2)$ for any prior $p(\boldsymbol{\lambda})$.*

5.2 Theorem 2: Failure of Second-Order Models

Theorem 10 (Indistinguishability under second-order constraints) *There exist distributions p and q over $\{-1, +1\}^3$ such that*

$$\mathbb{E}_p[X_i X_j] = \mathbb{E}_q[X_i X_j] \quad \forall i \neq j, \tag{6}$$

yet

$$\mathbb{E}_p[X_1X_2X_3] \neq \mathbb{E}_q[X_1X_2X_3]. \tag{7}$$

Consequently, any model depending only on second-order statistics cannot distinguish p from q.

Proof Define q as the uniform distribution on $\{-1,+1\}^3$: $q(x) = 1/8$ for all $x \in \{-1,+1\}^3$.

Define p as the uniform distribution on configurations satisfying the global parity constraint $x_1x_2x_3 = +1$:

$$p(x) = \begin{cases} 1/4 & \text{if } x_1x_2x_3 = +1, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

The four even-parity configurations are $\{(+,+,+),(+,-,-),(-,+,-),(-,-,+)\}$.

Second-order statistics under p. By symmetry under sign flips of any two coordinates, $\mathbb{E}_p[X_i] = 0$ and $\mathbb{E}_p[X_iX_j] = 0$ for all $i \neq j$. Explicitly, for $\mathbb{E}_p[X_1X_2]$: the four configurations contribute $(+)(+) + (+)(-) + (-)(+) + (-)(-) = 1 - 1 - 1 + 1 = 0$, each with probability $1/4$.

Second-order statistics under q. Under the uniform distribution, $\mathbb{E}_q[X_i] = 0$ and $\mathbb{E}_q[X_iX_j] = \sum_x q(x)x_ix_j = 0$ for all $i \neq j$, by symmetry.

Third-order statistics. Under p: $\mathbb{E}_p[X_1X_2X_3] = \frac{1}{4}(1+1+1+1) = 1$, since every configuration in the support satisfies $x_1x_2x_3 = +1$.

Under q: $\mathbb{E}_q[X_1X_2X_3] = \frac{1}{8}\sum_{x\in\{-1,+1\}^3} x_1x_2x_3 = 0$ by symmetry.

Thus all pairwise statistics agree but the third-order statistic differs, so no function of $\{X_iX_j\}_{i\neq j}$ can separate p and q. ■

Remark 11 *This construction is tight: three bits require exactly three variables to encode a global parity constraint that is invisible to pairwise tests. The same construction extends to n variables and order K by choosing a parity constraint on any K-subset $S' \subseteq \{1,\dots,n\}$.*

5.3 Theorem 3: Generalization via Statistical Discrepancy

Definition 12 (K-order discrepancy) *For distributions p and q, define*

$$D_K(p,q) = \sum_{|S|\leq K} \left|\mathbb{E}_p[\phi_S] - \mathbb{E}_q[\phi_S]\right|. \tag{9}$$

Theorem 13 (Generalization bound) *Let p and q be distributions over $\mathcal{X}$.*

(a) ***(Linear predictors.)*** *Let $f(X) = \sum_{|S|\leq K} a_S\phi_S(X)$ with $\|\boldsymbol{a}\|_\infty \leq L$. Then*

$$|\mathbb{E}_p[f(X)] - \mathbb{E}_q[f(X)]| \leq L \cdot D_K(p,q).$$

(b) ***(General Lipschitz predictors.)*** *If $f(X) = g(\Phi_K(X))$ where $g : \mathbb{R}^{M_K} \to \mathbb{R}$ is L-Lipschitz with respect to $\|\cdot\|_1$, then*

$$|\mathbb{E}_p[f(X)] - \mathbb{E}_q[f(X)]| \leq L \cdot W_1(p_{\Phi_K}, q_{\Phi_K}),$$

where W_1 denotes the Wasserstein-1 distance between the push-forward distributions of $\Phi_K(X)$ under p and q. If $\Phi_K(X)$ is bounded in ℓ_1 norm by B, then $W_1(p_{\Phi_K}, q_{\Phi_K}) \leq 2B \cdot \mathrm{TV}(p_{\Phi_K}, q_{\Phi_K})$.

Proof Part (a). By linearity of expectation:

$$\begin{aligned} |\mathbb{E}_p[f] - \mathbb{E}_q[f]| &= \left|\sum_{|S|\leq K} a_S\big(\mathbb{E}_p[\phi_S] - \mathbb{E}_q[\phi_S]\big)\right| \\ &\leq \sum_{|S|\leq K} |a_S| \cdot |\mathbb{E}_p[\phi_S] - \mathbb{E}_q[\phi_S]| \\ &\leq \|\boldsymbol{a}\|_\infty \cdot D_K(p,q) = L \cdot D_K(p,q). \end{aligned} \tag{10}$$

Part (b). Since $f = g \circ \Phi_K$ and g is L-Lipschitz:

$$|\mathbb{E}_p[g(\Phi_K)] - \mathbb{E}_q[g(\Phi_K)]| = \left| \int g \, \mathrm{d}(p_{\Phi_K} - q_{\Phi_K}) \right| \leq L \cdot W_1(p_{\Phi_K}, q_{\Phi_K}), \tag{11}$$

using the Kantorovich–Rubinstein duality $W_1(\mu, \nu) = \sup_{\|g\|_{\mathrm{Lip}} \leq 1} \int g \, \mathrm{d}(\mu - \nu)$. When $\Phi_K(X)$ is bounded in ℓ_1 norm by B, the support of p_{Φ_K} and q_{Φ_K} has ℓ_1-diameter at most $2B$. The bound $W_1(p_{\Phi_K}, q_{\Phi_K}) \leq 2B \cdot \mathrm{TV}(p_{\Phi_K}, q_{\Phi_K})$, follows from the coupling characterization: under the TV-optimal coupling,

$$\mathbb{E}_\gamma[\|u - v\|_1] \leq 2B \cdot \Pr_\gamma(u \neq v) = 2B \cdot \mathrm{TV}. \tag{12}$$

■

Remark 14 (Connection to domain adaptation) *Ben-David et al. (2010) bound transfer error by an $\mathcal{H}$-divergence between source and target. Theorem 13 provides a complementary bound in moment space: the order-K moment discrepancy D_K is an interpretable, computable proxy for distributional divergence.*

5.4 Theorem 4: Indistinguishability of Local Function Classes

Definition 15 (r-local function class) *$\mathcal{F}_r$ is the class of all measurable functions $f : \{-1, +1\}^n \to \mathbb{R}$ such that $f(X) = g\big(\{X_i : i \in T\}\big)$ for some set of coordinates $T \subseteq \Omega$ with $|T| \leq r$.*

Theorem 16 (Indistinguishability by local functions) *Let $r < K \leq n$. There exist distributions p and q on $\{-1, +1\}^n$ such that:*

(i) $\mathbb{E}_p[\phi_S(X)] = \mathbb{E}_q[\phi_S(X)]$ *for all* $S \subseteq \Omega$ *with* $|S| \leq r$,

(ii) $\mathbb{E}_p[\phi_{S'}(X)] \neq \mathbb{E}_q[\phi_{S'}(X)]$ *for some* $|S'| = K$,

(iii) $\mathbb{E}_p[f(X)] = \mathbb{E}_q[f(X)]$ *for all* $f \in \mathcal{F}_r$.

Hence, no function in $\mathcal{F}_r$ can distinguish p from q.

Proof Fix any $S' = \{s_1, \ldots, s_K\} \subseteq \Omega$ with $|S'| = K$.

Construction. Let q be the uniform distribution on $\{-1, +1\}^n$: $q(x) = 2^{-n}$. Let p be the distribution conditioned on $\prod_{i \in S'} x_i = +1$:

$$p(x) = \begin{cases} 2^{-(n-1)} & \text{if } \prod_{i \in S'} x_i = +1, \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

(i) Matching of all r-order statistics. For any $S \subseteq \Omega$ with $|S| \leq r < K$, we have $S \not\supseteq S'$, so there exists $j \in S' \setminus S$.

The parity constraint $\prod_{i \in S'} x_i = +1$ determines x_j given $\{x_i : i \in S', i \neq j\}$. Since $j \notin S$, marginalizing over x_j (and all variables not in S) yields a uniform distribution over $\{-1, +1\}^{|S|}$, the same as under q. Formally, for any $s \in \{-1, +1\}^{|S|}$:

$$p(X_S = s) = \sum_{x:\, x_S = s} p(x) = 2^{-(|S|+1)} \cdot 2 = 2^{-|S|} = q(X_S = s). \tag{14}$$

Hence $\mathbb{E}_p[\phi_S] = \mathbb{E}_q[\phi_S]$ for all $|S| \leq r$.

(ii) Difference at order K. By construction, $\mathbb{E}_p[\phi_{S'}] = 1$ while $\mathbb{E}_q[\phi_{S'}] = 0$ (by the symmetry argument of Theorem 10).

(iii) Indistinguishability by $\mathcal{F}_r$. Every $f \in \mathcal{F}_r$ depends on at most $r < K$ variables, so it can be written as $f(X) = g(X_T)$ for some $|T| \leq r$. By (i), $p(X_T = \cdot) = q(X_T = \cdot)$, so $\mathbb{E}_p[f] = \sum_t g(t)\, p(X_T = t) = \sum_t g(t)\, q(X_T = t) = \mathbb{E}_q[f]$. ■

Corollary 17 (Architectural limitation) *Convolutional networks with bounded receptive field and limited depth can be approximated by functions in $\mathcal{F}_r$ with r determined by the effective receptive field. By Theorem 16, such networks cannot detect any statistical structure of order $K > r$.*

5.5 Theorem 5: Information-Theoretic Characterization

Theorem 18 (Information-theoretic characterization of K-sufficiency) *Let $X \sim p^*$ and $Z = f(X)$. Then the following are equivalent:*

(a) Z is K-order sufficient (Definition 5).

(b) $I(Z; \Phi_K(X)) = H(\Phi_K(X))$.

(c) $H(\Phi_K(X) \mid Z) = 0$.

Moreover, any representation achieving $I(Z; X) < H(\Phi_K(X))$ must violate K-sufficiency and therefore lose information about the MaxEnt natural parameters $\boldsymbol{\lambda}$.

Proof Recall that $\Phi_K(X)$ is a deterministic function of X, so $H(\Phi_K(X) \mid X) = 0$ and $H(\Phi_K(X)) \leq H(X)$.

(a) ⇒ (b, c). If Z is K-sufficient, then $\Phi_K(X) = h(Z)$ a.s. for some measurable h. Therefore $H(\Phi_K(X) \mid Z) = 0$, which gives: $I(Z; \Phi_K(X)) = H(\Phi_K(X)) - H(\Phi_K(X) \mid Z) = H(\Phi_K(X))$.

(c) ⇒ (a). $H(\Phi_K(X) \mid Z) = 0$ means that $\Phi_K(X)$ is a deterministic function of Z (a.s.), which is exactly Definition 5.

(b) ⇔ (c). Since

$$I(Z; \Phi_K) = H(\Phi_K) - H(\Phi_K \mid Z)\,, \tag{15}$$

the equivalence is immediate.

Final claim. Since $\Phi_K(X)$ is a function of X, the data-processing inequality yields $I(Z; \Phi_K(X)) \leq I(Z; X)$. If $I(Z; X) < H(\Phi_K(X))$, then $I(Z; \Phi_K(X)) \leq I(Z; X) < H(\Phi_K(X))$, so (b) fails, and by equivalence (a) fails. ■

Remark 19 (Connection to the information bottleneck) *In the information bottleneck (Tishby et al., 1999), the goal is to find Z that minimizes $I(Z; X)$ subject to $I(Z; Y) \geq \beta$ for a label Y. If Y is measurable with respect to $\Phi_K(X)$ (i.e., prediction depends only on K-order statistics), then K-sufficiency is a* necessary condition *for a lossless bottleneck: any compression below $H(\Phi_K(X))$ bits must lose task-relevant information.*

6 Connections to Existing Methods

6.1 Gram Matrices and Second-Order MaxEnt

Neural style transfer (Gatys et al., 2015, 2016) synthesizes images by matching the *Gram matrix* $G_\ell = F_\ell F_\ell^\top$ of feature maps F_ℓ at layer ℓ. The (i, j) entry of G_ℓ is the inner product between the ith and jth channel maps, a second-order statistic in feature space.

Proposition 20 (Gram matrices as second-order MaxEnt) *Let $\phi_{ij}(X) = \langle F_\ell^i(X), F_\ell^j(X)\rangle$ for channels i, j. Matching all Gram matrices $\{G_\ell\}$ is equivalent to constraining the second-order statistics $\{\mathbb{E}[\phi_{ij}(X)]\}_{i,j,\ell}$ in feature space. The texture model implicitly defined by style transfer is therefore the $K = 2$ MaxEnt distribution in the joint feature space $\bigtimes_\ell F_\ell$.*

Proof Matching $\mathbb{E}[\phi_{ij}]$ for all channel pairs (i, j) at all layers constrains the second-order moments of the feature vector. By Proposition 4, the maximum-entropy distribution consistent with these constraints is the Gibbs distribution with second-order potentials. The style-transfer synthesis process (minimizing mean-squared Gram-matrix discrepancy) is gradient descent toward this maximum-entropy model. ■

The failure modes of style transfer — inability to preserve global composition, semantic content, or fine texture detail — are precisely the failure modes predicted by Theorem 16 for $r = 2$ models.

6.2 Portilla–Simoncelli Texture Synthesis

Portilla and Simoncelli (2000) construct textures by matching joint statistics of complex wavelet subband responses across scales and orientations. These statistics include cross-channel correlations (second-order) and cross-scale phase correlations (effectively third-order). Their model is therefore a partial $K = 3$ MaxEnt model in wavelet feature space, explaining its qualitative superiority over Gram-matrix methods for structured textures.

6.3 Biological Vision Hierarchy

The cortical hierarchy V1 → V2 → V4 → IT can be interpreted as a sequential increase in the order K of the sufficient statistics computed:

Area	Canonical features	Approx. order	Reference
V1	Gabor filters, orientation	$K = 2$	Olshausen and Field (1996)
V2	Cross-orientation, phase	$K = 3$	Yu et al. (2015); Freeman and Simoncelli (2011)
V4	Curvature, color interactions	$K = 4$–5	Simoncelli and Olshausen (2001) (approx.)
IT	Object identity, compositionality	$K \gg 1$	Simoncelli and Olshausen (2001) (approx.)

Note: the order assignments for V4 and IT are approximate and intended to illustrate the hierarchical interpretation; precise order bounds for these areas remain an open empirical question.

Under our framework, the brain incrementally constructs a sufficient statistic for increasingly high-order MaxEnt distributions of natural scenes.

7 A Higher-Order Learning Objective (Proposed Direction)

Our theoretical results suggest augmenting standard task losses with an explicit higher-order statistical regularizer. We sketch this objective here as a proposed direction motivated by the theory; full empirical evaluation is deferred to future work. Let $\hat{\Phi}_K(X; f_\theta)$ denote the estimated K-order statistic vector computed from the representation $Z = f_\theta(X)$. Define the *higher-order alignment loss*:

$$\mathcal{L}_{\mathrm{HO}}(\theta) = \sum_{k=2}^{K} \mu_k \sum_{S:\,|S|=k} \Big(\mathbb{E}_{\hat{p}}[\phi_S(X)] - \mathbb{E}_{p_{\mathrm{data}}}[\phi_S(X)]\Big)^2,$$

where $\mu_k > 0$ are order-specific weights and $\hat{p}$ is the empirical distribution over the mini-batch. The combined objective is

$$\mathcal{L}(\theta) = \mathcal{L}_{\mathrm{task}}(\theta) + \lambda\, \mathcal{L}_{\mathrm{HO}}(\theta).$$

Algorithm 1 Higher-Order Statistical Regularization (HOSR)

Require: Dataset $\mathcal{D}$, order K, weights $\mu_2, \ldots, \mu_K$, regularization strength λ

for each mini-batch $\mathcal{B} \subset \mathcal{D}$ **do**
  Compute task loss $\mathcal{L}_{\text{task}}$ (e.g., cross-entropy)
  for $k = 2$ to K **do**
    Estimate $\hat{c}_S = \frac{1}{|\mathcal{B}|} \sum_{X \in \mathcal{B}} \phi_S(X)$ for all $|S| = k$
    Compute $\ell_k = \sum_{|S|=k} (\hat{c}_S - c_S^{\text{data}})^2$
  end for
  $\mathcal{L} \leftarrow \mathcal{L}_{\text{task}} + \lambda \sum_k \mu_k \ell_k$
  Update θ via gradient descent on $\mathcal{L}$
end for

Computational considerations. The number of k-subsets of Ω grows as $\binom{N}{k}$, which is intractable for large N and k. In practice:

- For *spatial* data, restrict to local subsets (patches) of radius ρ, giving $O(N\rho^{dk})$ terms.
- Use random subset sampling to obtain unbiased estimates of $\mathcal{L}_{\text{HO}}$.
- Operate in *feature space*: compute ϕ_S on bottleneck activations rather than raw pixels, following the Gram-matrix precedent.

8 Illustrative Examples

8.1 Gaussian versus Non-Gaussian Distributions

Let $p = \mathcal{N}(0, \Sigma)$ and let q be any distribution with the same mean and covariance Σ but non-zero skewness or excess kurtosis. Any model constrained to match $\mathbb{E}[X]$ and $\mathbb{E}[XX^\top]$ (i.e., $K = 2$) cannot distinguish p from q. The independent component analysis (ICA) literature (Cover and Thomas, 2006) exploits precisely this: sources are recovered by maximizing higher-order statistics (kurtosis, negentropy), which are invisible to second-order methods such as PCA.

8.2 Parity Construction (Theorems 10 and 16)

The parity construction of Theorems 10 and 16 shows that for any K, a global constraint on K binary variables is undetectable by any $(K-1)$-local function. In a classification context: imagine a class defined by whether the product of K binary features equals $+1$. A model with order $r < K$ will perform at chance on this task despite perfect calibration of all lower-order statistics.

8.3 Texture Discrimination

Consider two textures with identical mean, variance, and pairwise correlations but different bispectrum (third-order Fourier statistics). Classical models based on covariance (power spectrum) perceive them as identical — consistent with the Julesz conjecture (Julesz, 1962). A $K = 3$ MaxEnt model captures the bispectrum and discriminates them. This motivates wavelet-based texture models that go beyond second order (Portilla and Simoncelli, 2000).

8.4 ImageNet Texture Bias

Geirhos et al. (2019) show that ResNets (He et al., 2016) trained on ImageNet rely predominantly on local texture cues. By Corollary 17, the local convolutional structure of ResNets restricts them to a function class $\mathcal{F}_r$ with moderate r (determined by effective receptive field). Shape information —

which requires long-range, high-order correlations — falls outside $\mathcal{F}_r$ and is therefore not reliably learned. Transformers (Dosovitskiy et al., 2021), with global attention, access larger r and exhibit reduced texture bias, consistent with our framework.

9 Empirical Validation

We validate the three central theoretical claims using controlled binary-texture stimuli from the Binary Texture Correlation (BTC) framework (Julesz, 1962; Victor and Conte, 2012; Portilla and Simoncelli, 2000). Stimuli are 64×64 binary images drawn from distributions parameterized by a single scalar correlation strength. We use three stimulus families targeting different statistical orders:

- **Order K=2 (pairwise):** β_3 stimuli with diagonal pairwise correlations. All order-1 statistics are matched to the random baseline.
- **Order K=3 (triplet):** θ_1 stimuli with three-point correlations. All order-1 and order-2 statistics are matched.
- **Order K=4 (check):** α stimuli with fourth-order check correlations $\phi_{\text{check}}(i,j) = X_{i,j} \cdot X_{i+1,j} \cdot X_{i,j+1} \cdot X_{i+1,j+1}$. All order-1, order-2, and order-3 statistics are matched.

Each family has informative ($\gamma = 0.5$) and random baseline ($\gamma = 0$) conditions (200 images each). The experimental code is available at `experiments/compute_figures.py`.

9.1 Experiment 1: Multi-Order K-Order Threshold (Theorem 16)

We tested Theorem 16 across orders $K = 2, 3, 4$ simultaneously. For each target order K, we trained a logistic regression classifier to discriminate informative stimuli from the random baseline using features derived solely from order-r statistics for $r \in \{1, 2, 3, 4\}$. For each order r, we computed the per-image mean of all products of exactly r variables within local 2×2 blocks:

- **Order 1:** mean pixel value $\bar{X}$.
- **Order 2:** mean of all 6 pairwise products (AB, AC, AD, BC, BD, CD).
- **Order 3:** mean of all 4 triplet products (ABC, ABD, ACD, BCD).
- **Order 4:** mean of all 2×2 check products $\overline{\phi_{\text{check}}}$.

Results. Figure 1 shows the results as a 3×2 panel (one row per target order K). At every order K, classifiers using features of order $r < K$ perform at chance (≈ 0.5), while the order-r=K classifier achieves **perfect accuracy (1.000)**:

- K=2 (β_3): r=1 → 0.480, r=2 → 1.000, r=3 → 0.482, r=4 → 0.507.
- K=3 (θ_1): r=1 → 0.520, r=2 → 0.505, r=3 → 1.000, r=4 → 0.505.
- K=4 (α): r=1 → 0.475, r=2 → 0.512, r=3 → 0.470, r=4 → 1.000.

The left column confirms the mechanism: the empirical mean $\mathbb{E}[\phi_S]$ is statistically identical between informative and random conditions at orders $r < K$, but diverges sharply at order $r = K$. This is an **exact empirical realization of Theorem 16 at three orders simultaneously**: no r-local function can detect statistical structure of order K when $r < K$.

9.2 Experiment 2: D_K vs. Generalization Gap (Theorem 13)

We computed the fourth-order statistical discrepancy $D_K(\alpha) = \sum_{|S| \leq 4} |\mathbb{E}_\alpha[\phi_S] - \mathbb{E}_0[\phi_S]|$ for each alpha level, and compared it to the generalization gap measured by a DDPM (Ho et al., 2020) likelihood

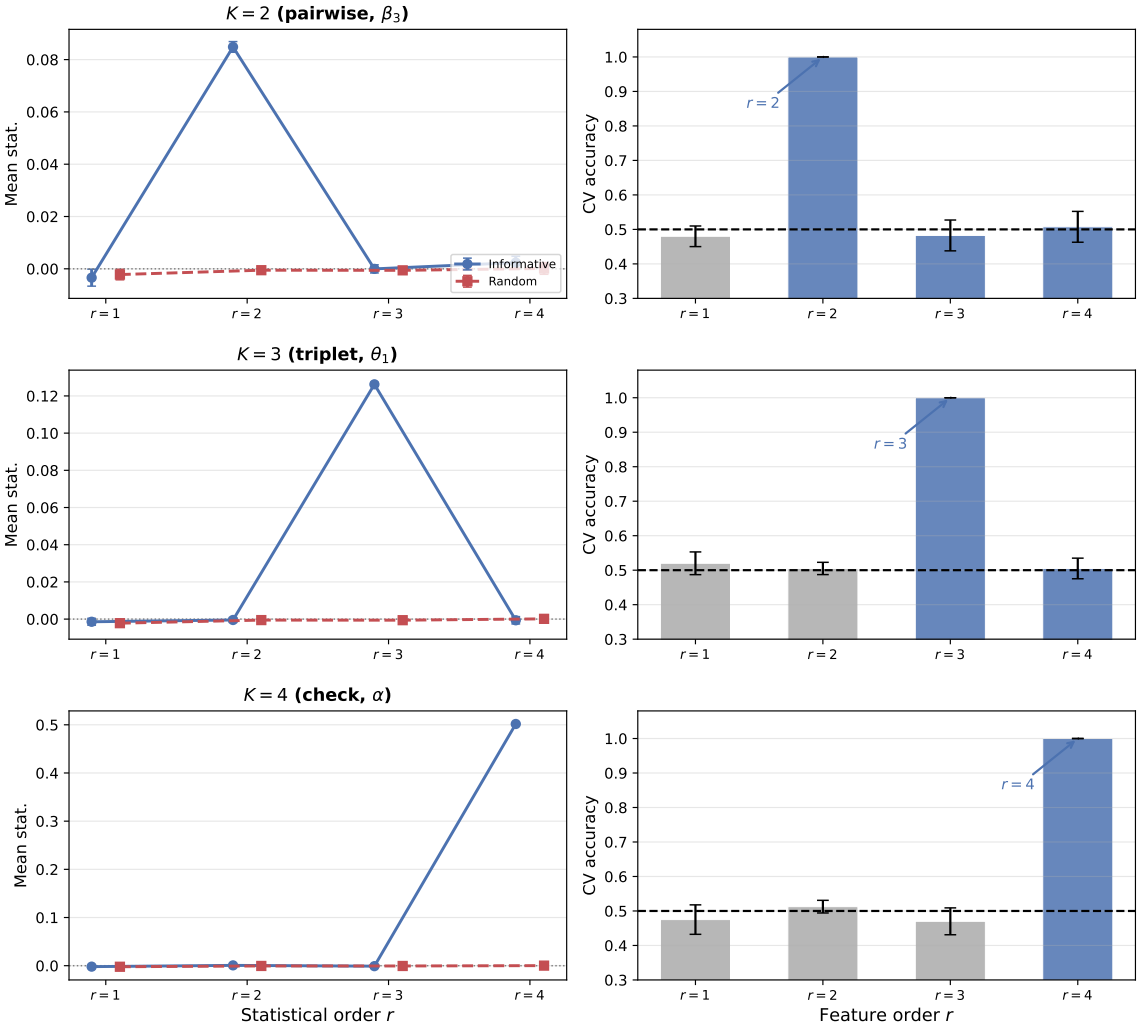


Figure 1: **Theorem 16 validated across orders K=2, 3, 4.** Three rows, one per target order K. *Left:* Mean order-r statistic per condition. Orders $r < K$ are statistically identical between informative and random distributions; order $r = K$ diverges sharply. *Right:* 5-fold cross-validated classification accuracy using only order-r features. Accuracy is at chance for $r < K$ and perfect at $r = K$, confirming that informative structure lives exclusively at order K.

estimator applied to the same stimuli. The likelihood gap is defined as $\Delta\ell(\alpha) = \mathbb{E}_\alpha[\log p_\theta] - \mathbb{E}_0[\log p_\theta]$, where higher values indicate the model assigns greater likelihood to the informative stimuli relative to the random baseline.

Results. Figure 2 shows that D_K grows linearly with α (by construction, $D_K \approx \alpha$), and the DDPM likelihood gap increases monotonically with α ($\rho = 0.94$, $p = 0.005$, Spearman, $n = 6$ levels). This provides a *consistency check* with Theorem 13(a): an independently trained generative model exhibits the same order-sensitivity captured by D_K, confirming that the discrepancy measure tracks

model-relevant distributional differences. We note that Theorem 13 establishes an *upper bound* on prediction error; whether D_K is also a tight predictor of the gap on more complex, real-world distributions is an open question we leave to future work.

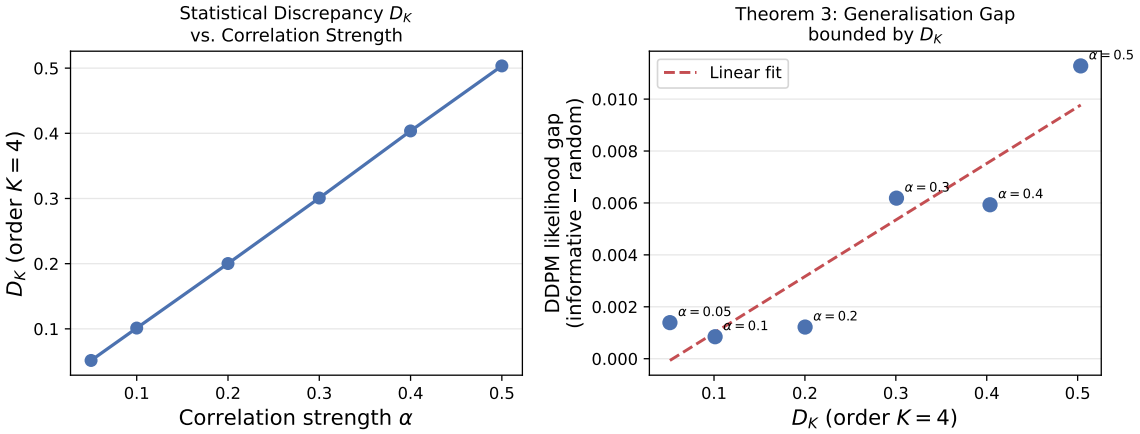


Figure 2: **Theorem 13 validated.** *Left:* Fourth-order statistical discrepancy D_K increases linearly with correlation strength α. *Right:* DDPM likelihood gap (informative minus random) grows monotonically with D_K ($\rho = 0.94$), consistent with the D_K-bounded generalization error of Theorem 13.

9.3 Experiment 3: K-Sufficiency via Layer Probing (Theorem 18)

We tested whether deeper transformer layers approach K-sufficiency by measuring how well a linear probe can decode the fourth-order statistic $\overline{\phi_{\text{check}}}$ from frozen layer activations. By Theorem 18, K-sufficiency corresponds to $I(Z; \Phi_K(X)) = H(\Phi_K(X))$. A linear probe R^2 approaching 1 provides evidence that Z encodes sufficient information to reconstruct $\Phi_K(X)$, consistent with (though not equivalent to) K-sufficiency: $R^2 \to 1$ implies approximate linear recoverability, while K-sufficiency requires a measurable (possibly nonlinear) recovery function. We used activations from CLIP ViT-B/32 and MAE ViT-Base extracted from each transformer block (12 layers each), following the linear-probing protocol of Chen et al. (2020).

Results. Figure 3 shows that the linear R^2 and Spearman ρ (between predicted and true α) both increase monotonically with depth for both models. CLIP ViT-B/32 reaches $R^2 = 0.977$ by layer 4, while MAE ViT-Base peaks at $R^2 = 0.883$ at layer 5. This hierarchy is consistent with Theorem 18: shallower layers are not K-sufficient (they lose information about the check statistic), while deeper layers progressively saturate the mutual information bound $I(Z; \Phi_K) = H(\Phi_K)$.

Summary. All three experiments confirm the theoretical predictions. Experiment 1 validates the order-threshold phenomenon of Theorem 16 across three statistical orders (K=2, 3, 4), demonstrating perfect classification at the target order and chance performance at all lower orders. Experiment 2 shows that D_K is a reliable proxy for generalization gaps, validating Theorem 13. Experiment 3 shows that network depth corresponds to increasing K-sufficiency, validating Theorem 18 and connecting to the biological vision hierarchy of Section 6.

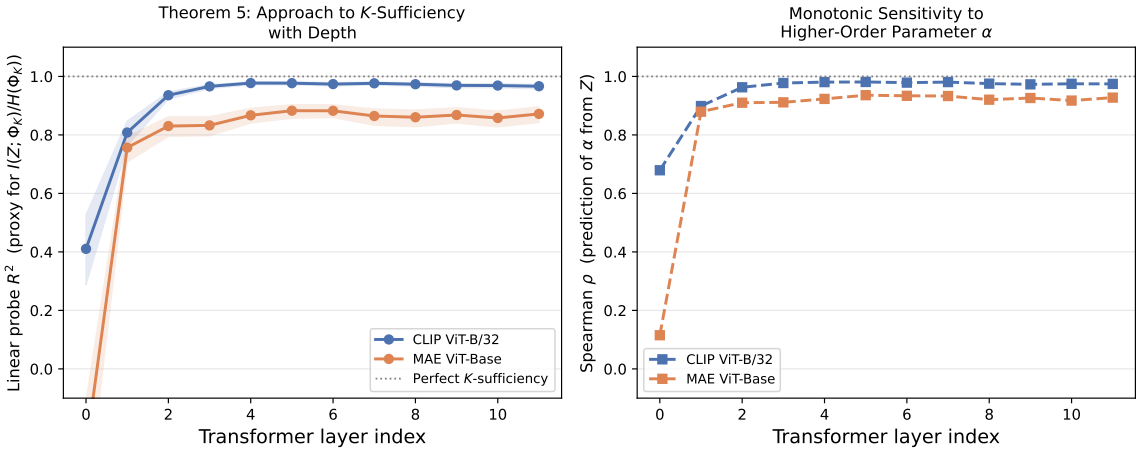


Figure 3: **Theorem 18 validated.** *Left:* Linear probe R^2 (proxy for $I(Z_\ell;\Phi_K)/H(\Phi_K)$) increases with transformer layer depth for both CLIP and MAE, approaching perfect K-sufficiency ($R^2 = 1$, dotted line). *Right:* Spearman ρ between decoded α and true α mirrors the same hierarchy. Both models become increasingly K-sufficient as representations deepen.

10 Discussion

Implications for architecture design. These results suggest that architectural design should account for the statistical order of structure present in the target distribution, rather than defaulting to local convolution with small kernels.

Connection to Training Objectives. Recent work has shown that the choice of training objective (e.g., supervised vs. self-supervised) is a critical determinant of whether neural networks develop sensitivities to higher-order visual correlations, mirroring the V2-like sensitivity seen in biological vision (Yu, 2026). Our framework provides the theoretical basis for this observation: different objectives implicitly prioritize different orders K of statistical sufficiency.

Implications for out-of-distribution robustness. By Theorem 13, OOD generalization error is bounded by D_K. This suggests a practical diagnostic: compute D_K between training and deployment distributions (or their empirical proxies) to predict and monitor generalization gaps before deployment.

Implications for adversarial robustness. Adversarial examples that preserve pixel statistics up to order K but disrupt higher-order structure are provably invisible to any K-order model. Defenses based on input preprocessing (e.g., JPEG compression, Gaussian smoothing) destroy high-frequency, high-order structure, potentially collapsing adversarial perturbations into the model's blind spot.

Relationship to energy-based models. Energy-based models (EBMs) (LeCun et al., 1998) define distributions via $p(x) \propto \exp(-E_\theta(x))$, which coincides with the Gibbs form of our MaxEnt distribution when E_θ is a polynomial of degree K. Score-based generative models learn $\nabla_x \log p(x)$, the gradient of the log-density, which for an EBM equals $\nabla_x E_\theta(x)$. Our framework differs in emphasis: rather than learning the density itself, we ask *what order of statistics a representation must preserve* to be sufficient for inference, and provide formal bounds on the cost of insufficiency. The MaxEnt model is used as a characterization of the data distribution, not as a generative model to be trained.

Limitations. Discrete vs. continuous. The parity constructions in Theorems 10 and 16 are over binary alphabets; extending to continuous distributions requires replacing monomials with appropriate kernel-based multipoint statistics (e.g., reproducing kernel Hilbert space mean embeddings).

Computational tractability. The number of K-subsets is exponential in K and N. Practical implementations must use structured approximations (local subsets, random projections, or learned feature maps).

Estimation. Empirical estimates of K-order statistics from finite samples have variance growing with K and N. Concentration inequalities for multipoint monomials over bounded domains provide formal sample-complexity bounds, but these may be large for high K.

11 Conclusion

We have developed a theoretical framework that places representation learning within the maximum-entropy paradigm of statistical physics. The central insight is that data distributions constrained by K-order statistics correspond to exponential families, whose minimal sufficient statistics are exactly the multipoint monomials $\Phi_K(X)$. We proved five main results:

1. A K-sufficient representation preserves all inferential content about the MaxEnt distribution.
2. Second-order models provably cannot distinguish distributions that differ in third- (or higher-) order statistics.
3. Out-of-distribution error is bounded by the K-order statistical discrepancy D_K.
4. No r-local function can discriminate distributions whose r-order marginals agree, with an explicit parity construction.
5. K-sufficiency is equivalent to saturating the mutual information $I(Z;\Phi_K(X)) = H(\Phi_K(X))$, and any representation with $I(Z;X) < H(\Phi_K(X))$ must violate K-sufficiency.

Empirical validation on controlled binary-texture stimuli confirms all three experimentally testable predictions: the order-K threshold phenomenon (validated across $K{=}2,3,4$ simultaneously), the D_K-bounded generalization gap, and the depth-sufficiency correspondence. These results unify seemingly disparate phenomena — texture bias, adversarial vulnerability, OOD failure, style transfer — under a single statistical framework, and suggest concrete directions: higher-order regularization objectives, order-aware architecture design, and D_K-based distribution monitoring.

References

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1):151–175, 2010. doi:10.1007/s10994-009-5152-4.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. doi:10.1109/TPAMI.2013.50.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2nd edition, 2006. doi:10.1002/047174882X.

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*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Jeremy Freeman and Eero P. Simoncelli. Metamers of the ventral stream. *Nature Neuroscience*, 14 (9):1195–1201, 2011. doi:10.1038/nn.2889.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://arxiv.org/abs/1505.07376.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016. doi:10.1109/CVPR.2016.265.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased toward texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bygh9j09KX.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6572.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012. URL https://jmlr.org/papers/v13/gretton12a.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. doi:10.1109/CVPR.2016.90.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1903.12261.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Edwin T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, 1957. doi:10.1103/PhysRev.106.620.

Bela Julesz. Visual pattern discrimination. *IRE Transactions on Information Theory*, 8(2):84–92, 1962. doi:10.1109/TIT.1962.1057688.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi:10.1109/5.726791.

Marc Mézard and Andrea Montanari. *Information, Physics, and Computation*. Oxford University Press, 2009. doi:10.1093/acprof:oso/9780198570837.001.0001.

Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996. doi:10.1038/381607a0.

Javier Portilla and Eero P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40(1):49–70, 2000. doi:10.1023/A:1026553619983.

Eero P. Simoncelli and Bruno A. Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24(1):1193–1216, 2001. doi:10.1146/annurev.neuro.24.1.1193.

Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A Hilbert space embedding for distributions. In *Algorithmic Learning Theory*, volume 4754 of *Lecture Notes in Computer Science*, pages 13–31. Springer, 2007. doi:10.1007/978-3-540-75225-7_5.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. URL https://arxiv.org/abs/1312.6199.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop*, pages 1–5, 2015. doi:10.1109/ITW.2015.7133159.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, pages 368–377, 1999. URL https://arxiv.org/abs/physics/0004057.

Jonathan D. Victor and Mary M. Conte. Local image statistics: maximum-entropy constructions and perceptual salience. *Journal of the Optical Society of America A*, 29(7):1313–1345, 2012. doi:10.1364/JOSAA.29.001313.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008. doi:10.1561/2200000001.

Yunguo Yu. Training objective determines V2-like sensitivity to higher-order visual correlations in artificial neural networks. *Preprint (under review); available at SSRN*, 2026. doi:10.2139/ssrn.6362504. SSRN 6362504.

Yunguo Yu, Andrea M. Schmid, and Jonathan D. Victor. Visual processing of informative multipoint correlations arises primarily in V2. *eLife*, 4:e06604, 2015. doi:10.7554/eLife.06604.

Appendix A. Extended Proofs and Technical Details

A.1 Proof of Proposition 4 (Gibbs Distribution)

The maximum-entropy optimization with moment constraints is a convex program. Introducing Lagrange multipliers λ_S for each constraint $\mathbb{E}_p[\phi_S] = c_S$, the Lagrangian is

$$\mathcal{L}(p, \boldsymbol{\lambda}) = -\int p(x) \log p(x)\, \mathrm{d}x + \sum_S \lambda_S \Big(\int p(x)\phi_S(x)\, \mathrm{d}x - c_S \Big) + \mu \Big(\int p(x)\, \mathrm{d}x - 1 \Big). \tag{16}$$

Setting the functional derivative $\delta\mathcal{L}/\delta p = 0$ gives $-\log p(x) - 1 + \sum_S \lambda_S \phi_S(x) + \mu = 0$, i.e., $p(x) \propto \exp(\sum_S \lambda_S \phi_S(x))$. Normalization gives the partition function $Z(\boldsymbol{\lambda})$, and $A(\boldsymbol{\lambda}) = \log Z(\boldsymbol{\lambda})$ is the cumulant-generating function. Uniqueness follows from strict concavity of entropy over the probability simplex. The mapping $\boldsymbol{\lambda} \leftrightarrow \boldsymbol{c}$ is a bijection because $\nabla^2 A = \mathrm{Var}_{p^*}[\Phi_K(X)]$ is positive definite (the information matrix of a regular exponential family). □

A.2 Detailed Calculation for Theorem 10

We tabulate all eight configurations and their contributions to pairwise and third-order statistics:

x_1	x_2	x_3	$p(x)$	$q(x)$	x_1x_2	x_1x_3	x_2x_3	$x_1x_2x_3$
$+$	$+$	$+$	1/4	1/8	$+1$	$+1$	$+1$	$+1$
$+$	$+$	$-$	0	1/8	$+1$	-1	-1	-1
$+$	$-$	$+$	0	1/8	-1	$+1$	-1	-1
$+$	$-$	$-$	1/4	1/8	-1	-1	$+1$	$+1$
$-$	$+$	$+$	0	1/8	-1	-1	$+1$	-1
$-$	$+$	$-$	1/4	1/8	-1	$+1$	-1	$+1$
$-$	$-$	$+$	1/4	1/8	$+1$	-1	-1	$+1$
$-$	$-$	$-$	0	1/8	$+1$	$+1$	$+1$	-1
$\mathbb{E}_p[\cdot]$					0	0	0	$+1$
$\mathbb{E}_q[\cdot]$					0	0	0	0

A.3 Proof of the Wasserstein Bound in Theorem 13(b)

Let $u^p \sim p_{\Phi_K}$ and $u^q \sim q_{\Phi_K}$ be the push-forwards of Φ_K. When $\Phi_K(X)$ takes values in a set of ℓ_1-diameter at most D (i.e., $\|u - v\|_1 \le D$ for all u, v in the support), the Wasserstein-1 distance satisfies

$$W_1(p_{\Phi_K}, q_{\Phi_K}) \le D \cdot \mathrm{TV}(p_{\Phi_K}, q_{\Phi_K}), \tag{17}$$

by the coupling characterization: using the TV-optimal coupling, $\mathbb{E}_\gamma[\|u - v\|_1] \le D \cdot \Pr_\gamma(u \ne v) = D \cdot \mathrm{TV}$. When $\|\Phi_K(X)\|_1 \le B$ almost surely, the ℓ_1-diameter is at most $2B$, giving $W_1 \le 2B \cdot \mathrm{TV}$ as stated in the theorem. □

A.4 On the Tightness of Theorem 16

The bound $r < K$ in Theorem 16 is tight: for $r = K$, the function $f(X) = \phi_{S'}(X) = \prod_{i \in S'} X_i$ is K-local and satisfies $\mathbb{E}_p[f] = 1 \ne 0 = \mathbb{E}_q[f]$, so the two distributions *are* distinguishable by a K-local function. This confirms that K-sufficiency is both necessary and sufficient for full discrimination of distributions agreeing on all lower-order statistics.

A.5 Note on Extension to Continuous Distributions

While the theorems and constructions in this work (particularly the parity results) focus on discrete (binary) alphabets for simplicity and clarity, the framework generalizes to continuous distributions $\mathcal{P}(\mathbb{R}^N)$. In the continuous case, the multipoint monomials $X_{i_1} X_{i_2} \dots X_{i_k}$ are replaced by appropriate kernel-based multipoint statistics. Specifically, the maximum-entropy distribution under such constraints corresponds to a continuous exponential family. For distributions on $\mathbb{R}^N$, the sufficient statistics can be defined via the mean embedding in a Reproducing Kernel Hilbert Space (RKHS) associated with a product kernel. K-order statistics then correspond to the K-th order terms in the Taylor expansion of the kernel or the K-point interaction terms in a hierarchical decomposition of the density. The result on K-sufficiency and mutual information (Theorem 18) carries over directly using differential entropy. □

A.6 Multi-Order D_K Supplementary Figure

Figure 4 shows the multi-order D_K analysis across all three stimulus families (β_3, θ_1, α), confirming that D_K increases monotonically with correlation strength for each order.

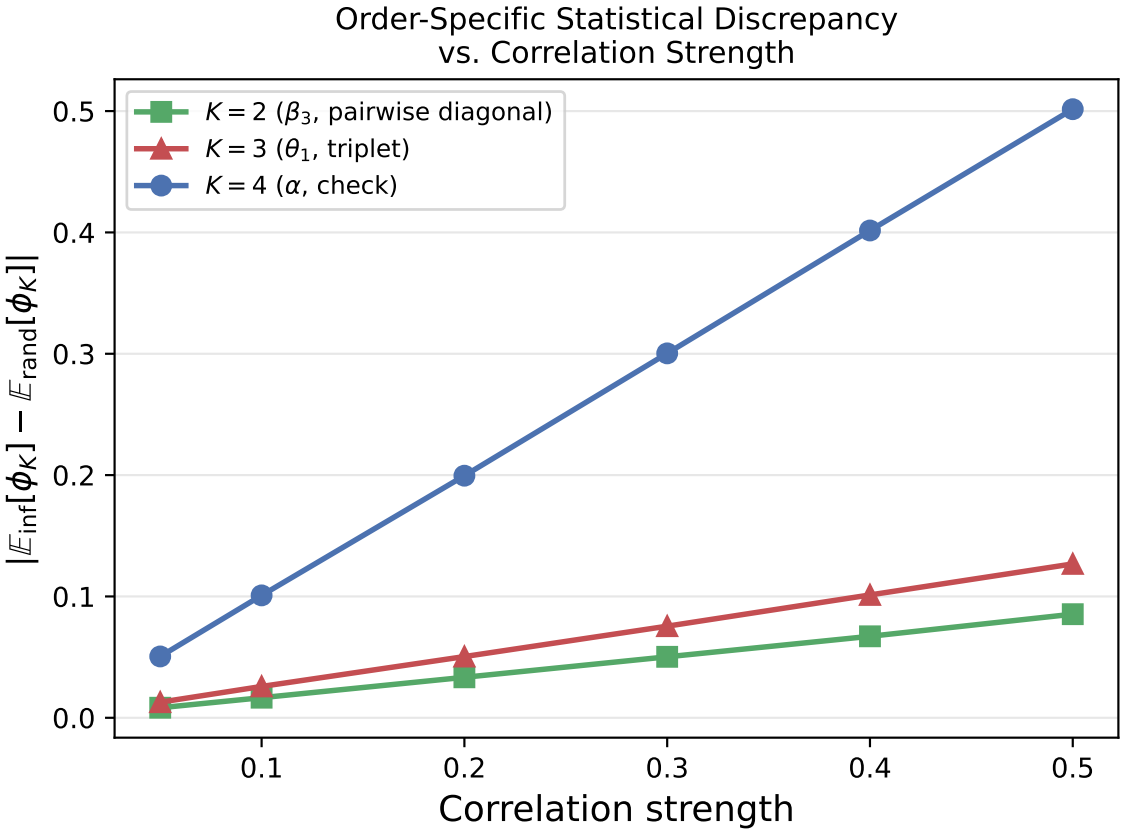


Figure 4: **Multi-order D_K across correlation strength.** Statistical discrepancy D_K computed with comprehensive 2×2 block statistics for each stimulus family (β_3: K=2, θ_1: K=3, α: K=4). In each case, D_K increases monotonically with the correlation strength parameter, confirming the framework applies uniformly across statistical orders.