VARIABLE BINDING IS NOT A PROPERTY OF REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Variable binding is the capacity of a network to hold each attribute of an object as a separable part of the object, so that attributes and objects can be recombined into novel arrangements. The field has long believed that binding is an intrinsic property of a representation and thus sought to measure and certify it. However, we demonstrate it is not such a property. We first prove that the extent to which an object's attributes can be separated is determined by the observation. This is done before any encoder is trained, so no optimizer or architecture can improve the separability. If the objects in an observation are separable, a representation can bind them, but it cannot do so when they are not. We then show that the leakage metric the field currently uses to certify that a network binds changes from zero to positive under a lossless relabeling, and thus no architecture-agnostic certificate can exist. Across seven architecture families, including the tensor-product representation, an information-preserving relabel changes each model's binding verdict, showing that variable binding is not an intrinsic property. Object entanglement also predicts the out-of-distribution gap on real photographs, and reducing this entanglement closes the binding gap.

1 INTRODUCTION

Variable binding is the capacity of a neural network to associate variables with a specific entity in a given context. Consider a scenario with colored cars and apples. During training, the neural network sees blue cars with red apples, red cars with blue apples and a variety of combinations of colors on the cars and apples. Assuming the network is queried for a red car and it returns an output such as a blank car or even a red car and a red apple with a very high confidence, it has failed to bind the color to the car. This is a general failure of models and it is such a failure that interests us. Models interpolate easily inside the combinations present in their training data and break on combinations that are individually familiar and jointly new. This is how a system can saturate a benchmark and fail on a trivially recombined version of it (Lake & Baroni, 2018; Kim & Linzen, 2020; Hupkes et al., 2020).

This ability to recombine familiar parts is central to the systematicity question raised by Fodor & Pylyshyn (1988). They argued that connectionist architectures have no principled mechanism to maintain bindings between attributes and objects and therefore no architectural guarantee that such bindings could be produced, unlike symbolic systems that bind through constituent structure. Researchers in the decades that have followed have investigated mechanisms that enable neural representations to bind (Smolensky, 1990; Plate, 1995; Von Der Malsburg, 1994; Hummel & Holyoak, 2003; Greff et al., 2020; Miyato et al., 2025). These investigations have often sought a measurable property of a representation that could also serve as a training objective and certify that the representation binds (Touska et al., 2026; Dhayalkar, 2025; Kori et al., 2024; Gopalakrishnan et al., 2024; Webb et al., 2024; 2020). The general belief has always been that if binding could be measured and thus induced, then compositional generalization would become engineerable rather than hoped for as a byproduct of training.

In this paper, we show that this search has been aimed at the wrong object. Binding is not an intrinsic property that a representation possesses, but rather a relation between a representation and the models that process it. Call the objects in a scene (i.e., *the car* and *the apple*) the roles, and their attributes (*the colours*) the fillers. We define the binding excess of a representation as the information one

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

role's representation carries about the other roles' fillers once the roles themselves are known, i.e., how much the car's representation says about the apple's colour beyond what the roles already say. Then, we show that the data determines the dependence that remains between the fillers after the whole representation is known and it's measured by a role-conditioned total correlation. Every representation that preserves all filler information inherits the entanglement of the observation and any other representation can carry less only by discarding at least as much joint filler information. We call this the *binding limit* and no architecture or optimizer can reduce it without losing information. We then construct an invertible information-preserving relabeling that mixes the role representations and raises the measured leakage on a representation from zero. Any reader of the original representation can recover it exactly from the relabeled one. Also, when a representation retains independent noise, an invertible relabeling can drive the measured leakage to zero while the representation remains entangled. This shows that binding cannot be an intrinsic property of a representation alone. It must instead be defined relative to the information available in the observation and the computations that can be performed on the resulting representation.

Contributions. We make several contributions;

1. **A binding limit:** We prove that every jointly sufficient representation has the same role-conditioned total correlation as the observation, and that any other representation can reduce it only by discarding joint filler information (Lemma 3.1, Corollary 3.2, Proposition B.7, Theorem B.8).
2. We prove that a lossless relabeling can raise role-local leakage from zero or lower it to zero on an entangled representation, even under role supervision (Theorems 4.3, B.11, and B.13).
3. Across language, vision–language, state-space, object-centric, oscillator, tensor-product, memory, and text-to-image models that bind, a certified information-preserving relabel flips the native binding readout far more often than matched random relabels.
4. **Practical tools:** We introduce the role-conditioned total correlation $\mathrm{TC}(\phi \mid S, r)$, the entanglement modulus η that bounds the optimal readout's transfer gap (Theorem B.16, Proposition B.17), a causal swap procedure, and a library that audits whether interpretability findings survive reparametrization.

Related Work. Recent work in mechanistic interpretability has identified binding subspaces and circuits responsible for binding in language models (Feng & Steinhardt, 2024; Dai et al., 2024; Prakash et al., 2024; 2025), with more recent work finding that binding arises from multiple positional and lexical mechanisms (Gur-Arieh et al., 2025) rather than localized to a single subspace. These results identify where and how binding is represented and read out in a particular representation, but leave open whether binding is an intrinsic property of a representation. We show that it is not because binding can change under an information-preserving reparameterization, making mechanistic binding claims necessarily relative to a particular coordinatization of the representation.

Object-centric identifiability gives sufficient conditions under which a slot map provably reconstructs every in-support filler (Locatello et al., 2020b; Brady et al., 2023; Wiedemer et al., 2023; 2024). These results guarantee zero *risk* on the combinations represented in the training support. But compositional generalization concerns precisely the combinations outside that support, and identifiability alone provides no guarantee about this *gap*. We decouple this gap from reconstruction risk and prove that, even with a perfect slot map, the role-conditioned total correlation cannot be reduced below the entanglement of the rendering.

2 Setup

We work on a common standard Borel probability space so that all conditional distributions below are well defined. We use the convention $0 \log 0 = 0$ and define mutual information as the divergence between a joint law and the product of its marginals. It is nonnegative and equals zero if and only if the corresponding variables are independent (Cover, 1999).

Let a finite set of roles (e.g., *car* and *apple*) and a Polish space of fillers (e.g., *blue*, *red*) be denoted as $\mathcal{R}$ and $\mathcal{F}$ respectively, with an arity k. A latent structure is a pair $(\boldsymbol{r}, \boldsymbol{\phi}) \in \mathcal{R}^k \times \mathcal{F}^k$, where ϕ_i is the filler bound to role r_i. A rendering kernel $\rho : \mathcal{R}^k \times \mathcal{F}^k \to \mathcal{P}(\mathcal{X})$ maps a latent structure to a distribution over observations with $X \sim \rho(\boldsymbol{r}, \boldsymbol{\phi})$. ρ is fixed across training and test, so distribution shift can arise only through the joint distribution of fillers. The rendering may be deterministic, as for symbolic inputs, or stochastic, as for photographs.

Assumption 2.1 (Role distinctness)**.** Every realization of $\boldsymbol{r}$ in the support has pairwise distinct components.

Assumption 2.2 (Per-role factorization under training)**.** $\mathbb{P}_{\mathrm{tr}}(\phi_1, \ldots, \phi_k \mid \boldsymbol{r}) = \prod_{i=1}^{k} \mathbb{P}_{\mathrm{tr}}(\phi_i \mid r_i)$.

Assumption 2.3 (Role identification)**.** The representation identifies each role exactly and independently of its position, so that $\hat{r}_i = \iota(r_i)$ almost surely for a measurable injection $\iota : \mathcal{R} \to \mathbb{R}^{d_R}$.

Exact role recovery (Assumption 2.3) is an idealization that can be relaxed. If roles are recovered only ε-accurately, our results degrade by a bound that vanishes with ε (Proposition B.14). The per-role factorization in Assumption 2.2 describes the clean compositional setting, but it does not hold on benchmarks that explicitly remove held-out combinations since each removal creates a structural zero in the training joint. The resulting distribution is no longer the product of its per-role marginals (Lake & Baroni, 2018; Kim & Linzen, 2020; Keysers et al., 2019). Theorem B.16 treats this setting without requiring Assumption 2.2.

This factorized setting captures the basic strict-substitutive shift we study. During training, colours are drawn independently for each object, so every filler occurs with every role and the joint distribution factorizes across roles. At test time, every per-role marginal remains unchanged while the joint law of the fillers changes. For example, the car and apple colours may become coupled in a combination absent from training. Held-out-combination benchmarks are more general because their training laws contain structural zeros. Definition 2.4 covers both cases.

Definition 2.4 (Strict-substitutive shift)**.** A test law $\mathbb{P}_{\mathrm{te}}$ is *strict-substitutive* relative to $\mathbb{P}_{\mathrm{tr}}$ if it preserves the role distribution (S1), the per-role filler marginals and supports (S2), and the rendering kernel (S3), while allowing the joint law of fillers given the roles to change otherwise freely (S4). The *factorization-preserving* subclass additionally requires the test filler law given roles to remain the product of its per-role marginals (S5).

We call the consumer of a representation its *reader*. A linear probe, a nonlinear decoder and a downstream head are all readers and a reader class is a family of such functions. Binding excess is necessarily defined relative to a reader, because how much one slot reveals about another slot's filler depends how much computation the reader has. The least powerful reader we consider reads each slot on its own coordinates and the most powerful, the unrestricted reader, may apply any measurable function of the whole representation.

A representation is binding-compatible when each slot keeps the information about its own filler, while retaining no information about the other fillers. We define these two requirements as sufficiency and minimality.

Definition 2.5 (Binding-compatibility and binding excess)**.** An encoder with role and filler readouts is binding-compatible relative to $\mathbb{P}_{\mathrm{tr}}$ if, for every i, it satisfies per-role sufficiency, $I(\hat{\phi}_i; \phi_i \mid \hat{r}_i) = I(X; \phi_i \mid \hat{r}_i)$, and per-role minimality, $I(\hat{\phi}_i; \phi_{-i} \mid \hat{\boldsymbol{r}}) = 0$, where $\phi_{-i} = (\phi_l)_{l \neq i}$. We define the local binding excess at slot i by $\beta_i := I(\hat{\phi}_i; \phi_{-i} \mid \hat{\boldsymbol{r}}) \in [0, \infty]$. Thus, β_i measures how much information about the other fillers remains in slot i once the roles are known.

Definition 2.6 (Joint sufficiency and joint minimality)**.** A code $S = E(X)$ is *per-filler sufficient* when $I(S; \phi_i \mid \boldsymbol{r}) = I(X; \phi_i \mid \boldsymbol{r})$ for every i, and *jointly sufficient* when $I(S; \boldsymbol{\phi} \mid \boldsymbol{r}) = I(X; \boldsymbol{\phi} \mid \boldsymbol{r})$. For a slot factorization, define the conditional excess $\widetilde{\beta}_i := I(\hat{\phi}_i; \boldsymbol{\phi}_{-i} \mid \hat{\boldsymbol{r}}, \phi_i)$. The factorization is *jointly minimal* when $\mathrm{Law}(\hat{\boldsymbol{\phi}} \mid \boldsymbol{r}, \boldsymbol{\phi}) = \bigotimes_i \kappa_i(\cdot \mid \hat{\boldsymbol{r}}, \phi_i)$ throughout the product of the per-role supports. Under Assumption 2.2, joint minimality implies $\widetilde{\beta}_i = 0$, and $\widetilde{\beta}_i = 0$ implies $\beta_i = 0$.

3 The Binding Limit

We first ask how much entanglement can remain even when a representation preserves all the information needed to recover each filler. Let $S = E(X)$ be per-filler sufficient, such that $I(S; \phi_i \mid \boldsymbol{r}) = I(X; \phi_i \mid \boldsymbol{r})$ for every i. We measure the residual dependence among fillers using the role-conditioned total correlation,

$$\mathrm{TC}(\boldsymbol{\phi} \mid S, \hat{\boldsymbol{r}}) = \sum_i H(\phi_i \mid S, \hat{\boldsymbol{r}}) - H(\boldsymbol{\phi} \mid S, \hat{\boldsymbol{r}}),$$

which we introduce as the role-conditioned version of Watanabe's total correlation (Watanabe, 1960). Conditioning on the roles is important because it removes the dependence due to uncertainty

about which role occupies which slot. What remains then, is dependence between fillers given the representation, and importantly, this quantity is invariant to how the roles are represented. We prove this invariance in Section 4.

For a per-filler sufficient code, the rate above the per-filler minimum decomposes as $I(S; X \mid \boldsymbol{r}) - \sum_i R_i^{\mathrm{suf}} = I(S; X \mid \boldsymbol{\phi}, \boldsymbol{r}) + \mathrm{TC}(\boldsymbol{\phi} \mid S, \boldsymbol{r})$ (Theorem B.4). Both terms are nonnegative. The first accounts for noise in the code, while the second is the residual dependence among fillers. For jointly sufficient codes, this residual dependence is determined entirely by the observation:

Lemma 3.1 (The binding limit)**.** *For any code $S = E(X)$, let $\delta_{\mathrm{J}}(S) := I(\boldsymbol{\phi}; X \mid S, \boldsymbol{r})$ and $\varepsilon_{\perp}(S) := \sum_i I(\phi_i; X \mid S, \boldsymbol{r})$. Let U be the tuple of marginal posteriors $P(\phi_i \in \cdot \mid X, \boldsymbol{r})$. Then*

$$\mathrm{TC}(\boldsymbol{\phi} \mid S, \boldsymbol{r}) + \delta_{\mathrm{J}}(S) = \mathrm{TC}(\boldsymbol{\phi} \mid X, \boldsymbol{r}) + \varepsilon_{\perp}(S).$$

Consequently, every jointly sufficient code satisfies $\mathrm{TC}(\boldsymbol{\phi} \mid S, \boldsymbol{r}) = \mathrm{TC}(\boldsymbol{\phi} \mid X, \boldsymbol{r})$. If S is per-filler sufficient, then U is per-filler sufficient and

$$\mathrm{TC}(\boldsymbol{\phi} \mid U, \boldsymbol{r}) \le \mathrm{TC}(\boldsymbol{\phi} \mid S, \boldsymbol{r}) = \mathrm{TC}(\boldsymbol{\phi} \mid X, \boldsymbol{r}) - \delta_{\mathrm{J}}(S) \le \mathrm{TC}(\boldsymbol{\phi} \mid X, \boldsymbol{r}).$$

Proof sketch. For every code Z, $\mathrm{TC}(\boldsymbol{\phi} \mid Z, \boldsymbol{r}) = \mathrm{TC}(\boldsymbol{\phi} \mid \boldsymbol{r}) + I(\boldsymbol{\phi}; Z \mid \boldsymbol{r}) - \sum_i I(\phi_i; Z \mid \boldsymbol{r})$. Subtracting the identity for $Z = S$ from the identity for $Z = X$ proves the first equality. Joint sufficiency makes both shortfalls zero. Under per-filler sufficiency, U is a function of $(S, \boldsymbol{r})$ and $\delta_{\mathrm{J}}(U) = \delta_{\mathrm{J}}(S) + I(\boldsymbol{\phi}; S \mid U, \boldsymbol{r})$, which proves the remaining inequalities. Appendix B.2 gives the full argument.

The residual dependence of a jointly sufficient code is therefore fixed by the rendering before any encoder is trained. More generally, $\mathrm{TC}(\boldsymbol{\phi} \mid S, \boldsymbol{r}) \ge \mathrm{TC}(\boldsymbol{\phi} \mid X, \boldsymbol{r}) - \delta_{\mathrm{J}}(S)$. A code can report less dependence than the observation carries only by discarding at least as much joint filler information.

The above also shows that two common routes to binding; rate compression which involves compressing a code toward minimal rate and slot-independent architecture which involve building an architecture that reads slots independently, are enforcing the same underlying condition on the rendering. For jointly sufficient codes, perfect binding requires $\mathrm{TC}(\boldsymbol{\phi} \mid X, \boldsymbol{r}) = 0$ and for per-filler sufficient codes, it requires $\mathrm{TC}(\boldsymbol{\phi} \mid U, \boldsymbol{r}) = 0$. The two routes differ in how they remove entanglement, not in what the rendering must permit.

Corollary 3.2 (The condition for perfect binding)**.** *A jointly sufficient code has $\mathrm{TC}(\boldsymbol{\phi} \mid S, \boldsymbol{r}) = 0$ if and only if $\mathrm{TC}(\boldsymbol{\phi} \mid X, \boldsymbol{r}) = 0$. A per-filler sufficient code with zero role-conditioned total correlation exists if and only if $\mathrm{TC}(\boldsymbol{\phi} \mid U, \boldsymbol{r}) = 0$, in which case U is such a code. Every code with $\mathrm{TC}(\boldsymbol{\phi} \mid S, \boldsymbol{r}) = 0$ satisfies $\delta_{\mathrm{J}}(S) \ge \mathrm{TC}(\boldsymbol{\phi} \mid X, \boldsymbol{r})$.*

The distinction between per-filler and joint sufficiency disappears in three common settings, namely, fixed-coupling finite models, jointly Gaussian models, and slotwise observation models whose conditional likelihood factorizes across roles (Proposition B.7). Joint minimality also requires zero role-conditioned total correlation (Theorem B.8), while a role-identifying slot-separable rendering satisfies this condition (Proposition B.10). Appendix B states and provides full proofs for all three results.

This condition is valid in the ideal setting but for a trained encoder, three caveats apply. First, training achieves sufficiency only approximately, and Lemma B.5 bounds the resulting deviation from the limit by the per-filler and joint sufficiency shortfalls. Second, zero binding excess does not imply correctness. A slot-equivariant encoder with an insufficient shared map has zero generalization gap but positive risk (Theorem B.19), which complements identifiability results that secure zero risk (Locatello et al., 2020b; Brady et al., 2023; Wiedemer et al., 2023; 2024; Lachapelle et al., 2023). Third, attaining the limit is an optimization question. Gradient descent reaches it at a linear rate in the linear-Gaussian family (Propositions B.21 and B.23), and an invertible warp of the observation leaves the limit unchanged (Theorem B.25). We discuss the general nonlinear case in Conjecture B.26.

4 Binding Certification Is Reparametrization-Relative

We have shown that there is a binding limit fixed by the data. The question is whether the usual measures of binding can detect that limit in a representation. In practice, binding is often certified by inspecting the role-local leakage $\sum_i \beta_i^{\pi}$ in the particular coordinates of the representation. The problem is that those coordinates are not unique. We show that this leakage can change under an

information-preserving reparametrization, even though the underlying representation has not lost or gained any information. By contrast, the role-conditioned total correlation is invariant to such changes of coordinates.

Definition 4.1 (Slot factorization and reparametrization)**.** A k-slot factorization is a measurable isomorphism $\pi : \mathcal{S} \to \mathcal{S}_1 \times \cdots \times \mathcal{S}_k$ together with the role readout. Writing $\hat{\phi}_i^\pi$ for the i-th slot of $\pi(S)$, the role-local excess is $\beta_i^\pi = I(\hat{\phi}_i^\pi; \phi_{-i} \mid \hat{\boldsymbol{r}})$. The β_i of Definition 2.5 is $\beta_i^{\pi_0}$ in the encoder's native factorization π_0. A reparametrization is a bimeasurable bijection $T : \mathcal{S} \to \mathcal{S}$, possibly depending on $\hat{\boldsymbol{r}}$, such that $\sigma(T(S), \hat{\boldsymbol{r}}) = \sigma(S, \hat{\boldsymbol{r}})$. These maps form a group $\mathcal{G}$. The subgroup $\mathcal{G}_{\text{slot}}^\pi$ consists of reparametrizations that, conditional on $\hat{\boldsymbol{r}}$, act separately and invertibly within each slot. Such maps do not move information between slots. Two representations are *informationally equivalent* if $S' = T(S)$ for some $T \in \mathcal{G}$.

A binding certificate is a functional F of the joint law of $(S, \hat{\boldsymbol{r}}, \boldsymbol{\phi})$. We call it *reparametrization-invariant* if it is unchanged under every $T \in \mathcal{G}$, so that informationally equivalent representations receive the same certificate. We can therefore test a certificate's invariance by applying an information-preserving reparametrization and asking whether its verdict changes.

Consider a code that binds, with the car slot carrying the car's colour and nothing about the apple's. Mix a multiple a of the apple slot into the car slot, $T(S) = (\hat{\phi}_{\text{car}} + a\,\hat{\phi}_{\text{apple}},\ \hat{\phi}_{\text{apple}})$. No information is lost because T is invertible; the original car slot is recovered by subtracting $a\,\hat{\phi}_{\text{apple}}$. Thus, $T \in \mathcal{G}$. Yet, the transformed car slot now contains the apple's colour. A slot-local reader therefore reports leakage where there was none before, while an unrestricted reader can invert T and recover the original representation exactly.

Lemma 4.2 (Invariance of the leakage and the total correlation)**.** *The role-local excess β_i^π is invariant under $\mathcal{G}_{\text{slot}}^\pi$ but not under $\mathcal{G}$. For example, if $S = (\phi_1, \phi_2)$ and the fillers are independent standard Gaussians, then $T(S) = (\phi_1 + a\phi_2, \phi_2)$ changes β_1^π from 0 to $\frac{1}{2}\log(1 + a^2) > 0$. In contrast, the role-conditioned total correlation $\mathrm{TC}(\boldsymbol{\phi} \mid S, \hat{\boldsymbol{r}})$ is invariant under every $T \in \mathcal{G}$.*

Proof sketch. A coordinatewise reparametrization preserves each $\sigma(\hat{\phi}_i^\pi, \hat{\boldsymbol{r}})$ and therefore leaves β_i^π unchanged. In the Gaussian example, $I(\phi_1 + a\phi_2; \phi_2 \mid \hat{\boldsymbol{r}}) = \frac{1}{2}\log(1 + a^2)$. For the total correlation, every $T \in \mathcal{G}$ satisfies $\sigma(T(S), \hat{\boldsymbol{r}}) = \sigma(S, \hat{\boldsymbol{r}})$, so the conditional entropies defining TC are unchanged. The full argument is in Appendix B.3.

Theorem 4.3 (Role-local binding is not determined by information)**.** *There are informationally equivalent representations S and $T(S)$ for which $B^\pi(S) \neq B^\pi(T(S))$. Consequently, no functional invariant under $\mathcal{G}$ agrees with the role-local binding indicator B^π on every representation.*

Proof sketch. Take $S = (\phi_1, \phi_2)$ for independent standard Gaussian fillers and set $T(S) = (\phi_1 + a\phi_2, \phi_2)$ with $a \neq 0$. The map is invertible, but the first slot changes from independent of ϕ_2 to informative about it. Thus B^π changes while every $\mathcal{G}$-invariant functional remains fixed. A full proof is in Appendix B.3.

The role-conditioned total correlation is invariant to every information-preserving reparametrization, and for jointly sufficient codes, the binding limit fixes it from the rendering.

Corollary 4.4 (The invariant log-loss excess)**.** *$\mathrm{TC}(\boldsymbol{\phi} \mid S, \hat{\boldsymbol{r}})$ is invariant under every $T \in \mathcal{G}$. When the fillers take values in finite alphabets, it equals the additional log-loss incurred by decoding them with separate role-wise heads rather than one joint head:*

$$\min_{q_1,\dots,q_k} \mathbb{E}\left[-\sum_i \log q_i(\phi_i \mid S, \hat{\boldsymbol{r}})\right] - \min_q \mathbb{E}[-\log q(\boldsymbol{\phi} \mid S, \hat{\boldsymbol{r}})].$$

For jointly sufficient codes it equals $\mathrm{TC}(\boldsymbol{\phi} \mid X, \hat{\boldsymbol{r}})$ and is positive exactly when the fillers remain conditionally dependent given $(X, \hat{\boldsymbol{r}})$.

Locatello et al. (2019) establish that disentangled factors are not identifiable without supervision. When roles are observed and fillers are recoverable, supervised identifiability can restrict the remaining factor ambiguity to separate within-slot reparametrizations (Locatello et al., 2020a; Khemakhem et al., 2020; Von Kügelgen et al., 2021). This restriction does not fix the coordinates of the learned representation because the cross-slot transformations in Theorem 4.3 leave the roles fixed and preserve the full information. Under a role-identifying slot-separable rendering, a jointly sufficient slot-equivariant encoder recovers each filler from its own slot exactly when the observation determines the full filler tuple. Appendix Theorem B.11 states the full result.

Reader classes differ in which changes of coordinates they can undo. Appendix Definition B.12 defines the corresponding excess B_H. If $H \subseteq H'$, then $B_{H'} \leq B_H$, so enlarging the reader class can only reduce the excess (Theorem B.13). Neither $B_{\mathcal{G}}$ nor the total correlation bounds the other. We report the total correlation separately because it equals the excess log-loss of separate role-wise decoding (Corollary 4.4).

5 Experiments

The theory makes three predictions. First, if binding is not intrinsic to a representation, an information-preserving relabel should be able to change a binding readout without changing the information available to the model (Section 5.1). Second, every jointly sufficient representation should attain the binding limit, and no representation should fall below it without losing joint filler information (Section 5.2). Third, if that limit governs compositional generalisation, the held-out gap should follow data entanglement rather than coordinate-level leakage (Section 5.3). We test these predictions experimentally.

5.1 Is binding certification reparametrisation-relative in trained models?

Prior work has identified subspaces that track which filler belongs to which role (Feng & Steinhardt, 2024; Dai et al., 2024; Prakash et al., 2024; 2025). For each model, we identify a binding subspace and exchange it between two roles while leaving the rest of the representation unchanged. The exchange is an involution, so applying it twice recovers the original representation exactly. No information is lost, and $\mathrm{TC}(\boldsymbol{\phi} \mid S, \hat{\boldsymbol{r}})$ is therefore unchanged. The downstream network is held fixed rather than composed with the inverse exchange. The intervention therefore tests whether the model's own reader depends on the coordinates in which the binding subspace is expressed. Subspaces are fit, selected, and scored on disjoint splits against twenty same-rank random controls (Appendix D). This isolates the question we care about. If the targeted swap changes the binding while the random swap does not, then the binding readout depends on the coordinates even though the information is unchanged.

On Pythia-2.8B, for example, the prompt `Henry is in Egypt. Jack is in Norway. ...Where is Henry?` places probability 0.520 on `Egypt` and 0.069 on `Norway`. Swapping the rank-11 binding subspace reverses the answer. `Norway` rises to 0.863 and `Egypt` falls to 0.017. The matched random swap leaves the answer unchanged. The intervention reconstructs to 3.55×10^{-15}, so the information has not been removed or corrupted; only the coordinates through which the binding is read have changed.

The same pattern holds broadly in Table 1. On nine autoregressive attention language models (Biderman et al., 2023; Yang et al., 2025; Grattafiori et al., 2024) with parameter sizes from 410M to 8B, the swap flips the answer on 0.733–1.000 of held-out examples and every control mean is at most 0.011. The four larger Mamba models reach 0.776–0.917 against controls at most 0.015, and the vision-language models reach 0.740–1.000 against zero control (Gu & Dao, 2023; Wang et al., 2024; Liu et al., 2024). Slot attention, ESBN, the abstractor, and the role–filler tensor-product and oscillator models show the same separation.

The tensor-product result is especially informative because role–filler binding is explicit in the architecture's outer products (Smolensky, 1990). Across the three qualifying runs, the targeted relabel flips 0.943–0.995 of eligible directions, compared with 0.087–0.151 for the same-rank controls. Because the relabel is invertible, the original code remains exactly recoverable. Even an architecture built to express binding explicitly therefore reads that binding in a coordinate-dependent way.

We test the same hypothesis at the text-conditioning interface in the generator study. CLIP-conditioned systems bind poorly and show little separation between the targeted and random swaps. T5-conditioned systems bind more reliably and show a much larger targeted effect. DeepFloyd IF combines T5 conditioning with a denoising UNet and it reaches 0.51 binding accuracy, and the targeted colour swap changes the bound colour on 0.42 of examples compared with 0.018 for the matched control. Table 2 therefore points to the conditioning representation rather than the sampling procedure as the source of the role–colour binding signal that the denoiser uses to place each colour on the correct object. We note that because the models we consider differ simultaneously in sampler, encoder, and conditioning streams, this is not a true causal ablation.

Architecture	Read site	Flip ↑	Control ↓
Pretrained foundation models, binding localized by intervention			
Autoregressive attention language models			
Pythia-410M	answer	0.733	0.011
Pythia-1.4B	answer	0.984	0.003
Pythia-2.8B	answer	0.997	0.002
Qwen2.5-0.5B	answer	1.000	0.000
Qwen2.5-1.5B	answer	1.000	0.000
Qwen2.5-3B	answer	1.000	0.000
Llama-3.1-8B	answer	0.995	0.001
Llama-3.2-1B	answer	0.912	0.006
Llama-3.2-3B†	entity	0.484	0.009
Non-attention state-space model, Mamba			
Mamba-130M	answer	0.357	0.050
Mamba-370M	answer	0.850	0.015
Mamba-790M	answer	0.828	0.007
Mamba-1.4B	answer	0.917	0.003
Mamba-2.8B‡	answer	0.776	0.007
Vision-language models			
Qwen2-VL-2B	answer	0.990	0.000
Qwen2-VL-7B	answer	1.000	0.000
LLaVA-1.5-7B	answer	0.740	0.000
Trained architectures with a known binding mechanism			
Tensor-product	block 1	0.943–0.995	0.087–0.151
Oscillator synchronisation (AKOrN)	block 1–2	0.955–0.999	0.001–0.075
Emergent-symbol memory (ESBN)	memory	0.932–0.970	0.018–0.021
Relational-bottleneck abstractor	encoder	0.999–1.000	0.000–0.001
Slot attention	slot	0.991–0.993	0.003–0.012
Architectures that fail the binding task (negative controls)			
CLIP ViT-L/14 (contrastive, CLIP)	image embed	0.101	0.011

Table 1: **Information-preserving relabels change the binding readout across architectures.** Flip is computed over directions whose unmodified prediction is correct and control is the mean over twenty same-rank random subspaces.

Sampler	Encoder	Model	acc	swap / ctrl
rectified-flow (MMDiT)	T5	FLUX.1-schnell	0.688	0.606 / 0.005
rectified-flow (MMDiT)	T5	FLUX.1-dev	0.677	0.642 / 0.001
rectified-flow (MMDiT)	CLIP+T5	SD3.5-large	0.781	0.628 / 0.010
rectified-flow (MMDiT)	CLIP+T5	SD3.5-medium	0.750	0.624 / 0.018
denoising (UNet)	**T5**	**DeepFloyd IF-I-XL**	**0.510**	**0.422 / 0.018**
denoising (UNet)	CLIP	SD-Turbo	0.115	0.132 / 0.031
denoising (UNet)	CLIP	SD-2.1-base	0.125	0.090 / 0.007
contrastive	CLIP	CLIP ViT-L/14	0.543	0.101 / 0.011

Table 2: **Binding and the targeted intervention follow the conditioning representation across sampler families.**

These results refine the interpretation of causal binding subspaces. Such subspaces can genuinely control a model's role–filler assignments, as prior work has shown (Feng & Steinhardt, 2024; Dai et al., 2024). What they do not provide is an intrinsic certificate of binding. An invertible change of coordinates can alter the same readout while preserving all of the information in the representation.

We note that the usual recommendations of scaling, different training procedures and architectures do not change the conclusion. Increasing model size cannot remove the obstruction because it comes from an information-preserving symmetry rather than limited capacity. It also appears across different training procedures, including post-trained Qwen2.5 and Llama-3 models as well as models without reinforcement-learning post-training.

Substrate	Axis	Filler	Spearman(gap, η) [95% CI]
Composite CLEVR	dialed (θ)	colour	$+0.983, [0.86, 1.00]$
Multi-MNIST	dialed (overlap)	digit	$+0.976, [0.88, 1.00]$
Raw-CLEVR occlusion	found	colour	$+0.604, [0.19, 0.87]$
Visual Genome occlusion	found, real images	colour	$+0.752, [0.57, 0.85]$

Table 3: **The generalisation gap rises with rendering entanglement across controlled and natural substrates.**

5.2 Does the binding limit hold when the ground truth is known?

We first test the binding limit in a setting where the rendering is known exactly. In a two-slot system, we vary the mixing angle from fully separable to strongly entangled and train sufficient codes throughout. Their total correlation closely follows the closed-form limit, with correlation 0.99998 and maximum deviation 0.0024 (Figure 8). An invertible rotation of a sufficient code leaves the value unchanged to 3.11×10^{-15}, as the theory predicts.

The result is stable across a broader sweep. Across 1128 configurations varying filler dimension, slot count, anisotropy, noise, and optimisation, the trained code matches the value predicted by Lemma 3.1 to a maximum relative error of 0.0258 (Appendix G). A noise-only ablation separates entanglement from failed fitting: when the code carries no filler information, the sufficiency shortfall is 1.449, but once sufficiency is recovered it falls to 0.0012 while the measured limit remains tied to the rendering. The total-rate decomposition in Theorem B.4 also holds numerically, with maximum residual 2.75×10^{-14} over 250 random encoders and renderings.

These experiments show that the limit is not an empirical regularity of a particular architecture or optimiser. Once the code is jointly sufficient, or per-filler sufficient, its residual binding entanglement is set by the rendering.

5.3 Does data entanglement, not leakage, govern generalisation?

The binding limit makes a direct prediction about generalisation. For a jointly sufficient code, the role-conditioned total correlation is fixed by the rendering and therefore changes across renderings, not across equivalent representations of the same rendering. Coordinate leakage can vary even when the represented information does not. If the limit governs compositional transfer, the generalisation gap should follow data entanglement and remain unchanged when only the coordinates are altered.

Across renderings, the gap follows entanglement. We measure rendering entanglement with the modulus η from Definition B.15: the part of the optimal squared error for predicting one filler from the observation that disappears when the other fillers are revealed. This L^2 quantity avoids the finite-sample ceiling for mutual-information estimators (McAllester & Stratos, 2020). Under the finite-support positivity condition of Proposition C.1, it vanishes exactly when the role-conditioned total correlation does. Proposition B.17 then relates it to the transfer gap without claiming that positive entanglement always forces a gap. On CLEVR, varying the mixing angle changes η continuously and produces a rank correlation of $0.983, [0.860, 1.000]$ with the held-out gap across five binding mechanisms (Johnson et al., 2017; Bahdanau et al., 2019). Varying overlap in Multi-MNIST gives $0.976, [0.882, 1.000]$ (LeCun et al., 1998). The same relationship appears when entanglement is not imposed analytically. Raw-CLEVR occlusion gives a count-controlled correlation of $0.604, [0.19, 0.87]$; on the ten-by-four re-binning in Appendix F, it falls to 0.299 with an interval containing zero. Visual Genome gives $0.752, [0.571, 0.850]$ over $439{,}336$ colour-binding objects. (Krishna et al., 2017) In Visual Genome, neighboring colors are nearly independent ($\rho_{\max} = 0.041$), the gap rises across the occlusion bins, and the separable regime forms the limit.

The effect is not explained by training co-occurrence. Entanglement can correlate with the combinations seen during training, so we separate the two factors directly. A pre-registered factorial independently varies occlusion, which controls η, and a position–colour shortcut c, which controls how strongly position predicts colour. Their realised rank correlation across the 49 conditions is 0.01. Holding co-occurrence fixed, the partial correlation between η and the gap is $0.635, [0.27, 0.91]$ on the 25-cell grid and $0.900, [0.808, 0.951]$ on the denser in-regime grid. Holding η fixed, co-occurrence contributes essentially nothing, with partial correlation -0.063 and an interval containing zero. The

Factor varied	Grid	Partial $\rho(\text{gap})$ [95% CI]
Entanglement η	25-cell	$+0.635, [+0.27, +0.91]$
Entanglement η	49-cell, in-regime	$+0.900, [+0.808, +0.951]$
Co-occurrence c	49-cell, in-regime	-0.063 (interval contains 0)
Co-occurrence c	49-cell, full	$+0.357$

Table 4: **When entanglement and co-occurrence are varied independently, the gap follows entanglement.**

only increase attributed to co-occurrence appears in the fully coupled $c = 1$ column, where the shortcut itself forces the rendering to become entangled. The gap therefore follows the entanglement of the observation rather than the training co-occurrence of its parts.

At fixed rendering, leakage can move without the gap. The complementary test holds the rendering fixed and changes only the representation's coordinates. On separable token tasks, neither leakage nor entanglement predicts the gap on SCAN (0.105 and 0.099), and nineteen configurations have more than 1.5 nats of role-local leakage while entanglement remains below 0.05 and accuracy stays near one. Under an information-preserving relabel, role-local leakage rises above 0.571 nats while total correlation changes by only 9.5×10^{-9} and the best-reader generalisation gap by 4.5×10^{-10}. The same dissociation appears in trained perceptual slots, where total correlation remains between 0.001 and 0.006 nats across eight capacities while coordinate leakage ranges from 2.54 to 3.31 nats (Figure 3). Leakage can therefore change substantially without changing either the invariant or its transfer consequence.

Changing the rendering changes the gap. We finally intervene on the quantity the theory identifies as causal. De-occluding the Visual Genome training corpus lowers rendering entanglement and reduces the held-out-combination gap by 79–85% at every intervention dose (Appendix E). Because de-occlusion also changes absolute position, we compare it with overlap-preserving relocation. The control reduces the gap from 0.081 to 0.041–0.063, whereas de-occlusion reaches about 0.014 at every dose. The larger reduction under de-occlusion shows that the positional change alone does not explain the full reduction. The prediction also holds out of sample. On six new mixing angles whose ordering was fixed before the corresponding gaps were computed, the realised gaps increase strictly with the mixing angle, giving rank correlation $1.000, [1.000, 1.000]$.

6 Conclusion

Prior work in mechanistic interpretability has found binding subspaces and circuits responsible for binding (Feng & Steinhardt, 2024; Dai et al., 2024; Prakash et al., 2024; 2025). Disentanglement and object-centric learning, on the other hand, ask when the underlying factors or slots can be recovered (Locatello et al., 2019; 2020a;b; Brady et al., 2023; Wiedemer et al., 2024). Together, these results identify how binding is expressed or how factors are recovered in a particular representation but they do not establish that binding is an intrinsic property of the representation itself. We show that it is not. An invertible relabeling can change the apparent binding of a representation while preserving all of the information in the representation, which makes binding relative to the reader used to interpret it. We also show that every jointly sufficient representation inherits the residual dependence already present in the observation, and any other representation can remove that dependence only by discarding joint information about the fillers. For mechanistic interpretability, this changes what a binding subspace can support as a claim. Existing results can still identify mechanisms that a model uses to recover role–filler assignments, but they cannot by themselves establish that binding resides in that subspace or in the representation more generally. Future work therefore has to state which reader makes the binding visible and whether the conclusion survives an information-preserving change of coordinates. For compositional generalization, improvements from architectures, objectives, or slot structure should be interpreted against the separability already available in the observation, and benchmarks should distinguish failures caused by the model from failures imposed by the data. More broadly, representation learning needs a theory of equivalence before it can define intrinsic representational quality. Otherwise, a criterion can rank equivalent descriptions of the same information differently and mistake a choice of coordinates for a learned capability.

AI USE DISCLOSURE

LLMs were used to help write research code and polish selected prose. The authors reviewed and verified the resulting code and text and remain responsible for the contents of this paper.

REFERENCES

Dzmitry Bahdanau, Harm de Vries, Timothy J O'Donnell, Shikhar Murty, Philippe Beaudoin, Yoshua Bengio, and Aaron Courville. Closure: Assessing systematic generalization of clevr models. *arXiv preprint arXiv:1912.05783*, 2019.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pp. 531–540. PMLR, 2018.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International conference on machine learning*, pp. 2397–2430. PMLR, 2023.

Jack Brady, Roland S Zimmermann, Yash Sharma, Bernhard Schölkopf, Julius Von Kügelgen, and Wieland Brendel. Provably learning object-centric representations. In *International Conference on Machine Learning*, pp. 3038–3062. PMLR, 2023.

Gal Chechik, Amir Globerson, Naftali Tishby, and Yair Weiss. Information bottleneck for gaussian variables. *Advances in Neural Information Processing Systems*, 16, 2003.

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, pp. 1779–1788. PMLR, 2020.

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

Qin Dai, Benjamin Heinzerling, and Kentaro Inui. Representational analysis of binding in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17468–17493, 2024.

Sahil Rajesh Dhayalkar. Attention as binding: A vector-symbolic perspective on transformer reasoning. *arXiv preprint arXiv:2512.14709*, 2025.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context? In *International Conference on Learning Representations*, volume 2024, pp. 36391–36413, 2024.

Jerry A Fodor and Zenon W Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988.

Pascal Fries. A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends in cognitive sciences*, 9(10):474–480, 2005.

Hans Gebelein. Das statistische problem der korrelation als variations-und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 21(6):364–379, 1941.

Anand Gopalakrishnan, Aleksandar Stanić, Jürgen Schmidhuber, and Michael C Mozer. Recurrent complex-weighted autoencoders for unsupervised object discovery. *Advances in Neural Information Processing Systems*, 37:140787–140811, 2024.

Anirudh Goyal, Alex Lamb, Jordan Hoffmann, Shagun Sodhani, Sergey Levine, Yoshua Bengio, and Bernhard Schölkopf. Recurrent independent mechanisms. *arXiv preprint arXiv:1909.10893*, 2019.

Anirudh Goyal, Aniket Didolkar, Nan Rosemary Ke, Charles Blundell, Philippe Beaudoin, Nicolas Heess, Michael C Mozer, and Yoshua Bengio. Neural production systems. *Advances in Neural Information Processing Systems*, 34:25673–25687, 2021.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Klaus Greff, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks. *arXiv preprint arXiv:2012.05208*, 2020.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Yoav Gur-Arieh, Mor Geva, and Atticus Geiger. Mixing mechanisms: How language models retrieve bound entities in-context. *arXiv preprint arXiv:2510.06182*, 2025.

Hermann O Hirschfeld. A connection between correlation and contingency. In *Mathematical proceedings of the cambridge philosophical society*, volume 31, pp. 520–524. Cambridge University Press, 1935.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

John E Hummel and Keith J Holyoak. A symbolic-connectionist theory of relational inference and generalization. *Psychological review*, 110(2):220, 2003.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795, 2020.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Olav Kallenberg. *Foundations of modern probability*. Springer, 1997.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 795–811. Springer, 2016.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, et al. Measuring compositional generalization: A comprehensive method on realistic data. *arXiv preprint arXiv:1912.09713*, 2019.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International conference on artificial intelligence and statistics*, pp. 2207–2217. PMLR, 2020.

Najoung Kim and Tal Linzen. Cogs: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)*, pp. 9087–9105, 2020.

Avinash Kori, Francesco Locatello, Ainkaran Santhirasekaram, Francesca Toni, Ben Glocker, and Fabio De Sousa Ribeiro. Identifiable object-centric representation learning via probabilistic slot attention. *Advances in Neural Information Processing Systems*, 37:93300–93335, 2024.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

Sébastien Lachapelle, Divyat Mahajan, Ioannis Mitliagkas, and Simon Lacoste-Julien. Additive decoders for latent variables identification and cartesian-product extrapolation. *Advances in Neural Information Processing Systems*, 36:25112–25150, 2023.

Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, pp. 2873–2882. PMLR, 2018.

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.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pp. 4114–4124. PMLR, 2019.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International conference on machine learning*, pp. 6348–6359. PMLR, 2020a.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in neural information processing systems*, 33:11525–11538, 2020b.

David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*, pp. 875–884. PMLR, 2020.

Takeru Miyato, Sindy Löwe, Andreas Geiger, and Max Welling. Artificial kuramoto oscillatory neurons. In *International Conference on Learning Representations*, volume 2025, pp. 44278–44322, 2025.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Tony A Plate. Holographic reduced representations. *IEEE Transactions on Neural networks*, 6(3): 623–641, 1995.

Boris T Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International conference on machine learning*, pp. 5171–5180. PMLR, 2019.

Nikhil Prakash, Tamar Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *International Conference on Learning Representations*, volume 2024, pp. 8057–8082, 2024.

Nikhil Prakash, Natalie Shapira, Arnab Sen Sharma, Christoph Riedl, Yonatan Belinkov, Tamar Rott Shaham, David Bau, and Atticus Geiger. Language models use lookbacks to track beliefs. *arXiv preprint arXiv:2505.14685*, 2025.

Alfréd Rényi. On measures of dependence. *Acta mathematica hungarica*, 10(3-4):441–451, 1959.

Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks. In *International conference on machine learning*, pp. 4433–4441. PMLR, 2018.

Imanol Schlag, Paul Smolensky, Roland Fernandez, Nebojsa Jojic, Jürgen Schmidhuber, and Jianfeng Gao. Enhancing the transformer with explicit relational encoding for math problem solving. *arXiv preprint arXiv:1910.06611*, 2019.

Wolf Singer and Charles M Gray. Visual feature integration and the temporal correlation hypothesis. *Annual review of neuroscience*, 18(1):555–586, 1995.

Paul Smolensky. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial intelligence*, 46(1-2):159–216, 1990.

Despoina Touska, Bastiaan Onne Fagginger Auer, Alexandru Onose, Tejaswi Kasarla, Luis Armando Pérez Rey, Maximilian Lipp, Lyubov Amitonova, Martin R Oswald, and Pascal Cerfontaine. Orthorf: Exploring orthogonality in object-centric representations. In *The Fourteenth International Conference on Learning Representations*, 2026.

Anne Treisman. The binding problem. *Current opinion in neurobiology*, 6(2):171–178, 1996.

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

Christoph Von Der Malsburg. The correlation theory of brain function. In *Models of neural networks: Temporal aspects of coding and information processing in biological systems*, pp. 95–119. Springer, 1994.

Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467, 2021.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Satosi Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of research and development*, 4(1):66–82, 1960.

Taylor W Webb, Ishan Sinha, and Jonathan D Cohen. Emergent symbols through binding in external memory. *arXiv preprint arXiv:2012.14601*, 2020.

Taylor W Webb, Steven M Frankland, Awni Altabaa, Simon Segert, Kamesh Krishnamurthy, Declan Campbell, Jacob Russin, Tyler Giallanza, Randall O'Reilly, John Lafferty, et al. The relational bottleneck as an inductive bias for efficient abstraction. *Trends in Cognitive Sciences*, 28(9): 829–843, 2024.

Thaddäus Wiedemer, Prasanna Mayilvahanan, Matthias Bethge, and Wieland Brendel. Compositional generalization from first principles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Thaddäus Wiedemer, Jack Brady, Alexander Panfilov, Attila Juhos, Matthias Bethge, and Wieland Brendel. Provable compositional generalization for object-centric learning. In *International Conference on Learning Representations (ICLR)*, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, volume 2024, pp. 1651–1678, 2024.

Feng Zhao. The statistical problem of correlation as variational and eigenvalue problem, including its connection with the curve fitting, 02 2023.

A PRELIMINARIES

Table 5 collects the notation used throughout the appendix. We otherwise follow the notation of the main paper.

Symbol	Meaning
$\boldsymbol{r}, \boldsymbol{\phi}$	roles and their fillers; $\phi_{-i} = (\phi_l)_{l \neq i}$
$X,\ S = E(X)$	observation and representation
$\hat{\boldsymbol{r}}, \hat{\phi}_i$	role readout and slot-i filler readout
$\rho(\boldsymbol{r}, \boldsymbol{\phi})$	rendering kernel
$\mathbb{P}_{\mathrm{tr}}, \mathbb{P}_{\mathrm{te}}$	training and strict-substitutive test laws
$\bigotimes_i$	product of per-role laws or kernels
β_i	role-local binding excess at slot i
$\mathrm{TC}(\boldsymbol{\phi} \mid S, \hat{\boldsymbol{r}})$	role-conditioned total correlation
η	entanglement modulus
$\mathcal{G}$	information-preserving reparametrizations of S
R_i^{suf}	per-filler sufficient rate
$\rho_{\max}$	Hirschfeld–Gebelein–Rényi maximal correlation

Table 5: Notation used in the appendix.

We use a number of standard information-theoretic facts in the proofs below. We state them here.

Lemma A.1 (Pinsker and Bretagnolle–Huber)**.** *For probability laws p and q,*

$$\mathrm{TV}(p, q)^2 \leq \frac{1}{2} \mathrm{KL}(p \| q), \qquad \mathrm{TV}(p, q) \leq \sqrt{1 - e^{-\mathrm{KL}(p \| q)}}.$$

These inequalities let us turn a KL divergence into a change in distribution.

We also use the standard comparison between KL and the chi-squared divergence.

Lemma A.2 ($\mathrm{KL} \leq \chi^2$)**.** *If $p \ll q$ and $r = dp/dq$, then*

$$\mathrm{KL}(p \| q) = \mathbb{E}_q[r \log r] \leq \mathbb{E}_q[r^2 - r] = \chi^2(p \| q).$$

For conditional joint and product laws, the leading dependence term in the corresponding chi-squared expansion is the squared Hirschfeld–Gebelein–Rényi maximal correlation (Hirschfeld, 1935; Gebelein, 1941; Zhao, 2023; Rényi, 1959).

Lemma A.3 (Standard probabilistic tools)**.** *Noise outsourcing represents any pair (W, V) as $V = h(W, U)$ almost surely for a measurable h and a uniform random variable $U \perp\!\!\!\perp W$. For a discrete target $T \in \mathcal{T}$ predicted from O with error probability $P_e = \Pr(\hat{T} \neq T)$, Fano's inequality yields $P_e \geq [H(T \mid O) - \log 2] / \log |\mathcal{T}|$. See Kallenberg (1997); Cover (1999).*

Lemma A.4 (Conditional additivity)**.** *If $Y_1, \ldots, Y_m$ are conditionally independent given Z, then*

$$I(X; Y_{1:m} \mid Z) \geq \sum_{l=1}^{m} I(X; Y_l \mid Z).$$

The reverse inequality requires conditional independence given (X, Z) instead.

We note that if Y_1 and Y_2 are independent fair bits and $X = Y_1 \oplus Y_2$, then $I(X; Y_1, Y_2) = 1$ while $I(X; Y_1) = I(X; Y_2) = 0$. This is why we use the forward direction and not the reverse (Watanabe, 1960; Cover, 1999). For every code S,

$$I(S; \boldsymbol{\phi} \mid \boldsymbol{r}) = \sum_i I(S; \phi_i \mid \boldsymbol{r}) + \mathrm{TC}(\boldsymbol{\phi} \mid S, \boldsymbol{r}) - \mathrm{TC}(\boldsymbol{\phi} \mid \boldsymbol{r}).$$

Under Assumption 2.2, $\mathrm{TC}(\boldsymbol{\phi} \mid \boldsymbol{r}) = 0$, and it becomes,

$$I(S; \boldsymbol{\phi} \mid \boldsymbol{r}) = \sum_i I(S; \phi_i \mid \boldsymbol{r}) + \mathrm{TC}(\boldsymbol{\phi} \mid S, \boldsymbol{r}).$$

B PROOFS

This appendix proves the theoretical results in Sections 3 and 4.

B.1 PRELIMINARY LEMMAS

Lemma B.1 (Minimality as independent noise). *For each slot i, let $\widetilde{\beta}_i = I(\hat{\phi}_i; \boldsymbol{\phi}_{-i} \mid \hat{\boldsymbol{r}}, \phi_i)$. Then $\widetilde{\beta}_i = 0$ if and only if $\hat{\phi}_i = h_i(\hat{\boldsymbol{r}}, \phi_i, \omega_i)$ almost surely for a measurable h_i and a uniform random variable $\omega_i \perp\!\!\!\perp (\boldsymbol{r}, \boldsymbol{\phi})$. The joint law of $(\boldsymbol{r}, \boldsymbol{\phi}, \hat{\boldsymbol{\phi}})$ admits such a representation with jointly independent noises if and only if the factorization is jointly minimal. Under Assumption 2.2, joint minimality implies $\widetilde{\beta}_i = 0$ for every i, and $\widetilde{\beta}_i = 0$ implies $\beta_i = 0$.*

Proof. Suppose first that $\hat{\phi}_i = h_i(\hat{\boldsymbol{r}}, \phi_i, \omega_i)$ with $\omega_i \perp\!\!\!\perp (\boldsymbol{r}, \boldsymbol{\phi})$. After conditioning on $(\hat{\boldsymbol{r}}, \phi_i)$, the slot depends only on noise that is independent of $\boldsymbol{\phi}_{-i}$. Hence $\widetilde{\beta}_i = 0$.

Conversely, $\widetilde{\beta}_i = 0$ is the conditional independence of $\hat{\phi}_i$ and $\boldsymbol{\phi}_{-i}$ given $(\hat{\boldsymbol{r}}, \phi_i)$. By conditional noise outsourcing, there is a uniform $\omega_i \perp\!\!\!\perp (\boldsymbol{r}, \boldsymbol{\phi})$ and a measurable h_i such that $\hat{\phi}_i = h_i(\hat{\boldsymbol{r}}, \phi_i, \omega_i)$ almost surely.

With jointly independent noises, $\mathrm{Law}(\hat{\boldsymbol{\phi}} \mid \boldsymbol{r}, \boldsymbol{\phi}) = \bigotimes_i \mathrm{Law}(h_i(\hat{\boldsymbol{r}}, \phi_i, \omega_i) \mid \boldsymbol{r}, \boldsymbol{\phi})$, and each factor depends on the fillers only through ϕ_i. This is joint minimality. Conversely, joint minimality supplies the corresponding kernels. They can be realized with independent uniform noises on a product extension. Finally, under Assumption 2.2,

$$\beta_i \le I(\hat{\phi}_i, \phi_i; \boldsymbol{\phi}_{-i} \mid \hat{\boldsymbol{r}}) = I(\phi_i; \boldsymbol{\phi}_{-i} \mid \hat{\boldsymbol{r}}) + \widetilde{\beta}_i = \widetilde{\beta}_i.$$

□

Lemma B.2 (Cross-slot independence and local preservation). *Under Assumption 2.2, joint minimality implies $I_{\mathbb{P}_{\mathrm{tr}}}(\hat{\phi}_i; \hat{\phi}_j \mid \hat{\boldsymbol{r}}) = 0$ for $i \neq j$. The same conclusion holds under every factorization-preserving test law. More generally, suppose $\mathrm{Law}(\hat{\phi}_i \mid \boldsymbol{r}, \boldsymbol{\phi}) = \kappa_i(\cdot \mid \hat{\boldsymbol{r}}, \phi_i)$ throughout the product of the per-role supports. Every strict-substitutive test law then preserves $\mathrm{Law}(\phi_i \mid \hat{\boldsymbol{r}}, \hat{\phi}_i)$ and $\mathbb{E}[g(r_i, \phi_i) \mid \hat{\boldsymbol{r}}, \hat{\phi}_i]$ for every bounded measurable g.*

Proof. Joint minimality and Assumption 2.2 imply

$$\mathrm{Law}(\boldsymbol{\phi}, \hat{\boldsymbol{\phi}} \mid \boldsymbol{r}) = \bigotimes_i \left[\mathbb{P}_{\mathrm{tr}}(\phi_i \mid r_i) \otimes \kappa_i(\cdot \mid \hat{\boldsymbol{r}}, \phi_i)\right].$$

The pairs $(\phi_i, \hat{\phi}_i)$ are independent given $\boldsymbol{r}$. A factorization-preserving test law has the same product form.

For the second claim, condition (S3) keeps κ_i fixed, while conditions (S1) and (S2) preserve the law of $\boldsymbol{r}$ and the law of ϕ_i given $\boldsymbol{r}$. Thus $(\hat{\boldsymbol{r}}, \phi_i, \hat{\phi}_i)$ has the same joint law under $\mathbb{P}_{\mathrm{tr}}$ and $\mathbb{P}_{\mathrm{te}}$. Its posterior and the stated conditional expectations are therefore preserved. □

The restriction to the own slot is necessary. A strict-substitutive shift may couple the fillers, so the posterior given $(\hat{\boldsymbol{r}}, \hat{\boldsymbol{\phi}})$ need not be preserved even when the posterior given $(\hat{\boldsymbol{r}}, \hat{\phi}_i)$ is preserved.

Lemma B.3 (Measurable recovery). *If Z is $\sigma(X)$-measurable and $Z \perp\!\!\!\perp X \mid W$, then Z is $\sigma(W)$-measurable.*

Proof. Because Z is determined by X, its conditional law given (X, W) is δ_Z. Conditional independence makes this law equal to the conditional law given W. Hence $\mathrm{Law}(Z \mid W) = \delta_Z$, which means that Z is W-measurable. □

B.2 Proofs for Section 3

Theorem B.4 (Total-rate decomposition). *Let $S = E(X)$ be per-filler sufficient and let $R_i^{\mathrm{suf}} = I(X; \phi_i \mid \boldsymbol{r})$. Then*

$$I(S; X \mid \boldsymbol{r}) - \sum_i R_i^{\mathrm{suf}} = I(S; X \mid \boldsymbol{\phi}, \boldsymbol{r}) + \mathrm{TC}(\boldsymbol{\phi} \mid S, \boldsymbol{r}),$$

and both terms on the right are nonnegative.

Proof. Under Assumption 2.2,

$$I(S; \boldsymbol{\phi} \mid \boldsymbol{r}) = \sum_i I(S; \phi_i \mid \boldsymbol{r}) + \mathrm{TC}(\boldsymbol{\phi} \mid S, \boldsymbol{r}).$$

Per-filler sufficiency replaces each mutual information on the right by R_i^{suf}. Since S is a function of X, $I(S; X \mid \boldsymbol{r}) = I(S; \boldsymbol{\phi} \mid \boldsymbol{r}) + I(S; X \mid \boldsymbol{\phi}, \boldsymbol{r})$. Substitution proves the identity. □

Proof of Lemma 3.1. For every code $Z = f(X)$,

$$\mathrm{TC}(\boldsymbol{\phi} \mid Z, \boldsymbol{r}) = \mathrm{TC}(\boldsymbol{\phi} \mid \boldsymbol{r}) + I(\boldsymbol{\phi}; Z \mid \boldsymbol{r}) - \sum_i I(\phi_i; Z \mid \boldsymbol{r}). \tag{1}$$

Subtracting (1) with $Z = S$ from (1) with $Z = X$ yields $\mathrm{TC}(\boldsymbol{\phi} \mid S, \boldsymbol{r}) + \delta_{\mathrm{J}}(S) = \mathrm{TC}(\boldsymbol{\phi} \mid X, \boldsymbol{r}) + \varepsilon_{\perp}(S)$. If S is jointly sufficient, then $\delta_{\mathrm{J}}(S) = 0$. Moreover, $I(\phi_i; X \mid S, \boldsymbol{r}) \le I(\boldsymbol{\phi}; X \mid S, \boldsymbol{r})$ for every i, so $\varepsilon_{\perp}(S) = 0$.

Now suppose that S is per-filler sufficient. Each posterior in U is measurable from $(S, \boldsymbol{r})$, so U is per-filler sufficient. The chain rule yields $\delta_{\mathrm{J}}(U) = \delta_{\mathrm{J}}(S) + I(\boldsymbol{\phi}; S \mid U, \boldsymbol{r}) \ge \delta_{\mathrm{J}}(S)$. Applying the first identity to S and U, with both marginal shortfalls equal to zero, proves

$$\mathrm{TC}(\boldsymbol{\phi} \mid U, \boldsymbol{r}) \le \mathrm{TC}(\boldsymbol{\phi} \mid S, \boldsymbol{r}) = \mathrm{TC}(\boldsymbol{\phi} \mid X, \boldsymbol{r}) - \delta_{\mathrm{J}}(S) \le \mathrm{TC}(\boldsymbol{\phi} \mid X, \boldsymbol{r}).$$

□

Lemma B.5 (Deviation from the binding limit). *Under Assumption 2.2,*

$$\mathrm{TC}(\boldsymbol{\phi} \mid S, \boldsymbol{r}) - \mathrm{TC}(\boldsymbol{\phi} \mid X, \boldsymbol{r}) = \varepsilon_{\perp}(S) - \delta_{\mathrm{J}}(S).$$

Both shortfalls are nonnegative. Per-filler sufficiency makes the left side nonpositive, while joint sufficiency makes it zero. In general,

$$|\mathrm{TC}(\boldsymbol{\phi} \mid S, \boldsymbol{r}) - \mathrm{TC}(\boldsymbol{\phi} \mid X, \boldsymbol{r})| \le \max\{\varepsilon_{\perp}(S), \delta_{\mathrm{J}}(S)\}.$$

Proof. Applying Equation (1) to S and X yields the stated identity. Since S is a function of X,

$$\varepsilon_{\perp}(S) = \sum_i I(\phi_i; X \mid S, \boldsymbol{r}) \ge 0, \qquad \delta_{\mathrm{J}}(S) = I(\boldsymbol{\phi}; X \mid S, \boldsymbol{r}) \ge 0.$$

The remaining claims follow immediately. □

Proof of Corollary 3.2. For a jointly sufficient code, Lemma 3.1 equates the two total correlations, so one vanishes exactly when the other does. If a per-filler sufficient code S has zero total correlation, the lemma's lower bound implies $\mathrm{TC}(\boldsymbol{\phi} \mid U, \boldsymbol{r}) = 0$. Conversely, U is itself per-filler sufficient. Finally, setting $\mathrm{TC}(\boldsymbol{\phi} \mid S, \boldsymbol{r}) = 0$ in the identity from the lemma yields

$$\delta_{\mathrm{J}}(S) = \mathrm{TC}(\boldsymbol{\phi} \mid X, \boldsymbol{r}) + \varepsilon_{\perp}(S) \ge \mathrm{TC}(\boldsymbol{\phi} \mid X, \boldsymbol{r}).$$

□

Lemma B.6 (Fixed coupling). *Let $h \ge 0$ be nonzero on the finite product $F_1 \times \cdots \times F_k$. Suppose*

$$p(\boldsymbol{\phi}) = \frac{h(\boldsymbol{\phi}) \prod_i a_i(\phi_i)}{Z}, \qquad p'(\boldsymbol{\phi}) = \frac{h(\boldsymbol{\phi}) \prod_i a_i'(\phi_i)}{Z'},$$

where the a_i and a_i' are strictly positive probability mass functions. If p and p' have the same one-dimensional marginals, then $p = p'$.

Proof. On the common support $\{h > 0\}$,

$$\log \frac{p}{p'} = \log \frac{Z'}{Z} + \sum_i \log \frac{a_i(\phi_i)}{a'_i(\phi_i)}.$$

The constant cancels after adding the two directed divergences. Hence

$$\mathrm{KL}(p\|p') + \mathrm{KL}(p'\|p) = \sum_i (\mathbb{E}_p - \mathbb{E}_{p'}) \left[\log \frac{a_i(\phi_i)}{a'_i(\phi_i)}\right] = 0,$$

because the marginals agree. Both divergences are nonnegative, so both vanish. □

Proposition B.7 (When per-filler sufficiency is joint)**.** *Suppose one of the following model conditions holds. For finite fillers, assume* $\log \rho(x \mid \boldsymbol{r}, \boldsymbol{\phi}) = a(x) + \sum_i g_i(x, \phi_i) + \psi(\boldsymbol{\phi})$*, where* ψ *does not depend on* x*. In the Gaussian case, assume that the fillers and noise are jointly Gaussian and* $X = A\boldsymbol{\phi} + e$*. In the slotwise case, assume Assumption 2.2, write* $X = (X_1, \ldots, X_k)$*, and require*

$$P(X \in dx \mid \boldsymbol{\phi}, \boldsymbol{r}) = \bigotimes_i P(X_i \in dx_i \mid \phi_i, r_i).$$

Then U *is jointly sufficient, every per-filler sufficient code is jointly sufficient, and every such code satisfies* $\mathrm{TC}(\boldsymbol{\phi} \mid S, \boldsymbol{r}) = \mathrm{TC}(\boldsymbol{\phi} \mid X, \boldsymbol{r})$*. The prior need not factorize in the first two cases.*

Proof. Fix the roles. In the finite-filler case, Bayes' rule yields

$$p(\boldsymbol{\phi} \mid x, \boldsymbol{r}) \propto h(\boldsymbol{\phi}) \prod_i e^{g_i(x, \phi_i)}, \qquad h(\boldsymbol{\phi}) = e^{\psi(\boldsymbol{\phi})} P(\boldsymbol{\phi} \mid \boldsymbol{r}).$$

Every posterior belongs to the same fixed-coupling exponential family. Lemma B.6 shows that its one-dimensional marginals determine the joint law. The inverse marginal map is Borel measurable on its image, so U determines the joint posterior.

In the Gaussian case, the posterior covariance does not depend on x, while the posterior mean is affine in x. The marginal posteriors in U determine every component of this mean and hence the joint posterior. In the slotwise case, Bayes' rule and the two factorizations imply

$$P(\boldsymbol{\phi} \in \cdot \mid X, \boldsymbol{r}) = \bigotimes_i P(\phi_i \in \cdot \mid X_i, \boldsymbol{r}) = \bigotimes_i P(\phi_i \in \cdot \mid X, \boldsymbol{r}).$$

Thus U determines the joint posterior in all three cases and is jointly sufficient.

If S is per-filler sufficient, each marginal posterior given $(S, \boldsymbol{r})$ equals the corresponding posterior given $(X, \boldsymbol{r})$. Hence U is measurable from $(S, \boldsymbol{r})$. Since U determines the joint posterior, that posterior is also measurable from $(S, \boldsymbol{r})$. Therefore S is jointly sufficient. Lemma 3.1 then establishes equality of the two total correlations. □

Theorem B.8 (Joint minimality requires conditional separability)**.** *Under Assumption 2.2, every jointly minimal factorization satisfies* $\mathrm{TC}(\boldsymbol{\phi} \mid S, \boldsymbol{r}) = 0$*. Thus, if* $\mathrm{TC}(\boldsymbol{\phi} \mid X, \boldsymbol{r}) > 0$*, no jointly sufficient code has a jointly minimal factorization. If* $\mathrm{TC}(\boldsymbol{\phi} \mid U, \boldsymbol{r}) > 0$*, no per-filler sufficient code has one.*

The per-slot equalities do not detect this obstruction. Let ϕ_1*,* ϕ_2*, and* K *be independent fair bits and set* $X = (\phi_1 \oplus \phi_2 \oplus K, K)$*. The jointly sufficient slots* $\hat{\phi}_1 = \phi_1 \oplus \phi_2 \oplus K$ *and* $\hat{\phi}_2 = K$ *satisfy* $\beta_i = \widetilde{\beta}_i = 0$*, although* $\mathrm{TC}(\boldsymbol{\phi} \mid X, \boldsymbol{r}) = \log 2$*.*

Proof. Joint minimality and Assumption 2.2 imply $\mathrm{Law}(\boldsymbol{\phi}, \hat{\boldsymbol{\phi}} \mid \boldsymbol{r}) = \bigotimes_i [\mathbb{P}_{\mathrm{tr}}(\phi_i \mid r_i) \otimes \kappa_i(\cdot \mid \hat{\boldsymbol{r}}, \phi_i)]$. The pairs $(\phi_i, \hat{\phi}_i)$ are independent given $\boldsymbol{r}$. The posterior of $\boldsymbol{\phi}$ given $(\hat{\boldsymbol{\phi}}, \boldsymbol{r})$ therefore factorizes across roles, so $\mathrm{TC}(\boldsymbol{\phi} \mid \hat{\boldsymbol{\phi}}, \boldsymbol{r}) = 0$. A slot factorization is bimeasurably invertible, so $\mathrm{TC}(\boldsymbol{\phi} \mid S, \boldsymbol{r}) = 0$. Lemma 3.1 proves both impossibility claims.

In the example, X determines $\phi_1 \oplus \phi_2 = X_1 \oplus X_2$. Given X, either filler remains uniform and determines the other, so $\mathrm{TC}(\boldsymbol{\phi} \mid X) = \log 2$. Each displayed slot is uniform and independent of the fillers, even after conditioning on its own filler. Hence $\beta_i = \widetilde{\beta}_i = 0$. □

Definition B.9 (Slot-separable rendering and slot-equivariant encoder)**.** A rendering is slot-separable when $X = \mathcal{C}\big(\rho_0(r_1, \phi_1), \ldots, \rho_0(r_k, \phi_k)\big)$, where $\rho_0 \colon \mathcal{R} \times \mathcal{F} \to \mathcal{Z}$ is shared across slots and $\mathcal{C} \colon \mathcal{Z}^k \to \mathcal{X}$ is permutation-invariant and injective on multisets. Each rendered component $\rho_0(r, \phi)$ in the support must identify its role r. An encoder is slot-equivariant when a shared map $e_0 \colon \mathcal{Z} \to \mathcal{S}_0$ satisfies $\hat{\phi}_i = e_0(\rho_0(r_i, \phi_i))$ for every i.

Proposition B.10 (The classical binding problem)**.** *Suppose the rendering in Definition B.9 is deterministic and Assumption 2.2 holds. Then* $\mathrm{TC}(\boldsymbol{\phi} \mid X, \boldsymbol{r}) = 0$*, and the slots* $\hat{\phi}_i = \rho_0(r_i, \phi_i)$ *form a jointly sufficient and jointly minimal code.*

Without role identification, the conclusion can fail. For two roles with independent uniform binary fillers and $\rho_0(r, \phi) = \phi$*, the observation is the multiset* $\{\phi_1, \phi_2\}$ *and* $\mathrm{TC}(\boldsymbol{\phi} \mid X, \boldsymbol{r}) = \frac{1}{2}\log 2$.

Proof. Write $z_i = \rho_0(r_i, \phi_i)$. Injectivity of $\mathcal{C}$ on multisets determines $\{z_1, \ldots, z_k\}$ from X, and role identification determines the ordered tuple. Conversely, that tuple determines X. Assumption 2.2 then implies

$$P(\boldsymbol{\phi} \in \cdot \mid X, \boldsymbol{r}) = \bigotimes_i P(\phi_i \in \cdot \mid z_i, \boldsymbol{r}), \qquad \mathrm{TC}(\boldsymbol{\phi} \mid X, \boldsymbol{r}) = 0.$$

The slots determine X and are functions of it, so the code is jointly sufficient. Their conditional law given the fillers is the product of the point masses at $\rho_0(r_i, \phi_i)$, which proves joint minimality.

Without role identification, observing $\{0, 1\}$ leaves the assignments $(0, 1)$ and $(1, 0)$ equally likely. This event has probability one half and contributes $\log 2$ of conditional mutual information. The other two multisets contribute zero, so the total correlation is $\frac{1}{2}\log 2$. □

B.3 Proofs for Section 4

Proof of Lemma 4.2. For $T \in \mathcal{G}^{\pi}_{\mathrm{slot}}$, each slot is transformed invertibly using only that slot and $\hat{\boldsymbol{r}}$. Hence $\sigma(\hat{\phi}_i, \hat{\boldsymbol{r}})$ and β^{π}_i are unchanged.

This fails for the full group $\mathcal{G}$. Let the fillers be independent standard Gaussians and take $S = (\phi_1, \phi_2)$. The map $T_a(S) = (\phi_1 + a\phi_2, \phi_2)$ has inverse $(u, v) \mapsto (u - av, v)$ and therefore belongs to $\mathcal{G}$. Its first slot satisfies

$$\beta^{\pi}_1(T_a(S)) = I(\phi_1 + a\phi_2; \phi_2 \mid \hat{\boldsymbol{r}}) = \frac{1}{2}\log(1 + a^2) > 0.$$

Thus the role-local excess is not $\mathcal{G}$-invariant.

Every $T \in \mathcal{G}$ preserves $\sigma(S, \hat{\boldsymbol{r}})$. It therefore preserves the conditional law of $\boldsymbol{\phi}$ and each conditional marginal law of ϕ_i given $(S, \hat{\boldsymbol{r}})$. The conditional entropies in $\mathrm{TC}(\boldsymbol{\phi} \mid S, \hat{\boldsymbol{r}})$ are unchanged. □

Proof of Theorem 4.3. Use the Gaussian code and the invertible map T_a from the preceding proof with $a \neq 0$. The original code has $B^{\pi}(S) = 1$, while the transformed code has $\beta^{\pi}_1(T_a(S)) = \frac{1}{2}\log(1 + a^2) > 0$ and therefore $B^{\pi}(T_a(S)) = 0$. The two codes are informationally equivalent.

Any $\mathcal{G}$-invariant functional must take the same value on S and $T_a(S)$. Since B^{π} takes different values on this pair, no such functional can agree with B^{π} on every representation. □

Theorem B.11 (Binding non-certifiability under role supervision)**.** *The transformations in Theorem 4.3 fix* $\hat{\boldsymbol{r}}$ *and preserve the information in the full representation, so the impossibility persists when the roles are observed. Now suppose that the rendering satisfies Definition B.9, Assumption 2.2 holds, and the encoder is slot-equivariant and jointly sufficient. Then every* ϕ_i *is measurable from* $(\hat{r}_i, \hat{\phi}_i)$ *if and only if* $\boldsymbol{\phi}$ *is measurable from* $(X, \boldsymbol{r})$*. When this condition holds, the remaining ambiguity is a separate invertible reparametrization within each slot. If* $(X, \boldsymbol{r})$ *does not determine* $\boldsymbol{\phi}$*, no encoder can identify the fillers.*

Proof. The transformations in Theorem 4.3 act only on S. They fix $\hat{\boldsymbol{r}}$ and preserve $\sigma(S, \hat{\boldsymbol{r}})$ while changing the role-local binding indicator. Observing the roles therefore does not remove the impossibility.

For the second claim, slot equivariance has the form $\hat{\phi}_i = e_0(\rho_0(r_i, \phi_i))$, which is a function of X once the roles have identified the rendered components. If every ϕ_i is measurable from $(\hat{r}_i, \hat{\phi}_i)$, then $\boldsymbol{\phi}$ is measurable from $(X, \boldsymbol{r})$.

Conversely, suppose that $\boldsymbol{\phi}$ is measurable from $(X, \boldsymbol{r})$. Joint sufficiency implies $\boldsymbol{\phi} \perp\!\!\!\perp X \mid (S, \boldsymbol{r})$. Lemma B.3 then shows that $\boldsymbol{\phi}$ is measurable from $(S, \boldsymbol{r})$ and hence from $(\hat{\boldsymbol{\phi}}, \boldsymbol{r})$.

Assumption 2.2 and slot equivariance make the pairs $(\phi_j, \hat{\phi}_j)$ independent across j given $\boldsymbol{r}$. In particular, $\phi_i \perp\!\!\!\perp \hat{\boldsymbol{\phi}}_{-i} \mid (\hat{\phi}_i, \boldsymbol{r})$. Since ϕ_i is measurable from $(\hat{\boldsymbol{\phi}}, \boldsymbol{r})$, Lemma B.3 applied conditionally makes it measurable from $(\hat{\phi}_i, \boldsymbol{r})$. The same assumptions imply $(\phi_i, \hat{\phi}_i) \perp\!\!\!\perp \boldsymbol{r} \mid r_i$. Applying the lemma again reduces the arguments to $(\hat{\phi}_i, r_i)$, and exact role recovery replaces r_i by $\hat{r}_i$.

Conditional on the role, the slot is a function of the filler and the filler is a function of the slot. Two such codes differ only by separate bimeasurable maps within the slots. If $(X, \boldsymbol{r})$ does not determine $\boldsymbol{\phi}$, no function of X can determine it. □

Definition B.12 (Reader-induced binding excess)**.** A reader maps $(S, \hat{\boldsymbol{r}})$ to one output for each role. A class $\mathcal{R}$ undoes a subgroup $H \leq \mathcal{G}$ when $\{f \circ T : f \in \mathcal{R}\} = \mathcal{R}$ for every $T \in H$. Role-local, linear, and unrestricted readers undo $\mathcal{G}^{\pi}_{\text{slot}}$, $\mathcal{G}_{\text{lin}}$, and $\mathcal{G}$, respectively. For $H \leq \mathcal{G}$, define

$$B_H(S) = \inf_{T \in H} \sum_i \beta^{\pi}_i(T(S)).$$

Theorem B.13 (The invariant reader-induced excess)**.** *If $H \leq H' \leq \mathcal{G}$, then*

$$B_{H'} \leq B_H \leq \sum_i \beta^{\pi}_i, \qquad B_{\mathcal{G}^{\pi}_{\text{slot}}} = \sum_i \beta^{\pi}_i.$$

The quantity $B_{\mathcal{G}}$ is $\mathcal{G}$-invariant. More generally, B_H is $\mathcal{G}$-invariant exactly when $B_H = B_{\mathcal{G}}$. Invariance alone does not certify binding. The quantities $B_{\mathcal{G}}$ and $\mathrm{TC}(\boldsymbol{\phi} \mid S, \hat{\boldsymbol{r}})$ do not bound one another, and $B_{\mathcal{G}}$ can vanish on an entangled representation.

Proof. The larger group has a larger feasible set, so $B_{H'} \leq B_H$. The identity belongs to every subgroup, so $B_H \leq \sum_i \beta^{\pi}_i$. Every slot-respecting transformation preserves each β^{π}_i, which proves the equality for $\mathcal{G}^{\pi}_{\text{slot}}$.

For $T_0 \in \mathcal{G}$,

$$B_{\mathcal{G}}(T_0(S)) = \inf_{T \in \mathcal{G}} \sum_i \beta^{\pi}_i(TT_0(S)) = B_{\mathcal{G}}(S).$$

Conversely, suppose B_H is $\mathcal{G}$-invariant. Then

$$B_H(S) = \inf_{T \in \mathcal{G}} \inf_{T' \in H} \sum_i \beta^{\pi}_i(T'T(S)) = B_{\mathcal{G}}(S),$$

because $\{T'T : T' \in H, T \in \mathcal{G}\} = \mathcal{G}$.

It remains to compare $B_{\mathcal{G}}$ with the total correlation. For independent fair bits, place $S = \phi_1 \oplus \phi_2$ in one slot and keep the other slot constant. The parity is independent of either filler, so $B_{\mathcal{G}}(S) = 0$, whereas $\mathrm{TC}(\boldsymbol{\phi} \mid S) = \log 2$. For the other direction, place $S = \phi_1 \wedge \phi_2$ in one slot. A bimeasurable change of this two-point code can only relabel its two values and choose which slot contains them. Hence

$$B_{\mathcal{G}}(S) = \frac{3}{2}\log 2 - \frac{3}{4}\log 3 > \frac{3}{4}\log 3 - \log 2 = \mathrm{TC}(\boldsymbol{\phi} \mid S).$$

□

Proof of Corollary 4.4. Every $T \in \mathcal{G}$ preserves $\sigma(S, \hat{\boldsymbol{r}})$, so it preserves the conditional laws and entropies that define $\mathrm{TC}(\boldsymbol{\phi} \mid S, \hat{\boldsymbol{r}})$.

For finite filler alphabets and any conditional distribution q_i,

$$\mathbb{E}[-\log q_i(\phi_i \mid S, \mathbf{h}_r)] = H(\phi_i \mid S, \mathbf{h}_r) + D_{\mathrm{KL}}(P(\cdot \mid S, \mathbf{h}_r) \,||\, q_i(\cdot \mid S, \mathbf{h}_r)).$$

The minimum is $H(\phi_i \mid S, \hat{\boldsymbol{r}})$. The corresponding joint minimum is $H(\boldsymbol{\phi} \mid S, \hat{\boldsymbol{r}})$. Their difference is exactly $\mathrm{TC}(\boldsymbol{\phi} \mid S, \hat{\boldsymbol{r}})$. For jointly sufficient codes, Lemma 3.1 implies $\mathrm{TC}(\boldsymbol{\phi} \mid S, \hat{\boldsymbol{r}}) = \mathrm{TC}(\boldsymbol{\phi} \mid X, \hat{\boldsymbol{r}})$, which vanishes exactly under conditional independence. □

Proposition B.14 (ε-accurate role recovery)**.** *Replace Assumption 2.3 by $\mathbb{P}(\hat{\boldsymbol{r}} \neq \boldsymbol{r}) \leq \varepsilon$. Let N_k be the number of role tuples in their common support. For $\varepsilon \leq \frac{1}{2}$, define*

$$\delta(\varepsilon) = \min\{H_b(\varepsilon) + \varepsilon \log(N_k - 1), k\big(H_b(\varepsilon) + \varepsilon \log(|\mathcal{R}| - 1)\big)\}.$$

Then $|\mathrm{TC}(\boldsymbol{\phi} \mid S, \hat{\boldsymbol{r}}) - \mathrm{TC}(\boldsymbol{\phi} \mid S, \boldsymbol{r})| \leq (k+1)\delta(\varepsilon)$. If an information-preserving relabel is constructed using $\hat{\boldsymbol{r}}$, its effect measured against the true roles is at most $2(k+1)\delta(\varepsilon)$. Both bounds vanish with ε.

Proof. Set $\delta = \max\{H(\boldsymbol{r} \mid \hat{\boldsymbol{r}}), H(\hat{\boldsymbol{r}} \mid \boldsymbol{r})\}$. For any random variable Y,

$$H(Y \mid S, \hat{\boldsymbol{r}}) - H(Y \mid S, \boldsymbol{r}) = I(Y; \boldsymbol{r} \mid S, \hat{\boldsymbol{r}}) - I(Y; \hat{\boldsymbol{r}} \mid S, \boldsymbol{r}).$$

Replacing $\boldsymbol{r}$ by $\hat{\boldsymbol{r}}$ changes one conditional entropy by at most δ. Total correlation contains k marginal entropies and one joint entropy, so the factor is $k+1$.

Applying Fano's inequality to the full role tuple yields

$$H(\boldsymbol{r} \mid \hat{\boldsymbol{r}}),\ H(\hat{\boldsymbol{r}} \mid \boldsymbol{r}) \leq H_b(\varepsilon) + \varepsilon \log(N_k - 1).$$

Applying it coordinatewise yields

$$H(\boldsymbol{r} \mid \hat{\boldsymbol{r}}) \leq \sum_i H(r_i \mid \hat{r}_i) \leq k\big(H_b(\varepsilon) + \varepsilon \log(|\mathcal{R}| - 1)\big),$$

and the same bound holds after interchanging $\boldsymbol{r}$ and $\hat{\boldsymbol{r}}$. Thus $\delta \leq \delta(\varepsilon)$.

If T is constructed from $\hat{\boldsymbol{r}}$, then $\mathrm{TC}(\boldsymbol{\phi} \mid T(S), \hat{\boldsymbol{r}}) = \mathrm{TC}(\boldsymbol{\phi} \mid S, \hat{\boldsymbol{r}})$. Applying the first bound before and after T proves the second bound. □

B.4 Entanglement and transfer

Fix an encoder, a reader f, a target $y = t(\boldsymbol{r}, \boldsymbol{\phi})$, and a loss ℓ. The conditional risk is

$$m(\boldsymbol{r}, \boldsymbol{\phi}) = \mathbb{E}[\ell(f(S), y) \mid \boldsymbol{r}, \boldsymbol{\phi}].$$

Condition (S3) keeps this function fixed between train and test. For fixed $\boldsymbol{r}$, let

$$Q = \bigotimes_i \mathbb{P}_{\mathrm{tr}}(\phi_i \in \cdot \mid \boldsymbol{r}), \qquad m_i = \mathbb{E}_Q[m \mid \phi_i], \qquad \bar{m} = \mathbb{E}_Q[m],$$

and define

$$m_{\mathrm{int}} = m - \sum_i m_i + (k-1)\bar{m}.$$

The interaction term is the part of the risk orthogonal in $L^2(Q)$ to constants and to functions of one filler.

Definition B.15 (Entanglement modulus)**.** For a square-integrable function h of filler ϕ_i, define

$$\eta_i(h)^2 = \mathbb{E}\Big[\big(\mathbb{E}[h(\phi_i) \mid X, \boldsymbol{\phi}_{-i}, \boldsymbol{r}] - \mathbb{E}[h(\phi_i) \mid X, \boldsymbol{r}]\big)^2\Big].$$

For finite filler sets, let $\eta_i^2 = \sum_a \eta_i(\mathbf{1}\{\phi_i = a\})^2$.

The modulus is the reduction in optimal squared error obtained by observing the other fillers in addition to the observation. It depends only on the posterior. It is therefore unchanged by jointly sufficient coding and by every reparametrization in $\mathcal{G}$. Proposition C.1 relates it to the role-conditioned total correlation.

Theorem B.16 (Exact transfer identity)**.** *For fixed $\boldsymbol{r}$, suppose the training and test filler laws have densities w_{tr} and w_{te} with respect to Q. Write $\mathbb{P}_{\mathrm{tr}}^{(i)}$ and $\mathbb{P}_{\mathrm{te}}^{(i)}$ for their ϕ_i marginals and let $\Delta = \mathbb{E}_{\mathbb{P}_{\mathrm{te}}}[m] - \mathbb{E}_{\mathbb{P}_{\mathrm{tr}}}[m]$. If $m \in L^2(Q)$, then*

$$\Delta = \sum_i \int m_i \, d(\mathbb{P}_{\mathrm{te}}^{(i)} - \mathbb{P}_{\mathrm{tr}}^{(i)}) + \mathbb{E}_Q[(w_{\mathrm{te}} - w_{\mathrm{tr}}) m_{\mathrm{int}}].$$

Under a strict-substitutive shift, the first term vanishes. The gap vanishes under every such shift that is absolutely continuous with respect to Q if and only if $m_{\mathrm{int}} = 0$ almost surely. When $\mathbb{P}_{\mathrm{tr}} = Q$,

$$|\Delta| \le \|m_{\mathrm{int}}\|_{L^2(Q)} \chi^2(\mathbb{P}_{\mathrm{te}} \| Q)^{1/2}.$$

If m_{int} is essentially bounded and

$$0 < \varepsilon < \frac{\|m_{\mathrm{int}}\|_2}{\|m_{\mathrm{int}}\|_\infty},$$

then the largest absolute gap among strict-substitutive shifts satisfying $\chi^2(\mathbb{P}_{\mathrm{te}} \| Q) \le \varepsilon^2$ is $\varepsilon \|m_{\mathrm{int}}\|_2$. Finally, suppose the law of slot i depends on the fillers only through ϕ_i and the reader for role i uses only $(\hat{\boldsymbol{r}}, \hat{\phi}_i)$. Then $m_{\mathrm{int}} = 0$ for every task on that role. Such a reader has zero gap under every strict-substitutive shift, regardless of its risk.

Proof. Under Q, the interaction term has zero conditional expectation given every ϕ_i. Decompose

$$m = \bar{m} + \sum_i (m_i - \bar{m}) + m_{\mathrm{int}}.$$

If $P \ll Q$ has density w, integrating the decomposition yields

$$\int m \, dP = \bar{m} + \sum_i \int (m_i - \bar{m}) \, dP^{(i)} + \mathbb{E}_Q[w m_{\mathrm{int}}].$$

Subtracting the training expression from the test expression proves the identity. A strict-substitutive shift preserves the one-dimensional marginals, so the marginal terms cancel.

If $m_{\mathrm{int}} = 0$, every such gap is zero. For the converse, suppose $m_{\mathrm{int}} \neq 0$. Truncate it at level M, then remove its constant and one-filler projections. Denote the resulting bounded interaction by g_M. It satisfies $\mathbb{E}_Q[g_M \mid \phi_i] = 0$ for every i. For sufficiently small $\alpha > 0$,

$$w_\alpha = 1 + \alpha g_M$$

is a positive density with the same one-dimensional marginals as Q. If Q already has a nonzero gap relative to training, it is the required shift. Otherwise,

$$\mathbb{E}_Q[(w_\alpha - 1) m_{\mathrm{int}}] = \alpha \mathbb{E}_Q[g_M m_{\mathrm{int}}] > 0$$

for some M. The last inequality holds because $\mathbb{E}_Q[g_M m_{\mathrm{int}}]$ converges to $\|m_{\mathrm{int}}\|_2^2$.

When $\mathbb{P}_{\mathrm{tr}} = Q$, Cauchy–Schwarz yields

$$|\Delta| \le \|m_{\mathrm{int}}\|_2 \|w_{\mathrm{te}} - 1\|_2 = \|m_{\mathrm{int}}\|_2 \chi^2(\mathbb{P}_{\mathrm{te}} \| Q)^{1/2}.$$

Under the boundedness condition, the density

$$w = 1 + \varepsilon \frac{m_{\mathrm{int}}}{\|m_{\mathrm{int}}\|_2}$$

is positive, preserves the one-dimensional marginals, has $\chi^2(\mathbb{P}_{\mathrm{te}} \| Q) = \varepsilon^2$, and attains equality.

Under the local-slot hypothesis, the conditional risk for a single-role task depends only on ϕ_i. Its interaction term is zero, which proves the final claim. □

Proposition B.17 (The entanglement modulus bounds transfer)**.** *Assume 2.2. Let S be jointly sufficient, let $|h| \le B$, and use the train-optimal unrestricted squared-loss readout for $y = h(\phi_i)$. Let m^* be the conditional risk of the corresponding readout that also observes $\boldsymbol{\phi}_{-i}$ and write $\kappa_i(h) = \|m^*_{\mathrm{int}}\|_{L^2(Q)}$. Then*

$$\|m_{\mathrm{int}}\|_2 \le \kappa_i(h) + 4B\eta_i(h), \qquad |\Delta| \le (\kappa_i(h) + 4B\eta_i(h)) \chi^2(\mathbb{P}_{\mathrm{te}} \| \mathbb{P}_{\mathrm{tr}})^{1/2}.$$

Thus $\eta_i(h) = \kappa_i(h) = 0$ implies zero gap under every strict-substitutive shift.

Outside the bounded-target setting, let $a > 0$ and $X = \phi_1 + a\phi_2$ for independent standard Gaussian fillers. For the identity target, $\eta = a/\sqrt{1+a^2}$ and the maximal gap is $2\eta^3\sqrt{1-\eta^2}$. It increases until $\eta^2 = 3/4$ and decreases thereafter. Positive entanglement is not sufficient for a gap. If $X = \phi_1 \oplus \phi_2$ for independent fair bits, then $\mathrm{TC}(\boldsymbol{\phi} \mid X, \boldsymbol{r}) = \log 2$ while $m_{\mathrm{int}} = 0$ for every single-filler task.

Proof. Joint sufficiency lets us write the two Bayes predictions as

$$\widehat{y} = \mathbb{E}[h(\phi_i) \mid X, \boldsymbol{r}], \qquad \widehat{y^*} = \mathbb{E}[h(\phi_i) \mid X, \boldsymbol{\phi}_{-i}, \boldsymbol{r}].$$

Both lie in $[-B, B]$, and therefore

$$\left|(h - \widehat{y})^2 - (h - \widehat{y^*})^2\right| \le 4B|\widehat{y} - \widehat{y^*}|.$$

Conditional Jensen's inequality yields $\|m - m^*\|_2 \le 4B\eta_i(h)$. Since $g \mapsto g_{\mathrm{int}}$ is an orthogonal projection in $L^2(Q)$,

$$\|m_{\mathrm{int}}\|_2 \le \|m^*_{\mathrm{int}}\|_2 + \|(m - m^*)_{\mathrm{int}}\|_2 \le \kappa_i(h) + 4B\eta_i(h).$$

The transfer bound follows from Theorem B.16 because Assumption 2.2 makes $\mathbb{P}_{\mathrm{tr}} = Q$ after conditioning on the roles.

For $X = \phi_1 + a\phi_2$, the two conditional means are $X/(1 + a^2)$ and ϕ_1. Hence $\eta^2 = a^2/(1 + a^2)$. The coupling $\phi_2 = -\phi_1$ attains the maximal gap

$$\frac{2a^3}{(1 + a^2)^2} = 2\eta^3\sqrt{1 - \eta^2},$$

whose derivative changes sign at $\eta^2 = 3/4$. For $X = \phi_1 \oplus \phi_2$, the observation is independent of either filler. The optimal prediction of a single-filler target is constant, so its conditional risk depends only on that filler and $m_{\mathrm{int}} = 0$. □

B.5 Generalization and risk

Proposition B.18 (Factorization-preserving shifts)**.** *Under Assumption 2.2, every fixed encoder and fixed readout have zero generalization gap under a factorization-preserving* $\mathbb{P}_{\mathrm{te}}$.

Proof. Under $\mathbb{P}_{\mathrm{tr}}$, Assumption 2.2 factorizes the conditional filler law across roles. Under a factorization-preserving $\mathbb{P}_{\mathrm{te}}$, (S1), (S2), and (S5) specify the same role distribution and the same product of per-role filler marginals. Thus $(\boldsymbol{r}, \boldsymbol{\phi})$ has the same law under training and test. Condition (S3) keeps the rendering kernel unchanged, so the joint law of the observation, the representation, and any fixed target is also unchanged. The squared loss of any fixed readout consequently has the same expectation under $\mathbb{P}_{\mathrm{tr}}$ and $\mathbb{P}_{\mathrm{te}}$, and $\Delta = 0$. This zero-gap conclusion does not require binding-compatibility or sufficiency. □

Theorem B.19 (Gap-risk decoupling)**.** *Suppose the rendering is slot-separable and the encoder is slot-equivariant, under Assumptions 2.1 and 2.3. Then each slot has the form* $\hat{\phi}_i = h(r_i, \phi_i)$ *for a common measurable map* h*. Under Assumption 2.2,* $\beta_i = 0$ *for every* i*.*

If each filler ϕ_i *is recoverable from* (X, r_i) *and the encoder is sufficient for that filler, then* ϕ_i *is recoverable from* $(\hat{r}_i, \hat{\phi}_i)$*. Every additive task therefore has zero train and test risk under every strict-substitutive* $\mathbb{P}_{\mathrm{te}}$*, including held-out-combination shifts.*

Conversely, if some ϕ_i *is not recoverable from* $(\hat{r}_i, \hat{\phi}_i)$*, then there exists a bounded additive task whose optimal role-local risk is positive. Under every factorization-preserving shift that risk is identical on train and test, so its generalization gap is zero despite the positive risk.*

Proof. Slot-equivariance has the form $\hat{\phi}_i = e_0(\rho_0(r_i, \phi_i))$. With $h = e_0 \circ \rho_0$, this is exactly $\hat{\phi}_i = h(r_i, \phi_i)$. Under Assumption 2.2, ϕ_i is conditionally independent of ϕ_{-i} given $\boldsymbol{r}$. Because $\hat{\boldsymbol{r}}$ identifies the roles and $\hat{\phi}_i$ depends only on (r_i, ϕ_i), $I(\hat{\phi}_i; \phi_{-i} \mid \hat{\boldsymbol{r}}) = 0$. Hence $\beta_i = 0$.

Now suppose ϕ_i is recoverable from (X, r_i). Exact role identification makes it recoverable from $(X, \hat{r}_i)$ as well. Since $\hat{\phi}_i$ is a function of X, by the chain rule, $I(\phi_i; X \mid \hat{r}_i) = I(\phi_i; \hat{\phi}_i \mid \hat{r}_i) + I(\phi_i; X \mid \hat{r}_i, \hat{\phi}_i)$. Per-role sufficiency makes the final term zero. Because ϕ_i is measurable with respect to $(X, \hat{r}_i)$ and $\phi_i \perp\!\!\!\perp X \mid (\hat{r}_i, \hat{\phi}_i)$, Lemma B.3 implies that ϕ_i is measurable with respect to $(\hat{r}_i, \hat{\phi}_i)$. Thus there is a measurable map q_i such that $\phi_i = q_i(\hat{r}_i, \hat{\phi}_i)$ almost surely under $\mathbb{P}_{\mathrm{tr}}$.

Under a strict-substitutive shift, (S1)–(S3) preserve the role, the per-role filler marginal and support, and the rendering kernel. The slot-equivariant map is fixed, so the joint law of $(\phi_i, \hat{r}_i, \hat{\phi}_i)$ is unchanged.

The same recovery relation therefore holds under $\mathbb{P}_{\mathrm{te}}$. Each term $g_i(r_i, \phi_i)$ can be computed exactly from $(\hat{r}_i, \hat{\phi}_i)$, and summing these terms reconstructs any additive target with zero error under both laws.

For the converse, suppose ϕ_i is not measurable with respect to $(\hat{r}_i, \hat{\phi}_i)$. Then some measurable event A determined by ϕ_i is not determined by $(\hat{r}_i, \hat{\phi}_i)$. Take the additive task with the single nonzero term $g_i(r_i, \phi_i) = \mathbf{1}_A$. Its optimal role-local squared-error risk is $\mathbb{E}_{\mathbb{P}_{\mathrm{tr}}}[\mathrm{Var}(\mathbf{1}_A \mid \hat{r}_i, \hat{\phi}_i)]$, which is strictly positive because $\mathbf{1}_A$ is not measurable with respect to $(\hat{r}_i, \hat{\phi}_i)$. Under a factorization-preserving shift, Proposition B.18 applies to this fixed optimal readout, so the same positive risk appears at test and $\Delta = 0$. □

These results separate sufficiency from generalization. Sufficiency determines whether each slot retains the information needed for the task, whereas the generalization gap measures only how the resulting error changes from train to test. Therefore, a representation can discard task-relevant information and still have zero generalization gap if that information loss affects the two distributions equally. That is why a zero generalization gap does not, by itself, imply that the representation is sufficient or that the task has been solved correctly.

B.6 Role-local Gaussian transfer rate

Theorem B.20 (Role-local gap rate in the Gaussian family)**.** *Let $a > 0$, take independent standard Gaussian fillers under $\mathbb{P}_{\mathrm{tr}}$, and define $\hat{\phi}_1 = \phi_1 + a\phi_2$ and $\hat{\phi}_2 = 0$. For the additive task $y = \phi_1$, the train-optimal role-local readout has binding excess $\beta(a) = \frac{1}{2}\log(1 + a^2)$ and worst-case strict-substitutive generalization gap $\Delta(a) = 2a^3/(1 + a^2)^2$. The maximum is attained by the coupling $\phi_2 = -\phi_1$. Equivalently, $\Delta(\beta) = 2e^{-4\beta}(e^{2\beta} - 1)^{3/2} = 4\sqrt{2}\,\beta^{3/2} - 10\sqrt{2}\,\beta^{5/2} + O(\beta^{7/2})$ as $\beta \to 0$.*

Proof. Write $z = \phi_1 + a\phi_2$. Since the second slot is constant, every role-local predictor for $y = \phi_1$ is a function of z up to an additive constant. Under $\mathbb{P}_{\mathrm{tr}}$, $\mathrm{Cov}(\phi_1, z) = 1$ and $\mathrm{Var}(z) = 1 + a^2$, and therefore the optimal squared-error readout is $\hat{y} = wz$ with $w = 1/(1 + a^2)$. Its training risk is $R_{\mathrm{tr}} = a^2/(1 + a^2)$.

Let $c = \mathbb{E}_{\mathbb{P}_{\mathrm{te}}}[\phi_1\phi_2]$. A strict-substitutive shift preserves both standard Gaussian marginals, and so $c \in [-1, 1]$. The test risk of the fixed train-optimal readout is $R_{\mathrm{te}}(c) = 1 - 2w(1 + ac) + w^2(1 + a^2 + 2ac)$. Since $\partial_c R_{\mathrm{te}}(c) = -2aw(1 - w) < 0$, the test risk is largest at $c = -1$, which is attained by the admissible coupling $\phi_2 = -\phi_1$. Under this coupling, $z = (1 - a)\phi_1$, and hence $R_{\mathrm{te}} = a^2(1 + a)^2/(1 + a^2)^2$. Subtracting R_{tr} and simplifying yields $\Delta = 2a^3/(1 + a^2)^2$.

Under $\mathbb{P}_{\mathrm{tr}}$, z and ϕ_2 are jointly Gaussian with squared correlation $a^2/(1 + a^2)$, and therefore $\beta(a) = I(z; \phi_2) = \frac{1}{2}\log(1 + a^2)$. Thus $a = \sqrt{e^{2\beta} - 1}$, and substitution yields the exact relation $\Delta(\beta) = 2e^{-4\beta}(e^{2\beta} - 1)^{3/2}$. Expanding around $\beta = 0$ yields $\Delta = 4\sqrt{2}\,\beta^{3/2} - 10\sqrt{2}\,\beta^{5/2} + O(\beta^{7/2})$. □

Figure 1 compares this closed form with Monte Carlo estimates. The fitted log-log slope is 1.4997, against the predicted $3/2$, and the fitted intercept is 1.7296, against $\log(4\sqrt{2}) \approx 1.733$.

B.7 Optimization and attainment of the binding limit

The binding limit fixes the total correlation of a jointly sufficient representation, but attaining that limit is an optimization question. In the linear-Gaussian setting, the sufficient statistic and the population optimization problem can both be characterized explicitly.

Proposition B.21 (Linear-Gaussian attainment of the binding limit)**.** *Condition on $\boldsymbol{r}$ and let $\boldsymbol{\phi} \sim \mathcal{N}(0, I)$ and $X = G\boldsymbol{\phi} + e$, where e is independent Gaussian noise and $\Sigma_X \succ 0$. Set $B = \mathrm{Cov}(\boldsymbol{\phi}, X)\Sigma_X^{-1}$. Then $BX = \mathbb{E}[\boldsymbol{\phi} \mid X, \boldsymbol{r}]$ is jointly sufficient for $\boldsymbol{\phi}$, and every jointly sufficient representation $S = E(X)$ determines BX. Every per-filler sufficient linear code $S = WX$ is therefore jointly sufficient. Moreover, the population squared-error objective $\mathcal{L}(W) = \frac{1}{2}\mathbb{E}\|\boldsymbol{\phi} - WX\|^2$ is strongly convex with unique minimizer B, and gradient descent with any step size $0 <$*

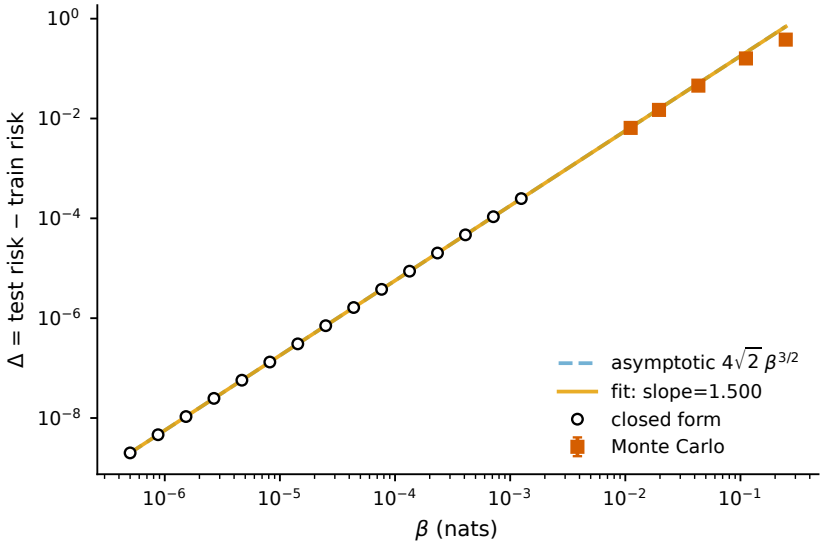


Figure 1: **Role-local gap rate.** Closed form and Monte Carlo estimate for Theorem B.20.

$\gamma < 2/\lambda_{\max}(\Sigma_X)$ *converges linearly to* B. *Consequently, the limiting representation satisfies* $\mathrm{TC}(\boldsymbol{\phi} \mid S, \boldsymbol{r}) = \mathrm{TC}(\boldsymbol{\phi} \mid X, \boldsymbol{r})$.

Proof. The Gaussian conditional law is $\boldsymbol{\phi} \mid X, \boldsymbol{r} \sim \mathcal{N}(BX, \Sigma_{\boldsymbol{\phi}|X,\boldsymbol{r}})$, with conditional covariance independent of X. Hence the posterior depends on X only through BX, and BX is jointly sufficient. Conversely, if $S = E(X)$ is jointly sufficient, then $\mathbb{E}[\boldsymbol{\phi} \mid S, \boldsymbol{r}] = \mathbb{E}[\boldsymbol{\phi} \mid X, \boldsymbol{r}] = BX$ almost surely, so BX is measurable with respect to $(S, \boldsymbol{r})$.

Now let $S = WX$ be per-filler sufficient. Since S is a function of X, per-filler sufficiency means $I(\phi_i; X \mid S, \boldsymbol{r}) = 0$ for every i. The joint law is Gaussian, so these equalities are equivalent to zero conditional cross-covariance between each ϕ_i and X. Stacking the rows yields $\mathrm{Cov}(\boldsymbol{\phi}, X \mid S, \boldsymbol{r}) = 0$, and therefore $\boldsymbol{\phi} \perp\!\!\!\perp X \mid (S, \boldsymbol{r})$. Thus the code is jointly sufficient, and Lemma 3.1 places its role-conditioned total correlation at the binding limit.

For the optimization claim, $\nabla\mathcal{L}(W) = W\Sigma_X - \mathrm{Cov}(\boldsymbol{\phi}, X)$, so the unique stationary point is $W = B$. The Hessian is $\Sigma_X \otimes I$, which is positive definite because $\Sigma_X \succ 0$. Thus $\mathcal{L}$ is strongly convex. Gradient descent satisfies $W_{t+1} - B = (W_t - B)(I - \gamma\Sigma_X)$, and $0 < \gamma < 2/\lambda_{\max}(\Sigma_X)$ makes the spectral radius of $I - \gamma\Sigma_X$ strictly smaller than one. Therefore W_t converges linearly to B from every initialization, and the limiting representation is jointly sufficient. □

Outside the linear-Gaussian setting, attainment depends on the optimization landscape. The joint sufficiency shortfall δ_{J} from Lemma B.5 connects convergence of the encoder directly to convergence of its total correlation.

Proposition B.22 (PL convergence to the binding limit)**.** *Suppose Assumption 2.2 holds,* $k \geq 2$, *and the encoder class contains a jointly sufficient representation. Let* $\delta_{\mathrm{J}}(\theta) = I(\boldsymbol{\phi}; X \mid S_\theta, \boldsymbol{r})$. *If* δ_{J} *is* L*-smooth and satisfies the Polyak–Łojasiewicz inequality* $\frac{1}{2}\|\nabla\delta_{\mathrm{J}}(\theta)\|^2 \geq \mu\,\delta_{\mathrm{J}}(\theta)$ *for some* $\mu > 0$, *then gradient descent with step size* $1/L$ *satisfies* $\delta_{\mathrm{J}}(\theta_t) \leq (1 - \mu/L)^t \delta_{\mathrm{J}}(\theta_0)$. *Moreover,* $|\,\mathrm{TC}(\boldsymbol{\phi} \mid S_{\theta_t}, \boldsymbol{r}) - \mathrm{TC}(\boldsymbol{\phi} \mid X, \boldsymbol{r})| \leq (k-1)(1 - \mu/L)^t \delta_{\mathrm{J}}(\theta_0)$.

Proof. For the update $\theta_{t+1} = \theta_t - L^{-1}\nabla\delta_{\mathrm{J}}(\theta_t)$, L-smoothness implies $\delta_{\mathrm{J}}(\theta_{t+1}) \leq \delta_{\mathrm{J}}(\theta_t) - \frac{1}{2L}\|\nabla\delta_{\mathrm{J}}(\theta_t)\|^2$. By the Polyak–Łojasiewicz inequality, $\delta_{\mathrm{J}}(\theta_{t+1}) \leq (1 - \mu/L)\delta_{\mathrm{J}}(\theta_t)$, and iteration proves the first bound (Polyak, 1963; Karimi et al., 2016).

For the total correlation, let $\varepsilon_\perp = \sum_i I(\phi_i; X \mid S_\theta, \boldsymbol{r})$ as in Lemma B.5. Since each ϕ_i is a function of $\boldsymbol{\phi}$, data processing implies $I(\phi_i; X \mid S_\theta, \boldsymbol{r}) \leq I(\boldsymbol{\phi}; X \mid S_\theta, \boldsymbol{r}) = \delta_{\mathrm{J}}(\theta)$, and therefore $0 \leq \varepsilon_\perp \leq k\delta_{\mathrm{J}}$. Lemma B.5 states that $\mathrm{TC}(\boldsymbol{\phi} \mid S_\theta, \boldsymbol{r}) - \mathrm{TC}(\boldsymbol{\phi} \mid X, \boldsymbol{r}) = \varepsilon_\perp - \delta_{\mathrm{J}}$, so this difference

lies in $[-\delta_{\mathrm{J}}, (k-1)\delta_{\mathrm{J}}]$. Since $k \geq 2$, $|\,\mathrm{TC}(\boldsymbol{\phi} \mid S_\theta, \boldsymbol{r}) - \mathrm{TC}(\boldsymbol{\phi} \mid X, \boldsymbol{r})| \leq (k-1)\delta_{\mathrm{J}}(\theta)$. Combining this inequality with the first bound proves the stated rate. □

Proposition B.23 (Explicit linear-Gaussian rate)**.** *In the setting of Proposition B.21, assume* $\Sigma_{\boldsymbol{\phi}|X,\boldsymbol{r}} \succ 0$. *Let* $\mathcal{L}(W) = \frac{1}{2}\mathbb{E}\|\boldsymbol{\phi} - WX\|^2$, *let* $\mathcal{L}^* = \min_W \mathcal{L}(W)$, *and let* $\lambda = \lambda_{\min}(\Sigma_{\boldsymbol{\phi}|X,\boldsymbol{r}})$. *Then*

$$\delta_{\mathrm{J}}(W) \leq \frac{\mathcal{L}(W) - \mathcal{L}^*}{\lambda}.$$

Gradient descent with step size $1/\lambda_{\max}(\Sigma_X)$ *satisfies*

$$|\mathrm{TC}(\boldsymbol{\phi} \mid W_t X, \mathbf{r}) - \mathrm{TC}(\boldsymbol{\phi} \mid X, \mathbf{r})| \leq (k-1)\left(1 - \frac{\lambda_{\min}(\Sigma_X)}{\lambda_{\max}(\Sigma_X)}\right)^t \frac{\mathcal{L}(W_0) - \mathcal{L}^*}{\lambda}$$

Proof. Let

$$D = \Sigma_{\boldsymbol{\phi}|WX,\boldsymbol{r}} - \Sigma_{\boldsymbol{\phi}|X,\boldsymbol{r}} \succeq 0, \qquad \Sigma = \Sigma_{\boldsymbol{\phi}|X,\boldsymbol{r}}.$$

The conditional mean based on WX cannot have greater error than the direct predictor WX. Therefore

$$\mathcal{L}(W) - \mathcal{L}^* \geq \frac{1}{2}\operatorname{tr} D.$$

Under Gaussianity,

$$\delta_{\mathrm{J}}(W) = \frac{1}{2}\log\det(I + \Sigma^{-1/2} D \Sigma^{-1/2}) \leq \frac{1}{2}\operatorname{tr}(\Sigma^{-1} D) \leq \frac{\operatorname{tr} D}{2\lambda}$$

This proves the first bound. The objective is $\lambda_{\min}(\Sigma_X)$-strongly convex and $\lambda_{\max}(\Sigma_X)$-smooth. Combining its standard gradient-descent contraction with the stability estimate in Proposition B.22 yields the displayed rate. □

Remark B.24. The same argument applies to a training objective $\mathcal{L}$ with minimum value $\mathcal{L}^\star$. Suppose $\mathcal{L}$ is L-smooth, satisfies $\frac{1}{2}\|\nabla\mathcal{L}(\theta)\|^2 \geq \mu(\mathcal{L}(\theta) - \mathcal{L}^\star)$, and dominates the shortfall as $\delta_{\mathrm{J}}(\theta) \leq \kappa(\mathcal{L}(\theta) - \mathcal{L}^\star)$ for some $\kappa > 0$. Then gradient descent on $\mathcal{L}$ with step size $1/L$ satisfies $\delta_{\mathrm{J}}(\theta_t) \leq \kappa(1 - \mu/L)^t(\mathcal{L}(\theta_0) - \mathcal{L}^\star)$, and the deviation of the total correlation from the binding limit satisfies the same bound multiplied by $k - 1$.

Invertible changes to the observation preserve the binding limit because they preserve exactly the information available about the fillers. When the inverse transformation is fixed before the trainable encoder, the optimization problem is preserved as well.

Theorem B.25 (Attainment under an invertible observation warp)**.** *Let* $z = G\boldsymbol{\phi} + e$ *satisfy Proposition B.21, and let* $X = g(z)$ *for a bimeasurable bijection* g. *Then* $\mathrm{TC}(\boldsymbol{\phi} \mid X, \boldsymbol{r}) = \mathrm{TC}(\boldsymbol{\phi} \mid z, \boldsymbol{r})$. *If the encoder uses the fixed inverse front end* g^{-1}, *with* $S_W(X) = W g^{-1}(X)$, *then its population objective, joint sufficiency shortfall, stationary points, and gradient-descent trajectory are identical to those of the unwarped problem. In particular, gradient descent converges to the same jointly sufficient representation and attains the same binding limit.*

Proof. Since g is a bimeasurable bijection, $\sigma(X, \boldsymbol{r}) = \sigma(z, \boldsymbol{r})$. The conditional laws of the fillers are therefore identical whether we condition on X or on z, and hence $\mathrm{TC}(\boldsymbol{\phi} \mid X, \boldsymbol{r}) = \mathrm{TC}(\boldsymbol{\phi} \mid z, \boldsymbol{r})$.

With the inverse front end fixed, $g^{-1}(X) = z$ almost surely, so $S_W(X) = Wz$ for every W. Therefore the warped and unwarped problems define the same representation at every parameter value. Their population objectives and joint sufficiency shortfalls coincide pointwise in W, and so their gradients, stationary points, and gradient-descent iterates coincide as well. Proposition B.21 then establishes convergence to the same jointly sufficient solution. □

Figure 2 measures the joint sufficiency shortfall δ_{J} directly. Across the raw representations, the deviation from the binding limit remains below the bound $(k-1)\delta_{\mathrm{J}}$ from Proposition B.22. With the inverse-warp front end, both quantities remain at numerical zero across all warp strengths. The largest δ_{J} is 2.7×10^{-15}, and the largest deviation from the binding limit is 1.8×10^{-15}. The warped and unwarped problems therefore agree to machine precision once the invertible warp is removed.

The remaining question is whether the same conclusion holds when a nonlinear backbone must learn the relevant representation rather than receiving an inverse warp explicitly.

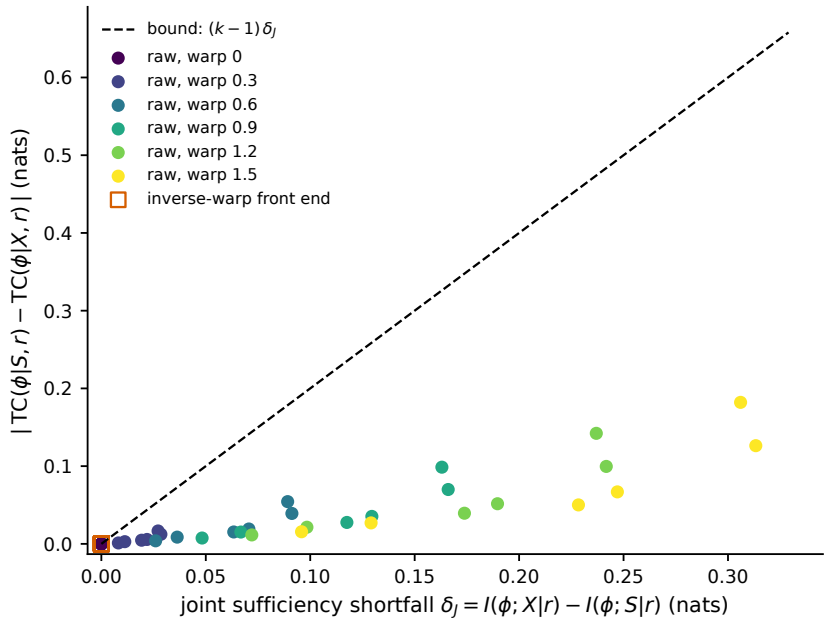


Figure 2: **Attainment of the binding limit.** Deviation from the limit against joint sufficiency shortfall.

Conjecture B.26 (Nonlinear attainment of the binding limit)**.** *For a slot-separable rendering with zero binding limit and a nonlinear encoder class that contains a jointly sufficient representation, gradient descent on the training objective of Appendix G, from an initialization drawn from the distribution used there, drives* $\delta_J(\theta_t) \to 0$*. Therefore,* $\mathrm{TC}(\boldsymbol{\phi} \mid S_{\theta_t}, \boldsymbol{r}) \to \mathrm{TC}(\boldsymbol{\phi} \mid X, \boldsymbol{r}) = 0$.

Proposition B.22 establishes this conclusion whenever the joint sufficiency shortfall satisfies a Polyak–Łojasiewicz condition. The robustness sweep of Appendix G provides empirical evidence that training reaches the binding limit across the nonlinear families tested there, but it does not establish a general property of nonlinear optimization. Known global convergence results for neural networks require additional structure such as overparameterization, while even two-layer ReLU objectives can contain spurious local minima (Du et al., 2018; Safran & Shamir, 2018).

C Estimator validation and finite-sample limits

Several empirical results in the paper depend on quantities that must be estimated from samples, including the role-conditioned total correlation, the information-based sufficiency quantities, and the entanglement modulus η. This appendix checks that the estimators recover known values in controlled settings and states the finite-sample limitation that applies to mutual-information estimation. Table 6 summarizes the validation results.

The conditional mutual informations and total correlations are estimated under the Gaussian parametric family used in the controlled experiments (Watanabe, 1960; Cover, 1999). Gaussian total correlation is computed from the log-determinant form after marginal Gaussianization, with a Wishart correction for finite-sample bias. The corresponding rate and sufficiency quantities are evaluated against the Gaussian information-bottleneck reduction (Chechik et al., 2003), with the contrastive log-ratio bound of Cheng et al. (2020) used where a variational estimate is required. Across settings in which the target quantity is available in closed form or through an independent estimator, the estimates closely match the reference values in Table 6.

The main limitation concerns nonparametric lower bounds on mutual information. Such bounds cannot in general certify arbitrarily large mutual information from a finite sample; the McAllester–Stratos bound limits them to order $\log n$ (McAllester & Stratos, 2020). At $n = 12{,}000$, this scale is

9.39 nats. Therefore a small nonparametric estimate of $\mathrm{TC}(\boldsymbol{\phi} \mid S, \boldsymbol{r})$ is not by itself a certificate of conditional independence. In the controlled Gaussian experiments, the conditional law is specified parametrically, so the total correlation is evaluated within that family.

In Section 5.1, the relabel is an exact invertible transformation, so the role-conditioned total correlation is invariant by construction. The entanglement modulus η is an L^2 second moment and is therefore not subject to the mutual-information lower-bound ceiling.

Proposition C.1 (Entanglement modulus and total correlation)**.** *Suppose the filler sets are finite, and let $q_{\min} > 0$ bound from below every positive value of $P(\phi_i = a \mid X, \boldsymbol{r})$, almost surely and for all i, a. Then*

$$\frac{1}{2} \max_i \eta_i^2 \le \mathrm{TC}(\boldsymbol{\phi} \mid X, \boldsymbol{r}) \le q_{\min}^{-1} \sum_{i=2}^{k} \eta_i^2.$$

For general fillers and bounded h, $\mathrm{TC}(\boldsymbol{\phi} \mid X, \boldsymbol{r}) \ge \eta_i(h)^2/(2\|h\|_\infty^2)$. Every jointly sufficient code satisfies the same bounds with X replaced by that code.

Proof. Let $p = P(\phi_i \in \cdot \mid X, \boldsymbol{\phi}_{-i}, \boldsymbol{r})$ and $q = P(\phi_i \in \cdot \mid X, \boldsymbol{r})$. Then $I(\phi_i; \boldsymbol{\phi}_{-i} \mid X, \boldsymbol{r}) = \mathbb{E}[\mathrm{KL}(p\|q)]$ and $\eta_i^2 = \mathbb{E}\|p - q\|_2^2$. Pinsker's inequality implies $\mathrm{KL}(p\|q) \ge \frac{1}{2}\|p - q\|_2^2$. Moreover,

$$\mathrm{TC}(\boldsymbol{\phi} \mid X, \boldsymbol{r}) - I(\phi_i; \boldsymbol{\phi}_{-i} \mid X, \boldsymbol{r}) = \mathrm{TC}(\boldsymbol{\phi}_{-i} \mid X, \boldsymbol{r}) \ge 0,$$

which proves the lower bound. For the upper bound, $\mathrm{KL}(p\|q) \le \chi^2(p\|q) \le q_{\min}^{-1}\|p - q\|_2^2$. By the chain rule,

$$\mathrm{TC}(\boldsymbol{\phi} \mid X, \boldsymbol{r}) = \sum_{i=2}^{k} I(\phi_i; \phi_{1:i-1} \mid X, \boldsymbol{r}) \le q_{\min}^{-1} \sum_{i=2}^{k} \eta_i^2.$$

For bounded h, Pinsker's inequality also implies $|\mathbb{E}_p h - \mathbb{E}_q h| \le \|h\|_\infty (2\,\mathrm{KL}(p\|q))^{1/2}$. Squaring and averaging proves the stated lower bound. Joint sufficiency preserves the joint and marginal posteriors, so the same argument applies to the code. □

Quantity	Reference	Error	Rank correlation
Gaussian total correlation	closed form and nearest-neighbour estimate	0.005 nats (max.)	1.000
Categorical token estimate	closed form and estimator-family swap	0.0014	0.92
Entanglement modulus η	closed form and independent estimator	0.0012 (relative)	0.99984
Trained-model leakage	estimator-family swap	—	0.949 (SCAN), 0.906 (COGS)

Table 6: **Estimator validation against reference values.**

D CAUSAL INTERVENTION PROTOCOL ACROSS MODEL FAMILIES

Tables 1 and 2 report the causal intervention results. This appendix records the held-out intervention procedure, the architecture-specific choices needed to reproduce those results, and the pattern obtained across model families.

Held-out intervention protocol. FIT estimates the centered class-mean span, whose rank is fixed at $C - 1$. Where candidate sites are compared, a disjoint SELECT split chooses only the site and layer, or the site and candidate label, by target flip minus the mean of five same-rank controls. Slot attention, CLIP, and the generators have fixed intervention sites and no SELECT split. The frozen configuration is scored once on disjoint EVAL data against twenty same-rank random subspaces. The fit, selection, and evaluation examples are therefore disjoint. This separation is important because conclusions from activation patching can depend on the localization metric and intervention procedure (Zhang & Nanda, 2024).

Let P denote the orthogonal projector onto the selected subspace. For two role representations E_0 and E_1, we interchange only their projected components,

$$T(E_0, E_1) = (E_0 - PE_0 + PE_1,\ E_1 - PE_1 + PE_0).$$

The map satisfies $T \circ T = \mathrm{id}$, so applying it twice recovers the original representation. It is therefore an information-preserving relabel in $\mathcal{G}$. Because it preserves $\sigma(S, \hat{\boldsymbol{r}})$, the role-conditioned total correlation is unchanged (Lemma 4.2). Flip is measured only on directions whose unmodified answer is correct, and it requires the intervened answer to become the specified foreign filler. The controls use twenty random subspaces of the same rank. Float64 involution errors range from roughly 10^{-15} to 10^{-12}; the relative round-trip errors in the reduced-precision model tensors are about 10^{-4} to 10^{-3} for the pretrained networks.

Pretrained language and multimodal models. SELECT chooses the layer and token position separately for each pretrained model. It selects the answer position for every Pythia and Qwen2.5 model, Llama-3.1-8B, Llama-3.2-1B, and every Mamba model. Llama-3.2-3B instead selects an entity position, although that run lies below the binding-accuracy floor. For the vision-language models, SELECT chooses a late layer at the answer position. The intervention therefore follows each model's native binding readout instead of imposing one common site across architectures.

Architectures with an explicit binding mechanism. The same protocol is applied to tensor-product representations (Smolensky, 1990), Kuramoto-oscillator binders (Miyato et al., 2025), slot attention, the emergent-symbol binding network (Webb et al., 2020), and the relational-bottleneck abstractor (Webb et al., 2024; Goyal et al., 2019; 2021). The latter two models are trained on the role–filler task with twelve of the sixty-four role–filler pairs held out. The intervention acts on ESBN's encoded pair representations and on the abstractor's encoder filler-position states. FIT estimates the candidate class-mean spans. SELECT chooses the site and label and selects the rank-three role-label span in all six runs. EVAL scores the frozen choice once. The tensor-product and oscillator experiments use the same split and intervene at the selected encoder block. Table 1 reports all qualifying runs.

Text-to-image generators. For the generator experiments, the intervention acts on the text conditioning rather than on the generated image. Prompts have the form "a c_1 circle and a c_2 square." FIT computes the rank-seven class-mean span of the eight colours separately in every token-level text stream, and the same involution interchanges the two tokens' projections. The noise seed is shared across the base, target, round-trip, and twenty control generations for each evaluation prompt. The comparison therefore changes the conditioning intervention while holding the sampling randomness fixed.

Each condition is evaluated on $N = 96$ prompts in the attribute-binding format used by compositional text-to-image benchmarks (Huang et al., 2023). The reported endpoint is the majority vote of Qwen2-VL-2B, Qwen2-VL-7B, and LLaVA-1.5-7B, with unreadable images counted as non-flips. CLIP is not used as a reader because it is itself one of the conditioning systems under study. For DeepFloyd IF, the relabeled T5 sequence conditions both denoising stages. Separate reader results agree on the relabel–control separation. Table 2 reports the base binding accuracy and intervention scores.

Across the runs that meet the protocol's binding-accuracy criterion, the relabel changes the model's own answer far more often than a same-rank random relabel. The qualifying attention language models flip on 0.733–1.000 of eligible directions, with control means at most 0.011. The four larger Mamba models flip on 0.776–0.917, with controls at most 0.015, and the vision-language models flip on 0.740–1.000 with zero control. Slot attention reaches 0.991–0.993 against controls of at most 0.012; ESBN reaches 0.932–0.970 against 0.018–0.021; and the abstractor reaches 0.999–1.000 against at most 0.001. On the held-out role–filler task, the qualifying tensor-product and oscillator models reach 0.943–0.995 and 0.955–0.999, respectively. The four generators with reliable base binding reach 0.606–0.642 against controls no larger than 0.018; DeepFloyd IF, whose base accuracy is only 0.510, reaches 0.422 against 0.018. These comparisons support the same scoped conclusion across architectures: when a model already binds, its native reader depends on the coordinates in which that binding information is expressed.

The scoring rule is shared across families: FIT never uses EVAL data, the intervention exchanges only the fitted span, invertibility is checked numerically, and EVAL compares the target with twenty equal-rank controls. Families differ in whether a SELECT split is needed and, for role–filler models, whether SELECT compares role and filler labels.

E TRAINED MODELS AND TRAINING SETTINGS

The token-task experiments use seven encoder–decoder architectures with a shared decoder. Table 7 lists their binding mechanisms and sizes. The encoders include vanilla and tensor-product Transformers (Vaswani et al., 2017; Schlag et al., 2019), two slot-attention variants (Locatello et al., 2020b), a capacity-starved control, a Kuramoto oscillator (Miyato et al., 2025), and a bidirectional LSTM. Table 8 gives the training settings. The held-out splits have no input overlap with training, and the datasets are checked against their canonical SHA-256 digests. Before role conditioning, a fixed Hungarian matching assigns permutation-symmetric slots to the ground-truth roles.

Encoder	Binding mechanism	Trainable params
Vanilla Transformer	self-attention	1.39M
Tensor-product Transformer	per-head role binding	1.54M
Slot attention (16 slots, 3 iters)	iterative softmax-over-slots	1.58M
Slot attention + anchor	slot binding, symmetry broken	1.58M
Capacity-starved object-centric	2 layers, 1 slot iteration	1.25M
Oscillator (Kuramoto, 2 groups, 4 steps)	synchronization	1.46M
Bidirectional LSTM	recurrent memory	0.997M

Table 7: **Token-task encoder configurations.** Shared decoder, pre-layer-norm, hidden dimension 128, eight heads, feed-forward dimension 512, three encoder and three decoder layers, dropout 0.1.

Setting	Value
Optimizer	Adam, $\beta_1 = 0.9$, $\beta_2 = 0.98$
Peak learning rate / schedule	10^{-3}, 4000 warmup steps, inverse-square-root decay
Batch size / gradient clip	256 / L2 norm 1
SCAN steps / floor	40k / exact-match 0.97 (primitive-jump, length splits held out)
COGS, CFQ steps / floor	60k / 0.95 (generalization, MCD splits held out)
Model selection	best held-out-accuracy checkpoint
Seeds per configuration	≥ 20

Table 8: **Training and seed settings** (Lake & Baroni, 2018; Kim & Linzen, 2020; Keysers et al., 2019).

The two architectures designed for variable binding are trained from scratch on the same role–filler task with twelve of the sixty-four pairs held out. The emergent-symbol binding network (Webb et al., 2020) stores encoded pairs in external memory, while the relational-bottleneck abstractor (Webb et al., 2024) represents fillers at its encoder states. FIT estimates the candidate class-mean spans. SELECT chooses the site and label and selects the rank-three role-label span in all six runs. EVAL scores the frozen choice once. ESBN flips on 0.932–0.970 of eligible directions against controls of 0.018–0.021, and the abstractor flips on 0.999–1.000 against controls no larger than 0.001. All six runs are decisive. Table 9 gives the architectures.

Architecture	Binding mechanism	Read site	Trainable params
ESBN (Webb et al., 2020)	key–value memory	encoded pair memory values	82K
Relational-bottleneck abstractor (Webb et al., 2024)	relational cross-attention	encoder filler-position states	110K

Table 9: **Binding-architecture configurations.** Each model has hidden dimension 64 and is trained for 8k steps with Adam at 3×10^{-4} and batch size 256. Training uses a synthetic task with four roles, sixteen fillers, three pairs per episode, and twelve held-out role–filler pairs. Table 1 reports results from the held-out protocol. ESBN uses an LSTM controller and external key–value memory. The abstractor uses a two-layer, four-head Transformer encoder whose query position retrieves the bound filler through a relation-weighted pointer.

The perceptual experiment uses CLEVR and CLOSURE (Johnson et al., 2017; Bahdanau et al., 2019) with a convolutional slot front end. Slots are read at the ground-truth pixel positions, and the foreign

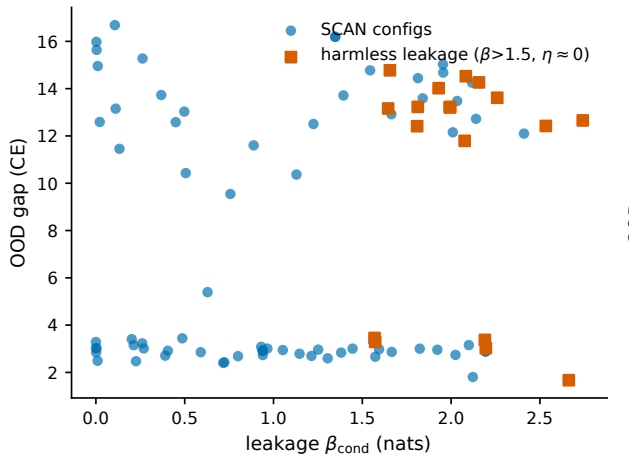


Figure 3: **At fixed rendering, the gap does not track leakage.** On SCAN, leakage spans both gap bands, including high-leakage configurations with near-zero entanglement and near-perfect accuracy.

filler is the nearest object's colour. We vary separability by mixing the slots from cleanly separable to entangled while keeping the front end fixed, as in the linear setting of Theorem B.20. We measure the maximal correlation between the own and foreign fillers and subtract the corresponding slack.

Visual Genome de-occlusion. We test whether reducing occlusion in the Visual Genome training images reduces the generalization gap. At dose $D \in [0, 1]$, each target object is selected for relocation with probability D. For each selected object, we sample forty candidate locations uniformly and choose the one with the smallest intersection over union with the other annotated objects. We paste the original object crop at that location and fill its former bounding box with the image's median RGB colour, removing evidence of the original position. This procedure introduces visible artifacts and removes semantic context, so improved visual coherence cannot explain the smaller gap. Because de-occlusion also changes absolute position, we compare it with overlap-preserving relocation. The control reduces the gap from 0.081 to 0.041–0.063, whereas de-occlusion reaches about 0.014 at every dose. The larger reduction under de-occlusion shows that the positional change alone does not explain the full reduction.

F Separating coordinate leakage from rendering entanglement

The main results distinguish two quantities that can otherwise appear together. Role-local leakage depends on the coordinates chosen for a representation, whereas role-conditioned total correlation depends on the information available in the observation. This appendix tests that distinction with controlled renderings, a factorial design, natural images, and fixed-rendering interventions. By varying the representation and the rendering separately, the experiments determine whether a change in coordinate leakage reflects a change in the entanglement imposed by the data.

F.1 Controlled changes in entanglement

Token tasks. On SCAN over ninety configurations the rank correlations of the gap with β and with η are 0.105 and 0.099, and the gap is split-driven, larger on primitive-jump than on length. On COGS, the correlations are 0.501 for β and 0.368 for η. Because both quantities vary with the split, this experiment does not isolate the proposed dissociation. Nineteen configurations carry a conditioned leakage above 1.5 nats with conditioned entanglement below 0.05 at accuracy near one. The collinearity between leakage and entanglement is 0.32 on SCAN and 0.48 on COGS, so they are distinct quantities.

Perception. The composite primary across all five mechanisms gives rank correlation 0.983, interval $[0.860, 1.000]$. At the largest entanglement value, the identity-reader gap is 0.659, and every other

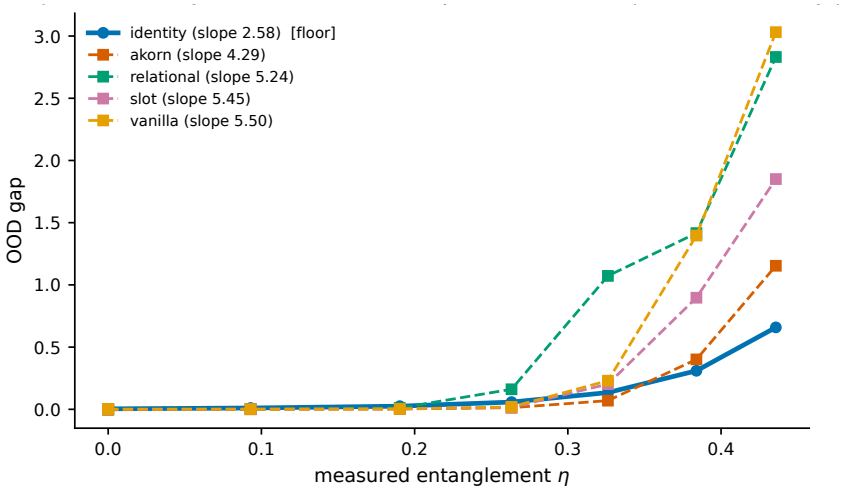


Figure 4: **Across renderings, the gap rises with entanglement.** The gap has rank correlation 0.983 with η, and all measured mechanisms lie at or above the identity-reader curve.

mechanism has a gap at least this large. The log–log slopes are 2.58 for the identity reader, 4.29 for the oscillator, 5.24 for the relational Transformer, 5.45 for slot attention, and 5.50 for the vanilla Transformer. Superimposed multi-MNIST adds two digit images placed at separate positions. As the digits move from separated to concentric, η increases from 0 to 8.351, the gap increases from 0.072 to 1.042, and their rank correlation is 0.976 with interval $[0.882, 1.000]$ (Figure 7a). An earlier experiment with alpha-blended solid colours did not show this relationship because the colours remained linearly separable after blending. We preregistered the replacement with multi-MNIST and report both outcomes. These controlled interventions establish that the gap increases with η. The remaining experiments test whether training co-occurrence explains the association and whether it also holds for naturally occurring entanglement.

F.2 Separating entanglement from co-occurrence

We test whether training co-occurrence explains the association in two ways: a controlled factorial experiment and an overlap-preserving relocation control on natural images. Neither shows an independent co-occurrence effect.

Factorial experiment. An observational dataset cannot distinguish the association with η from position–colour co-occurrence when the two variables covary. We therefore use a preregistered factorial experiment with two independently varied quantities: occlusion, which determines η, and shortcut strength c. With probability c, an object's colour is determined by its horizontal position, $\text{colour} = \lfloor c_x V/\text{IMG} \rfloor$; otherwise it is sampled uniformly. Colour is independent of position at $c = 0$ and fully determined by position at $c = 1$. Because the two objects have fixed separation, position can also predict the other object's colour. Under the marginal-preserving held-out-pair split, the marginal distribution of the target colour is unchanged, but its conditional distribution given the foreign colour differs between training and test. The factorial experiment tests whether this shortcut, rather than η, explains the gap.

Across the forty-nine cells, the rank correlation between η and c is $+0.0105$, compared with 0.399 between η and object count in Visual Genome. Thus η and c vary almost independently in this experiment. On a coarse five-by-five grid the gap's partial rank correlation with η, holding c fixed, is $+0.635\,[+0.27, +0.91]$; on a dense seven-by-seven grid, away from the degenerate fully-coupled column, it is $+0.900\,[+0.808, +0.951]$, decisive. Holding η fixed, co-occurrence is inert: its in-regime partial is -0.063 with an interval straddling zero, so once the entanglement is known the training pairing of colour with position adds nothing to the gap. The full-grid co-occurrence partial is $+0.357$, but this increase occurs entirely in the degenerate $c = 1$ column. In that column, position

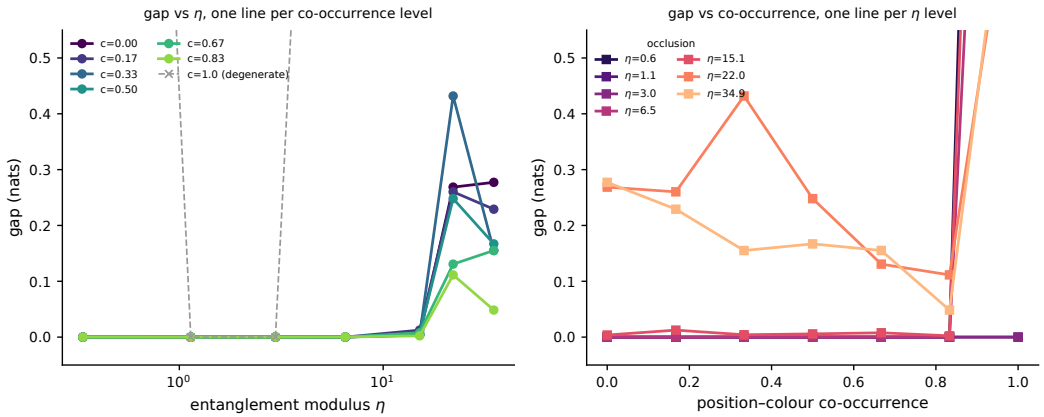


Figure 5: **Gap as a function of entanglement and co-occurrence.** For $c < 1$, the gap rises across η levels and remains nearly constant across co-occurrence levels, while the fully coupled column also changes the rendering.

determines colour and also increases rendering entanglement, so the comparison does not isolate a memorisation effect. Excluding $c = 1$ returns the in-regime value of -0.063.

The relationship is thresholded rather than monotone. The separable bin and the $o = 0$ cells at $\eta = 0.6$ have gaps near zero. The gap remains near zero at moderate η and increases when occlusion becomes large (Figure 5). At the highest value, the mean gap decreases from 0.24 at $\eta = 22.0$ to 0.17 at $\eta = 34.9$ because the target object becomes difficult to observe. Visual Genome and COCO show the same behavior under near-total occlusion. The verdict therefore relies on the partial correlation and the separable condition; monotonicity is claimed only below the total-occlusion regime. We use the same measurements as in the other experiments: η is the held-out L^2 residual attributable to the foreign filler, and the gap is held-out-combination transfer. Both are computed from the same SlotCNN front end.

Overlap-preserving relocation (real photographs). We use Visual Genome to vary absolute position while preserving the mutual occlusion of each object–neighbour pair. At dose $d \in \{0, \frac{1}{3}, \frac{2}{3}, 1\}$, each selected object and its most-occluding different-colour neighbour are translated together to a uniformly sampled location. This rigid translation preserves their relative offset and bounding-box overlap. Each pair moves at most once, and vacated boxes are filled with the image median. The intervention changes absolute position and position–colour correlation while preserving the pair's occlusion. By contrast, the de-occlusion intervention of Appendix E changes both position and occlusion. After each dose, we recompute occlusion and the foreign-neighbour assignment, train a SlotCNN front end with the same Visual Genome recipe, freeze it, and measure η and the gap.

Pooled occlusion changes only from 0.505 to 0.543, and η remains between 1.29 and 1.50. By comparison, de-occlusion reduces occlusion from 0.50 to 0.29. Under relocation, the gap changes from 0.081 at baseline to 0.041 at dose $\frac{1}{3}$ and 0.063 at dose 1. These values are three to five times larger than the value of about 0.014 under de-occlusion. The partial correlation $\rho(\text{gap}, \eta \mid \text{count})$ remains between $+0.762$ and $+0.874$ across doses (Table 10, Figure 6). Overlap-preserving relocation therefore reduces but does not close the gap, and the positional change alone does not explain the full reduction.

F.3 Natural-image results

Visual Genome. Controlled datasets manipulate η, whereas Raw-CLEVR and Visual Genome measure naturally occurring occlusion. Raw-CLEVR gives a moderate estimate of 0.604 with interval $[0.194, 0.866]$ over thirty-nine thousand objects, but after re-binning its estimate is 0.299 and the interval includes zero. Visual Genome provides the stronger natural-image test.

Table 10: **Overlap-preserving relocation changes position while keeping occlusion approximately fixed.** Each pair is translated rigidly, preserving its mutual overlap. Across doses, pooled occlusion ranges from 0.505 to 0.543, η ranges from 1.287 to 1.498, and the gap remains three to five times larger than the value of about 0.014 under de-occlusion. The partial correlation $\rho(\text{gap}, \eta \mid \text{count})$ ranges from $+0.762$ to $+0.874$.

Dose d	Mean occlusion	Pooled η	Pooled gap	Cell law $\rho(\text{gap}, \eta \mid \text{count})$
0	0.505	1.287	0.0811	$+0.874$
1/3	0.529	1.498	0.0414	$+0.828$
2/3	0.538	1.413	0.0565	$+0.808$
1	0.543	1.387	0.0632	$+0.762$

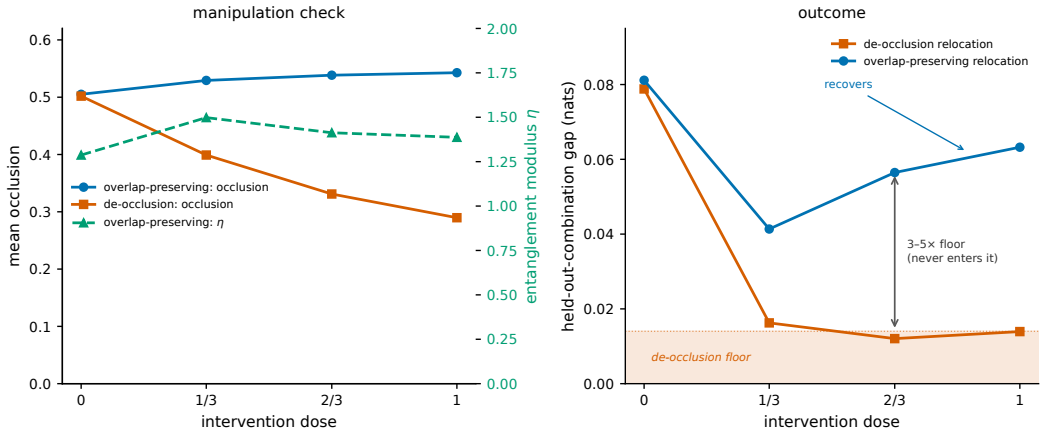


Figure 6: **De-occlusion reduces the gap more than overlap-preserving relocation.** Relocation keeps occlusion approximately fixed and leaves a gap of 0.041–0.063, whereas de-occlusion lowers the gap to about 0.014.

We train a convolutional slot front end to predict object colour from the feature map at each object's bounding-box centroid and then freeze it. For each object, the foreign filler is the colour of its most-occluding neighbour with a different colour. Occlusion is their box overlap, η is the held-out residual attributable to the foreign colour, and the gap is measured by a frozen linear readout trained on one half of the foreign colours and evaluated on the other. The analysis uses a five-by-four grid of occlusion and object-count cells, each containing at least three hundred objects.

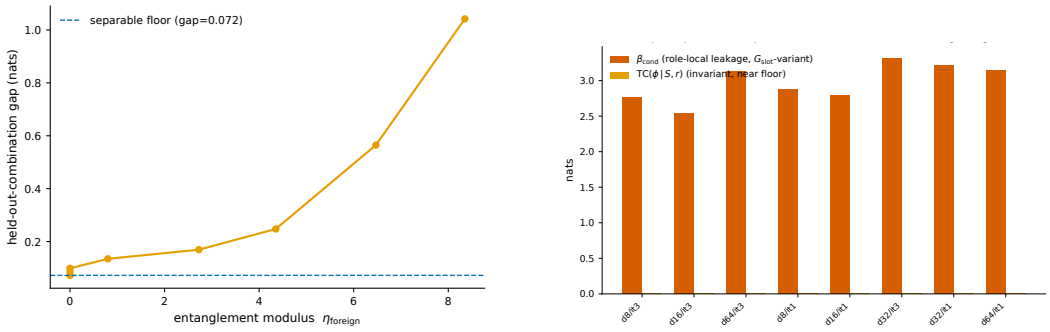
Across 861,516 objects in 97,860 photographs, the annotated colours are nearly independent across neighbouring objects ($\rho_{\max} = 0.036$). On a fifty-thousand-image sample with 439,336 objects, the count-controlled partial rank correlation between the gap and η is 0.752, with interval $[0.571, 0.850]$ and $\rho_{\max} = 0.041$. The gap rises across the five occlusion bins: $0.019, 0.050, 0.076, 0.101, 0.120$. The association with η is distinct from object count, whose correlation with η is 0.399.

The estimate remains stable under re-binning, endpoint exclusion, and leave-one-cell-out analysis; no omitted cell changes it by more than 0.15.

F.4 Total correlation and coordinate leakage

Capacity sweep. Across eight slot-attention configurations, held-out accuracy ranges from 0.90 to 0.994, and every configuration exceeds the binding threshold. Role-conditioned total correlation ranges from 0.001 to 0.006 nats and remains close to the rendering value, whereas coordinate leakage ranges from 2.54 to 3.31 nats (Figure 7b).

Fixed-rendering relabel. At the largest mixing angle, the relabel changes leakage from 0.038 to 0.609 nats. Total correlation and the best-reader gap change by only 9.5×10^{-9} and 4.5×10^{-10},



(a) Superimposed multi-MNIST. (b) Trained perceptual slots.

Figure 7: **Perceptual experiments.** The multi-MNIST gap rises with η, while the slot-attention sweep changes coordinate leakage by 0.77 nats with total correlation remaining between 0.001 and 0.006 nats.

respectively, while η is unchanged. Across the angle sweep, leakage and η are correlated at 0.99, so their separate correlations with the gap cannot distinguish them. The fixed-rendering relabel does: only leakage changes.

G Attaining the binding limit

The binding limit is fixed by the observation, but a trained encoder need not attain it. This appendix examines that optimization question. We train codes under a sufficiency objective across controlled linear and nonlinear families, then compare their role-conditioned total correlation with the value predicted from each observation. The sweep tests when training reaches the theoretical limit without conflating that limit with coordinate leakage.

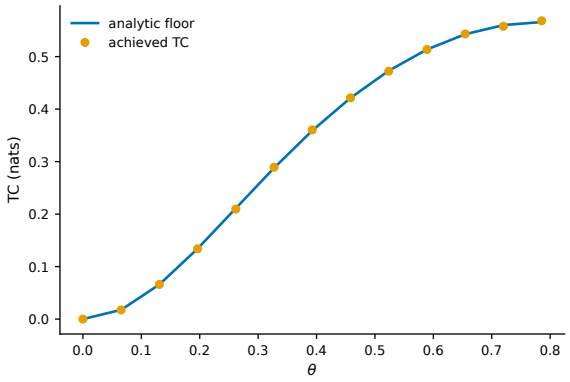


Figure 8: **Binding limit with observed ground truth.** The trained sufficient code matches the closed-form value with correlation 0.99998 and maximum deviation 0.0024.

We train a plain backbone without slot structure using a rate penalty and a per-filler sufficiency objective. We measure total correlation, the noise term, and the worst-case substitutive gap on matched separable and mixing renderings, together with a noise-only ablation. In the linear-Gaussian family, the objective is convex and gradient descent reaches a gap of 9.3×10^{-9}, consistent with Proposition B.21. On the nonlinear family, the separable rendering reaches the floor, with total correlation 2×10^{-4} and sufficiency preserved on held-out data. The mixing rendering reaches 0.5465 nats against the closed-form value 0.5659, a relative error of 0.034. By Lemma 3.1, the 0.0194-nat deficit requires at least 0.0194 nats of lost joint filler information, which is larger than the greatest estimator error of 0.005 nats in Table 6. No run reduces the total correlation further while preserving joint sufficiency. In these runs, reconstruction gap is not a reliable proxy for total

correlation: the two quantities have correlation -0.97 on the mixing rendering, and the reconstruction gap vanishes on the separable rendering.

The sweep contains 1128 configuration–seed records with paired noise controls. It varies the filler dimension, slot count, rendering anisotropy, mixing angle, noise, optimization settings, and encoder size. Every configuration has at least twenty seeds. On mixing renderings, the maximum relative error from the predicted floor is 0.0258 and the ninety-fifth percentile is 0.0073; on separable renderings, the ninety-fifth-percentile total correlation is 6×10^{-5}. These results establish robustness across the tested family, not a general nonlinear convergence theorem.

H IMPLICATIONS

The role-conditioned total correlation provides a coordinate-invariant way to compare representations across renderings. Its empirical value depends on reliable conditional-information estimates, so finite-sample bias remains an important limitation (McAllester & Stratos, 2020). Variational and contrastive bounds may improve these estimates, but they do not by themselves certify conditional independence (Belghazi et al., 2018; Oord et al., 2018; Poole et al., 2019; Cheng et al., 2020). The same distinction may also clarify neural accounts of binding: synchrony and interaction-information measurements become meaningful only after the represented roles and the reader are specified (Von Der Malsburg, 1994; Singer & Gray, 1995; Treisman, 1996; Fries, 2005). Across these settings, the central question is whether the observation preserves separable information and which reader can use it.