FROM COOPERATION TO WINNER-TAKES-ALL: IDENTIFIABLE SLOT DISENTANGLEMENT UNDER MAX-POOLING VIA ANNEALED LOG-SUM-EXP

Anonymous authors
Paper under double-blind review

ABSTRACT

Unsupervised disentangled representation learning enables interpretable concept discovery in domains ranging from images to biological systems. Recent advances establish provable identifiability, but they rely on strict structural assumptions like sparsity or the linear independence of first- and second-order derivatives in additive decoders. Violating these conditions can degrade model performance in real-world settings. For example, datasets with overlapping objects or missing colour channels can cause a single representational slot to cover multiple objects. We study max-pooled decoders to account for scenarios where concepts mask one another. We also provide a formal proof of identifiability for these generative processes requiring fewer smoothness assumptions than previous work. We propose a temperature-annealed Log-Sum-Exp (LSE) pooling framework, which approaches the identifiable max pooling with better optimisation behaviour. At low inverse temperature, this function behaves as an easily optimisable additive decoder. As the inverse temperature increases, it smoothly converges to strict max-pooling, where our guarantees hold. This mechanism dynamically routes backward gradients to the specific representational slots most responsible for the reconstruction. On synthetic datasets with occlusion, our annealed LSE decoders outperform standard additive models and recover the latents better than naive max-pooling. Our theorems also cover the softmax masks of object-centric decoders in their max limit, and on grayscale CLEVR6 such decoders separate objects more reliably than a joint decoder. Finally, we demonstrate that annealing improves the alignment of latents with annotated cell types in six of eleven single-cell transcriptomics atlases.

1 INTRODUCTION

Recovering unobserved sources of variation behind an observed mixture is a classical problem. Disentanglement applications range from separating voices in a recording to decomposing objects in a scene or discovering biological processes in cellular populations. However, unsupervised disentanglement is impossible in general. Without proper inductive bias in the data and model, a nonlinear mixing of independent sources can be undone in infinitely many ways under a rotation-invariant prior, resulting in the same model loss (Hyvärinen & Pajunen, 1999; Locatello et al., 2018). Three general approaches overcome this challenge: changing the prior, using supervised signals (Khemakhem et al., 2019), or relying on inductive bias to limit the search space to the intended solution. This paper focuses on the third approach.

Architectures for object-centric learning already encode such biases. For instance, convolutional decoders tie a latent to a local region of the image (Watters et al., 2019). Similarly, slot attention allocates exchangeable slots that compete for pixels (Locatello et al., 2020b), while additive decoders let each slot write its own image and sum the results (Lachapelle et al., 2023). For several of these models, the inductive bias provably provides identifiability. Specifically, additive decoders identify their slots up to permutation and per-slot reparametrisation under a condition on the first- and second-order derivatives of the per-slot maps (Lachapelle et al., 2023). Compositionality and generator irreducibility provide the same conclusion for another class of decoders (Brady et al., 2023). Furthermore, asymmetry between within-object and between-object interactions provides

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

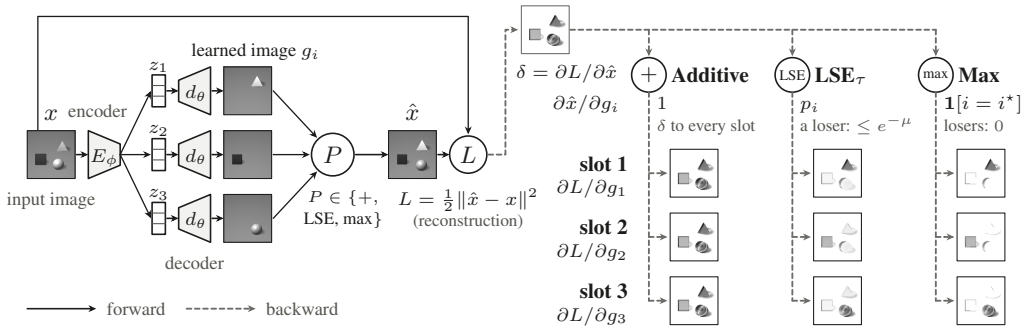


Figure 1: *Pooling decides which slot learns from each pixel.* The encoder E_ϕ maps the input image x to one latent z_i per slot, per-slot decoders d_θ produce one image g_i from each latent alone, and pooling combines the g_i into a learned image $\hat{x}$. *The backward pass.* The error $\delta = \partial L/\partial \hat{x}$ reaches slot i at feature j as $\partial L/\partial g_{i,j} \propto \partial \hat{x}_j/\partial g_{i,j}$. The additive decoder passes the same gradient to every slot (cooperation), the log-sum-exp (LSE) passes the softmax weight $p_{i,j}$ of equation 2, and max-pooling passes everything to the winner and nothing to the rest.

another route to provable identifiability (Brady et al., 2024). In temporal data, assuming every observed variable has a single latent parent identifies both the latents and their causal graph (Brouillard et al., 2024). These theoretical results help users select models whose assumptions hold and give researchers design ideas.

However, these assumptions remain hard to meet or verify in certain domains. Overlapping objects are the clearest example. Two objects at the same pixel do not add, because the nearer object hides parts of the farther one. Thus, a sum cannot express the scene. Models handle occlusion outside this theory by compositing objects in depth order with a learned mask (Yuan et al., 2019; Engelcke et al., 2019). The field has used this design for years (Yuan et al., 2023), yet identifiability results either exclude occlusion (Lachapelle et al., 2023; Brady et al., 2023) or cover partial occlusion only for differentiable decoders (Matthes et al., 2025). Grayscale images present another example. Colour is a strong cue for how slots divide a scene. For instance, the interaction asymmetry model performs well on colour images but becomes unreliable on their grayscale equivalents.

Here, we investigate the pooling operator rather than the architecture. Replacing the sum by a maximum over slots yields a decoder that directly supports occlusion. We prove that such a decoder is identifiable up to permutation and per-slot local homeomorphism under simple assumptions, which is a reparametrisation when slots are one-dimensional or the latents vary independently. Previous results assume a differentiable decoder, and several also need second-order conditions, whereas this proof needs neither, making it widely applicable. Nevertheless, optimising a max-pooled decoder is difficult because the gradient is zero for all but the winning slot. To solve this, we parametrise the pooling. Writing the decoder as a log-sum-exp of the per-slot outputs at inverse temperature τ creates a continuous family:

$$f_{\tau,l}(\mathbf{z}) \;=\; \frac{1}{\tau}\log\frac{1}{k}\sum_{i=1}^{k} e^{\tau g_{i,l}(z_i)}, \qquad f_{\tau,l} \xrightarrow[\tau\to 0]{} \frac{1}{k}\sum_{i=1}^{k} g_{i,l}(z_i), \qquad f_{\tau,l} \xrightarrow[\tau\to\infty]{} \max_i g_{i,l}(z_i). \tag{1}$$

This family contains an additive decoder at $\tau \to 0$ and a maximum at $\tau \to \infty$. Soft-to-hard vector quantisation (Agustsson et al., 2017) and Gumbel-softmax training (Jang et al., 2016) already use log-sum-exp as a differentiable maximum. We evaluate this approach on two toy datasets, the CLEVR6 benchmark, and single-cell transcriptomics data, where a log-sum-exp decoder is already in use at a fixed temperature (Moinfar & Theis, 2024).

The Mechanism. Write the softmax weights and the logit margins:

$$p_{i,l}(\mathbf{z}) = \frac{e^{\tau g_{i,l}(z_i)}}{\sum_j e^{\tau g_{j,l}(z_j)}}, \qquad \mu_{i,l}(\mathbf{z}) = \tau g_{i,l}(z_i) - \tau \max_{j\neq i} g_{j,l}(z_j). \tag{2}$$

Differentiating equation 1 gives $\nabla_{z_i} f_{\tau,l} = p_{i,l}\nabla g_{i,l}(z_i)$, so a slot's influence on a feature is its own derivative scaled by a weight controlled by the other slots. If slot i leads feature l by a margin

$\mu_{i,l} > 0$, every other slot receives a gradient bounded by $e^{-\mu_{i,l}} \sup\|\nabla g\|$, and the feature behaves as a private function of z_i up to this exponential accuracy. A sum fixes the weight at one, while a maximum routes it only to the winner. Log-sum-exp at finite τ is where this routing decays smoothly with the margin (Figure 1). The same argument applies to masked object-centric decoders, where a softmax over a learned mask feature mixes slots and routes gradients. We prove identifiability when this softmax is replaced by a maximum, so the softmax-masked decoder stands to it as log-sum-exp pooling stands to max pooling.

Contributions.

Identifiability of max-mask decoders. We prove that decoders in which a per-feature score selects the slot that supplies each feature are identifiable up to a permutation of slots and a per-slot local homeomorphism (Theorem 8). This score can resemble depth in images, which results in native support for occlusion. Hard-max decoders follow as a special case of max-mask decoders.

Why smooth decoders exclude occlusion. We show that a C^1 decoder can switch between slots only where both slots agree in value and have a critical point (Proposition 6), which prevents prior theories from applying to such settings. Accordingly, our identifiability results for max-mask and hard-max decoders provide a unique advantage compared to prior work, since we require neither second-order conditions, nor differentiability of the decoder, nor differentiability of the mixing map.

A trainable relaxation. We train hard-max decoders through log-sum-exp pooling and max-mask decoders through a softmax mask, each with an increasing inverse temperature, and evaluate them on synthetic occlusion data, grayscale CLEVR6 and single-cell atlases.

2 Background & Literature Review

We begin by formalising the observation space, the latent space, the ground-truth decoder, the learned decoder and disentanglement. These are the concepts our theory is built on, and we follow the setup of Brady et al. (2024) and Lachapelle et al. (2023).

Let $\mathcal{Z} \subseteq \mathbb{R}^K$ be a K-dimensional latent space carrying k slots, which correspond to the underlying concepts in the data, and let $[k] := \{1, \dots, k\}$. Slot i carries a block of latents $z_i \in \mathbb{R}^{n_i}$, so the true latents are $\mathbf{z} = (z_1, \dots, z_k) \in \mathcal{Z}$ with $K := \sum_i n_i$, and $\mathcal{Z}^\circ$ is the interior of $\mathcal{Z}$. A *decoder* maps a latent space to the observation space $\mathbb{R}^d$, and the ground-truth decoder is $\mathbf{f} : \mathcal{Z} \to \mathbb{R}^d$. Its d outputs are called *features* and are indexed by $l \in [d]$. The learned model also has k slots, indexed by m, with blocks $y_m \in \mathbb{R}^{\hat{n}_m}$ and latents $\boldsymbol{y} \in \widehat{\mathcal{Z}}$, and its decoder $\widehat{\mathbf{f}} : \widehat{\mathcal{Z}} \to \mathbb{R}^d$ maps them to the same observation space. All blocks have positive dimension, $n_i \geq 1$ and $\hat{n}_m \geq 1$, and the interiors $\mathcal{Z}^\circ$ and $\widehat{\mathcal{Z}}^\circ$ are nonempty.

Definition 1 (Box and mixing map)**.** A *box* is a product $N = \prod_i N_i \subseteq \mathcal{Z}^\circ$ of nonempty products $N_i \subseteq \mathbb{R}^{n_i}$ of open intervals, and N_i is its i-th side. A *mixing map* relates the two models, $\mathbf{h} = (\mathbf{h}_1, \dots, \mathbf{h}_k) : \mathcal{Z}^\circ \to \widehat{\mathcal{Z}}^\circ$ with $\mathbf{h}_m(\mathbf{z}) \in \mathbb{R}^{\hat{n}_m}$ and $\mathbf{f} = \widehat{\mathbf{f}} \circ \mathbf{h}$.

In the next definition, $\mathrm{Perm}(k)$ is the set of permutations of $[k]$.

Definition 2 (Disentangled boxes and regions)**.** $\mathbf{h}$ is **block-disentangled on** a box N **with** $\pi \in \mathrm{Perm}(k)$ if there are continuous injective $\sigma_i : N_i \to \mathbb{R}^{\hat{n}_{\pi(i)}}$ with

$$\mathbf{h}_{\pi(i)}(\mathbf{z}) = \sigma_i(z_i) \qquad \forall\, \mathbf{z} \in N,\ i \in [k].$$

$$D_\pi := \{\mathbf{z} :\ \exists \text{ box } N \text{ with } \mathbf{z} \in N \text{ on which } \mathbf{h} \text{ is block-disentangled with } \pi\},$$
$$D := \textstyle\bigcup_{\pi \in \mathrm{Perm}(k)} D_\pi.$$

Such an N is a **disentangled box** (with π). D_π is the **region** of π.

Definition 3 (Global disentanglement)**.** $\mathbf{h}$ is *globally disentangled* if $D_\pi = \mathcal{Z}^\circ$ for some $\pi \in \mathrm{Perm}(k)$, that is, if a single permutation π has a disentangled box around every point of $\mathcal{Z}^\circ$.

2.1 Literature Review

Identifiability of latent variable models. Our work concerns the identifiability of max-pooled decoders. Identifiability of an unconstrained decoder is known to be impossible in general settings

(Hyvärinen & Pajunen, 1999; Locatello et al., 2018; Horan et al., 2021). Besides auxiliary and supervised information, restrictions on the latent distribution and on the decoder have been used to overcome this. Since our work belongs to the last family and is orthogonal to the rest, we cover them in the extended literature review in Appendix J.

Structure on the decoder. There is a wide range of identifiable function classes. Some impose assumptions and regularisations on the Jacobian of the generative process (Buchholz et al., 2022; Gresele et al., 2021; Reizinger et al., 2022). A parallel line restricts sparsity, either of the decoder (Moran et al., 2021; Zheng et al., 2022; Zheng & Zhang, 2023) or of the mechanism relating latents (Lachapelle & Lacoste-Julien, 2022; Lachapelle et al., 2026). Brouillard et al. (2024) assume that each observed variable is affected by exactly one latent, and identify both the latents and the causal graph over them. Here, in contrast, we let the owned feature change from one region of the latent space to another.

Additivity, compositionality and interaction asymmetry. Peebles et al. (2020) regularise the decoder Hessian towards diagonal and observe improved disentanglement without a supporting theory. Brady et al. (2023) call a decoder compositional when each output coordinate locally depends on at most one slot. Lachapelle et al. (2023) prove identifiability of additive decoders and show that additivity and a diagonal Hessian are the same property. Brady et al. (2024) unify these works by requiring interactions within a slot to be of higher order than interactions across slots, and prove identifiability and compositional generalisation. Matthes et al. (2025) introduce a broad class of identifiability criteria, whose formulations rely on the Jacobian and the Hessian and therefore need a differentiable decoder. Altogether, none of these conditions can be evaluated on the nondifferentiable decoders we study here.

3 Max Poolings, Max Masking, and Log-Sum-Exp Approximation

Definition 4 (Decoders)**.** For each slot i and feature l, let $g_{i,l} : \mathbb{R}^{n_i} \to \mathbb{R}$ be a *slot map*. A pooled decoder $\mathbf{f}$ combines the slot maps of each feature by an operation over slots. We consider four types of decoders (hard-max is max-mask with $s_{i,l} = g_{i,l}$):
(i) Additive decoder (Lachapelle et al., 2023): $f_l^{\text{add}}(\mathbf{z}) = \sum_{i=1}^{k} g_{i,l}(z_i)$.
(ii) LSE$_\tau$ decoder: The slots are combined by a smooth Log-Sum-Exp pooling with inverse temperature $\tau > 0$, $f_{\tau,l}(\mathbf{z}) = \frac{1}{\tau} \log \frac{1}{k} \sum_{i=1}^{k} e^{\tau g_{i,l}(z_i)}$, as in equation 1.
(iii) Hard-max decoder: The slots compete independently per feature, $f_l^{\max}(\mathbf{z}) = \max_i g_{i,l}(z_i)$.
(iv) Max-mask decoder: The winner of a feature is chosen by a separate *score* $s_{i,l} : \mathbb{R}^{n_i} \to \mathbb{R}$ rather than by the output itself. The *winners* of feature l and the decoder are

$$W_l(\mathbf{z}) := \arg\max_i \; s_{i,l}(z_i), \qquad w := \min W_l(\mathbf{z}), \quad f_l(\mathbf{z}) := g_{w,l}(z_w), \qquad l \in [d]. \tag{3}$$

Here $s_{i,l}$ intuitively corresponds to the mask logit of object-centric decoders (Burgess et al., 2019; Engelcke et al., 2019; Locatello et al., 2020b). Below, $\mathbf{f}$ denotes a ground-truth decoder and $\widehat{\mathbf{f}}$ a learned one, with slot maps $\widehat{g}_{m,l}$ and scores $\hat{s}_{m,l}$ defined in the same way as $g_{i,l}$ and $s_{i,l}$. No correspondence between the true and the learned slots is assumed.

We give identifiability guarantees for the max-mask decoder, which is more general than the hard-max decoder. To the best of our knowledge, this is the first identifiability guarantee for a decoder that need not be differentiable and in which a separate mask selects the winning slot.

3.1 Identifiability of Max-Mask Decoders

Definition 5 (Ownership matrix and owned features)**.** The features owned by true slot i at $\mathbf{z}$ are $O_i(\mathbf{z}) := \{l \in [d] : \; W_l(\mathbf{z}) = \{i\}, \; \nabla g_{i,l}(z_i) \neq 0\}$. The ownership matrix $G(\mathbf{z}) \in \mathbb{R}^{K \times d}$ has block rows $i \in [k]$ of heights n_i, with $G_{i,l}(\mathbf{z}) := \nabla g_{i,l}(z_i) \in \mathbb{R}^{n_i}$ if $W_l(\mathbf{z}) = \{i\}$ and $G_{i,l}(\mathbf{z}) := 0$ otherwise, and its column l is written $G_{\cdot,l}(\mathbf{z}) \in \mathbb{R}^K$.

To prove disentanglement theorems, we assume the following. A1 to A4 give local disentanglement (Theorem 7), and A5 and A6 let us prove global disentanglement (Theorem 8).
Assumption A1 (Continuous learned slot maps)**.** Every $g_{i,l}$ is C^1, every $s_{i,l}$ is continuous, every $\widehat{g}_{m,l}$ is continuous. Nothing is assumed about $\hat{s}_{m,l}$.

Assumption A2 (Homeomorphic mixing)**.** $\mathbf{h} : \mathcal{Z}^\circ \to \widehat{\mathcal{Z}}^\circ$ is a homeomorphism and $\mathbf{f} = \widehat{\mathbf{f}} \circ \mathbf{h}$ on $\mathcal{Z}^\circ$.

Assumption A3 (Full rank)**.** $\operatorname{rank} G(\mathbf{z}) = K$ for every $\mathbf{z} \in \mathcal{Z}^\circ$.

Assumption A4 (Irreducible ownership)**.** For $\mathbf{z}$ in a dense subset of $\mathcal{Z}^\circ$, every i and every split $O_i(\mathbf{z}) = T \sqcup T'$ with $T, T' \neq \emptyset$: $\operatorname{rank}_i(T, \mathbf{z}) + \operatorname{rank}_i(T', \mathbf{z}) > n_i$, where $\operatorname{rank}_i(T, \mathbf{z})$ is the rank of $\big(G_{i,l}(\mathbf{z})\big)_{l \in T}$.

Assumption A5 (Connected block slices)**.** Every block slice $\mathcal{Z}_{i,c} := \{\mathbf{z} \in \mathcal{Z}^\circ : \ z_i = c\}$, $i \in [k]$, $c \in \mathbb{R}^{n_i}$, is connected.

Assumption A6 (Connected domain)**.** $\mathcal{Z}^\circ$ is connected.

Notes on assumptions. A2 gives the dimension sum $\sum_m \hat{n}_m = K$, since $\mathcal{Z}^\circ$ and $\widehat{\mathcal{Z}}^\circ = \mathbf{h}(\mathcal{Z}^\circ)$ are open and nonempty and SR3 (Appendix D) applies to $\mathbf{h}$ and to $\mathbf{h}^{-1}$. The individual learned block dimensions $\hat{n}_m$ are not assumed to match the true ones. A3 says that at every point the gradients of the features slot i owns span $\mathbb{R}^{n_i}$, so these features determine z_i locally. A4 is the irreducibility of Brady et al. (2023), stated on G.

Decoders do not need to be differentiable. A1 asks the true slot maps to be C^1 and the true scores to be continuous, but the learned slot maps only to be continuous. So, $\mathbf{f}$ may jump where the winner of a feature changes. This is deliberate, and the next proposition says what a decoder gives up by being differentiable instead.

Proposition 6 (A differentiable decoder cannot switch visibly)**.** *Assume A1 (continuous learned slot maps) and fix a feature l. For a slot i put $U_i := \{\mathbf{z} \in \mathcal{Z}^\circ : \ W_l(\mathbf{z}) = \{i\}\}$, which is open, and call $\mathbf{z}^0 \in \mathcal{Z}^\circ$ a* switching point of l between i and j *if $i \neq j$ and $\mathbf{z}^0 \in \overline{U_i} \cap \overline{U_j}$. At every such point the following hold.*

1. *If f_l is continuous at $\mathbf{z}^0$, then $g_{i,l}(z_i^0) = g_{j,l}(z_j^0)$.*
2. *If f_l is C^1 on a neighbourhood of $\mathbf{z}^0$, then also $\nabla g_{i,l}(z_i^0) = 0$ and $\nabla g_{j,l}(z_j^0) = 0$.*

Proof in Appendix E. A switching point of feature l is a point where the slot that supplies l changes, which is what occlusion does at the contour of an object. Part (1) says that a decoder that is continuous there has to give both objects the same value, and part (2) says that a C^1 decoder has to place a critical point of both slot maps exactly on it. Occlusion in a real scene does neither, since the surface in front generally differs in value from the one behind and both keep varying across the contour. Additive decoders (Lachapelle et al., 2023), compositional decoders (Brady et al., 2023) and interaction asymmetry (Brady et al., 2024) all assume a differentiable decoder and therefore exclude it. A max-mask decoder allows the switch, and no proof below requires $\mathbf{f}$ or $\mathbf{h}$ to be differentiable.

Theorem 7 (Every open set contains a disentangled box, and true and learned block dimensions match)**.** *Assume A1 (continuous learned slot maps), A2 (homeomorphic mixing), A3 (full rank), A4 (irreducible ownership). Then*

1. *D is open in $\mathcal{Z}^\circ$.*
2. *Every nonempty open $U \subseteq \mathcal{Z}^\circ$ contains a box N on which $\mathbf{h}$ is block-disentangled with some $\pi \in \operatorname{Perm}(k)$, and $\hat{n}_{\pi(i)} = n_i$ for all i.*
3. *D is dense in $\mathcal{Z}^\circ$.*

The proof is in Appendix F. Intuitively, the proof first uses that owning a feature is an open condition (by A1) to find a box N inside U on which neither the owner of a feature nor the learned slot that reproduces it changes. At one point of N it then counts directions, as in Brady et al. (2023). By A3, the features slot i owns move in n_i independent directions. A learned block y_m can carry at most $\hat{n}_m$ of them, since the homeomorphism of A2 cannot lower dimension, and A4 forbids reducible ownership. This gives the permutation π with $\hat{n}_{\pi(i)} = n_i$. Finally, as with the anchor features of Moran et al. (2021), n_i of the features slot i owns are local coordinates of z_i, so $\mathbf{h}_{\pi(i)}$ depends on z_i alone on a smaller box.

Theorem 8 (Each learned block is a function of one true block, with one permutation on the whole domain)**.** *Assume A1 (continuous learned slot maps), A2 (homeomorphic mixing), A3 (full rank), A4 (irreducible ownership), A5 (connected block slices), A6 (connected domain). Let $C_i := \{z_i : \mathbf{z} \in \mathcal{Z}^\circ\} \subseteq \mathbb{R}^{n_i}$ be the values that block i takes. Then:*

1. *There is one $\pi \in \operatorname{Perm}(k)$ with $D_\pi = \mathcal{Z}^\circ$, and $\hat{n}_{\pi(i)} = n_i$ for every i.*

2. *For every i there is a continuous $\sigma_i : C_i \to \mathbb{R}^{\hat{n}_{\pi(i)}} = \mathbb{R}^{n_i}$ with $\mathbf{h}_{\pi(i)}(\mathbf{z}) = \sigma_i(z_i)$ for every $\mathbf{z} \in \mathcal{Z}^\circ$.*
3. *Every σ_i is a local homeomorphism: each point of C_i has an open neighbourhood that σ_i maps homeomorphically onto an open set.*
4. *σ_i is injective if $n_i = 1$, or if $\mathcal{Z}^\circ = \prod_j C_j$.*

Proof in Appendix G. It first shows that the regions D_π are open and pairwise disjoint, that is, if $D_\pi \cap D_{\pi'} \neq \emptyset$ then $\pi = \pi'$. Then it shows that every point has a disentangled box, since otherwise it would be touched by at least two regions (in $\overline{D_\pi} \cap \overline{D_{\pi'}}$ with $\pi \neq \pi'$), which finally contradicts the injectivity of $\mathbf{h}$. So $\mathcal{Z}^\circ$ is the disjoint union of the open D_π regions, and by the connected domain assumption (A6) one permutation holds. Finally, connected block slices (A5) glue the local maps into the global σ_i.

3.2 Identifiability of Hard-Max Decoders

Theorem 9 (Hard-max decoders are globally disentangled)**.** *Let $\mathbf{f}$ and $\widehat{\mathbf{f}}$ be hard-max decoders, and assume A1 (continuous learned slot maps), A2 (homeomorphic mixing), A3 (full rank), A4 (irreducible ownership), A5 (connected block slices) and A6 (connected domain). Then $\mathbf{h}$ is globally disentangled (Definition 3): there is one $\pi \in \mathrm{Perm}(k)$ with $D_\pi = \mathcal{Z}^\circ$, and $\hat{n}_{\pi(i)} = n_i$ for every i. Moreover, with $C_i := \{z_i : \mathbf{z} \in \mathcal{Z}^\circ\}$, for every i there is a continuous local homeomorphism $\sigma_i : C_i \to \mathbb{R}^{n_i}$ with $\mathbf{h}_{\pi(i)}(\mathbf{z}) = \sigma_i(z_i)$ for every $\mathbf{z} \in \mathcal{Z}^\circ$.*

Theorem 9 is the special case $s_{i,l} = g_{i,l}$ and $\hat{s}_{m,l} = \widehat{g}_{m,l}$ of Theorem 8 (Appendix H). Each σ_i is injective when $n_i = 1$ or when $\mathcal{Z}^\circ = \prod_j C_j$ (Theorem 8). Under A1 to A4 alone, Theorem 7 still gives local disentanglement.

3.3 Finite Temperature

For a fixed τ, an LSE_τ decoder is an additive decoder after a known entrywise bijection, since $e^{\tau f_{\tau,l}(\mathbf{z})} = \frac{1}{k}\sum_i e^{\tau g_{i,l}(z_i)}$. If $\mathbf{f}$ and $\widehat{\mathbf{f}}$ are LSE_τ decoders with the same known τ and $\mathbf{f} = \widehat{\mathbf{f}} \circ \mathbf{h}$, then $e^{\tau\mathbf{f}} = e^{\tau\widehat{\mathbf{f}}} \circ \mathbf{h}$, so the results of Lachapelle et al. (2023) apply through their Remark 1, with their sufficient nonlinearity condition imposed on the maps $e^{\tau g_{i,l}}$ and their smoothness and support assumptions. Without it, no known guarantee applies at a finite τ.

Proposition 10 (A finite τ alone does not give identifiability)**.** *For every $\tau > 0$ and $d \geq k \geq 2$, with one-dimensional slots, there are LSE_τ decoders with affine slot maps and Jacobians of rank k, and a real-analytic diffeomorphism $\mathbf{h}$ with $\mathbf{f} = \widehat{\mathbf{f}} \circ \mathbf{h}$ and $\partial_{z_i}\mathbf{h}_j > 0$ for all i, j. Take $A \in \mathbb{R}^{k\times k}$ invertible and $N \in \mathbb{R}^{d\times k}$ of rank k, both entrywise positive, let $M := NA$, and set*

$$\mathbf{f}(\mathbf{z}) := \tfrac{1}{\tau}\log\big(Me^{\tau\mathbf{z}}\big), \qquad \widehat{\mathbf{f}}(\boldsymbol{y}) := \tfrac{1}{\tau}\log\big(Ne^{\tau\boldsymbol{y}}\big), \qquad \mathbf{h}(\mathbf{z}) := \tfrac{1}{\tau}\log\big(Ae^{\tau\mathbf{z}}\big), \tag{4}$$

with exp *and* log *taken entrywise.*

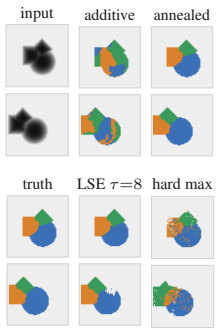
The proof is in Appendix I. In this example the first and second derivatives of $e^{\tau g_{i,l}}$ are collinear, which the additive theory does not support. Bounding how much a learned LSE_τ decoder mixes the true latents, as a function of τ and of the margins in equation 2, is left for future work.

4 Experiments

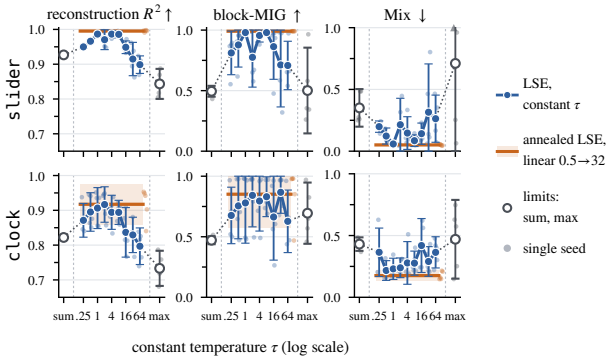
We test the utility of the hard-max and LSE_τ decoders in three settings, ordered from most to least controlled: *(i) Synthetic images* whose generative process is a max-pool over objects, *(ii) CLEVR6 (Greff et al., 2019; Locatello et al., 2020b)*, a standard benchmark for object-centric representation learning, and *(iii) single-cell RNA sequencing (scRNA-seq)*, a biological readout on the multiple cellular atlases (Sikkema et al., 2023; Xu et al., 2023). Further details about the experimental setup are provided in Appendix K.

4.1 Synthetic Images

When objects overlap, a shared pixel often shows a single object rather than a sum of objects, and which object it shows depends on the depth. The additive and compositional identifiability



(a) Slot ownership on slider, two held-out scenes per model.



(b) The family at constant τ against annealing.

Figure 2: Synthetic images. (a) The input, the true ownership map and, for each decoder, the learned slot with the largest per-slot output at each foreground pixel, coloured by the true slot it is matched to. The additive decoder and the hard max split shapes across slots, while LSE at $\tau = 8$ and annealed LSE nearly reproduce the true map. (b) Reconstruction R^2, block-MIG and Mix (lower is better) on slider (top) and clock (bottom). Filled points are LSE at a constant τ (mean over 5 seeds, 95% t interval, single seeds as small dots), and open circles are the two limits, the sum ($\tau \to 0$) and the hard max ($\tau \to \infty$). The orange line and band are annealed LSE, linear $0.5 \to 32$ over the first 70% of training. One hard-max seed on slider has Mix $= 1.60$ and is drawn at the top edge.

guarantees do not cover this case: additive decoders cannot represent it (Lachapelle et al., 2023), and compositionality fails at the borders of occluding objects (Brady et al., 2023). In their image experiments, Lachapelle et al. (2023) move their balls in separate columns or reject images with occlusion, and Wiedemer et al. (2023) exclude images with occluding sprites to keep their generator compositional. Similar to the column images in Lachapelle et al. (2023), our datasets feature few objects with one scalar latent each, but exhibit frequent occlusion and overlap.

Setup. Two 64×64 datasets with $k = 3$ slots and a scalar latent each: *slider*—a circle, a square and a diamond sliding along their own rows, and *clock*—three hands of different length. We compare the additive decoder, LSE at a fixed τ, LSE annealed from 0.5 to τ_{end} over the first 70% of training, and the hard maximum, with two references without slot structure: a joint MLP over all latents and a linear decoder. For all these modes encoder and decoder maps are MLPs since CNNs are already shown to impose an inductive bias. Disentanglement is measured by *block-MIG* and *Mix* (Appendix K.1).

Table 1: Synthetic images. Block-MIG (mean $\pm$ 95% t half-width over 5 seeds; seeds with block-MIG ≥ 0.9), Mix (lower is better) and reconstruction R^2. The annealed arm ramps τ linearly from 0.5 to 32 over the first 70% of training. Bold marks the best value per column, and also the second best when their intervals overlap.

	slider			clock		
decoder	block-MIG	Mix↓	R^2	block-MIG	Mix↓	R^2
linear	0.14 ± 0.01 (0/5)	2.977	0.564	0.03 ± 0.02 (0/5)	2.699	0.356
joint MLP	0.65 ± 0.39 (0/5)	0.421	0.986	0.08 ± 0.04 (0/5)	2.408	0.825
additive ($\tau \to 0$)	0.49 ± 0.05 (0/5)	0.356	0.927	0.51 ± 0.11 (0/5)	0.414	0.826
LSE, $\tau = 8$	$\mathbf{0.98 \pm 0.01}$ (5/5)	0.081	0.988	0.82 ± 0.27 (3/5)	0.259	0.886
LSE, $\tau = 32$	0.71 ± 0.35 (2/5)	0.316	0.915	$\mathbf{0.87 \pm 0.17}$ (3/5)	0.290	0.829
LSE, τ: $0.5 \to 32$ (annealed)	$\mathbf{0.99 \pm 0.00}$ (5/5)	**0.051**	**0.995**	$\mathbf{0.85 \pm 0.27}$ (3/5)	**0.176**	**0.917**
hard max ($\tau \to \infty$)	0.56 ± 0.32 (1/5)	0.558	0.843	0.60 ± 0.35 (1/5)	0.728	0.717

LSE pooling recovers the latents where both ends of the family fail. Figure 2b traces the whole family at constant τ: on slider every metric is poor at both ends and best for $1 \leq \tau \leq 8$. At a matched budget, the pooling alone moves slider block-MIG from 0.49 (additive) to 0.98 (LSE,

Table 2: CLEVR6 after 500k iterations: test J-ARI and JIS ($\times 100$; mean $\pm$ sd over three seeds, worst seed in brackets) on the first 500 of the 2,000 test scenes. The softmax masks pool raw presence logits, and the annealed mask ends at $\tau = 8$. IA's published colour result is 96.5 / 83.8. Bold marks the best mean per column, and also the second best when their mean $\pm$ sd intervals overlap.

	colour		grayscale	
decoder	J-ARI	JIS	J-ARI	JIS
IA	95.9 ± 1.2 [94.6]	83.8 ± 0.9	77.1 ± 29.9 [42.7]	89.6 ± 0.5
linear mask	98.1 ± 0.2 [98.0]	94.3 ± 1.4	97.5 ± 0.8 [96.6]	95.1 ± 0.3
softmax mask, $\tau = 1$	98.2 ± 0.4 [97.8]	94.6 ± 1.2	$\mathbf{98.1} \pm 0.1$ [98.1]	$\mathbf{96.2} \pm 0.1$
softmax mask, $\tau = 8$	$\mathbf{98.4} \pm 0.3$ [98.1]	$\mathbf{96.4} \pm 0.8$	$\mathbf{98.3} \pm 0.2$ [98.0]	$\mathbf{96.3} \pm 0.5$
softmax mask, $\tau : 0.5 \to 8$	$\mathbf{98.3} \pm 0.3$ [98.0]	$\mathbf{95.9} \pm 1.1$	97.8 ± 0.3 [97.5]	96.1 ± 0.3

$\tau = 8$), and Mix from 0.36 to 0.08 (Table 1). Figure 2a illustrates this: the additive decoder fits the scenes ($R^2 = 0.93$) with slots that each cover parts of two shapes, whereas LSE reproduces the true ownership map. The hard maximum fits worst of the slot decoders with R^2 0.84 on slider and 0.72 on clock. We attribute this failure to the optimisation problem that LSE is designed to remove: each pixel's gradient only reaches its current winner, so a wrong early partition is rarely corrected.

4.2 Object-Centric Learning on CLEVR6

In rendered images the nearest object determines each pixel. So a scene is approximately a max-mask decoder with depth as the score (Definition 4). Interaction Asymmetry (IA, Brady et al., 2024) recovers CLEVR6 objects by minimizing the interactions across slots. Colour makes this problem easier, because it tells the model which pixels belong together. Here, we ask whether decoding each slot separately and combining the slots only through a mask is more robust when colour is removed.

Setup. CLEVR6 contains 53,483 rendered scenes of partially occluding objects at 128×128 pixels, and grayscale replaces every pixel by its luminance. We keep IA's architecture and training recipe, vary only how the slots meet, and drop IA's interaction penalty (Appendix K.2). The IA decoder lets every pixel cross-attend to all slots and penalises interactions between slots. Our arms run the same decoder on one slot at a time, so slot i yields an image $\mathbf{u}_i$ and a presence logit a_i from z_i alone, and combine the slots only at the end, $\hat{x} = \sum_i w_i \mathbf{u}_i$ with per-pixel weights $w = \mathrm{softmax}_i(\tau a_i)$. We train $\tau \in \{1, 8\}$, τ annealed linearly from 0.5 to 8, and a *linear mask* without an exponential, for 500k iterations, and at $\tau = 1$ this is the mixture decoder of Slot Attention (Locatello et al., 2020b). Following IA, ***J-ARI*** scores whether the slot that drives each foreground pixel, measured by the decoder Jacobian, partitions the pixels into the objects, and ***JIS*** how concentrated each pixel's dependence is on one slot (Appendix K.2). We train three seeds per arm and score the first 500 of the 2,000 test scenes.

Per-slot decoding with a mask keeps objects apart in grayscale. In colour every decoder recovers the objects in all runs, and IA reproduces its published result (Table 2). Removing colour separates the designs. Without colour, decoding each slot alone and pooling with a mask keeps the objects apart. All twelve per-slot runs reach 96.6–98.5 J-ARI, while IA's three runs end at 96.5, 92.2 and 42.7. IA still concentrates each pixel on one slot (grayscale JIS 89.6), but without colour its partition can merge objects. In Figure 5 (Appendix K.2), its worst seed gives three objects of the first scene to one slot and splits the large sphere of the last scene between two. Both per-slot decoders shown give each object its own slot. The per-slot decoders also improve on IA in colour. Their J-ARI is 97.8–98.6 against 94.6–96.8, and their JIS is 92.7–97.2 against 82.8–84.4.

4.3 Disentanglement of Cell Types in scRNA-seq

Single-cell RNA sequencing measures RNA abundance for thousands of genes in individual cells. Differences between cells reflect multiple sources of biological variation, including cell type and developmental stage. Disentangled representation learning aims to separate these sources into individual latent dimensions whose biological meaning can be examined. DRVI (Moinfar & Theis, 2024) pursues this goal with a variational autoencoder whose decoder maps each latent dimension separately to per-gene log-scale contributions and combines them through log-sum-exp pooling.

Table 3: Disentanglement performance (LMS-SMI, mean ± sd over three seeds) of different temperature settings on scRNA-seq atlases. Bold marks the best mean per dataset, and also the second best when their mean ± sd intervals overlap. Cells is the size of each dataset, in thousands, and types the number of cell types in its annotation (HLCA's finest level, the tissues' curated one). A type with fewer than 10 cells in a seed's 50,000-cell scoring sample is left out of that seed's scores, up to three types in HLCA and at most one elsewhere. The tissues are listed by size. Per-seed values are in Appendix K.3.

	HLCA	Bone marrow	Pancreas	Lymph node	Kidney	Spleen	Skeletal muscle	Liver	Intestine	Lung	Blood
cells	585k	67k	122k	124k	195k	201k	202k	260k	263k	318k	336k
types	61	45	14	32	41	34	22	29	44	49	27
mean ($\tau \to 0$)	.458 ±.008	.378 ±.005	.514 ±.027	.536 ±.009	.377 ±.018	.426 ±.009	.511 ±.005	.475 ±.009	.444 ±.004	.433 ±.014	.514 ±.013
DRVI ($\tau = 1$)	**.670** ±.016	.550 ±.004	.660 ±.004	**.750** ±.006	**.653** ±.002	**.680** ±.026	**.636** ±.015	**.653** ±.010	.597 ±.014	**.671** ±.014	**.725** ±.002
LSE, τ:0.5→8	**.699** ±.014	**.598** ±.009	**.671** ±.002	**.747** ±.005	**.639** ±.018	**.710** ±.017	**.625** ±.012	**.656** ±.006	**.672** ±.006	**.671** ±.017	**.716** ±.013

DRVI thus represents a $\tau = 1$ member of the LSE_τ pooling family, allowing us to test whether annealing this parameter improves alignment with annotated cell types.

Setup. We retain DRVI's 128 scalar latent dimensions and compare three pooling configurations: mean pooling ($\tau \to 0$), the published DRVI architecture corresponding to LSE pooling at $\tau = 1$, and LSE with τ annealed geometrically from 0.5 to 8. We evaluate the Human Lung Cell Atlas (Sikkema et al., 2023, HLCA) and ten additional tissue atlases (Table 3). We use expert cell-type annotations for evaluating latent separation. We summarise associations between latent dimensions and cell types using the latent matching score *LMS* (Lachapelle et al., 2023), which averages these associations under a one-to-one matching of cell types to latent dimensions. Following Moinfar & Theis (2024), we use LMS on the scaled mutual information (SMI) for evaluation (Table 3).

Annealing τ improves disentanglement in most atlases. Replacing LSE at $\tau = 1$ by mean pooling lowers LMS-SMI in all eleven datasets, by 20–42% (32% on HLCA). Annealing to $\tau = 8$ approximates a hard maximum: the leading latent's mean margin is 11.6 on HLCA, against 2.2 at $\tau = 1$ (Table 13). It raises LMS-SMI in six of the eleven datasets, and the means differ by less than one standard deviation on the rest.

5 Conclusion

Here, we study the use of max pooling and max masking in the decoder of unsupervised representation learning models for latent variable identification. The proposed family supports occlusion and overlap by design, which are important phenomena in real datasets. We provide identifiability results for this class of decoders under fewer smoothness assumptions than previous work. Specifically, we do not assume second-order derivative conditions or differentiability of the pooled decoder. This matters, as a differentiable decoder cannot switch visibly between slots at occlusion boundaries. In addition, our formulation of max-mask decoders aligns with the commonly used softmax masking practice in object-centric representation learning, and to the best of our knowledge our identifiability results are the closest to it. The optimisation of max-pooled decoders is hard. Accordingly, we suggest using log-sum-exp pooling together with annealing of a temperature parameter. This allows for recovery of the true latents on synthetic occlusion data more often than additive or hard-max decoders. We further studied masked and log-sum-exp decoders on CLEVR6 and single-cell omics data, where masked decoders separate objects more reliably in grayscale and annealing improves the alignment with cell types in six of eleven atlases.

We show that a finite τ alone does not give exact identifiability. So our disentanglement results hold only at $\tau = \infty$, whereas the trained models use a finite τ. Providing bounds for these smooth decoders based on τ and the regularity of decoders remains an open question for future work.

We anticipate this work is a step towards a general theory of identifiability for architectures like transformers, where nonlinear poolings like softmax combine multiple predictions unevenly.

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks. We used them to aid or polish writing, specifically to improve the writing, to shorten the text, to format the LaTeX source, and to align concepts and notation across the paper. For the theory, we used them to help formulate mathematical claims, to help improve the accuracy of proofs, to provide additional ingredients for these proofs, to assist in writing and simplifying the proofs, and to check them. For the experiments, we used them to implement code and to generate synthetic data sets. The synthetic data sets were conceptualised and the experiments designed by the authors. We have not used generative AI tools to develop theoretical models or conceptual frameworks, to propose or refine hypotheses, to design the research methodology or experiments, to provide feedback on the experiments, to interpret results, or to clean or reformat data sets, and translation and qualitative or thematic data analysis are not applicable to this work. The problem formulation, the mathematical directions and the model architectures were developed entirely by the authors. We have reviewed all AI-assisted work. Every proof touched by these tools was checked line by line by at least one author, in addition to checks with these tools, and all AI-written code was reviewed and tested by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is methodological. The single-cell experiments use only publicly available, de-identified human atlases released by their original studies under their own consent and ethics approvals (Sikkema et al., 2023; Xu et al., 2023), and we generate no new human data. Disentangled representations can support the interpretation of biological data, and conclusions drawn from learned latents should be validated experimentally before any clinical use. We are not aware of other ethical concerns.

REPRODUCIBILITY STATEMENT

The synthetic data generators, architectures, training schedules and evaluation protocols are described in Appendix K.1, the CLEVR6 setup and every deviation from the recipe of Brady et al. (2024) in Appendix K.2, and the single-cell data, preprocessing and metrics in Appendix K.3. Anonymised source code for all experiments is provided as supplementary material. All theoretical assumptions are stated in Section 3.1 and restated in Appendix C, and complete proofs of every theorem and proposition are in Appendices E to I.

ACKNOWLEDGMENTS

Acknowledgments are omitted to preserve anonymity for double-blind review.

REFERENCES

Eirikur Agustsson, Fabian Mentzer, Michael Tschannen, Lukas Cavigelli, Radu Timofte, Luca Benini, and Luc Van Gool. Soft-to-hard vector quantization for end-to-end learning compressible representations. *arXiv [cs.LG]*, April 2017.

Kartik Ahuja, Jason Hartford, and Yoshua Bengio. Weakly supervised representation learning with sparse perturbations. In S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho, and A Oh (eds.), *Advances in Neural Information Processing Systems 35*, volume 35, pp. 15516–15528, San Diego, California, USA, 2022. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International Conference on Machine Learning*, pp. 372–407. PMLR, July 2023.

Jack Brady, Roland S Zimmermann, Yash Sharma, Bernhard Schölkopf, Julius von Kügelgen, and Wieland Brendel. Provably learning object-centric representations. *arXiv [cs.LG]*, May 2023.

Jack Brady, Julius von Kügelgen, Sébastien Lachapelle, Simon Buchholz, Thomas Kipf, and Wieland Brendel. Interaction asymmetry: A general principle for learning composable abstractions. *arXiv [cs.LG]*, November 2024.

Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco Cohen. Weakly supervised causal representation learning. In S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho, and A Oh (eds.), *Advances in Neural Information Processing Systems 35*, volume 35, pp. 38319–38331, San Diego, California, USA, 2022. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Philippe Brouillard, Sébastien Lachapelle, Julia Kaltenborn, Yaniv Gurwicz, Dhanya Sridhar, Alexandre Drouin, Peer Nowack, Jakob Runge, and David Rolnick. Causal representation learning in temporal data via single-parent decoding. *arXiv [cs.LG]*, October 2024.

Simon Buchholz, Michel Besserve, and Bernhard Schölkopf. Function classes for identifiable nonlinear independent component analysis. In S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho, and A Oh (eds.), *Advances in Neural Information Processing Systems 35*, volume 35, pp. 16946–16961, San Diego, California, USA, 2022. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. In A Oh, T Naumann, A Globerson, K Saenko, M Hardt, and S Levine (eds.), *Advances in Neural Information Processing Systems 36*, volume 36, pp. 45419–45462, San Diego, California, USA, 2023. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. MONet: Unsupervised scene decomposition and representation. *arXiv [cs.CV]*, January 2019.

Ricky T Q Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. *arXiv [cs.LG]*, February 2018.

Eric Crawford and Joelle Pineau. Spatially invariant unsupervised object detection with convolutional neural networks. *Proc. Conf. AAAI Artif. Intell.*, 33(01):3412–3420, July 2019.

Martin Engelcke, Adam R Kosiorek, Oiwi Parker Jones, and Ingmar Posner. GENESIS: Generative scene inference and sampling with object-centric latent representations. *arXiv [cs.LG]*, July 2019.

S M Ali Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Koray Kavukcuoglu, and Geoffrey E Hinton. Attend, infer, repeat: Fast scene understanding with generative models. In D Lee, M Sugiyama, U Luxburg, I Guyon, and R Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. In *International Conference on Machine Learning*, pp. 2424–2433. PMLR, May 2019.

Luigi Gresele, Paul K Rubenstein, Arash Mehrjou, Francesco Locatello, and Bernhard Schölkopf. The incomplete rosetta stone problem: Identifiability results for multi-view nonlinear ICA. In *Uncertainty in Artificial Intelligence*, pp. 217–227. PMLR, August 2020.

Luigi Gresele, Julius von Kügelgen, Vincent Stimper, Bernhard Schölkopf, and Michel Besserve. Independent mechanism analysis, a new concept? In M Ranzato, A Beygelzimer, Y Dauphin, P S Liang, and J Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 28233–28248. Curran Associates, Inc., 2021.

Daniel P Horan, Eitan Richardson, and Yair Weiss. When is unsupervised disentanglement possible? 34:5150–5161, 2021.

Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In D Lee, M Sugiyama, U Luxburg, I Guyon, and R Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Aapo Hyvarinen and Hiroshi Morioka. Nonlinear ICA of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pp. 460–469. PMLR, April 2017.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 859–868. PMLR, April 2019.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Netw.*, 12(3):429–439, April 1999.

Hermanni Hälvä and Aapo Hyvarinen. Hidden markov nonlinear ICA: Unsupervised learning from nonstationary time series. In *Conference on Uncertainty in Artificial Intelligence*, pp. 939–948. PMLR, August 2020.

Hermanni Hälvä, Sylvain Le Corff, Luc Lehéricy, Jonathan So, Yongjie Zhu, Elisabeth Gassiat, and Aapo Hyvarinen. Disentangling identifiable features from noisy data with structured nonlinear ICA. *arXiv [stat.ML]*, June 2021.

Hermanni Hälvä, Jonathan So, Richard E Turner, and Aapo Hyvärinen. Identifiable feature learning for spatial data with nonlinear ICA. In *International Conference on Artificial Intelligence and Statistics*, pp. 3331–3339. PMLR, April 2024.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv [stat.ML]*, November 2016.

Yibo Jiang and Bryon Aragam. Learning nonparametric latent causal graphs with unknown interventions. In A Oh, T Naumann, A Globerson, K Saenko, M Hardt, and S Levine (eds.), *Advances in Neural Information Processing Systems 36*, volume 36, pp. 60468–60513, San Diego, California, USA, 2023. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Ilyes Khemakhem, Diederik P Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. *arXiv [stat.ML]*, July 2019.

Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Identifiability of deep generative models without auxiliary information. In S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho, and A Oh (eds.), *Advances in Neural Information Processing Systems 35*, volume 35, pp. 15687–15701, San Diego, California, USA, 2022. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. *arXiv [stat.ML]*, July 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. In *Advances in Neural Information Processing Systems 37*, pp. 93300–93335, San Diego, California, USA, 2024. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Avinash Kori, Francesca Toni, and Ben Glocker. Identifiable object representations under spatial ambiguities. *arXiv [cs.LG]*, June 2025.

Julius Von Kügelgen, Michel Besserve, Liang Wendong, Luigi Gresele, Armin Kekic, Elias Bareinboim, David Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. In A Oh, T Naumann, A Globerson, K Saenko, M Hardt, and S Levine (eds.), *Advances in Neural Information Processing Systems 36*, volume 36, pp. 48603–48638, San Diego, California, USA, 2023. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Sébastien Lachapelle and Simon Lacoste-Julien. Partial disentanglement via mechanism sparsity. *arXiv [stat.ML]*, July 2022.

Sébastien Lachapelle, Divyat Mahajan, Ioannis Mitliagkas, and Simon Lacoste-Julien. Additive decoders for latent variables identification and cartesian-product extrapolation. In *Advances in Neural Information Processing Systems 36*, pp. 25112–25150, San Diego, California, USA, 2023. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Nonparametric partial disentanglement via mechanism sparsity: Sparse actions, interventions and sparse temporal dependencies. *J. Mach. Learn. Res.*, 27 (71):1–90, 2026.

Zhixuan Lin, Yi-Fu Wu, Skand Vishwanath Peri, Weihao Sun, Gautam Singh, Fei Deng, Jindong Jiang, and Sungjin Ahn. SPACE: Unsupervised object-oriented scene representation via spatial attention and decomposition. *arXiv [cs.LG]*, January 2020.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. Causal representation learning for instantaneous and temporal effects in interactive systems. *arXiv [cs.LG]*, June 2022a.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Stratis Gavves. CITRIS: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning*, pp. 13557–13603. PMLR, June 2022b.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *arXiv [cs.LG]*, November 2018.

Francesco Locatello, Ben Poole, Gunnar Raetsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, pp. 6348–6359. PMLR, November 2020a.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In H Larochelle, M Ranzato, R Hadsell, M F Balcan, and H Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 11525–11538. Curran Associates, Inc., 2020b.

Stefan Matthes, Zhiwei Han, and Hao Shen. Mechanistic independence: A principle for identifiable disentangled representations. *arXiv [cs.LG]*, September 2025.

Amir Ali Moinfar and Fabian J Theis. Unsupervised deep disentangled representation of single-cell omics. *bioRxiv*, pp. 2024.11.06.622266, November 2024.

Gemma E Moran, Dhanya Sridhar, Yixin Wang, and David M Blei. Identifiable deep generative models via sparse decoding. *arXiv [stat.ML]*, October 2021.

William Peebles, John Peebles, Jun-Yan Zhu, Alexei Efros, and Antonio Torralba. The hessian penalty: A weak prior for unsupervised disentanglement. In *Lecture Notes in Computer Science*, Lecture Notes in Computer Science, pp. 581–597. Springer International Publishing, Cham, 2020.

Patrik Reizinger, Luigi Gresele, Jack Brady, Julius von Kügelgen, Dominik Zietlow, Bernhard Schölkopf, Georg Martius, Wieland Brendel, and Michel Besserve. Embrace the gap: VAEs perform independent mechanism analysis. *arXiv [stat.ML]*, June 2022.

Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, and Francesco Locatello. Bridging the gap to real-world object-centric learning. *arXiv [cs.CV]*, September 2022.

Lisa Sikkema, Ciro Ramírez-Suástegui, Daniel C Strobl, Tessa E Gillett, Luke Zappia, Elo Madissoon, Nikolay S Markov, Laure-Emmanuelle Zaragosi, Yuge Ji, Meshal Ansari, Marie-Jeanne Arguel, Leonie Apperloo, Martin Banchero, Christophe Bécavin, Marijn Berg, Evgeny Chichelnitskiy, Mei-I Chung, Antoine Collin, Aurore C A Gay, Janine Gote-Schniering, Baharak Hooshiar Kashani, Kemal Inecik, Manu Jain, Theodore S Kapellos, Tessa M Kole, Sylvie Leroy, Christoph H Mayr, Amanda J Oliver, Michael von Papen, Lance Peter, Chase J Taylor, Thomas Walzthoeni, Chuan Xu, Linh T Bui, Carlo De Donno, Leander Dony, Alen Faiz, Minzhe Guo,

Austin J Gutierrez, Lukas Heumos, Ni Huang, Ignacio L Ibarra, Nathan D Jackson, Preetish Kadur Lakshminarasimha Murthy, Mohammad Lotfollahi, Tracy Tabib, Carlos Talavera-López, Kyle J Travaglini, Anna Wilbrey-Clark, Kaylee B Worlock, Masahiro Yoshida, Lung Biological Network Consortium, Maarten van den Berge, Yohan Bossé, Tushar J Desai, Oliver Eickelberg, Naftali Kaminski, Mark A Krasnow, Robert Lafyatis, Marko Z Nikolic, Joseph E Powell, Jayaraj Rajagopal, Mauricio Rojas, Orit Rozenblatt-Rosen, Max A Seibold, Dean Sheppard, Douglas P Shepherd, Don D Sin, Wim Timens, Alexander M Tsankov, Jeffrey Whitsett, Yan Xu, Nicholas E Banovich, Pascal Barbry, Thu Elizabeth Duong, Christine S Falk, Kerstin B Meyer, Jonathan A Kropski, Dana Pe'er, Herbert B Schiller, Purushothama Rao Tata, Joachim L Schultze, Sara A Teichmann, Alexander V Misharin, Martijn C Nawijn, Malte D Luecken, and Fabian J Theis. An integrated cell atlas of the lung in health and disease. *Nat. Med.*, 29(6):1563–1577, June 2023.

Gautam Singh, Yi-Fu Wu, and Sungjin Ahn. Simple unsupervised object-centric learning for complex and naturalistic videos. In S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho, and A Oh (eds.), *Advances in Neural Information Processing Systems 35*, volume 35, pp. 18181–18196, San Diego, California, USA, 2022. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Chandler Squires, Anna Seigal, Salil S Bhate, and Caroline Uhler. Linear causal disentanglement via interventions. In *International Conference on Machine Learning*, pp. 32540–32560. PMLR, July 2023.

Burak Varici, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. General identifiability and achievability for causal representation learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 2314–2322. PMLR, April 2024.

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. *arXiv [stat.ML]*, June 2021.

Nicholas Watters, Loic Matthey, Christopher P Burgess, and Alexander Lerchner. Spatial broadcast decoder: A simple architecture for learning disentangled representations in VAEs. *arXiv [cs.LG]*, January 2019.

Thaddäus Wiedemer, Jack Brady, Alexander Panfilov, Attila Juhos, Matthias Bethge, and Wieland Brendel. Provable compositional generalization for object-centric learning. *arXiv [cs.LG]*, October 2023.

Pengzhou Wu. Beyond ICA: Identifiability by symmetry breaking. *arXiv [stat.ML]*, July 2026.

Chuan Xu, Martin Prete, Simone Webb, Laura Jardine, Benjamin J Stewart, Regina Hoo, Peng He, Kerstin B Meyer, and Sarah A Teichmann. Automatic cell-type harmonization and integration across human cell atlas datasets. *Cell*, 186(26):5876–5891.e20, December 2023.

Jinyang Yuan, Bin Li, and X Xue. Generative modeling of infinite occluded objects for compositional scene representation. *ICML*, 97:7222–7231, May 2019.

Jinyang Yuan, Tonglin Chen, Bin Li, and Xiangyang Xue. Compositional scene representation learning via reconstruction: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(10):11540–11560, October 2023.

Jiaqi Zhang, Kristjan Greenewald, Chandler Squires, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. In A Oh, T Naumann, A Globerson, K Saenko, M Hardt, and S Levine (eds.), *Advances in Neural Information Processing Systems 36*, volume 36, pp. 50254–50292, San Diego, California, USA, 2023. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Yujia Zheng and Kun Zhang. Generalizing nonlinear ICA beyond structural sparsity. In *Advances in Neural Information Processing Systems 36*, pp. 13326–13355, San Diego, California, USA, 2023. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Yujia Zheng, Ignavier Ng, and Kun Zhang. On the identifiability of nonlinear ica: Sparsity and beyond. *Advances in neural information processing systems*, 35:16411–16422, 2022.

Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pp. 12979–12990. PMLR, July 2021.

APPENDICES

A Abbreviations **17**

B Notation and Terminology **17**

C Definitions and Assumptions **18**

D Standard Results **19**

E Switching Points and Occlusion **19**

F Local Disentanglement **20**

F.1 Proof of Theorem 7 21

G From Local to Global **23**

G.1 Proof of Theorem 8 28

H Proof of Theorem 9 **29**

I Proof of Proposition 10 **29**

J Extended Literature Review **30**

K Experimental Details and Additional Results **30**

K.1 Synthetic images 30

K.2 CLEVR6 31

K.3 Single-cell transcriptomics 35

A Abbreviations

Table 4: Abbreviations.

Abbreviation	Meaning
Models and pooling	
LSE, LSE$_\tau$	log-sum-exp pooling at inverse temperature τ (Definition 4)
IA	Interaction Asymmetry, the decoder and training recipe of Brady et al. (2024)
DRVI	Disentangled Representation Variational Inference (Moinfar & Theis, 2024)
IODINE	Iterative Object Decomposition Inference Network (Greff et al., 2019)
VAE	variational autoencoder
MLP	multilayer perceptron
CNN	convolutional neural network
SiLU, ELU	sigmoid-weighted linear unit, exponential linear unit (activations)
KL	Kullback–Leibler divergence
Data	
CLEVR6	CLEVR scenes with at most six objects (Greff et al., 2019; Locatello et al., 2020b)
HLCA	Human Lung Cell Atlas (Sikkema et al., 2023)
scRNA-seq	single-cell RNA sequencing
GT	ground truth
Metrics	
block-MIG	block mutual-information gap, a variant of the MIG of Chen et al. (2018): recovery score on synthetic images (Appendix K.1)
Mix	exclusivity score on synthetic images: largest off-diagonal over smallest diagonal Jacobian entry after matching (Appendix K.1)
R^2	coefficient of determination of the reconstruction
ARI	adjusted Rand index
J-ARI	Jacobian ARI: ARI between objects and the slot with the largest decoder-Jacobian norm per pixel (Brady et al., 2024)
JIS	Jacobian interaction score: rescaled share of the decoder-Jacobian norm held by the top slot (Brady et al., 2024)
SMI	scaled mutual information: latent–cell-type association score (Moinfar & Theis, 2024)
LMS	latent matching score: mean association under a one-to-one matching of cell types to latents (Lachapelle et al., 2023; Moinfar & Theis, 2024)
Theory	
A1–A6	assumptions of Section 3.1, restated in Appendix C
SR1–SR6	standard results of Appendix D

B Notation and Terminology

Table 5: Notation and terminology of Section 3 and the appendix. Hatted symbols belong to the learned model, unhatted ones to the ground truth.

Symbol	Meaning
$[k]$	the slots $\{1,\dots,k\}$, for the true and for the learned model. i,j,q,w index true slots, m,μ learned slots
$n_i \geq 1$, $K := \sum_i n_i$	dimension of true slot i, total dimension
$\mathbf{z} = (z_1,\dots,z_k)$, $z_i \in \mathbb{R}^{n_i}$	true latents, domain $\mathcal{Z} \subseteq \mathbb{R}^K$ with nonempty interior $\mathcal{Z}^\circ$. We write $\mathbf{z} = (z_i, z_{-i})$ with z_{-i} the other blocks
$\hat{n}_m \geq 1$, $\boldsymbol{y} = (y_1,\dots,y_k)$, $y_m \in \mathbb{R}^{\hat{n}_m}$	learned latents, domain $\widehat{\mathcal{Z}} \subseteq \mathbb{R}^{\sum_m \hat{n}_m}$ with interior $\widehat{\mathcal{Z}}^\circ$
$l \in [d]$	the d features (for example pixels)
$s_{i,l},\ g_{i,l} : \mathbb{R}^{n_i} \to \mathbb{R}$	true score and true slot map of slot i on feature l
$\hat{s}_{m,l},\ \widehat{g}_{m,l} : \mathbb{R}^{\hat{n}_m} \to \mathbb{R}$	learned score and learned slot map
$\mathbf{h} = (\mathbf{h}_1,\dots,\mathbf{h}_k)$, $\mathbf{h}_m(\mathbf{z}) \in \mathbb{R}^{\hat{n}_m}$	mixing map $\mathcal{Z}^\circ \to \widehat{\mathcal{Z}}^\circ$ from the true decoder $\mathbf{f} = (f_l)_{l\in[d]}$ to the learned decoder $\widehat{\mathbf{f}} = (\widehat{f}_l)_{l\in[d]}$ (Definition 4), $\mathbf{f} = \widehat{\mathbf{f}} \circ \mathbf{h}$
box	a product $N = \prod_i N_i \subseteq \mathcal{Z}^\circ$ of nonempty products $N_i \subseteq \mathbb{R}^{n_i}$ of open intervals. N_i is its i-th side. Other names of boxes, such as $M, N', N'', N''', N^0, N^\dagger$, give sides $M_i, N'_i, \dots$ in the same way. A **learned box** is a product $\prod_\mu V_\mu$ of nonempty products $V_\mu \subseteq \mathbb{R}^{\hat{n}_\mu}$ of open intervals
$\mathbf{g}_{i,T}$, $\widehat{\mathbf{g}}_{m,T}$	for a feature set $T \subseteq [d]$, $\mathbf{g}_{i,T}(z_i) := (g_{i,l}(z_i))_{l\in T}$ and $\widehat{\mathbf{g}}_{m,T}(y_m) := (\widehat{g}_{m,l}(y_m))_{l\in T}$
∂u	the Jacobian matrix of a map u. ∇ is the gradient of a scalar map, written as a column
$\mathrm{Perm}(k)$	permutations of $[k]$

C Definitions and Assumptions

This section repeats the definitions and assumptions of Sections 2 and 3, with their original numbers, so that the appendix can be read on its own. $\mathrm{Perm}(k)$ is the set of permutations of $[k]$.

Definition 1 (Box and mixing map)**.** A *box* is a product $N = \prod_i N_i \subseteq \mathcal{Z}^\circ$ of nonempty products $N_i \subseteq \mathbb{R}^{n_i}$ of open intervals, and N_i is its i-th side. A *mixing map* relates the two models, $\mathbf{h} = (\mathbf{h}_1,\dots,\mathbf{h}_k) : \mathcal{Z}^\circ \to \widehat{\mathcal{Z}}^\circ$ with $\mathbf{h}_m(\mathbf{z}) \in \mathbb{R}^{\hat{n}_m}$ and $\mathbf{f} = \widehat{\mathbf{f}} \circ \mathbf{h}$.

Definition 2 (Disentangled boxes and regions)**.** $\mathbf{h}$ is **block-disentangled on** a box N **with** $\pi \in \mathrm{Perm}(k)$ if there are continuous injective $\sigma_i : N_i \to \mathbb{R}^{\hat{n}_{\pi(i)}}$ with

$$\mathbf{h}_{\pi(i)}(\mathbf{z}) = \sigma_i(z_i) \qquad \forall\, \mathbf{z} \in N,\ i \in [k].$$

$$D_\pi := \{\mathbf{z} :\ \exists \text{ box } N \text{ with } \mathbf{z} \in N \text{ on which } \mathbf{h} \text{ is block-disentangled with } \pi\},$$
$$D := \textstyle\bigcup_{\pi\in\mathrm{Perm}(k)} D_\pi.$$

Such an N is a **disentangled box** (with π). D_π is the **region** of π.

Definition 3 (Global disentanglement)**.** $\mathbf{h}$ is *globally disentangled* if $D_\pi = \mathcal{Z}^\circ$ for some $\pi \in \mathrm{Perm}(k)$, that is, if a single permutation π has a disentangled box around every point of $\mathcal{Z}^\circ$.

Definition 4 (Decoders)**.** For each slot i and feature l, let $g_{i,l} : \mathbb{R}^{n_i} \to \mathbb{R}$ be a *slot map*. A pooled decoder $\mathbf{f}$ combines the slot maps of each feature by an operation over slots. We consider four types of decoders (hard-max is max-mask with $s_{i,l} = g_{i,l}$):

(i) Additive decoder (Lachapelle et al., 2023): $f_l^{\mathrm{add}}(\mathbf{z}) = \sum_{i=1}^k g_{i,l}(z_i)$.

(ii) LSE$_\tau$ decoder: The slots are combined by a smooth Log-Sum-Exp pooling with inverse temperature $\tau > 0$, $f_{\tau,l}(\mathbf{z}) = \frac{1}{\tau}\log\frac{1}{k}\sum_{i=1}^k e^{\tau g_{i,l}(z_i)}$, as in equation 1.

(iii) Hard-max decoder: The slots compete independently per feature, $f_l^{\max}(\mathbf{z}) = \max_i g_{i,l}(z_i)$.

(iv) Max-mask decoder: The winner of a feature is chosen by a separate *score* $s_{i,l} : \mathbb{R}^{n_i} \to \mathbb{R}$ rather than by the output itself. The *winners* of feature l and the decoder are

$$W_l(\mathbf{z}) := \arg\max_i\ s_{i,l}(z_i), \qquad w := \min W_l(\mathbf{z}), \quad f_l(\mathbf{z}) := g_{w,l}(z_w), \qquad l \in [d]. \tag{3}$$

In the appendix $\mathbf{f}$ and $\widehat{\mathbf{f}}$ are max-mask decoders, with $\widehat{g}_{m,l}$ and $\hat{s}_{m,l}$ defined in the same way as $g_{i,l}$ and $s_{i,l}$, and no correspondence between the true and the learned slots is assumed. A hard-max decoder lets the output itself decide which slot wins a feature. Object-centric decoders do not work that way. They emit a separate mask, and the slot with the largest mask supplies the feature,

whatever its value. Writing that mask as a score $s_{i,l}$ gives equation 3. Sharing one score across the three colour features of a pixel recovers the z-buffer, in which the winning slot supplies the whole pixel, but nothing in the results requires it.

Definition 5 (Ownership matrix and owned features)**.** The features owned by true slot i at $\mathbf{z}$ are $O_i(\mathbf{z}) := \{l \in [d] : \ W_l(\mathbf{z}) = \{i\},\ \nabla g_{i,l}(z_i) \neq 0\}$. The ownership matrix $G(\mathbf{z}) \in \mathbb{R}^{K\times d}$ has block rows $i \in [k]$ of heights n_i, with $G_{i,l}(\mathbf{z}) := \nabla g_{i,l}(z_i) \in \mathbb{R}^{n_i}$ if $W_l(\mathbf{z}) = \{i\}$ and $G_{i,l}(\mathbf{z}) := 0$ otherwise, and its column l is written $G_{\cdot,l}(\mathbf{z}) \in \mathbb{R}^K$.

Properties of the ownership matrix. For a feature set T, $\operatorname{rank}_i(T, \mathbf{z})$ is the rank of the $n_i \times |T|$ matrix $\big(G_{i,l}(\mathbf{z})\big)_{l\in T}$ (Definition 5). Every column of G has at most one nonzero block, so

$$\operatorname{rank} G(\mathbf{z}) = \textstyle\sum_i \operatorname{rank}_i\big(O_i(\mathbf{z}), \mathbf{z}\big), \qquad \operatorname{rank} G(\mathbf{z}) = K \iff \operatorname{rank}_i\big(O_i(\mathbf{z}), \mathbf{z}\big) = n_i \ \forall\, i.$$

Owning is an open condition: $l \in O_i(\mathbf{z})$ means $s_{i,l}(z_i) > s_{j,l}(z_j)$ for all $j \neq i$ and $\nabla g_{i,l}(z_i) \neq 0$, finitely many strict conditions on continuous functions.

Assumptions. A1 to A4 give local disentanglement (Appendix F), and A5 and A6 are needed only to make it global (Appendix G).

Assumption A1 (Continuous learned slot maps)**.** Every $g_{i,l}$ is C^1, every $s_{i,l}$ is continuous, every $\widehat{g}_{m,l}$ is continuous. Nothing is assumed about $\hat{s}_{m,l}$.

Assumption A2 (Homeomorphic mixing)**.** $\mathbf{h} : \mathcal{Z}^\circ \to \widehat{\mathcal{Z}}^\circ$ is a homeomorphism and $\mathbf{f} = \widehat{\mathbf{f}} \circ \mathbf{h}$ on $\mathcal{Z}^\circ$.

Assumption A3 (Full rank)**.** $\operatorname{rank} G(\mathbf{z}) = K$ for every $\mathbf{z} \in \mathcal{Z}^\circ$.

Assumption A4 (Irreducible ownership)**.** For $\mathbf{z}$ in a dense subset of $\mathcal{Z}^\circ$, every i and every split $O_i(\mathbf{z}) = T \sqcup T'$ with $T, T' \neq \emptyset$: $\operatorname{rank}_i(T, \mathbf{z}) + \operatorname{rank}_i(T', \mathbf{z}) > n_i$, where $\operatorname{rank}_i(T, \mathbf{z})$ is the rank of $\big(G_{i,l}(\mathbf{z})\big)_{l\in T}$.

Assumption A5 (Connected block slices)**.** Every block slice $\mathcal{Z}_{i,c} := \{\mathbf{z} \in \mathcal{Z}^\circ : \ z_i = c\}$, $i \in [k]$, $c \in \mathbb{R}^{n_i}$, is connected.

Assumption A6 (Connected domain)**.** $\mathcal{Z}^\circ$ is connected.

A2 gives the dimension sum $\sum_m \hat{n}_m = K$, since $\mathcal{Z}^\circ$ and $\widehat{\mathcal{Z}}^\circ = \mathbf{h}(\mathcal{Z}^\circ)$ are open and nonempty and SR3 applies to $\mathbf{h}$ and to $\mathbf{h}^{-1}$. We say that A3 holds at $\mathbf{z}$ if $\operatorname{rank} G(\mathbf{z}) = K$, and that A4 holds at $\mathbf{z}$ if its split condition holds at $\mathbf{z}$ for every i. The split condition of A4 holds automatically for every i with $n_i = 1$, since two nonempty sets of owned features have rank at least 1 each.

D Standard Results

The proofs reference the following standard results by label. In these results, Ω is an open set and u a map.

SR1 (finite cover). If a box is a finite union of relatively closed sets, one of them contains a box.

SR2 (invariance of domain). If $\Omega \subseteq \mathbb{R}^p$ is open and $u : \Omega \to \mathbb{R}^p$ is continuous and injective, then $u(\Omega)$ is open and u is a homeomorphism onto $u(\Omega)$.

SR3 (no squeezing). If $\Omega \subseteq \mathbb{R}^p$ is open and nonempty and $u : \Omega \to \mathbb{R}^{p'}$ is continuous and injective, then $p \leq p'$.

SR4 (inverse function theorem). If $\Omega \subseteq \mathbb{R}^p$ is open, $u : \Omega \to \mathbb{R}^p$ is C^1 and $\det \partial u(\omega) \neq 0$ at some $\omega \in \Omega$, then there is an open $\Omega_0 \subseteq \Omega$ with $\omega \in \Omega_0$ such that $u(\Omega_0)$ is open and u is a C^1 diffeomorphism of Ω_0 onto $u(\Omega_0)$. In particular u is injective on Ω_0.

SR5 (rank of a union). The rank of the columns indexed by a finite union of feature sets is at most the sum of their ranks.

SR6 (a product of intervals minus a point). If $\Omega \subseteq \mathbb{R}^p$ with $p \geq 2$ is a nonempty product of open intervals and $\omega \in \Omega$, then $\Omega \setminus \{\omega\}$ is connected.

E Switching Points and Occlusion

Proposition 6 (A differentiable decoder cannot switch visibly)**.** *Assume A1 (continuous learned slot maps) and fix a feature l. For a slot i put $U_i := \{\mathbf{z} \in \mathcal{Z}^\circ : \ W_l(\mathbf{z}) = \{i\}\}$, which is open, and call*

$\mathbf{z}^0 \in \mathcal{Z}^\circ$ *a* switching point of l between i and j *if* $i \neq j$ *and* $\mathbf{z}^0 \in \overline{U_i} \cap \overline{U_j}$. *At every such point the following hold.*

1. *If f_l is continuous at $\mathbf{z}^0$, then $g_{i,l}(z_i^0) = g_{j,l}(z_j^0)$.*
2. *If f_l is C^1 on a neighbourhood of $\mathbf{z}^0$, then also $\nabla g_{i,l}(z_i^0) = 0$ and $\nabla g_{j,l}(z_j^0) = 0$.*

Proof of Proposition 6. U_i is open, since $W_l(\mathbf{z}) = \{i\}$ means $s_{i,l}(z_i) > s_{i',l}(z_{i'})$ for every $i' \neq i$, finitely many strict inequalities between continuous functions (A1). On U_i the winner set is the singleton $\{i\}$, so $f_l(\mathbf{z}) = g_{i,l}(z_i)$ there (Definition 4).

Part (1). Take $\mathbf{z}^k \in U_i$ with $\mathbf{z}^k \to \mathbf{z}^0$, which exist as $\mathbf{z}^0 \in \overline{U_i}$. Then $f_l(\mathbf{z}^k) = g_{i,l}(z_i^k) \to g_{i,l}(z_i^0)$, since $g_{i,l}$ is continuous, and $f_l(\mathbf{z}^k) \to f_l(\mathbf{z}^0)$ by the assumed continuity of f_l, so $f_l(\mathbf{z}^0) = g_{i,l}(z_i^0)$. The same argument on U_j gives $f_l(\mathbf{z}^0) = g_{j,l}(z_j^0)$.

Part (2). U_i is open and f_l agrees on it with $\mathbf{z} \mapsto g_{i,l}(z_i)$, so at every $\mathbf{z} \in U_i$ the gradient $\nabla f_l(\mathbf{z})$ has block i equal to $\nabla g_{i,l}(z_i)$ and all its other blocks zero. Let $\mathbf{z}^k \in U_i$ with $\mathbf{z}^k \to \mathbf{z}^0$. Both ∇f_l and $\nabla g_{i,l}$ are continuous, the first because f_l is C^1 near $\mathbf{z}^0$, so taking $k \to \infty$ blockwise gives

$$\nabla_{z_i} f_l(\mathbf{z}^0) = \nabla g_{i,l}(z_i^0), \qquad \nabla_{z_j} f_l(\mathbf{z}^0) = 0.$$

The same argument on U_j gives $\nabla_{z_j} f_l(\mathbf{z}^0) = \nabla g_{j,l}(z_j^0)$ and $\nabla_{z_i} f_l(\mathbf{z}^0) = 0$. Comparing the two pairs gives $\nabla g_{i,l}(z_i^0) = 0$ and $\nabla g_{j,l}(z_j^0) = 0$. □

A switching point of feature l is where the slot that supplies l changes, which is what occlusion does at the contour of an object: on one side the pixel shows the object in front, on the other side the one behind. Part (1) says that a decoder that is merely continuous there has to give both objects the same value, so the contour is invisible. Part (2) says that a C^1 decoder has to place a critical point of both slot maps exactly on the contour.

F Local Disentanglement

Definition 11 (Agreement sets)**.** For a learned slot m and a feature l,

$$R_{m,l} := \big\{\mathbf{z} \in \mathcal{Z}^\circ : \ f_l(\mathbf{z}) = \widehat{g}_{m,l}\big(\mathbf{h}_m(\mathbf{z})\big)\big\}, \qquad \bigcup_{m\in[k]} R_{m,l} = \mathcal{Z}^\circ,$$

where the cover holds because $f_l(\mathbf{z}) = \widehat{f}_l(\mathbf{h}(\mathbf{z})) \in \{\widehat{g}_{m,l}(\mathbf{h}_m(\mathbf{z}))\}_m$ (A2 (homeomorphic mixing) and Definition 4). If $W_l \equiv \{i\}$ on a box, then $f_l = g_{i,l}(z_i)$ there, and $R_{m,l}$ meets that box in a relatively closed set when $g_{i,l}$, $\widehat{g}_{m,l}$ and $\mathbf{h}_m$ are continuous.

Overview. Freeze which true slot owns each feature and which learned slot reproduces it. At one point, count directions. If features that are a function of y_m have ρ independent gradients, then $\mathbf{h}_m$ is injective on a ρ-dimensional affine slice, so $\rho \leq \hat{n}_m$ by invariance of domain (Lemma 12). The dimensions add up to K on both sides, so no direction is wasted, and A4 (irreducible ownership) forbids splitting the features of a true slot. This gives the permutation. Finally $\mathbf{h}$ maps a box onto an open set, which contains a learned box, so one learned block can move while the others stay fixed. So a learned block that reproduces local coordinates of z_i is a function of z_i (Lemma 13). Theorem 7 uses ranks at one point only and needs the learned slot maps only to be continuous.

Steps 4 to 7 of the proof of Theorem 7 are the counting argument of Brady et al. (2023, App. A), done at one point and without the chain rule for $\mathbf{h}$. Lemma 13 is the anchor-feature inversion of Moran et al. (2021) for blocks.

Lemma 12 (A learned block cannot raise rank)**.** *Assume A2 (homeomorphic mixing). Let N be a box, $m \in [k]$ and $T \subseteq [d]$ a feature set such that $\widehat{\mathbf{g}}_{m,T} \circ \mathbf{h}_m$ is C^1 on N. Then*

$$\operatorname{rank} \partial\big(\widehat{\mathbf{g}}_{m,T} \circ \mathbf{h}_m\big)(\mathbf{z}) \leq \hat{n}_m \qquad \forall\, \mathbf{z} \in N.$$

Proof. Step 1, independent rows. Fix $\mathbf{z}^0 \in N$ and let ρ be the rank at $\mathbf{z}^0$. If $\rho = 0$ there is nothing to prove. Some ρ rows of the Jacobian are independent, so there is $T_0 \subseteq T$ with $|T_0| = \rho$ and

$$J := \partial\big(\widehat{\mathbf{g}}_{m,T_0} \circ \mathbf{h}_m\big)(\mathbf{z}^0) \in \mathbb{R}^{\rho\times K}, \qquad \operatorname{rank} J = \rho.$$

Step 2, an injective affine slice. The map $e : \mathbb{R}^\rho \to \mathbb{R}^K$, $e(t) := \mathbf{z}^0 + J^\top t$, is the affine slice through $\mathbf{z}^0$ along the ρ independent gradients. On the open set $e^{-1}(N)$, which contains 0, the map $\varphi := \widehat{\mathbf{g}}_{m,T_0} \circ \mathbf{h}_m \circ e$ is C^1, and by the chain rule

$$\partial\varphi(0) = JJ^\top, \quad \text{invertible as } J \text{ has independent rows.}$$

SR4 gives an open $P \subseteq e^{-1}(N)$ with $0 \in P$ and φ injective on P.

Step 3, no squeezing. For $t, t' \in P$,

$$\mathbf{h}_m(e(t)) = \mathbf{h}_m(e(t')) \implies \varphi(t) = \varphi(t') \overset{\text{Step 2}}{\implies} t = t'.$$

So $\mathbf{h}_m \circ e : P \to \mathbb{R}^{\hat{n}_m}$ is injective. It is continuous, since $\mathbf{h}$ is (A2 (homeomorphic mixing)). SR3 gives $\rho \le \hat{n}_m$. □

Lemma 12 assumes nothing about the smoothness of $\mathbf{h}$. Its hypothesis is on the composite $\widehat{\mathbf{g}}_{m,T} \circ \mathbf{h}_m$ only, and the proof differentiates only this composite. Step 4 of the proof of Theorem 7 applies it on a box where the composite equals true features, $\widehat{\mathbf{g}}_{m,l}(\mathbf{h}_m(\mathbf{z})) = g_{i,l}(z_i)$ by (E1) in Step 2 of the proof of Theorem 7, which are C^1 by A1 (continuous learned slot maps).

Lemma 13 (Sufficient conditions for local block disentanglement)**.** *Assume A2 (homeomorphic mixing). Let M be a box, $\pi \in \mathrm{Perm}(k)$, and for every $i \in [k]$ let $\alpha_i : M_i \to \mathbb{R}^{n_i}$ be injective and $\beta_i : \mathbb{R}^{\hat{n}_{\pi(i)}} \to \mathbb{R}^{n_i}$ any map, with*

$$\beta_i\big(\mathbf{h}_{\pi(i)}(\mathbf{z})\big) = \alpha_i(z_i) \qquad \forall\, \mathbf{z} \in M,\ i \in [k].$$

Then every $\mathbf{z}^0 \in M$ lies in a box $N \subseteq M$ on which $\mathbf{h}$ is block-disentangled with π.

Proof. Step 1, a learned box. $\mathbf{h}(M)$ is open, since $\mathbf{h}$ is a homeomorphism onto the open set $\widehat{\mathcal{Z}}^\circ$ (A2 (homeomorphic mixing)). So there is a learned box $V := \prod_{\mu\in[k]} V_\mu$ with $\mathbf{h}(\mathbf{z}^0) \in V \subseteq \mathbf{h}(M)$. $\mathbf{h}$ is continuous, so some box $N \subseteq M$ with $\mathbf{z}^0 \in N$ has $\mathbf{h}(N) \subseteq V$.

Step 2, an injective map on V. Put $\alpha(\mathbf{z}) := (\alpha_i(z_i))_{i\in[k]}$ for $\mathbf{z} \in M$, injective as every α_i is, and $\psi(\boldsymbol{y}) := (\beta_i(y_{\pi(i)}))_{i\in[k]}$ for $\boldsymbol{y} \in \mathbb{R}^{\sum_\mu \hat{n}_\mu}$. The hypothesis says $\psi \circ \mathbf{h} = \alpha$ on M. Every $\boldsymbol{y} \in V$ is $\mathbf{h}(\mathbf{z})$ for exactly one $\mathbf{z} \in M$ (Step 1, A2 (homeomorphic mixing)), so

$$\psi = \alpha \circ \mathbf{h}^{-1} \text{ on } V \qquad \implies \qquad \psi \text{ is injective on } V.$$

Step 3, the candidate block. Fix i and put $\sigma_i(z_i) := \mathbf{h}_{\pi(i)}(z_i, z^0_{-i})$ for $z_i \in N_i$, a continuous map. For $\mathbf{z} \in N$ let $\boldsymbol{y}'$ be $\mathbf{h}(\mathbf{z})$ with block $\pi(i)$ replaced by $\sigma_i(z_i)$. Then $\boldsymbol{y}' \in V$, since $\mathbf{h}(\mathbf{z}) \in V$, $\mathbf{h}(z_i, z^0_{-i}) \in V$ (as $(z_i, z^0_{-i}) \in N$) and V is a product.

Step 4, equal values of ψ. π is injective, and the hypothesis holds at $\mathbf{z}$ and at (z_i, z^0_{-i}), so

$$\begin{aligned} &y'_{\pi(j)} = \mathbf{h}_{\pi(j)}(\mathbf{z})\ \ \forall\, j \neq i, \qquad \beta_i\big(y'_{\pi(i)}\big) = \alpha_i(z_i) = \beta_i\big(\mathbf{h}_{\pi(i)}(\mathbf{z})\big) \\ &\implies \quad \psi(\boldsymbol{y}') = \psi(\mathbf{h}(\mathbf{z})). \end{aligned}$$

Step 5. Step 2 gives $\boldsymbol{y}' = \mathbf{h}(\mathbf{z})$, so $\mathbf{h}_{\pi(i)}(\mathbf{z}) = \sigma_i(z_i)$. σ_i is injective, since $\beta_i \circ \sigma_i = \alpha_i$ on N_i. □

F.1 Proof of Theorem 7

Theorem 7 (Every open set contains a disentangled box, and true and learned block dimensions match)**.** *Assume A1 (continuous learned slot maps), A2 (homeomorphic mixing), A3 (full rank), A4 (irreducible ownership). Then*

1. *D is open in $\mathcal{Z}^\circ$.*
2. *Every nonempty open $U \subseteq \mathcal{Z}^\circ$ contains a box N on which $\mathbf{h}$ is block-disentangled with some $\pi \in \mathrm{Perm}(k)$, and $\hat{n}_{\pi(i)} = n_i$ for all i.*
3. *D is dense in $\mathcal{Z}^\circ$.*

Proof of Theorem 7. Step 0, proves (1). Let $\mathbf{z} \in D_\pi$, with a box N around $\mathbf{z}$ on which $\mathbf{h}$ is block-disentangled with π (Definition 2). The same box N serves every point of N, so $N \subseteq D_\pi$. So D_π contains an open box around each of its points, and D_π and D are open.

Step 1, a box with fixed ownership (L). Let $U \subseteq \mathcal{Z}^\circ$ be open and nonempty. Take $\mathbf{z}^{\max} \in U$ where the integer $\sum_i |O_i(\mathbf{z})| \in \{0, \dots, d\}$ is largest on U. Put $L_i := O_i(\mathbf{z}^{\max})$ and $L := \bigcup_i L_i$. Owning is an open condition (A1 (continuous learned slot maps)), so there is a box $N' \subseteq U$ with $\mathbf{z}^{\max} \in N'$ and $L_i \subseteq O_i(\mathbf{z})$ for all $\mathbf{z} \in N'$. With maximality,

$$L_i \subseteq O_i(\mathbf{z}) \ \forall i, \qquad \textstyle\sum_i |O_i(\mathbf{z})| \le \sum_i |L_i| \qquad \Longrightarrow \qquad O_i(\mathbf{z}) = L_i \qquad \forall\, \mathbf{z} \in N',\ i \in [k].$$

The sets L_i are disjoint, since $l \in L_i$ means $W_l(\mathbf{z}^{\max}) = \{i\}$.

Step 2, a box with a fixed learned slot for each owned feature (λ). For $l \in L_i$, $W_l \equiv \{i\}$ on N' (Step 1), so every $R_{m,l} \cap N'$ is relatively closed in N' (Definition 11, A1 (continuous learned slot maps), A2 (homeomorphic mixing)). Every $\mathbf{z} \in N'$ lies in some $R_{m,l}$ for each l (Definition 11), so

$$N' = \bigcup_{\lambda : L \to [k]} \ \bigcap_{l \in L} \big(R_{\lambda(l),l} \cap N'\big),$$

a union of $k^{|L|}$ relatively closed sets. SR1 gives $\lambda : L \to [k]$ and a box $N'' \subseteq N'$ with $N'' \subseteq R_{\lambda(l),l}$ for every $l \in L$. So

$$g_{i,l}(z_i) \overset{\text{Step 1}}{=} f_l(\mathbf{z}) \overset{\text{Definition 11}}{=} \widehat{g}_{\lambda(l),l}\big(\mathbf{h}_{\lambda(l)}(\mathbf{z})\big) \qquad \forall\, \mathbf{z} \in N'',\ i \in [k],\ l \in L_i. \tag{E1}$$

Step 3, the point $\mathbf{z}^0$ and full rank there. A4 (irreducible ownership) holds on a dense subset of $\mathcal{Z}^\circ$ and N'' is a nonempty open box, so we fix $\mathbf{z}^0 \in N''$ at which A4 (irreducible ownership) holds. It is used at $\mathbf{z}^0$ in Step 6. As $\mathbf{z}^0 \in N'' \subseteq N'$, Step 1 gives $O_i(\mathbf{z}^0) = L_i$, and A3 (full rank) with the properties of G gives

$$\operatorname{rank}_i(L_i, \mathbf{z}^0) = \operatorname{rank}_i\big(O_i(\mathbf{z}^0), \mathbf{z}^0\big) = n_i \qquad \forall\, i \in [k].$$

For Steps 4 to 6 put

$$L_{i,m} := \{l \in L_i :\ \lambda(l) = m\}, \qquad r_{i,m} := \operatorname{rank}_i(L_{i,m}, \mathbf{z}^0),$$

the features of true slot i given to learned slot m and their rank, with $r_{i,m} = 0$ if $L_{i,m} = \emptyset$.

Step 4, upper bound. Fix m and put $L_{\cdot,m} := \bigcup_i L_{i,m}$. By (E1) from Step 2, $\widehat{\mathbf{g}}_{m,L_{\cdot,m}} \circ \mathbf{h}_m = (g_{i,l}(z_i))_{l \in L_{\cdot,m}}$ on N'', with i the slot with $l \in L_i$, which is C^1 (A1 (continuous learned slot maps)). Its row $l \in L_{i,m}$ at $\mathbf{z}^0$ is $G_{\cdot,l}(\mathbf{z}^0)^\top$, since $W_l(\mathbf{z}^0) = \{i\}$ (Step 1), which is nonzero only in block i. So the ranks of the groups of rows add, and

$$\begin{aligned} \textstyle\sum_i r_{i,m} &= \operatorname{rank} \partial\big(g_{i,l}(z_i)\big)_{l \in L_{\cdot,m}}(\mathbf{z}^0) \\ &\overset{\text{(E1)}}{=} \operatorname{rank} \partial\big(\widehat{\mathbf{g}}_{m,L_{\cdot,m}} \circ \mathbf{h}_m\big)(\mathbf{z}^0) \overset{\text{Lemma 12}}{\le} \hat{n}_m. \end{aligned}$$

Step 5, the count is tight. SR5 for $L_i = \bigsqcup_m L_{i,m}$ and Step 3 give $n_i \le \sum_m r_{i,m}$. With Step 4 and the dimension sum,

$$K = \textstyle\sum_i n_i \le \sum_i \sum_m r_{i,m} \le \sum_m \hat{n}_m = K.$$

Both inequalities are sums of termwise inequalities in one direction, so every term is an equality:

$$\textstyle\sum_m r_{i,m} = n_i \ \ \forall i, \qquad \sum_i r_{i,m} = \hat{n}_m \ \ \forall m. \tag{E2}$$

Step 6, no split. Suppose $\emptyset \ne L_{i,m} \ne L_i$ for some i, m. A4 (irreducible ownership) at $\mathbf{z}^0$ for the split $O_i(\mathbf{z}^0) = L_i = L_{i,m} \sqcup (L_i \setminus L_{i,m})$, and SR5 for $L_i \setminus L_{i,m} = \bigsqcup_{\mu \ne m} L_{i,\mu}$, give

$$n_i < r_{i,m} + \operatorname{rank}_i(L_i \setminus L_{i,m}, \mathbf{z}^0) \le \textstyle\sum_\mu r_{i,\mu} \overset{\text{(E2)}}{=} n_i,$$

a contradiction. So $L_{i,m} \in \{\emptyset, L_i\}$ for all i, m. Also $L_i \ne \emptyset$, as $\operatorname{rank}_i(L_i, \mathbf{z}^0) = n_i \ge 1$. So each i has exactly one learned slot $\pi(i)$ with

$$L_i = L_{i,\pi(i)}, \qquad r_{i,\pi(i)} = n_i, \qquad r_{i,m} = 0 \ \ \forall\, m \ne \pi(i).$$

Step 7, the permutation. (E2) from Step 5 and Step 6 give

$$\hat{n}_m = \textstyle\sum_{i:\ \pi(i)=m} n_i \qquad \forall\, m \in [k].$$

Every $\hat{n}_m \geq 1$, so this sum is not empty: every learned slot m is $\pi(i)$ for some true slot i. A map of $[k]$ onto $[k]$ is a bijection, so $\pi \in \mathrm{Perm}(k)$. Then the sum for $m = \pi(i)$ has the single term n_i, so $\hat{n}_{\pi(i)} = n_i$.

Step 8, local coordinates. $\mathrm{rank}_i(L_i, \mathbf{z}^0) = n_i$ (Step 3) gives $B_i \subseteq L_i$ with $|B_i| = n_i$ and $\det \partial \mathbf{g}_{i,B_i}(z_i^0) \neq 0$. SR4 gives, for every i, an open set around z_i^0 on which $\mathbf{g}_{i,B_i}$ is injective. So there is a box $N''' \subseteq N''$ with $\mathbf{z}^0 \in N'''$ whose sides N_i''' lie in these sets. Since $B_i \subseteq L_i = L_{i,\pi(i)}$, (E1) from Step 2 gives

$$\mathbf{g}_{i,B_i}(z_i) = \widehat{\mathbf{g}}_{\pi(i),B_i}\big(\mathbf{h}_{\pi(i)}(\mathbf{z})\big) \qquad \forall\, \mathbf{z} \in N''',\ i \in [k].$$

Step 9, disentanglement. Lemma 13 with $M := N'''$, $\mathbf{z}^0$, $\alpha_i := \mathbf{g}_{i,B_i}$ on N_i''' and $\beta_i := \widehat{\mathbf{g}}_{\pi(i),B_i}$ gives a box $N \subseteq N''' \subseteq U$ on which $\mathbf{h}$ is block-disentangled with π. With $\hat{n}_{\pi(i)} = n_i$ (Step 7) this is (2). So $N \subseteq D \cap U$. As U was any nonempty open set, D is dense, which is (3). □

G From Local to Global

Overview. Theorem 7 gives a disentangled box inside every open set, but the permutation may change from box to box. This section shows that every point has a disentangled box (Lemma 22). So one permutation holds on a connected domain, and the local formulas glue along connected block slices (Theorem 8). Lemma 22 argues by contradiction at a bad point touched by the fewest regions, in learned coordinates, and ends with two different learned points that get the same true local coordinates. The learned slot maps are used only through continuity.

Lemma 14 (Near any point, each true block is determined by as many of its owned features as its dimension)**.** *Assume A1 (continuous learned slot maps) and A3 (full rank). Let $\mathbf{z}^0 \in \mathcal{Z}^\circ$ and let $U \subseteq \mathcal{Z}^\circ$ be open with $\mathbf{z}^0 \in U$. Then for every i there is $B_i \subseteq O_i(\mathbf{z}^0)$ with $|B_i| = n_i$, and there is a box $N \subseteq U$ with $\mathbf{z}^0 \in N$, such that for every i:*

1. *$W_l(\mathbf{z}) = \{i\}$ for all $\mathbf{z} \in N$ and $l \in B_i$.*
2. *$\mathbf{g}_{i,B_i}$ is a C^1 diffeomorphism of N_i onto an open subset of $\mathbb{R}^{n_i}$.*

Proof. Step 1, independent gradients. A3 (full rank) and the properties of G give $\mathrm{rank}_i(O_i(\mathbf{z}^0), \mathbf{z}^0) = n_i$. So there is $B_i \subseteq O_i(\mathbf{z}^0)$ with $|B_i| = n_i$ and $\det \partial \mathbf{g}_{i,B_i}(z_i^0) \neq 0$.

Step 2. Owning is an open condition, so $W_l = \{i\}$ for $l \in B_i$ near $\mathbf{z}^0$. SR4 makes $\mathbf{g}_{i,B_i}$ a C^1 diffeomorphism of a neighbourhood of z_i^0 onto an open set. A small enough box $N \subseteq U$ with $\mathbf{z}^0 \in N$ has both properties. □

Lemma 15 (Disentangled boxes match block dimensions, and the regions are open and disjoint)**.** *Assume A2 (homeomorphic mixing).*

1. *If $\mathbf{h}$ is block-disentangled on a box N with π and maps σ_i, then for every $i \in [k]$, $\hat{n}_{\pi(i)} = n_i$ and σ_i is a homeomorphism of N_i onto an open subset of $\mathbb{R}^{n_i}$. In particular $\hat{n}_{\pi(i)} = n_i$ for all i if $D_\pi \neq \emptyset$.*
2. *The sets D_π, $\pi \in \mathrm{Perm}(k)$, are open and pairwise disjoint.*

Proof. Step 1, dimensions. Each σ_i is continuous and injective on the open set $N_i \subseteq \mathbb{R}^{n_i}$, so SR3 gives $n_i \leq \hat{n}_{\pi(i)}$. With the dimension sum,

$$K = \textstyle\sum_i n_i \leq \sum_i \hat{n}_{\pi(i)} = \sum_m \hat{n}_m = K,$$

so $n_i = \hat{n}_{\pi(i)}$ for every i.

Step 2, homeomorphisms. Now σ_i is a continuous injective map of the open set $N_i \subseteq \mathbb{R}^{n_i}$ into $\mathbb{R}^{n_i}$, and SR2 gives the rest of (1).

Step 3, open. As in Step 0 of the proof of Theorem 7, D_π is a union of boxes.

Step 4, disjoint. Let $\mathbf{z} \in D_\pi \cap D_{\pi'}$ with $\pi \neq \pi'$. Take boxes N and N' containing $\mathbf{z}$ on which $\mathbf{h}$ is block-disentangled with π and maps σ_i, and with π' and maps σ'_i. $N \cap N'$ is a box. Take i with $\pi(i) \neq \pi'(i)$, and put $m := \pi(i)$ and $j := \pi'^{-1}(m)$, so $j \neq i$. Then

$$\sigma_i(z'_i) = \mathbf{h}_m(\mathbf{z}') = \sigma'_j(z'_j) \qquad \forall\, \mathbf{z}' \in N \cap N'.$$

Moving z'_i with z'_j fixed shows that σ_i is constant on the side $(N \cap N')_i$, which has more than one point. This contradicts the injectivity of σ_i. □

Lemma 16 (On the boundary of D_π, the mapping π still holds for owned features)**.** *Assume A1 (continuous learned slot maps) and A2 (homeomorphic mixing). Let N be a box, $i \in [k]$, $\pi \in \mathrm{Perm}(k)$ and $T \subseteq [d]$ a feature set with*

$$W_l(\mathbf{z}) = \{i\}, \qquad \nabla g_{i,l}(z_i) \neq 0 \qquad \forall\, \mathbf{z} \in N,\ l \in T.$$

Then

$$\mathbf{g}_{i,T}(z_i) = \widehat{\mathbf{g}}_{\pi(i),T}\big(\mathbf{h}_{\pi(i)}(\mathbf{z})\big) \qquad \forall\, \mathbf{z} \in N \cap \overline{D_\pi}.$$

Proof. Step 1, a disentangled box inside N. Let $\mathbf{z}^0 \in N \cap D_\pi$. Intersecting a box of Definition 2 with N gives a box $N^0 \subseteq N$ with $\mathbf{z}^0 \in N^0$ on which $\mathbf{h}$ is block-disentangled with π and maps σ_j.

Step 2, a closed cover. Fix $l \in T$. $W_l \equiv \{i\}$ on N^0, so $f_l(\mathbf{z}) = g_{i,l}(z_i)$ there, and the sets $R_{\mu,l} \cap N^0$, $\mu \in [k]$, are relatively closed in N^0 and cover it (Definition 11, A1 (continuous learned slot maps), A2 (homeomorphic mixing)).

Step 3, other learned slots hold on no box. Let $\mu \neq \pi(i)$ and $j := \pi^{-1}(\mu)$, so $j \neq i$. If $R_{\mu,l} \cap N^0$ contained a box N^1, then

$$g_{i,l}(z_i) = f_l(\mathbf{z}) = \widehat{g}_{\mu,l}\big(\sigma_j(z_j)\big) \qquad \forall\, \mathbf{z} \in N^1,$$

and moving z_i with z_j fixed would make $g_{i,l}$ constant on N^1_i, against $\nabla g_{i,l} \neq 0$.

Step 4, the equation on $N \cap D_\pi$. If $N^0 \not\subseteq R_{\pi(i),l}$, the open set $N^0 \setminus R_{\pi(i),l}$ (Step 2) contains a box N^2, covered by the relatively closed sets $R_{\mu,l} \cap N^2$, $\mu \neq \pi(i)$ (Step 2). SR1 gives a box inside one of them, against Step 3. So $N^0 \subseteq R_{\pi(i),l}$, and in particular

$$g_{i,l}(z^0_i) = f_l(\mathbf{z}^0) = \widehat{g}_{\pi(i),l}\big(\mathbf{h}_{\pi(i)}(\mathbf{z}^0)\big).$$

This holds for every $l \in T$.

Step 5, the closure. Both sides of the equation are continuous on N (A1 (continuous learned slot maps), A2 (homeomorphic mixing)). N is open, so every point of $N \cap \overline{D_\pi}$ is a limit of points of $N \cap D_\pi$, where the equation holds by Step 4. □

Definition 17 (Bad points and touching regions)**.** A point is **bad** if no disentangled box, for any permutation, contains it, and **good** otherwise. The permutations whose regions touch a point $\mathbf{z}$ form $\mathcal{Q}(\mathbf{z})$:

$$\Gamma := \mathcal{Z}^\circ \setminus D \quad \text{(the bad points)}, \qquad \mathcal{Q}(\mathbf{z}) := \big\{\pi \in \mathrm{Perm}(k) :\ \mathbf{z} \in \overline{D_\pi}\big\} \quad \text{for } \mathbf{z} \in \mathcal{Z}^\circ.$$

Closures are taken in $\mathcal{Z}^\circ$.

Definition 18 (Setting at a bad point touched by the fewest regions)**.** Assume A1 (continuous learned slot maps), A2 (homeomorphic mixing), A3 (full rank), A4 (irreducible ownership) and $\Gamma \neq \emptyset$. The setting is a bad point touched by the fewest regions, a small box around it, and its true local coordinates as functions of the learned latents.

The point. Fix $\mathbf{z}^\dagger \in \Gamma$ with $|\mathcal{Q}(\mathbf{z}^\dagger)|$ smallest on Γ, and put

$$\mathcal{Q}^\dagger := \mathcal{Q}(\mathbf{z}^\dagger), \qquad U^\dagger := \mathcal{Z}^\circ \setminus \textstyle\bigcup_{\pi \notin \mathcal{Q}^\dagger} \overline{D_\pi}.$$

$U^\dagger$ is open and contains $\mathbf{z}^\dagger$, as the complement of finitely many closed sets none of which contains $\mathbf{z}^\dagger$.

The box. Lemma 14 at $\mathbf{z}^\dagger$ with $U^\dagger$ gives a box $N^\dagger \subseteq U^\dagger$ with $\mathbf{z}^\dagger \in N^\dagger$ and feature sets B_i, $i \in [k]$. Put

$$\mathbf{g}^\dagger(\mathbf{z}) := \big(\mathbf{g}_{i,B_i}(z_i)\big)_{i\in[k]} \in \mathbb{R}^K \qquad \text{for } \mathbf{z} \in N^\dagger,$$

which is injective, as every $\mathbf{g}_{i,B_i}$ is injective on $N_i^\dagger$ (Lemma 14). [$N^\dagger$ and the B_i satisfy the hypothesis of Lemma 16 with $N := N^\dagger$ and $T := B_i$: $W_l = \{i\}$ on $N^\dagger$ for $l \in B_i$ (Lemma 14), and the rows $\nabla g_{i,l}^\top$, $l \in B_i$, of the invertible Jacobian $\partial \mathbf{g}_{i,B_i}$ are nonzero.]

The learned box. $\mathbf{h}(N^\dagger)$ is open (A2 (homeomorphic mixing)), so it contains a learned box $\hat{N} := \prod_{m\in[k]} \hat{N}_m$ with $\mathbf{h}(\mathbf{z}^\dagger) \in \hat{N}$. On $\hat{N}$, $\mathbf{h}^{-1}$ takes values in $N^\dagger$.

The true local coordinates. On $\hat{N}$ we define

$$\mathbf{x} := \mathbf{g}^\dagger \circ \mathbf{h}^{-1},$$

and for every $\pi \in \mathrm{Perm}(k)$

$$\hat{\mathbf{x}}^\pi(\boldsymbol{y}) := \big(\widehat{\mathbf{g}}_{\pi(i),B_i}(y_{\pi(i)})\big)_{i\in[k]} \qquad \forall\, \boldsymbol{y} \in \hat{N}.$$

$\mathbf{x}(\boldsymbol{y})$ are the true local coordinates of $\mathbf{h}^{-1}(\boldsymbol{y})$, and $\mathbf{x}$ is continuous and injective, as $\mathbf{h}^{-1}$ (A2 (homeomorphic mixing)) and $\mathbf{g}^\dagger$ are. $\hat{\mathbf{x}}^\pi(\boldsymbol{y})$ are the values that the learned slot maps give these coordinates when π assigns true slot i to learned slot $\pi(i)$.

Blocks. Block i of $\mathbf{x}$ and of $\hat{\mathbf{x}}^\pi$, in $\mathbb{R}^{n_i}$, is written $\mathbf{x}_i$ and $\hat{\mathbf{x}}_i^\pi$. The block $\hat{\mathbf{x}}_i^\pi(\boldsymbol{y})$ depends only on $y_{\pi(i)}$. For $m \in [k]$ put $\hat{N}_{-m} := \prod_{\mu\neq m} \hat{N}_\mu$ and write $\boldsymbol{y} = (y_m, y_{-m})$ with $y_{-m} \in \hat{N}_{-m}$.

Further definitions.

- A point $\boldsymbol{y} \in \hat{N}$ is **bad** if $\mathbf{h}^{-1}(\boldsymbol{y})$ is bad, and **good** otherwise.
- For $\pi \in \mathrm{Perm}(k)$, $\hat{D}_\pi := \hat{N} \cap \mathbf{h}(D_\pi)$ and $\hat{D}_\pi^{\mathrm{cl}} := \hat{N} \cap \mathbf{h}(\overline{D_\pi})$. So $\hat{D}_\pi \subseteq \hat{D}_\pi^{\mathrm{cl}}$.
- A learned slot m is **switched** if $\pi^{-1}(m) \neq \pi'^{-1}(m)$ for some $\pi, \pi' \in \mathcal{Q}^\dagger$.

Lemma 19 (Properties of the setting at a bad point)**.** *In the setting of Definition 18:*

1. *For all $\pi \in \mathrm{Perm}(k)$: $\mathbf{x} = \hat{\mathbf{x}}^\pi$ on $\hat{D}_\pi^{\mathrm{cl}}$.*
2. *Every bad point of $\hat{N}$ lies in $\hat{D}_\pi^{\mathrm{cl}}$ for every $\pi \in \mathcal{Q}^\dagger$.*
3. *The set of good points in $\hat{N}$ is the union of the sets $\hat{D}_\pi$, $\pi \in \mathcal{Q}^\dagger$, that is $\{\boldsymbol{y} \in \hat{N} : \boldsymbol{y}$ good$\} = \bigcup_{\pi\in\mathcal{Q}^\dagger} \hat{D}_\pi$. These sets are open and pairwise disjoint.*
4. *For every nonempty connected set $S \subseteq \hat{N}$ of good points there is $\pi \in \mathcal{Q}^\dagger$ with $S \subseteq \hat{D}_\pi$.*
5. *$\hat{D}_\pi \neq \emptyset$ for every $\pi \in \mathcal{Q}^\dagger$, and $|\mathcal{Q}^\dagger| \geq 2$.*

Proof. Step 1. $N^\dagger \subseteq U^\dagger$ misses $\overline{D_\pi}$ for every $\pi \notin \mathcal{Q}^\dagger$, so

$$D \cap N^\dagger \subseteq \textstyle\bigcup_{\pi\in\mathcal{Q}^\dagger} D_\pi, \qquad \mathcal{Q}(\mathbf{z}) \subseteq \mathcal{Q}^\dagger \quad \forall\, \mathbf{z} \in N^\dagger.$$

Step 2, proves (1). Let $\boldsymbol{y} \in \hat{D}_\pi^{\mathrm{cl}}$ and $\mathbf{z} := \mathbf{h}^{-1}(\boldsymbol{y})$, so $\mathbf{z} \in N^\dagger \cap \overline{D_\pi}$. Lemma 16 with $T := B_i$, for every i, gives

$$\mathbf{x}_i(\boldsymbol{y}) = \mathbf{g}_{i,B_i}(z_i) = \widehat{\mathbf{g}}_{\pi(i),B_i}\big(y_{\pi(i)}\big) = \hat{\mathbf{x}}_i^\pi(\boldsymbol{y}) \qquad \forall\, i \in [k].$$

Step 3, proves (2). Let $\boldsymbol{y}$ be bad and $\mathbf{z} := \mathbf{h}^{-1}(\boldsymbol{y}) \in \Gamma \cap N^\dagger$. Step 1 gives $\mathcal{Q}(\mathbf{z}) \subseteq \mathcal{Q}^\dagger$, and the choice of $\mathbf{z}^\dagger$ gives $|\mathcal{Q}(\mathbf{z})| \geq |\mathcal{Q}^\dagger|$. So $\mathcal{Q}(\mathbf{z}) = \mathcal{Q}^\dagger$, that is $\mathbf{z} \in \overline{D_\pi}$ for every $\pi \in \mathcal{Q}^\dagger$.

Step 4, proves (3). Every $\boldsymbol{y} \in \hat{N}$ has $\mathbf{h}^{-1}(\boldsymbol{y}) \in N^\dagger$ (Definition 18), and $\boldsymbol{y}$ is good exactly when $\mathbf{h}^{-1}(\boldsymbol{y}) \in D$ (Definition 17). With Step 1 and $D_\pi \subseteq D$,

$$\{\boldsymbol{y} \in \hat{N} :\ \boldsymbol{y} \text{ good}\} = \hat{N} \cap \mathbf{h}(D \cap N^\dagger) = \textstyle\bigcup_{\pi\in\mathcal{Q}^\dagger} \hat{D}_\pi.$$

The D_π are open and pairwise disjoint (Lemma 15 (2)). $\mathbf{h}$ is injective and maps open sets onto open sets (A2 (homeomorphic mixing)), so the $\hat{D}_\pi$ are open and pairwise disjoint.

Step 5, proves (4). Let S be a nonempty connected set of good points. The sets $S \cap \hat{D}_\pi$, $\pi \in \mathcal{Q}^\dagger$, are open in S, pairwise disjoint and cover S. If two of them were nonempty, S would split into two disjoint nonempty open parts, one of them and the union of the others, against connectedness. So S lies in one $\hat{D}_\pi$.

Step 6, proves the first half of (5). Let $\pi \in \mathcal{Q}^\dagger$. $\mathbf{h}^{-1}(\hat{N})$ is an open set containing $\mathbf{z}^\dagger \in \overline{D_\pi}$, so it meets D_π, and $\hat{D}_\pi \neq \emptyset$.

Step 7, proves the second half of (5). $D \cap N^\dagger$ is dense in $N^\dagger$ (Theorem 7 (3)) and lies in $\bigcup_{\pi \in \mathcal{Q}^\dagger} D_\pi$ (Step 1). If $|\mathcal{Q}^\dagger| \leq 1$, then $\mathcal{Q}^\dagger = \{\pi\}$, as $D \cap N^\dagger \neq \emptyset$, and $N^\dagger \subseteq \overline{D_\pi}$, and Lemma 16 with $T := B_i$ gives

$$\mathbf{g}_{i,B_i}(z_i) = \widehat{\mathbf{g}}_{\pi(i),B_i}\big(\mathbf{h}_{\pi(i)}(\mathbf{z})\big) \qquad \forall\, \mathbf{z} \in N^\dagger,\ i \in [k].$$

Lemma 13 with $M := N^\dagger$, the point $\mathbf{z}^\dagger$, $\alpha_i := \mathbf{g}_{i,B_i}$ on $N_i^\dagger$ and $\beta_i := \widehat{\mathbf{g}}_{\pi(i),B_i}$ gives a box around $\mathbf{z}^\dagger$ on which $\mathbf{h}$ is block-disentangled with π. So $\mathbf{z}^\dagger \in D_\pi$, against $\mathbf{z}^\dagger \in \Gamma$. □

Lemma 20 (Two bad points that differ only in a switched learned block are equal)**.** *In the setting of Definition 18, let $m \in [k]$ be switched. If $\boldsymbol{y}, \boldsymbol{y}' \in \hat{N}$ are bad and $y_{-m} = y'_{-m}$, then $\boldsymbol{y} = \boldsymbol{y}'$.*

Proof. As m is switched, take $\pi, \pi' \in \mathcal{Q}^\dagger$ with $\pi^{-1}(m) \neq \pi'^{-1}(m)$. For $\tau \in \{\pi, \pi'\}$ and every $i \neq \tau^{-1}(m)$, so that $y_{\tau(i)} = y'_{\tau(i)}$,

$$\mathbf{x}_i(\boldsymbol{y}) \overset{\text{Lemma 19 (1), (2)}}{=} \hat{\mathbf{x}}_i^\tau(\boldsymbol{y}) \overset{\text{Definition 18}}{=} \hat{\mathbf{x}}_i^\tau(\boldsymbol{y}') \overset{\text{Lemma 19 (1), (2)}}{=} \mathbf{x}_i(\boldsymbol{y}').$$

Every i differs from $\pi^{-1}(m)$ or from $\pi'^{-1}(m)$, so $\mathbf{x}(\boldsymbol{y}) = \mathbf{x}(\boldsymbol{y}')$, and $\boldsymbol{y} = \boldsymbol{y}'$ as $\mathbf{x}$ is injective. □

Lemma 21 (In the setting at a bad point, some line along a one-dimensional learned block passes from one region to another through a single bad point)**.** *In the setting of Definition 18 there are $m \in [k]$ with $\hat{n}_m = 1$, a bad point $\boldsymbol{y}^0 \in \hat{N}$ and $\pi^-, \pi^+ \in \mathcal{Q}^\dagger$ with $(\pi^-)^{-1}(m) \neq (\pi^+)^{-1}(m)$ such that*

$$\boldsymbol{y}^0 + t\,\boldsymbol{\epsilon}_m \in \hat{D}_{\pi^-}\ \ \forall\, t \in (t^{\min}, 0), \qquad \boldsymbol{y}^0 + t\,\boldsymbol{\epsilon}_m \in \hat{D}_{\pi^+}\ \ \forall\, t \in (0, t^{\max}).$$

Here $\boldsymbol{\epsilon}_m \in \mathbb{R}^K$ is the unit vector of the scalar learned block y_m (the learned space is $\mathbb{R}^K$ by the dimension sum), and $(t^{\min}, t^{\max}) := \{t \in \mathbb{R} :\ \boldsymbol{y}^0 + t\,\boldsymbol{\epsilon}_m \in \hat{N}\}$.

Proof. Step 1, a switched slot. By Lemma 19 (5) there are $\pi \neq \pi'$ in $\mathcal{Q}^\dagger$, so some $m \in [k]$ has $\pi^{-1}(m) \neq \pi'^{-1}(m)$. Fix π, π', m.

Step 2, lines and labels. For $c \in \hat{N}_{-m}$ put $\mathcal{L}_c := \{(a, c) :\ a \in \hat{N}_m\} \subseteq \hat{N}$, a homeomorphic copy of $\hat{N}_m$. On the good points put

$$\chi(\boldsymbol{y}) := \tau^{-1}(m) \qquad \text{for } \boldsymbol{y} \in \hat{D}_\tau,\ \tau \in \mathcal{Q}^\dagger,$$

well defined by Lemma 19 (3). Each $\mathcal{L}_c$ has at most one bad point (Lemma 20), so it has good points.

Step 3, a line with two labels. Suppose χ is constant, say $\bar{\chi}(c)$, on the good points of every $\mathcal{L}_c$. Then $\bar{\chi}$ is locally constant on $\hat{N}_{-m}$:

$$(a, c) \in \hat{D}_\tau \overset{\text{Lemma 19 (3)}}{\Longrightarrow} \exists \text{ open } Y \subseteq \hat{N}_{-m} \text{ with } c \in Y :\ (a, c') \in \hat{D}_\tau\ \ \forall\, c' \in Y$$
$$\Longrightarrow\ \bar{\chi} = \tau^{-1}(m) \text{ on } Y.$$

$\hat{N}_{-m}$ is a product of intervals, so it is connected and $\bar{\chi}$ is constant. Then χ is constant on all good points, against $\chi = \pi^{-1}(m)$ on $\hat{D}_\pi \neq \emptyset$ and $\chi = \pi'^{-1}(m)$ on $\hat{D}_{\pi'} \neq \emptyset$ (Lemma 19 (5)). So some line $\mathcal{L} := \mathcal{L}_c$ carries two labels.

Step 4, one bad point, and $\hat{n}_m = 1$.

1. The good points of $\mathcal{L}$ are not connected, since they carry two labels and so meet two disjoint regions (Lemma 19 (3), (4)).
2. $\mathcal{L}$ has a bad point, since $\mathcal{L} \cong \hat{N}_m$ is connected and its good points are not.
3. $\mathcal{L}$ has exactly one bad point $\boldsymbol{y}^0$ (Lemma 20), so $\mathcal{L} \setminus \{\boldsymbol{y}^0\}$ is not connected (by 1).
4. $a \mapsto (a, c)$ maps $\hat{N}_m \setminus \{y^0_m\}$ homeomorphically onto $\mathcal{L} \setminus \{\boldsymbol{y}^0\}$, so $\hat{N}_m$ minus a point is not connected, and SR6 gives $\hat{n}_m = 1$.

Step 5, the two halves. As $\hat{n}_m = 1$ and $\boldsymbol{y}^0 \in \mathcal{L}$ (so $c = y^0_{-m}$), we parametrise $\mathcal{L}$ by $t := a - y^0_m$, so that $t = 0$ is the bad point $\boldsymbol{y}^0$:

$$\mathcal{L} = \{\boldsymbol{y}^0 + t\,\boldsymbol{\epsilon}_m :\ t^{\min} < t < t^{\max}\}, \qquad t^{\min} < 0 < t^{\max}\ \ (\hat{N} \text{ is open}).$$

By Step 4, $\boldsymbol{y}^0$ is the only bad point of $\mathcal{L}$, so its halves $t < 0$ and $t > 0$ are nonempty connected sets of good points, so Lemma 19 (4) gives $\pi^{\pm} \in \mathcal{Q}^{\dagger}$ with

$$\{\boldsymbol{y}^0 + t\,\boldsymbol{\epsilon}_m :\ t^{\min} < t < 0\} \subseteq \hat{D}_{\pi^-}, \qquad \{\boldsymbol{y}^0 + t\,\boldsymbol{\epsilon}_m :\ 0 < t < t^{\max}\} \subseteq \hat{D}_{\pi^+}.$$

$\mathcal{L}$ carries two labels, so $(\pi^-)^{-1}(m) \neq (\pi^+)^{-1}(m)$. □

Lemma 22 (Every point has a disentangled box around it)**.** *Assume A1 (continuous learned slot maps), A2 (homeomorphic mixing), A3 (full rank), A4 (irreducible ownership). Then $\Gamma = \emptyset$, that is $D = \mathcal{Z}^{\circ}$.*

Proof. Step 1. Suppose $\Gamma \neq \emptyset$. Take the setting of Definition 18, and m, $\boldsymbol{y}^0$, $\pi^{\pm}$, $t^{\min}$, $t^{\max}$ from Lemma 21.

Step 2, a second switched slot. $(\pi^{\pm})^{-1}$ are bijections of $[k]$, so

$$(\pi^-)^{-1}(m) \neq (\pi^+)^{-1}(m) \implies \exists\, m' \neq m :\ (\pi^-)^{-1}(m') \neq (\pi^+)^{-1}(m').$$

If they agreed on $[k] \setminus \{m\}$, they would also agree at m, the one value left. So m' is switched.

Step 3, a second bad point. Fix $t^{\min} < t^- < 0 < t^+ < t^{\max}$. As $\hat{D}_{\pi^{\pm}}$ are open (Lemma 19 (3)) and $\hat{N}$ is a product,

$$\begin{aligned}&\exists\, \boldsymbol{\xi} \neq 0 \text{ supported in block } m' :\\ &\gamma(t) := \boldsymbol{y}^0 + t\,\boldsymbol{\epsilon}_m + \boldsymbol{\xi} \in \hat{N}\ \ \forall\, t \in [t^-, t^+], \qquad \gamma(t^{\pm}) \in \hat{D}_{\pi^{\pm}}.\end{aligned}$$

$\gamma([t^-, t^+])$ is connected and meets the disjoint sets $\hat{D}_{\pi^-}$ and $\hat{D}_{\pi^+}$, so

$$\gamma([t^-, t^+]) \not\subseteq \text{good points (Lemma 19 (4))} \qquad \Longrightarrow \qquad \exists\, t^1 \in (t^-, t^+) :\ \boldsymbol{y}^1 := \gamma(t^1) \text{ bad.}$$

Step 4, $t^1 \neq 0$.

$$t^1 = 0 \implies \boldsymbol{y}^1 = \boldsymbol{y}^0 + \boldsymbol{\xi},\ \ y^1_{-m'} = y^0_{-m'},\ \ \boldsymbol{y}^0, \boldsymbol{y}^1 \text{ bad} \overset{\text{Lemma 20}}{\implies} \boldsymbol{\xi} = 0,$$

a contradiction. We treat $t^1 > 0$. The case $t^1 < 0$ is the same after replacing $\boldsymbol{\epsilon}_m$ by $-\boldsymbol{\epsilon}_m$, which exchanges π^- and π^+.

Step 5, two corners in different regions. Put $\boldsymbol{y}^+ := \boldsymbol{y}^0 + t^1\boldsymbol{\epsilon}_m$ and $\boldsymbol{y}^- := \boldsymbol{y}^0 + \boldsymbol{\xi}$. Lemma 21 gives $\boldsymbol{y}^+ \in \hat{D}_{\pi^+}$. For $\boldsymbol{y}^-$: $\gamma(t)_{-m} = y^1_{-m}$ for all t, so $\gamma(t)$ is good for $t \neq t^1$ (Lemma 20), and

$$\begin{aligned}&\gamma([t^-, t^1)) \text{ connected, good, contains } \gamma(t^-) \in \hat{D}_{\pi^-}\\ &\quad \overset{\text{Lemma 19 (3), (4)}}{\implies} \gamma([t^-, t^1)) \subseteq \hat{D}_{\pi^-},\ \text{so } \boldsymbol{y}^- = \gamma(0) \in \hat{D}_{\pi^-}.\end{aligned}$$

Step 6, which true blocks move. Put $j^{\pm} := (\pi^{\pm})^{-1}(m)$ and $q^{\pm} := (\pi^{\pm})^{-1}(m')$. The four points differ as follows:

$$y^+_{-m} = y^0_{-m}, \qquad y^+_{-m'} = y^1_{-m'}, \qquad y^-_{-m'} = y^0_{-m'}, \qquad y^-_{-m} = y^1_{-m}.$$

At $\boldsymbol{y}^0, \boldsymbol{y}^1, \boldsymbol{y}^\pm$ we have $\mathbf{x} = \hat{\mathbf{x}}^{\pi^\pm}$ (Lemma 19 (1), (2)), and $\hat{\mathbf{x}}_i^\pi$ depends only on $y_{\pi(i)}$ (Definition 18). So

$$\mathbf{x}_i(\boldsymbol{y}^+) = \mathbf{x}_i(\boldsymbol{y}^0) \ \forall\, i \neq j^+, \qquad \mathbf{x}_i(\boldsymbol{y}^+) = \mathbf{x}_i(\boldsymbol{y}^1) \ \forall\, i \neq q^+,$$
$$\mathbf{x}_i(\boldsymbol{y}^-) = \mathbf{x}_i(\boldsymbol{y}^0) \ \forall\, i \neq q^-, \qquad \mathbf{x}_i(\boldsymbol{y}^-) = \mathbf{x}_i(\boldsymbol{y}^1) \ \forall\, i \neq j^-.$$

Step 7, the two corners collide. $j^+ \neq q^+$ and $q^- \neq j^-$ as $m \neq m'$, $j^+ \neq j^-$ by Lemma 21, and $q^- \neq q^+$ by Step 2. So

$$\{j^+, q^-\} \cap \{q^+, j^-\} = \emptyset.$$

Every $i \in [k]$ lies outside $\{j^+, q^-\}$ or outside $\{q^+, j^-\}$, so Step 6 gives

$$i \notin \{j^+, q^-\}: \quad \mathbf{x}_i(\boldsymbol{y}^+) = \mathbf{x}_i(\boldsymbol{y}^0) = \mathbf{x}_i(\boldsymbol{y}^-),$$
$$i \notin \{q^+, j^-\}: \quad \mathbf{x}_i(\boldsymbol{y}^+) = \mathbf{x}_i(\boldsymbol{y}^1) = \mathbf{x}_i(\boldsymbol{y}^-),$$
$$\text{so} \quad \mathbf{x}(\boldsymbol{y}^+) = \mathbf{x}(\boldsymbol{y}^-) \overset{\mathbf{x} \text{ injective}}{\Longrightarrow} \boldsymbol{y}^+ = \boldsymbol{y}^-.$$

This contradicts $y^+_{m'} = y^0_{m'} \neq y^0_{m'} + \xi_{m'} = y^-_{m'}$. So $\Gamma = \emptyset$. □

G.1 Proof of Theorem 8

Proposition 6 is the reason this proof works with a homeomorphism rather than with a differentiable map. Lachapelle et al. (2023) and Brady et al. (2023) assume a differentiable decoder, so their mixing map is C^1 and the corresponding step is short and direct: one differentiates $\mathbf{h}$, and the chain rule reads the block structure off the Jacobian. That route is closed here, because we want the wider class of decoders in which a mask chooses the winner, and there $\mathbf{f}$ may jump exactly where the mask switches (Proposition 6). We therefore assume only that $\mathbf{h}$ is a homeomorphism (A2), which is weaker than assuming a C^1 diffeomorphism. The price is the longer argument below, which replaces the chain rule by invariance of domain and by connectedness.

Theorem 8 (Each learned block is a function of one true block, with one permutation on the whole domain)**.** *Assume A1 (continuous learned slot maps), A2 (homeomorphic mixing), A3 (full rank), A4 (irreducible ownership), A5 (connected block slices), A6 (connected domain). Let* $C_i := \{z_i : \mathbf{z} \in \mathcal{Z}^\circ\} \subseteq \mathbb{R}^{n_i}$ *be the values that block* i *takes. Then:*

1. *There is one* $\pi \in \mathrm{Perm}(k)$ *with* $D_\pi = \mathcal{Z}^\circ$*, and* $\hat{n}_{\pi(i)} = n_i$ *for every* i*.*
2. *For every* i *there is a continuous* $\sigma_i : C_i \to \mathbb{R}^{\hat{n}_{\pi(i)}} = \mathbb{R}^{n_i}$ *with* $\mathbf{h}_{\pi(i)}(\mathbf{z}) = \sigma_i(z_i)$ *for every* $\mathbf{z} \in \mathcal{Z}^\circ$*.*
3. *Every* σ_i *is a local homeomorphism: each point of* C_i *has an open neighbourhood that* σ_i *maps homeomorphically onto an open set.*
4. σ_i *is injective if* $n_i = 1$*, or if* $\mathcal{Z}^\circ = \prod_j C_j$*.*

Proof of Theorem 8. Step 1, one permutation, proves (1).

$$\mathcal{Z}^\circ \overset{\text{Lemma 22}}{=} D = \bigsqcup_\pi D_\pi \ \text{(open, disjoint, Lemma 15 (2))},$$
$$\mathcal{Z}^\circ \text{ connected (A6 (connected domain))} \implies \exists!\, \pi: \ D_\pi = \mathcal{Z}^\circ.$$

Lemma 15 (1) gives $\hat{n}_{\pi(i)} = n_i$.

Step 2, a formula near every point. By Step 1, Definition 2 and Lemma 15 (1),

$$\forall\, \mathbf{z}^0 \in \mathcal{Z}^\circ \ \exists \text{ box } N \text{ with } \mathbf{z}^0 \in N, \text{ homeomorphisms } \sigma_i^N \text{ of } N_i \text{ onto open sets}:$$
$$\mathbf{h}_{\pi(i)}(\mathbf{z}) = \sigma_i^N(z_i) \ \forall\, \mathbf{z} \in N, \ i \in [k].$$

Step 3, one value on each block slice. Fix i and $c \in C_i$. By Step 2, $\mathbf{h}_{\pi(i)} = \sigma_i^N(c)$ on $N \cap \mathcal{Z}_{i,c}$, so

$$\mathbf{h}_{\pi(i)}\big|_{\mathcal{Z}_{i,c}} \text{ locally constant}, \quad \mathcal{Z}_{i,c} \text{ connected (A5 (connected block slices))}$$
$$\implies \mathbf{h}_{\pi(i)}\big|_{\mathcal{Z}_{i,c}} \equiv: \sigma_i(c).$$

So $\mathbf{h}_{\pi(i)}(\mathbf{z}) = \sigma_i(z_i)$ for all $\mathbf{z} \in \mathcal{Z}^\circ$.

Step 4, local homeomorphism, proves (2) and (3). C_i is open, as a projection of an open set. For $c = z_i^0 \in C_i$ take N from Step 2 around $\mathbf{z}^0$. As $(z_i, z_{-i}^0) \in N$ for $z_i \in N_i$,

$$\sigma_i(z_i) \overset{\text{Step 3}}{=} \mathbf{h}_{\pi(i)}(z_i, z_{-i}^0) \overset{\text{Step 2}}{=} \sigma_i^N(z_i) \qquad \forall\, z_i \in N_i.$$

So σ_i maps the open neighbourhood N_i of c homeomorphically onto an open set, and σ_i is continuous.

Step 5, injectivity for $n_i = 1$. C_i is an open interval (the projection of the connected open set $\mathcal{Z}^\circ$, A6 (connected domain)), and σ_i is continuous and injective near every point (Step 4). Suppose $\sigma_i(c) = \sigma_i(c')$ with $c < c'$. Then

$$\sigma_i \text{ is not constant on } [c, c'] \implies \exists\, c'' \in (c, c') :\ \sigma_i(c'') = \max_{[c,c']} \sigma_i > \sigma_i(c)$$

(or the same with $\min$).

Take $c_1'' < c'' < c_2''$ in an open interval $I \subseteq (c, c')$ around c'' on which σ_i is injective, and put $v := \max\{\sigma_i(c_1''), \sigma_i(c_2'')\}$, which is $< \sigma_i(c'')$ as $\sigma_i(c'')$ is the maximum and σ_i is injective on I. The intermediate value theorem gives v at a point of $[c_1'', c'')$ and at a point of $(c'', c_2'']$, two different points of I, against injectivity on I.

Step 6, injectivity for a product domain. Let $\mathcal{Z}^\circ = \prod_j C_j$, and fix $z_j \in C_j$ for $j \neq i$. Then

$$\sigma_i(c) = \sigma_i(c') \overset{\text{Step 3}}{\implies} \mathbf{h}(c, z_{-i}) = \mathbf{h}(c', z_{-i}) \overset{\text{A2 (homeomorphic mixing)}}{\implies} c = c'.$$

With Step 5 this is (4). □

H Proof of Theorem 9

Theorem 9 (Hard-max decoders are globally disentangled)**.** *Let $\mathbf{f}$ and $\widehat{\mathbf{f}}$ be hard-max decoders, and assume A1 (continuous learned slot maps), A2 (homeomorphic mixing), A3 (full rank), A4 (irreducible ownership), A5 (connected block slices) and A6 (connected domain). Then $\mathbf{h}$ is globally disentangled (Definition 3): there is one $\pi \in \mathrm{Perm}(k)$ with $D_\pi = \mathcal{Z}^\circ$, and $\hat{n}_{\pi(i)} = n_i$ for every i. Moreover, with $C_i := \{z_i :\ \mathbf{z} \in \mathcal{Z}^\circ\}$, for every i there is a continuous local homeomorphism $\sigma_i : C_i \to \mathbb{R}^{n_i}$ with $\mathbf{h}_{\pi(i)}(\mathbf{z}) = \sigma_i(z_i)$ for every $\mathbf{z} \in \mathcal{Z}^\circ$.*

Proof of Theorem 9. A hard-max decoder is the max-mask decoder with $s_{i,l} = g_{i,l}$ and $\hat{s}_{m,l} = \widehat{g}_{m,l}$. The true scores $s_{i,l} = g_{i,l}$ are continuous, since the true slot maps are C^1, and A1 (continuous learned slot maps) assumes nothing about the learned scores. So the hypotheses of Theorem 8 hold, and its parts (1) to (3) are the claims. □

I Proof of Proposition 10

Proposition 10 (A finite τ alone does not give identifiability)**.** *For every $\tau > 0$ and $d \geq k \geq 2$, with one-dimensional slots, there are LSE_τ decoders with affine slot maps and Jacobians of rank k, and a real-analytic diffeomorphism $\mathbf{h}$ with $\mathbf{f} = \widehat{\mathbf{f}} \circ \mathbf{h}$ and $\partial_{z_i} \mathbf{h}_j > 0$ for all i, j. Take $A \in \mathbb{R}^{k \times k}$ invertible and $N \in \mathbb{R}^{d \times k}$ of rank k, both entrywise positive, let $M := NA$, and set*

$$\mathbf{f}(\mathbf{z}) := \tfrac{1}{\tau} \log\big(Me^{\tau \mathbf{z}}\big), \qquad \widehat{\mathbf{f}}(\boldsymbol{y}) := \tfrac{1}{\tau} \log\big(Ne^{\tau \boldsymbol{y}}\big), \qquad \mathbf{h}(\mathbf{z}) := \tfrac{1}{\tau} \log\big(Ae^{\tau \mathbf{z}}\big), \tag{4}$$

with exp *and* log *taken entrywise.*

Proof. Feature l of $\mathbf{f}$ is $\frac{1}{\tau} \log \sum_i M_{li} e^{\tau z_i}$, an LSE_τ decoder with the affine slot maps $g_{i,l}(z_i) = z_i + \frac{1}{\tau} \log(k M_{li})$, and likewise $\widehat{\mathbf{f}}$ with N. Its Jacobian $\mathrm{diag}(Me^{\tau \mathbf{z}})^{-1} M \,\mathrm{diag}(e^{\tau \mathbf{z}})$ has rank k, as $M = NA$ does. From $e^{\tau \mathbf{h}(\mathbf{z})} = Ae^{\tau \mathbf{z}}$,

$$\widehat{\mathbf{f}}\big(\mathbf{h}(\mathbf{z})\big) = \tfrac{1}{\tau} \log\big(NAe^{\tau \mathbf{z}}\big) = \tfrac{1}{\tau} \log\big(Me^{\tau \mathbf{z}}\big) = \mathbf{f}(\mathbf{z}), \qquad \partial_{z_i} \mathbf{h}_j(\mathbf{z}) = \frac{A_{ji} e^{\tau z_i}}{(Ae^{\tau \mathbf{z}})_j} > 0.$$

$\mathbf{h}$ composes $e^{\tau \cdot}$, the invertible A, which maps the positive orthant onto an open cone inside it, and $\frac{1}{\tau} \log$, so it is a real-analytic diffeomorphism onto its open image. □

J Extended Literature Review

Employing an auxiliary variable spans works using temporal structure (Hyvarinen & Morioka, 2016; 2017; Hälvä & Hyvarinen, 2020; Klindt et al., 2020), spatial structure (Hälvä et al., 2021; 2024), several views of the same source (Gresele et al., 2020; Locatello et al., 2020a; Zimmermann et al., 2021; von Kügelgen et al., 2021; Brehmer et al., 2022; Ahuja et al., 2022), or conditioning on an observed auxiliary variable (Hyvarinen et al., 2019; Khemakhem et al., 2019; Lippe et al., 2022b;a; Ahuja et al., 2023; Squires et al., 2023; Buchholz et al., 2023; Zhang et al., 2023; Jiang & Aragam, 2023; Kügelgen et al., 2023; Varici et al., 2024).

Structure on the latent distribution. Another line of work assumes a parametric mixture prior (Kivva et al., 2022; Wu, 2026) to prevent free rotation of the latent factors and gain identifiability guarantees. Closest to our setting, Kori et al. (2024; 2025) use it to obtain identifiable object-centric representations from probabilistic slot attention.

Object-centric representation learning. One class of object-centric methods predicts objects by spatial attention on the input space (Eslami et al., 2016; Crawford & Pineau, 2019; Lin et al., 2020; Seitzer et al., 2022). Another line of work presents an image as a mixture of masked scenes (Burgess et al., 2019; Greff et al., 2019; Engelcke et al., 2019; Locatello et al., 2020b; Singh et al., 2022). The usual justification for a mask is occlusion (Yuan et al., 2019; 2023), and it is usually performed by combining the predicted concepts with a softmax mask. While additive decoders are a step towards explaining why this disentangles, they do not model masking, which is the main mechanism. To our knowledge, no prior work gives an identifiability guarantee for such mask-based decoders without assuming differentiability, as we do here.

K Experimental Details and Additional Results

Within the synthetic and single-cell settings the arms share the encoder, the per-slot networks, the widths, the budget and the seeds, and only the pooling changes. On CLEVR6 the interaction asymmetry (IA) baseline (Brady et al., 2024) instead uses its own decoder, which attends to all slots at once, and its interaction penalty. We report a *recovery score*, which asks whether each learned slot tracks one factor, object or cell type (block-MIG, J-ARI, LMS), and an *exclusivity score*, which asks whether each output depends on a single slot (Mix, JIS).

K.1 Synthetic images

Data. Both datasets render 64×64 images on a regular grid over $[0,1]^2$. Each of the three objects has one latent, drawn independently from $U(0.15, 0.85)$ on slider and from $U(0.05, 0.95)$ on clock. On slider, object i is a circle, a square or a diamond (the unit ball of the ℓ_2, ℓ_∞ or ℓ_1 norm) centred at (z_i, y_i) on its own row, $y = 0.4$, 0.5 or 0.6, with log-intensity field $u_i(p) = 3 - 25\,(\|p - c_i\|_{(i)}/r_i)^3$ and radii $r = (0.45,\ 0.332,\ 0.376)$, chosen to equalise the areas. On clock, hand i points from the centre $(0.5, 0.5)$ at the angle $-\pi/2 + 1.75\pi z_i$, with lengths 0.18, 0.30 and 0.42; its field is the maximum of a bump at the tip t_i, $3 - \|p - t_i\|^2/(2\sigma_i^2)$ with $\sigma = (0.075,\ 0.065,\ 0.057)$, and a needle along the segment from the centre to the tip, $3 - d(p, [\text{centre}, t_i])^2/(2 \cdot 0.009^2)$. The image is $x_\ell = \max(\max_i u_{i,\ell}, 0)$, with values in $[0, 3]$. Over $4{,}000$ images, 74% (slider) and 77% (clock) of the pixels are background, and more than one object has a positive field at 24.0% and 12.6% of the lit pixels. A run with seed s renders $100{,}000$ training, $8{,}000$ validation and $10{,}000$ evaluation images from seeds derived from s; every reported score is computed on the evaluation images.

Models. The encoder is an MLP $4096 \to 1024 \to 1024 \to 1024 \to 3$ with SiLU activations, followed by a batch normalisation of the three latents without learned scale or shift. Each slot decoder is an MLP $1 \to 256 \to 256 \to 4096$ with SiLU activations that sees only its own latent. LSE and the hard maximum pool a learned per-pixel background field, initialised at 0, as a fourth competitor, so that pixels can stay dark; LSE is the mean-normalised form $\frac{1}{\tau}\log\frac{1}{4}\sum_j e^{\tau F_j}$ over the four fields F_j, which tends to their mean as $\tau \to 0$ and to their maximum as $\tau \to \infty$. The additive decoder sums the three fields and uses its output bias as background; up to a scale and an offset, which the slot decoders can absorb, it is the $\tau \to 0$ end of the family. The joint MLP ($3 \to 256 \to 256 \to 4096$) decodes all latents together, and the linear decoder is a single affine map. The loss is the mean squared error between the pooled output and the image.

Training and temperature schedules. We use Adam with learning rate 10^{-3}, decayed to 0 by a cosine schedule over 100,000 steps, batch size 256, single precision, and seeds 0–4; a run takes 8–18 minutes on one GPU. An annealed τ is set at every step from the training progress $p \in [0, 1]$ and a warm-up fraction w: with $t = \min(p/w, 1)$, a linear ramp from τ_0 to τ_1 uses $\tau = \tau_0 + t(\tau_1 - \tau_0)$, and a geometric ramp $\tau = \tau_0(\tau_1/\tau_0)^t$. Each dataset has three groups of runs with the same base configuration: a decoder sweep (linear, joint MLP, additive, LSE at $\tau = 8$, hard maximum), a temperature sweep (additive, LSE at nine constant values of τ from 0.25 to 64, hard maximum) and two annealing arms (a linear ramp from 0.5 to 32 and a geometric ramp from 0.5 to 8, the schedule of the single-cell experiments and of the CLEVR6 colour-pooling runs, both with $w = 0.7$).

Scores. Reconstruction R^2 is $1 - \sum(x - \hat{x})^2 / \sum(x - \bar{x})^2$, with $\bar{x}$ the per-pixel mean. Block-MIG discretises each true factor into 8 quantile bins, estimates the mutual information between the bins and every learned latent with a nearest-neighbour estimator, and takes the information of the most informative latent minus the mean over the other two, divided by the factor's entropy; we average over the three factors. It uses no matching, so one latent can score for two factors. True latents score 1.00, and random rotations of them 0.27 on average. Mix takes the Jacobian of $\mathrm{enc} \circ \mathrm{render}$ by central differences (step 0.02 standard deviations of each latent) at 256 evaluation points; at every point it matches learned to true latents to maximise the diagonal and divides the largest off-diagonal entry by the smallest diagonal one, and we report the median over points. Coverage is the fraction of evaluation images in which every learned slot owns at least one pixel, winning it by a margin above 0.5 against the other slots and the background.

Results. Table 6 summarises every configuration and Table 7 lists every run; Figures 3 and 4 show the ownership maps and the mutual information of every decoder on both datasets. Three observations support the main text. First, failed runs are not starved of pixels: coverage is at least 0.9875 in every run of a slot decoder except one (0.931, clock, $\tau = 0.25$), and 1 in every hard-max run. Second, failures show early. After 8% of training, the failed hard-max runs on slider are at block-MIG 0.19–0.40 and the successful ones at 0.86 and 0.90; runs at a constant $\tau = 32$ are at 0.16–0.94 (median 0.34), and the linear ramp to 32, then at $\tau \approx 4.2$, at 0.65–0.97 (median 0.93). Third, a fixed seed does not fix the outcome: runs that repeat a configuration and seed in a second sweep do not always reproduce, and LSE at $\tau = 8$ on clock with seed 2, for example, reaches block-MIG 0.98 in the decoder sweep and 0.79 in the temperature sweep.

K.2 CLEVR6

Data. We use the CLEVR6 release of IA (Brady et al., 2024): 53,483 images of 128×128 pixels with instance masks, split into 49,483 training, 2,000 validation and 2,000 test images. Pixels are scaled to $[0, 1]$ without augmentation. The grayscale copy replaces every pixel by its luminance $0.299R + 0.587G + 0.114B$ in all three channels, so only colour is removed. The masks are used only for scoring.

Models. All decoders share IA's encoder, a CNN followed by a Transformer that outputs seven slots of 64 dimensions sampled from a Gaussian posterior, and IA's decoder network: Fourier features of the pixel coordinates as queries, a two-layer cross-attention Transformer, and an output MLP. In IA's decoder every pixel query attends to all seven slots, and IA's interaction penalty is added with weight $\alpha = 0.05$, ramped in over the first 25k iterations. The per-slot decoders apply the same network to one slot at a time and emit a fourth channel, the presence logit a_i, at every pixel; they have no interaction penalty. The softmax masks use $w = \mathrm{softmax}_i(\tau a_i)$ at every pixel. The linear mask uses $w_i = |a_i| / \sum_m |a_m|$. The annealed mask raises τ linearly from 0.5 to 8 over the first 70% of training.

Training. We follow IA's recipe: Adam, with the learning rate warmed up linearly to $5 \cdot 10^{-4}$ over 2,500 iterations, held until 30k and then set to $10^{-4} \cdot 0.1^{(t-30\mathrm{k})/500\mathrm{k}}$ at iteration t; batch size 32 with all 16,384 pixels of each image; and the loss $\|x - \hat{x}\|^2$ plus 0.05 times the KL divergence of the slot posterior, plus the interaction penalty for IA. The budget is 500k iterations, with seeds 1–3.

Differences from IA's implementation. We train in bfloat16 with compiled kernels and evaluate in single precision; a 500k run takes about 17 hours on one H100 GPU. IA's decoder runs 1.85 times

Table 6: Synthetic images, configurations of the three groups of runs: block-MIG (mean $\pm$ 95% t half-width over five seeds), Mix and R^2. w is the fraction of training over which τ is annealed. Configurations that appear in two sweeps (same configuration and seeds) were trained twice. Bold marks the best value per column, and also the second best when their intervals overlap.

	slider			clock		
configuration	block-MIG	Mix ↓	R^2	block-MIG	Mix ↓	R^2
Decoder sweep						
linear	0.14 ± 0.01	2.977	0.564	0.03 ± 0.02	2.699	0.356
joint MLP	0.65 ± 0.39	0.421	0.986	0.08 ± 0.04	2.408	0.825
additive	0.49 ± 0.05	0.356	0.927	0.51 ± 0.11	0.414	0.826
LSE, $\tau = 8$	$\mathbf{0.98 \pm 0.01}$	0.081	0.988	0.82 ± 0.27	0.259	0.886
hard max	0.56 ± 0.32	0.558	0.843	0.60 ± 0.35	0.728	0.717
Temperature sweep						
additive	0.49 ± 0.04	0.351	0.927	0.47 ± 0.03	0.431	0.822
LSE, $\tau = 0.25$	0.81 ± 0.18	0.199	0.950	0.68 ± 0.23	0.365	0.871
LSE, $\tau = 0.5$	0.88 ± 0.18	0.123	0.966	0.76 ± 0.27	0.217	0.895
LSE, $\tau = 1$	0.98 ± 0.00	0.058	0.987	0.78 ± 0.33	0.230	0.906
LSE, $\tau = 2$	0.78 ± 0.25	0.215	0.970	0.84 ± 0.26	0.239	0.917
LSE, $\tau = 4$	0.96 ± 0.05	0.147	0.986	0.80 ± 0.25	0.278	0.894
LSE, $\tau = 8$	0.98 ± 0.01	0.085	0.986	0.83 ± 0.16	0.277	0.894
LSE, $\tau = 16$	0.86 ± 0.27	0.142	0.949	0.66 ± 0.36	0.419	0.837
LSE, $\tau = 32$	0.71 ± 0.35	0.316	0.915	$\mathbf{0.87 \pm 0.17}$	0.290	0.829
LSE, $\tau = 64$	0.71 ± 0.20	0.263	0.899	0.63 ± 0.26	0.364	0.797
hard max	0.50 ± 0.35	0.711	0.843	0.69 ± 0.25	0.470	0.733
Annealing						
LSE, linear $0.5 \to 32$, $w = 0.7$	$\mathbf{0.99 \pm 0.00}$	**0.051**	**0.995**	0.85 ± 0.27	0.176	0.917
LSE, geometric $0.5 \to 8$, $w = 0.7$	0.90 ± 0.24	0.072	**0.995**	$\mathbf{0.88 \pm 0.27}$	**0.170**	**0.932**

Table 7: Synthetic images: block-MIG of every run, seeds 0–4.

	slider					clock				
configuration	0	1	2	3	4	0	1	2	3	4
Decoder sweep										
linear	0.15	0.14	0.14	0.13	0.14	0.01	0.04	0.04	0.04	0.01
joint MLP	0.74	0.10	0.84	0.87	0.71	0.13	0.07	0.09	0.04	0.08
additive	0.51	0.52	0.52	0.43	0.50	0.46	0.66	0.52	0.46	0.47
LSE, $\tau = 8$	0.98	0.99	0.97	0.99	0.99	0.50	0.97	0.98	0.69	0.97
hard max	0.31	0.35	0.58	0.96	0.61	0.97	0.42	0.26	0.57	0.78
Temperature sweep										
additive	0.51	0.52	0.52	0.43	0.50	0.46	0.48	0.45	0.46	0.52
LSE, $\tau = 0.25$	0.87	0.55	0.88	0.89	0.86	0.97	0.52	0.71	0.50	0.68
LSE, $\tau = 0.5$	0.90	0.62	0.96	0.96	0.96	0.98	0.65	0.98	0.47	0.70
LSE, $\tau = 1$	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.49	0.48	0.97
LSE, $\tau = 2$	0.99	0.99	0.65	0.58	0.67	0.97	0.98	0.97	0.79	0.49
LSE, $\tau = 4$	0.97	0.99	0.88	0.95	0.98	0.71	0.84	0.96	0.49	0.97
LSE, $\tau = 8$	0.98	0.99	0.97	0.99	0.98	0.74	0.97	0.79	0.69	0.97
LSE, $\tau = 16$	0.96	0.97	0.48	0.98	0.93	0.92	0.56	0.96	0.26	0.60
LSE, $\tau = 32$	0.98	0.59	0.70	0.98	0.32	0.68	0.97	0.97	0.97	0.76
LSE, $\tau = 64$	0.71	0.65	0.54	0.97	0.66	0.96	0.44	0.61	0.66	0.47
hard max	0.26	0.35	0.58	0.97	0.34	0.97	0.46	0.57	0.65	0.82
Annealing										
LSE, linear $0.5 \to 32$, $w = 0.7$	0.99	0.99	0.99	0.99	0.99	0.98	0.98	0.98	0.47	0.85
LSE, geometric $0.5 \to 8$, $w = 0.7$	0.99	0.55	0.99	0.99	0.99	0.98	0.98	0.98	0.48	0.98

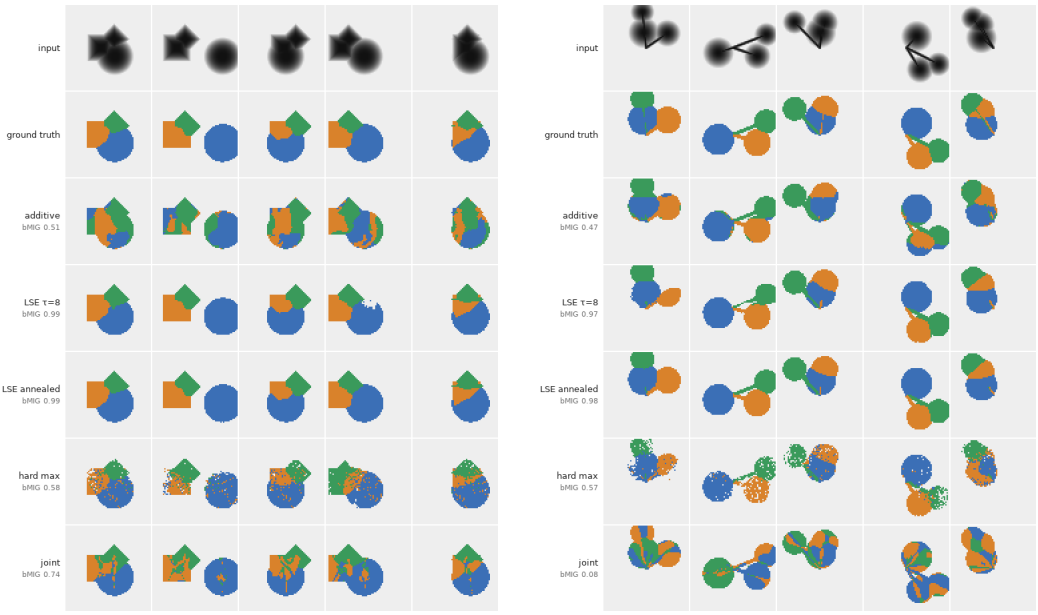


Figure 3: Ownership on five held-out scenes of slider (left) and clock (right): the input, the true ownership map and, for each decoder except the linear one, the learned slot that wins each lit pixel, coloured by its matched true slot; median seed of each decoder, with its block-MIG. The annealed row is the linear ramp from 0.5 to 32 ($w = 0.7$) on both datasets.

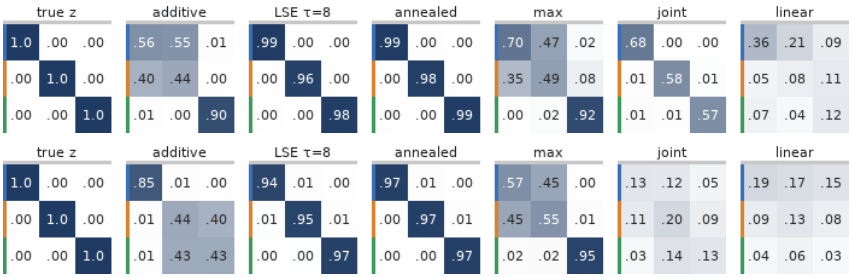


Figure 4: Normalised mutual information between true factors (rows) and learned latents (columns, ordered by the matching) on slider (top) and clock (bottom), for the true latents and every decoder; same runs as Figure 3.

more iterations per second than the per-slot decoders, which evaluate the Transformer once per slot, so budgets are matched in iterations, not in time. We score the first 500 of the 2,000 test images, whereas IA scores all of them. Runs with the same seed are not bitwise reproducible, and their outcomes can differ.

Scores. J-ARI and JIS follow IA. For every test image, the decoder Jacobian with respect to the 7×64 slot coordinates is taken at a sample from the slot posterior; its absolute values are summed over each slot's coordinates and the three colour channels and normalised over slots at every pixel. J-ARI is the adjusted Rand index, over the ground-truth foreground pixels, between object identity and the slot with the largest normalised Jacobian, averaged over images. JIS is the share m of the largest slot, averaged over foreground pixels and rescaled as $(m - 1/7)/(1 - 1/7)$.

Results. Table 8 lists every run, Table 9 the range of the presence logits, and Figure 5 shows slot assignments in grayscale.

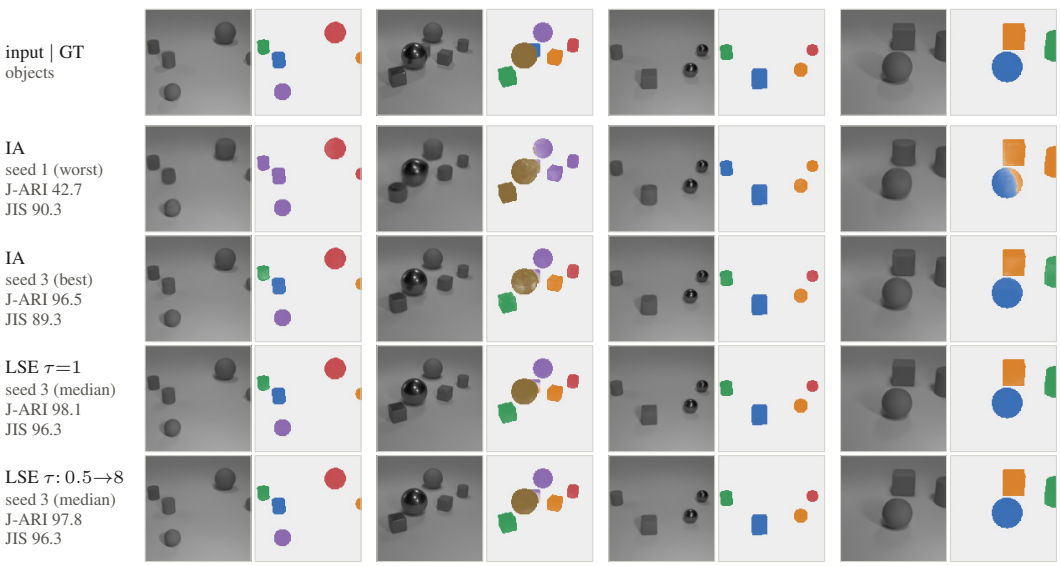


Figure 5: Slot assignments on four grayscale CLEVR6 test scenes after 500k iterations. *Top row:* the input and the ground-truth object masks. Other rows: the reconstruction and, at each foreground pixel, the slot with the largest normalised decoder Jacobian (the assignment scored by J-ARI), coloured by the ground-truth object it is matched to.

Table 8: CLEVR6 after 500k iterations: test J-ARI and JIS ($\times 100$) on the first 500 of the 2,000 test scenes, mean $\pm$ sd over three seeds with seeds 1 / 2 / 3 below. The softmax masks pool raw presence logits, and the annealed mask ends at $\tau = 8$. IA's published colour result is 96.5 / 83.8.

	colour		grayscale	
decoder	J-ARI	JIS	J-ARI	JIS
IA	95.9 ± 1.2 96.4 / 94.6 / 96.8	83.8 ± 0.9 82.8 / 84.2 / 84.4	77.1 ± 29.9 42.7 / 92.2 / 96.5	89.6 ± 0.5 90.3 / 89.4 / 89.3
linear mask	98.1 ± 0.2 98.3 / 98.1 / 98.0	94.3 ± 1.4 92.7 / 95.0 / 95.2	97.5 ± 0.8 98.0 / 97.9 / 96.6	95.1 ± 0.3 94.8 / 95.2 / 95.2
softmax mask, $\tau = 1$	98.2 ± 0.4 98.6 / 97.8 / 98.3	94.6 ± 1.2 94.9 / 93.3 / 95.7	98.1 ± 0.1 98.3 / 98.1 / 98.1	96.2 ± 0.1 96.2 / 96.2 / 96.3
softmax mask, $\tau = 8$	98.4 ± 0.3 98.6 / 98.1 / 98.6	96.4 ± 0.8 95.5 / 97.2 / 96.6	98.3 ± 0.2 98.5 / 98.0 / 98.2	96.3 ± 0.5 96.9 / 95.8 / 96.3
softmax mask, $\tau : 0.5 \to 8$	98.3 ± 0.3 98.0 / 98.6 / 98.4	95.9 ± 1.1 94.6 / 96.5 / 96.5	97.8 ± 0.3 98.1 / 97.5 / 97.8	96.1 ± 0.3 95.8 / 96.1 / 96.3

Table 9: CLEVR6, 500k iterations: range of the presence logits across slots (largest minus smallest, averaged over test pixels), per seed. The network shrinks the range when τ is large: τ times the range, the margin that sets the softmax weights, lies between 1.7 and 5.2 in every softmax-mask run except seed 2 of the colour $\tau = 8$ mask (47).

	colour			grayscale		
decoder	1	2	3	1	2	3
linear mask	0.96	1.14	1.09	0.68	0.69	0.70
softmax mask, $\tau = 1$	2.95	1.68	3.35	2.91	3.91	3.04
softmax mask, $\tau = 8$	0.42	5.92	0.52	0.36	0.40	0.50
softmax mask, $\tau : 0.5 \to 8$	0.42	0.65	0.49	0.37	0.55	0.63

Table 10: CLEVR6, 500k iterations: decoders that pool the slots' colour values directly, per pixel and channel, followed by a small per-pixel head. Test J-ARI per seed ($\times 100$), mean JIS, and, for LSE, the mean largest pooling weight $\max_i \partial\hat{x}/\partial u_i$ over pixels, before the head (uniform over 7 slots: 0.14).

	colour					grayscale				
pooling	J-ARI			JIS	weight	J-ARI			JIS	weight
sum	89.2	94.4	93.9	84.1	–	41.1	96.6	96.2	78.4	–
LSE, $\tau = 1$	92.6	93.2	75.5	73.0	0.15	94.0	61.7	50.1	67.8	0.15
LSE, $\tau = 8$	96.9	95.7	68.3	93.3	0.73	30.9	62.8	31.9	52.6	0.41
LSE, geometric $0.5 \to 8$	92.7	92.4	90.9	72.6	0.15	97.5	90.9	52.4	75.6	0.18
hard max	96.2	94.7	92.0	99.5	–	65.7	68.9	68.4	88.5	–

K.2.1 Pooling colour values instead of a presence score

The decoders of the synthetic experiments pool the slots' outputs themselves. Their CLEVR6 counterpart decodes each slot to an RGB image $\mathbf{u}_i$, with no presence logit, and pools the seven images per pixel and channel, by a sum, by LSE, $\frac{1}{\tau}\log\frac{1}{7}\sum_i e^{\tau u_i}$, or by a maximum; a small per-pixel head with two residual blocks ($3 \to 10 \to 3$ channels) follows the pooling. Everything else is as for the per-slot decoders, at 500k iterations; the annealed LSE follows a geometric ramp from 0.5 to 8 over the first 70% of training. Table 10 gives the results. With colour, 13 of 15 runs recover the objects (J-ARI ≥ 85); in grayscale 5 of 15 do, none of them among the six with a constant $\tau = 8$ or a hard maximum. Under annealing, the network shrinks the spread of the slot outputs as τ rises, and the largest pooling weight ends near the uniform $1/7$ (0.15 in five of six runs), as at $\tau = 1$, whereas at a constant $\tau = 8$ it stays sharp (0.39–0.79). Here the pooled value is the colour, whose scale the network sets freely; in the per-slot decoders of the main text, a separate presence logit decides which slot a pixel shows.

K.3 Single-cell transcriptomics

Data. The core of the Human Lung Cell Atlas (Sikkema et al., 2023) has 584,944 cells, of which we use 1,996 highly variable genes, with the sample (166) as batch covariate and the cell-type annotations at levels 1–3 and the finest level. The ten tissue atlases are the CellHint collections of Xu et al. (2023), each combining four source studies: bone marrow (66,613 cells), pancreas (121,916), lymph node (124,440), kidney (194,504), spleen (200,664), skeletal muscle (201,575), liver (259,678), intestine (263,275), lung (318,426) and blood (335,916). Each has 4,000 highly variable genes, selected with the batch-aware procedure of scIB using the source study as batch, the donor (17–65 per atlas) as batch covariate and one curated cell-type annotation. The lung atlas shares two of its four source studies, about 30% of its cells, with the HLCA core. All models are trained on raw counts.

Model and training. We use DRVI's published configuration: 128 latent dimensions, each decoded separately into per-gene contributions to the mean logit, encoder and decoder MLPs with two hidden layers of 256 units, layer normalisation and ELU activations, dropout 0.1 in the encoder, and the batch as a one-hot covariate of the decoder. The three configurations differ only in how the 128 contributions x_k to a gene are pooled: their mean ($\tau \to 0$), $\log\frac{1}{128}\sum_k e^{x_k}$ (LSE at $\tau = 1$, the published model), or $\frac{1}{\tau}\log\frac{1}{128}\sum_k e^{\tau x_k}$ with $\tau = 0.5 \cdot 16^{\min(e/280,\,1)}$ at epoch e, a geometric ramp from 0.5 to 8 over the first 280 of 400 epochs. We train with Adam (learning rate 10^{-3}, weight decay 10^{-6}), batch size 128 and a KL weight warmed up linearly over all 400 epochs, without early stopping, and use the model of the last epoch; seeds are 1–3, and a run takes 0.7 to 6.5 hours on one GPU.

Scores. We use DRVI's benchmark for discrete targets. Scores are computed on the posterior means of 50,000 cells drawn with the run's seed, so the three configurations of a seed share their cells; cell types with fewer than 10 cells in the subsample are dropped (Table 3 gives the number scored). For a latent k and a cell type c, SMI is the mutual information between latent k, discretised into 10 equal-width bins, and the indicator of type c, divided by the entropy of the indicator. This

gives a $128 \times C$ score matrix S for C cell types. LMS (Lachapelle et al., 2023) averages S over an optimal one-to-one assignment of the types to distinct latents (a maximum-weight bipartite matching). The pooling margin γ is the gap between the largest and the second-largest contribution to a gene's pooled logit, averaged over genes and the first $2{,}048$ cells, and $\tau\gamma$ is the margin that sets the weights of equation 2.

Results. Table 11 gives LMS-SMI with its standard deviation, Table 12 the LMS-SMI of every run, and Table 13 the pooling margins.

Table 11: scRNA-seq: LMS-SMI (mean $\pm$ sd over three seeds). Bold marks the best mean per row, and also the second best when their mean $\pm$ sd intervals overlap.

dataset	mean ($\tau \to 0$)	LSE, $\tau = 1$	LSE, $\tau : 0.5 \to 8$
HLCA finest	$.458 \pm .008$	$\mathbf{.670} \pm .016$	$\mathbf{.699} \pm .014$
Bone marrow	$.378 \pm .005$	$.550 \pm .004$	$\mathbf{.598} \pm .009$
Pancreas	$.514 \pm .027$	$.660 \pm .004$	$\mathbf{.671} \pm .002$
Lymph node	$.536 \pm .009$	$\mathbf{.750} \pm .006$	$\mathbf{.747} \pm .005$
Kidney	$.377 \pm .018$	$\mathbf{.653} \pm .002$	$\mathbf{.639} \pm .018$
Spleen	$.426 \pm .009$	$\mathbf{.680} \pm .026$	$\mathbf{.710} \pm .017$
Skeletal muscle	$.511 \pm .005$	$\mathbf{.636} \pm .015$	$\mathbf{.625} \pm .012$
Liver	$.475 \pm .009$	$\mathbf{.653} \pm .010$	$\mathbf{.656} \pm .006$
Intestine	$.444 \pm .004$	$.597 \pm .014$	$\mathbf{.672} \pm .006$
Lung	$.433 \pm .014$	$\mathbf{.671} \pm .014$	$\mathbf{.671} \pm .017$
Blood	$.514 \pm .013$	$\mathbf{.725} \pm .002$	$\mathbf{.716} \pm .013$

Table 12: scRNA-seq: LMS-SMI of every run, seeds 1–3.

	mean ($\tau \to 0$)			LSE, $\tau = 1$			LSE, $\tau : 0.5 \to 8$		
dataset	1	2	3	1	2	3	1	2	3
HLCA finest	0.449	0.465	0.459	0.677	0.681	0.651	0.684	0.703	0.711
Bone marrow	0.381	0.381	0.371	0.547	0.547	0.555	0.598	0.589	0.607
Pancreas	0.534	0.483	0.524	0.658	0.657	0.664	0.672	0.673	0.669
Lymph node	0.525	0.540	0.542	0.743	0.753	0.755	0.752	0.746	0.742
Kidney	0.377	0.358	0.395	0.652	0.653	0.655	0.641	0.621	0.657
Spleen	0.420	0.437	0.421	0.707	0.680	0.655	0.729	0.694	0.706
Skeletal muscle	0.510	0.516	0.507	0.621	0.638	0.650	0.617	0.639	0.620
Liver	0.467	0.472	0.485	0.646	0.665	0.648	0.658	0.661	0.650
Intestine	0.445	0.440	0.447	0.584	0.595	0.612	0.666	0.673	0.677
Lung	0.417	0.441	0.440	0.683	0.674	0.655	0.670	0.688	0.654
Blood	0.505	0.507	0.528	0.723	0.725	0.726	0.703	0.718	0.728

Table 13: scRNA-seq: pooling hardness. $\tau\gamma$ is the mean margin, in nats, between the leading and the second latent of a gene's pooled logit (first $2{,}048$ cells, one value per model). Means over three seeds.

	$\tau\gamma$ (nats)	
dataset	$\tau = 1$	$\tau : 0.5 \to 8$
HLCA finest	2.2	11.6
Bone marrow	2.3	17.6
Pancreas	1.1	5.7
Lymph node	1.4	4.0
Kidney	1.4	6.7
Spleen	1.8	9.5
Skeletal muscle	2.2	10.3
Liver	2.0	16.4
Intestine	1.5	8.8
Lung	3.5	21.0
Blood	2.1	11.5