Provably Learning Object-Centric Representations

Jack Brady [*] [1] [2] **Roland S. Zimmermann** [*] [1] [2] **Yash Sharma** [2] [3] **Bernhard Schölkopf** [1]
Julius von Kügelgen [†] [1] [4] **Wieland Brendel** [†] [1] [2]

Abstract

Learning structured representations of the visual world in terms of objects promises to significantly improve the generalization abilities of current machine learning models. While recent efforts to this end have shown promising empirical progress, a theoretical account of when unsupervised object-centric representation learning is possible is still lacking. Consequently, understanding the reasons for the success of existing object-centric methods as well as designing new theoretically grounded methods remains challenging. In the present work, we analyze when object-centric representations can provably be learned without supervision. To this end, we first introduce two assumptions on the generative process for scenes comprised of several objects, which we call *compositionality* and *irreducibility*. Under this generative process, we prove that the ground-truth object representations can be identified by an invertible and compositional inference model, even in the presence of dependencies between objects. We empirically validate our results through experiments on synthetic data. Finally, we provide evidence that our theory holds predictive power for existing object-centric models by showing a close correspondence between models' compositionality and invertibility and their empirical identifiability.[1]

1 Introduction

Human intelligence exhibits an unparalleled ability to generalize from a limited amount of experience to a wide range of novel situations (Tenenbaum et al., 2011). To build machines with similar capabilities, a fundamental question is what types of abstract representations of sensory inputs enable such generalization (Goyal & Bengio, 2022). Research in cognitive psychology suggests that one key abstraction is the ability to represent visual scenes in terms of individual objects (Spelke, 2003; Spelke & Kinzler, 2007; Dehaene, 2020; Peters & Kriegeskorte, 2021). Such *object-centric representations* are thought to facilitate core cognitive abilities such as compositional generalization (Fodor & Pylyshyn, 1988; Lake et al., 2017; Battaglia et al., 2018; Greff et al., 2020) and causal reasoning over discrete concepts (Marcus, 2001; Gopnik et al., 2004; Gerstenberg & Tenenbaum, 2017; Gerstenberg et al., 2021).

Significant effort has thus gone into endowing machine learning models with the capacity to learn object-centric representations from raw visual input. While initial approaches were mostly supervised (Ronneberger et al., 2015; He et al., 2017; Chen et al., 2017), a recent wave of new methods explore learning object-centric representations without direct supervision (Greff et al., 2019; Burgess et al., 2019; Lin et al., 2020; Kipf et al., 2020; Locatello et al., 2020; Weis et al., 2021; Biza et al., 2023). These methods have begun exhibiting impressive results, showing potential to scale to complex visual scenes (Caron et al., 2021; Singh et al., 2022a; Sajjadi et al., 2022; Seitzer et al., 2023) and real-world video datasets (Kipf et al., 2022; Singh et al., 2022b; Elsayed et al., 2022).

Yet, despite this empirical progress, we still lack a *theoretical* understanding of when unsupervised object-centric representation learning is possible. This makes it challenging to isolate the reasons underlying the success and failure of existing object-centric models and to develop principled ways to improve them. Furthermore, it is currently not possible to design novel object-centric methods that are theoretically grounded and not solely based on heuristics, many of which break down in more realistic settings (Karazija et al., 2021; Papa et al., 2022; Yang & Yang, 2022).

In the present work, we aim to address this deficiency by investigating when object-centric representations can *provably* be learned without any supervision. To this end, we first specify a data-generating process for multi-object scenes as

[*]Equal contribution [†]Shared last author [1]MPI for Intelligent Systems, Tübingen [2]Tübingen AI Center, Tübingen [3]University of Tübingen, Tübingen, Germany [4]Department of Engineering, University of Cambridge, Cambridge, United Kingdom. Correspondence to: Jack Brady, Wieland Brendel <first.last@tue.mpg.de>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

[1] Code/Website: brendel-group.github.io/objects-identifiability

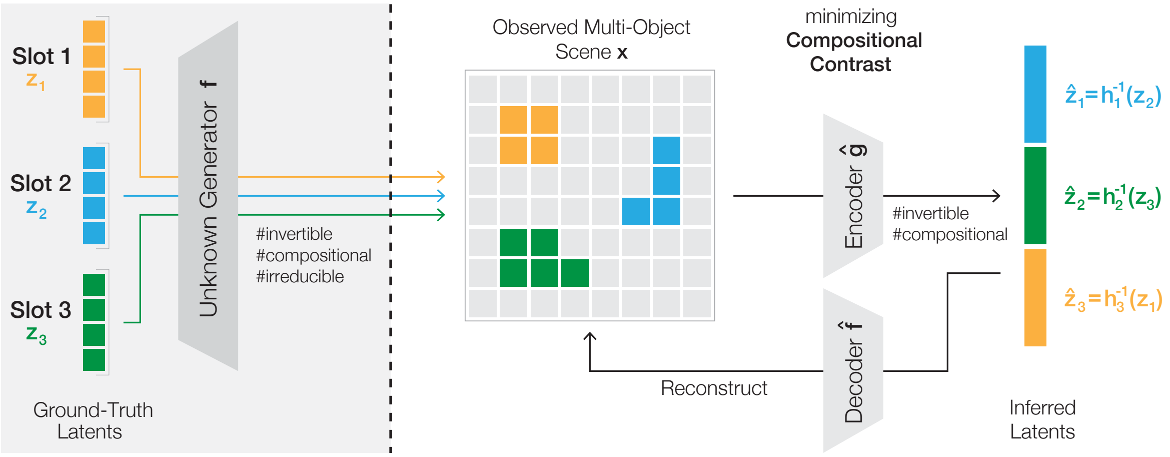

Figure 1. **When can unsupervised object-centric representations provably be learned?** We assume that observed scenes $\mathbf{x}$ comprising K objects are rendered by an unknown generator $\mathbf{f}$ from multiple ground-truth latent slots $\mathbf{z}_1, ..., \mathbf{z}_K$ (here, $K = 3$). We assume that this generative model has two key properties, which we call *compositionality* (Defn. 1) and *irreducibility* (Defn. 5). Under this model, we prove (Thm. 1): An invertible inference model with a compositional inverse yields latent slots $\hat{\mathbf{z}}_i$ which identify the ground-truth slots up to permutation and slot-wise invertible functions $\mathbf{h}_i$ (*slot identifiability*, Defn. 6). To measure violations of compositionality in practice, we introduce a contrast function (Defn. 7) which is zero if and only if a function is compositional, while to measure invertibility, we rely on the reconstruction loss in an auto-encoder framework.

a structured latent variable model in which each object is described by a subset of latents, or a latent *slot*. We then study the *identifiability* of object-centric representations under this model, i.e., we investigate under which conditions an inference model will be guaranteed to recover the subset of ground-truth latents for each object.

Because identifying the ground-truth latent variables is impossible without further assumptions on the generative process (Hyvärinen & Pajunen, 1999; Locatello et al., 2019), previous identifiability results primarily rely on distributional assumptions on the latents (Hyvärinen & Morioka, 2016; 2017; Hyvärinen et al., 2019; Khemakhem et al., 2020a;b; Klindt et al., 2021; Zimmermann et al., 2021). In contrast, we make no such assumptions, thus allowing for arbitrary statistical and causal dependencies between objects.

Structure and Main Contributions. In the present work, we instead take the position that the object-centric nature of the problem imposes a very specific *structure* on the *generator function* that renders scenes from latent slots (§ 2). Specifically, we define two key properties that this function should satisfy: *compositionality* (Defn. 1) and *irreducibility* (Defn. 5). Informally, these properties imply that every pixel can only correspond to one object and that information is shared across different parts of the same object but not between parts of different objects—inspired by the principle of independent causal mechanisms (Peters et al., 2017). Under this generative model, we then prove in § 3 our *main theoretical result*: the ground-truth latent slots can be identified without supervision by an invertible inference model with a compositional inverse (Thm. 1). To quantify compositionality, we introduce a *contrast function* (Defn. 7) that is zero if and only if a function is compositional; to quantify invertibility, we rely on reconstruction error. We validate on synthetic data that inference models which maximize invertibility and compositionality indeed identify the ground-truth latent slots, even with dependencies between latents (§ 5.1). Finally, we examine existing object-centric learning models on image data and find a close correspondence between models' compositionality and invertibility and their success in identifying the ground-truth latent slots (§ 5.2).

To the best of our knowledge, the present work provides the first identifiability result for object-centric representations. We hope that this lays the groundwork for a better understanding of success and failure in unsupervised object-centric learning, and that future work can build on these insights to develop more effective learning methods.

Notation. Bold lowercase $\mathbf{z}$ denotes vectors, bold uppercase $\mathbf{J}$ denotes matrices. For $n \in \mathbb{N}$, let $[n]$ denote the set $\{1, \dots, n\}$. Additionally, if $\mathbf{f}$ is a function with n component functions, let $\mathbf{f}_S$ denote the restriction of $\mathbf{f}$ to the component functions indexed by $S \subseteq [n]$, i.e., $\mathbf{f}_S := (f_s)_{s \in S}$.

2 Generative Model

While humans have a clear intuition for what constitutes an object, formalizing this notion mathematically is not straightforward. Indeed, there is no universally agreed-upon definition of an object; various formalizations based upon distinct criteria co-exist (Green, 2019; Spelke, 1990; Koffka, 1936; Greff et al., 2020). We approach the problem by

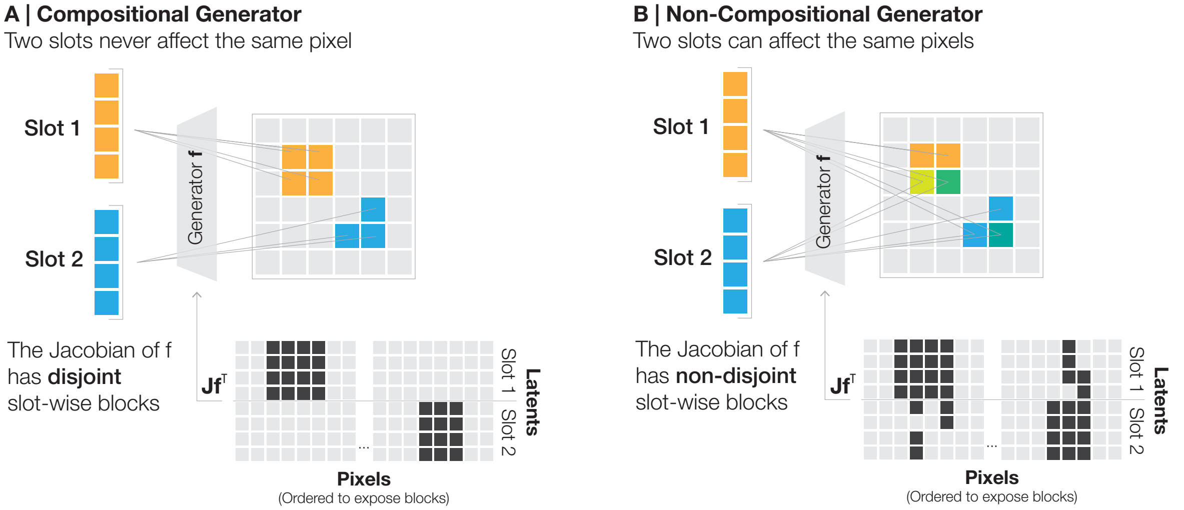
A | Compositional Generator
Two slots never affect the same pixel
Slot 1
Slot 2
Generator f
The Jacobian of f has disjoint slot-wise blocks
Jf^T
Latents
Slot 1
Slot 2
Pixels
(Ordered to expose blocks)
B | Non-Compositional Generator
Two slots can affect the same pixels
Slot 1
Slot 2
Generator f
The Jacobian of f has non-disjoint slot-wise blocks
Jf^T
Latents
Slot 1
Slot 2
Pixels
(Ordered to expose blocks)

Figure 2. **Difference between a compositional and a non-compositional generator. (A)** For a compositional generator $\mathbf{f}$, every pixel is affected by at most one latent slot. As a result, there always exists an ordering of the pixels such that the generator's Jacobian $\mathbf{Jf}$ consists of disjoint blocks, one for each latent slot *(bottom)*. Note that both the pixel ordering and the specific structure of the Jacobian are not fixed across scenes and might depend on the latent input $\mathbf{z}$. **(B)** For a non-compositional generator, there exists no pixel ordering that exposes such a structure in the Jacobian, since the same pixel can be affected by more than one latent slot.

defining multi-object scenes in terms of a latent variable model (see Fig. 1 for an overview) and argue that the object-centric nature of the problem necessitates a very specific structure on the generator, which we leverage in § 3 to prove our identifiability result.

As a starting point, we assume that observed data samples $\mathbf{x}$ of multi-object scenes are generated from a set of latent random vectors $\mathbf{z}$ through a diffeomorphism[2] $\mathbf{f} : \mathcal{Z} \to \mathcal{X}$, mapping from a *latent* space $\mathcal{Z}$ to an *observation* space $\mathcal{X}$,

$$\mathbf{z} \sim p_{\mathbf{z}}, \qquad \mathbf{x} = \mathbf{f}(\mathbf{z}). \tag{1}$$

The only assumption we place on $p_{\mathbf{z}}$ is that it is fully supported on $\mathcal{Z}$. In particular, we do not require independence and allow for arbitrary dependencies between components of $\mathbf{z}$, motivated by the fact that the presence or properties of certain objects may be correlated with those of other objects.

2.1 Slots and Compositionality

We think of an object in a scene as being encoded not by a single latent component z_i but instead by a group of latents $\mathbf{z}_k$ which specify its properties. For a scene comprised of K objects, we thus assume that the latent space $\mathcal{Z}$ factorizes into K subspaces $\mathcal{Z}_k$, which we refer to as *slots*. Each slot is assumed to have dimension M, representing, e.g., M distinct object properties. More precisely, $\forall k \in [K]$: $\mathcal{Z}_k = \mathbb{R}^M$, and $\mathcal{Z} = \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_K = \mathbb{R}^{KM}$.

Let $\mathbf{z}_k$ be the latent vector of the k^{th} slot. The full KM-dimensional latent *scene representation* vector $\mathbf{z}$ is then given by the concatenation of the latents from all slots,

$$\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_K). \tag{2}$$

We would like to ensure that each latent slot $\mathbf{z}_k$ is responsible for encoding a distinct object in a scene. To this end, the latent scene representation $\mathbf{z}$ should be rendered by $\mathbf{f}$ such that each slot generates *exactly* one object (see Fig. 1). If $\mathbf{f}$ is an arbitrary function with no additional constraints, however, this will generally not be the case.

First, $\mathbf{f}$ lacks any structure which ensures that an object is not generated by more than one latent slot. To see this, let $I_k(\mathbf{z}) \subseteq [N]$ denote the subset of pixels in an image generated from scene representation $\mathbf{z}$ that functionally depend on slot k,

$$I_k(\mathbf{z}) := \left\{ n \in [N] : \frac{\partial f_n}{\partial \mathbf{z}_k}(\mathbf{z}) \neq \mathbf{0} \right\}. \tag{3}$$

Note the dependence on $\mathbf{z}$, which encodes that an object may appear in different places across different scenes.

Without further constraints on $\mathbf{f}$, the pixel subsets $I_k(\mathbf{z})$ and $I_j(\mathbf{z})$ can overlap for any $k \neq j$ such that latent slots k, j can affect the same pixels and thus contribute to generating the same object (see Fig. 2B, top). To avoid this, we impose the following structure on $\mathbf{f}$, which we call *compositionality*.

Definition 1 (Compositionality)**.** *Let* $\mathbf{f} : \mathcal{Z} \to \mathcal{X}$ *be differentiable.* $\mathbf{f}$ *is said to be* compositional *if*

$$\forall \mathbf{z} \in \mathcal{Z} : \qquad k \neq j \implies I_k(\mathbf{z}) \cap I_j(\mathbf{z}) = \varnothing. \tag{4}$$

Compositionality implies that each pixel is a function of at most one latent slot and thus imposes a local sparsity

[2] a differentiable bijection with differentiable inverse

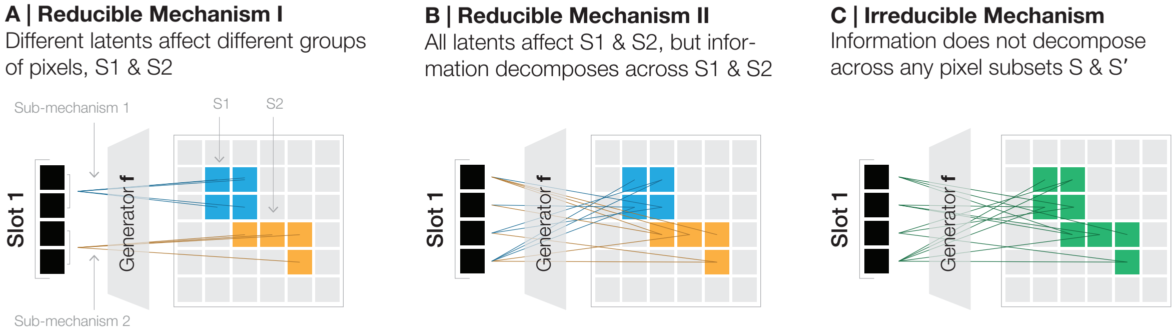
A | Reducible Mechanism I
Different latents affect different groups of pixels, S1 & S2
Sub-mechanism 1
S1
S2
Slot 1
Generator f
Sub-mechanism 2
B | Reducible Mechanism II
All latents affect S1 & S2, but information decomposes across S1 & S2
Slot 1
Generator f
C | Irreducible Mechanism
Information does not decompose across any pixel subsets S & S′
Slot 1
Generator f

Figure 3. **(Ir)reducible mechanisms. (A)** A simple example of a *reducible mechanism* is one for which disjoint subsets of latents from the same slot render pixel groups S_1 and S_2 separately such that they form *independent sub-mechanisms* according to Defn. 4. This independence between sub-mechanisms is indicated by the difference in colors. **(B)** Not all reducible mechanisms look as simple as panel A: here, S_1 and S_2 depend on every latent component in the slot, but the information in $S_1 \cup S_2$ still decomposes across S_1 and S_2 as sub-mechanisms 1 and 2 are independent. **(C)** In contrast, for an irreducible mechanism, the information does not decompose across any pixel partition $S, S^{'}$, and so it is impossible to separate it into independent sub-mechanisms.

structure on the *Jacobian* matrix $\mathbf{Jf} = \left(\frac{\partial f_i}{\partial z_j}\right)_{ij}$ of $\mathbf{f}$, which is visualized in Fig. 2, bottom. Intuitively, the Jacobian of a compositional generator can always be brought into block structure through an appropriate permutation of the pixels. However, this block structure is local in that the required permutation may differ across scene representations $\mathbf{z}$.

2.2 Mechanisms and Irreducibility

While compositionality ensures that different latent slots do not generate the same object, we need an additional constraint on $\mathbf{f}$ to ensure that each slot generates only one object, rather than something humans would regard as multiple objects. To see this, consider the example depicted in Fig. 3A, where $\mathbf{f}$ maps the first half of the latent slot to the pixels denoted S_1 and the second half to S_2. It is clear that for humans, these groups of pixels would likely be considered as distinct objects. On the other hand, it is not immediately clear what formal criteria would give rise to such a distinction.

Intuitively, the issue with the two "sub-objects" S_1 and S_2 in Fig. 3A appears to be that they are *independent* of each other in some sense. To avoid such splitting of objects within slots, we would thus like to enforce that pixels belonging to the same object are *dependent* on one another. But what is a meaningful notion of such *instance-level* independence of objects? Since we are dealing with a single scene sampled according to Eq. (1), it cannot be statistical in nature. Instead, our intuition is more aligned with the notion of *algorithmic independence* of objects (Janzing & Schölkopf, 2010), a formalization[3] of the principle of independent causal mechanisms (ICM) which posits that physical generative processes consist of "autonomous modules that do not inform or influence each other" (Peters et al., 2017). The two subsets of pixels S_1 and S_2 in Fig. 3A are independent of each other in precisely this sense: they arise from autonomous processes that do not share information.

In the following, we therefore draw inspiration from prior implementations of the ICM principle (Daniusis et al., 2010; Janzing et al., 2012; Gresele et al., 2021, see § 4 for more details) to formalize our intuitions about independence of objects. First, we define the mapping which locally renders information from the k^{th} latent slot to the affected pixels $I_k(\mathbf{z})$ which we refer to as a *mechanism.*

Definition 2 (Mechanism)**.** *$\forall \mathbf{z} \in \mathcal{Z}, k \in [K]$, we define the k^{th}* mechanism *of* $\mathbf{f}$ *at* $\mathbf{z}$ *as the Jacobian matrix* $\mathbf{Jf}_{I_k}(\mathbf{z})$.

The k^{th} mechanism can be understood as the sub-matrix of the Jacobian of $\mathbf{f}$ whose rows correspond to the pixels $I_k(\mathbf{z})$ affected by slot k. Further, we define a *sub-mechanism* as the restriction to a *subset* of the affected pixels.

Definition 3 (Sub-Mechanism)**.** $\mathbf{Jf}_S(\mathbf{z})$ *is said to be a* sub-mechanism *of* $\mathbf{Jf}_{I_k}(\mathbf{z})$*, if $S \subseteq I_k(\mathbf{z})$ and S is nonempty.*

In light of these definitions, Fig. 3A consists of two sub-mechanism, $\mathbf{Jf}_{S_1}(\mathbf{z})$ and $\mathbf{Jf}_{S_2}(\mathbf{z})$, which generate pixels S_1 and S_2. To characterize the level of dependence between sets of pixels and their associated sub-mechanisms, we propose to use the matrix *rank*, which can be seen as a non-statistical measure of information as it locally characterizes the latent capacity used to generate the corresponding pixels.

Definition 4 (Independent/Dependent Sub-Mechanisms)**.** *Let $S_1, S_2 \subseteq [N]$ and $\mathbf{z} \in \mathcal{Z}$. The sub-mechanisms $\mathbf{Jf}_{S_1}(\mathbf{z})$ and $\mathbf{Jf}_{S_2}(\mathbf{z})$ are said to be* independent *if:*

$$\operatorname{rank}\left(\mathbf{Jf}_{S_1 \cup S_2}(\mathbf{z})\right) = \operatorname{rank}\left(\mathbf{Jf}_{S_1}(\mathbf{z})\right) + \operatorname{rank}\left(\mathbf{Jf}_{S_2}(\mathbf{z})\right). \tag{5}$$

Conversely, they are said to be dependent *if:*

$$\operatorname{rank}\left(\mathbf{Jf}_{S_1 \cup S_2}(\mathbf{z})\right) < \operatorname{rank}\left(\mathbf{Jf}_{S_1}(\mathbf{z})\right) + \operatorname{rank}\left(\mathbf{Jf}_{S_2}(\mathbf{z})\right).$$

[3] albeit an impractical one formulated in terms of Kolmogorov complexity (algorithmic information), which is not computable

Intuitively, two sub-mechanisms $\mathbf{J}\mathbf{f}_{S_1}(\mathbf{z})$ and $\mathbf{J}\mathbf{f}_{S_2}(\mathbf{z})$ are independent according to Defn. 4 if the information content of pixels $S_1 \cup S_2$ decomposes across S_1 and S_2 in the sense that the latent capacity required to *jointly* generate $S_1 \cup S_2$ (LHS of Eq. (5)) is the same as that required to generate S_1 and S_2 *separately* (RHS of Eq. (5)). Such a decomposition will occur when the rows of the sub-mechanism $\mathbf{J}\mathbf{f}_{S_1}(\mathbf{z})$ do not lie in the row-space of the sub-mechanism $\mathbf{J}\mathbf{f}_{S_2}(\mathbf{z})$ and vice-versa. This will be the case in Fig. 3A where $\mathbf{J}\mathbf{f}_{S_1}(\mathbf{z})$ and $\mathbf{J}\mathbf{f}_{S_2}(\mathbf{z})$ affect different pixels since the rows of the Jacobian for pixels S_1 and S_2 will never have non-zero entries for the same column. As shown in Fig. 3B, however, it could also be the case that all latents within a slot affect pixels in both S_1 and S_2, yet the information content of $S_1 \cup S_2$ still decomposes across S_1 and S_2 since the rows of $\mathbf{J}\mathbf{f}_{S_1}(\mathbf{z})$ and $\mathbf{J}\mathbf{f}_{S_2}(\mathbf{z})$ could span linearly independent subspaces.

To enforce that each slot generates only one object, we now finally place the condition on the mechanisms of $\mathbf{f}$ that they cannot be partitioned into independent sub-mechanisms (see Fig. 3C). We refer to this property as *irreducibility*.

Definition 5 (Irreducibility)**.** $\mathbf{f}$ *is said to have* irreducible mechanisms*, or is* irreducible*, if for all* $\mathbf{z} \in \mathcal{Z}$*,* $k \in [K]$ *and any partition of* $I_k(\mathbf{z})$ *into* S_1 *and* S_2*, the sub-mechanisms* $\mathbf{J}\mathbf{f}_{S_1}(\mathbf{z})$ *and* $\mathbf{J}\mathbf{f}_{S_2}(\mathbf{z})$ *are dependent in the sense of Defn. 4.*

3 Theory: Slot Identifiability

Given multi-object scenes sampled from the generative model outlined in § 2, we now seek to understand under what conditions an *inference model* $\hat{\mathbf{g}} : \mathcal{X} \to \mathcal{Z}$ will provably identify the ground-truth object representations. Ideally, we would like $\hat{\mathbf{g}}$ to recover the true inverse $\mathbf{g} := \mathbf{f}^{-1}$, but that is generally only possible up to certain irresolvable ambiguities. In our multi-object setting, the objective is to separate the object representations such that each inferred slot captures *one and only one* ground-truth slot. We refer to this notion as *slot identifiability* and define it as follows.

Definition 6 (Slot Identifiability)**.** *Let* $\mathbf{f} : \mathcal{Z} \to \mathcal{X}$ *be a diffeomorphism. An inference model* $\hat{\mathbf{g}} : \mathcal{X} \to \mathcal{Z}$ *is said to* slot-identify $\mathbf{z} = \mathbf{g}(\mathbf{x})$ *via* $\hat{\mathbf{z}} = \hat{\mathbf{g}}(\mathbf{x}) = \hat{\mathbf{g}}(\mathbf{f}(\mathbf{z}))$ *if for all* $k \in [K]$ *there exist a unique* $j \in [K]$ *and a diffeomorphism* $\mathbf{h}_k : \mathcal{Z}_k \to \mathcal{Z}_j$ *such that* $\hat{\mathbf{z}}_j = \mathbf{h}_k(\mathbf{z}_k)$ *for all* $\mathbf{z} \in \mathcal{Z}$*.*

We are now in a position to state our main theoretical result (all complete proofs are provided in Appx. A).

Theorem 1. *Let* $\mathbf{f} : \mathcal{Z} \to \mathcal{X}$ *be a diffeomorphism that is compositional (Defn. 1) with irreducible mechanisms (Defn. 5). If an inference model* $\hat{\mathbf{g}} : \mathcal{X} \to \mathcal{Z}$ *is* (i) *a diffeomorphism with* (ii) *compositional inverse* $\hat{\mathbf{f}} = \hat{\mathbf{g}}^{-1}$*, then* $\hat{\mathbf{g}}$ *slot-identifies* $\mathbf{z} = \mathbf{g}(\mathbf{x})$ *in the sense of Defn. 6.*

Proof Sketch. Irreducibility of $\mathbf{f}$ ensures that information is shared across different parts of an object, and compositionality of $\mathbf{f}$ that this information is not shared with other objects. This creates an asymmetry in the latent capacity required to encode the entirety of one object compared to parts of different objects. When $\hat{\mathbf{g}}$ satisfies (i) and (ii), this asymmetry can be leveraged to show that each inferred slot $\hat{\mathbf{z}}_j$ maps to *one* and *only* ground-truth slot $\mathbf{z}_k$ by a *proof by contradiction*. Namely, suppose that $\hat{\mathbf{g}}$ maps pixels of two distinct objects to the same slot j. If $\hat{\mathbf{g}}$ were to encode all latent information required to generate these pixels in slot j, there would not be sufficient total latent capacity to recover the entire scene, leading to a violation of (i) invertibility. Hence, information for at least one of the pixels needs to be distributed across multiple slots, violating (ii) compositionality of $\hat{\mathbf{f}} = \hat{\mathbf{g}}^{-1}$.

Implications for Object-Centric Learning. Thm. 1 highlights important conceptual points for object-centric representation learning. First, it shows that distributional assumptions on the latents $\mathbf{z}$ are not necessary for slot identifiability; instead, it suffices to enforce structure on the generator $\mathbf{f}$. This falls in line with state-of-the-art (SOTA) object-centric learning methods (Locatello et al., 2020; Singh et al., 2022b; Seitzer et al., 2023; Elsayed et al., 2022), which are based on an auto-encoding framework, thus imposing no additional structure on $p_{\mathbf{z}}$. However, while these models directly enforce invertibility through the reconstruction objective, it is less clear whether and to what extent they also enforce compositionality. Specifically, compositionality is not explicitly optimized in any object-centric methods. Yet, the success of SOTA models in practice suggests that it may be implicitly enforced to some extent through additional inductive biases in the model. We explore this point empirically (see Fig. 6) and leave a more theoretical exploration for future work.

Thm. 1 also emphasizes that using a restricted latent bottleneck plays an important role in achieving slot identifiability. Specifically, Thm. 1 is predicated on $\dim(\mathbf{z}){=}\dim(\hat{\mathbf{z}})$ and would no longer hold in its current form if $\dim(\mathbf{z}){<}\dim(\hat{\mathbf{z}})$. The importance of restricting the latent capacity of object-centric models was emphasized empirically by Engelcke et al. (2020a). Yet, the most successful object-centric models in practice often use $\dim(\mathbf{z}){<}\dim(\hat{\mathbf{z}})$ (Dittadi et al., 2022; Locatello et al., 2020; Sajjadi et al., 2022). A potential explanation for this discrepancy is that SOTA object-centric models do encode information from multiple objects in each latent slot, but this additional information is ignored by the decoder during reconstruction such that image-level segmentations remain accurate. We provide some evidence for this hypothesis through experiments with existing object-centric models in § 5.2.

Measuring Compositionality. While Thm. 1 reveals properties an inference function should satisfy to achieve slot identifiability, it presents these properties in an abstract mathematical form. If we seek to leverage Thm. 1 to assess the performance of existing object-centric models or inform new training objectives for object-centric learning, we require a way to quantify whether an inference model is (i) a diffeomorphism and (ii) compositional. Regarding (i), one clear choice is to train an auto-encoder with differentiable encoder $\hat{\mathbf{g}}$ and decoder $\hat{\mathbf{f}}$ and minimize reconstruction loss to enforce invertibility. Regarding (ii), on the other hand, it is much less obvious how to quantify compositionality. To this end, we introduce the following contrast function, which we prove to be zero if and only if a function is compositional:

Definition 7 (Compositional Contrast)**.** *Let* $\mathbf{f} : \mathcal{Z} \to \mathcal{X}$ *be differentiable. The* compositional contrast *of* $\mathbf{f}$ *at* $\mathbf{z}$ *is*

$$C_{\mathrm{comp}}(\mathbf{f}, \mathbf{z}) = \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{j=k+1}^{K} \left\| \frac{\partial f_n}{\partial \mathbf{z}_k}(\mathbf{z}) \right\| \left\| \frac{\partial f_n}{\partial \mathbf{z}_j}(\mathbf{z}) \right\| . \quad (6)$$

For a given scene representation $\mathbf{z}$ and generator $\mathbf{f}$, the contrast function in Eq. (6) computes the sum over all pixels n of all pairwise products of the (L2) norms of those pixels' gradients with respect to any two distinct slots $k \neq j$. As such, it is a non-negative quantity that can only be zero if every pixel is affected by at most one slot (i.e., $\mathbf{f}$ is *compositional*), for otherwise there would be a pair of slots $k \neq j$ for which the gradient norms are both non-zero resulting in their product being non-zero.

We leverage this characterization of compositionality to provide our second result, which can be viewed as an optimization-based perspective on Thm. 1.

Theorem 2. *Let* $\mathbf{f} : \mathcal{Z} \to \mathcal{X}$ *be a diffeomorphism that is compositional (Defn. 1) with irreducible mechanisms (Defn. 5). If an encoder* $\hat{\mathbf{g}} : \mathcal{X} \to \mathcal{Z}$ *and decoder* $\hat{\mathbf{f}} : \mathcal{Z} \to \mathcal{X}$ *are both differentiable and solve the following functional equation*

$$\mathbb{E}_{\mathbf{x} \sim p_{\mathbf{x}}} \left[\left\| \hat{\mathbf{f}}(\hat{\mathbf{g}}(\mathbf{x})) - \mathbf{x} \right\|_2^2 + \lambda C_{\mathrm{comp}} \left(\hat{\mathbf{f}}, \hat{\mathbf{g}}(\mathbf{x}) \right) \right] = 0, \quad (7)$$

for $\lambda > 0$*, then* $\hat{\mathbf{g}}$ *slot-identifies* $\mathbf{z}$ *in the sense of Defn. 6.*

4 Related Work

Object-Centric Generative Models. Prior works have also formulated generative models for multi-object scenes based on latent slots (Roux et al., 2011; Heess, 2012; Greff et al., 2015; 2017; 2019; van Steenkiste et al., 2018; von Kügelgen et al.; Engelcke et al., 2020b; 2021), though without studying identifiability. Our assumptions on the generative model (§ 2) bear intuitive similarity to some of these prior works, but they also differ in several fundamental ways. First, compositionality (Defn. 1) is stated as a desideratum for nearly all object-centric generative models. Yet, this constraint is not actually enforced by most existing approaches, particularly those based on spatial mixture models in which every slot may affect every pixel (Greff et al., 2015; 2017; 2019; van Steenkiste et al., 2018; Engelcke et al., 2020b; 2021). More closely related is a dead-leaves model approach, in which a scene is sequentially generated by layering objects such that each pixel is affected by at most one slot (Roux et al., 2011; von Kügelgen et al.; Tangemann et al., 2023). In contrast, we define compositionality directly through assumptions on the structure of the (Jacobian of the) generator. Second, our irreducibility criterion (Defns. 4 and 5) bears conceptual similarity to prior works, which assume that different objects do not share information whereas parts of the same object do (Hyvärinen & Perkiö, 2006; Greff et al., 2015; 2017; van Steenkiste et al., 2018). Importantly, however, these works formalize this intuition using statistical criteria such as *statistical independence* between pixels from different objects and dependence between pixels from the same object. However, this leads to an incorrect characterization of objects: e.g., the presence of a coffee cup should increase the likelihood that a table is also present, despite these being separate objects (Träuble et al., 2021; Schölkopf et al., 2021). Here, we instead formulate independence/dependence between objects in a *non-statistical* sense, inspired by algorithmic independence of mechanisms.

Objects and Causal Mechanisms. In causal modelling (Spirtes et al., 2001; Pearl, 2009), a *mechanism* typically refers to a function that determines the value of an effect variable from its direct causes and possibly a noise term, leading to a conditional distribution of effect given causes. Thus, we could view objects as the effects of the latent variables that cause them. While the causal variables are generally not independent, it has been argued that the mechanisms producing them should be (Schölkopf et al., 2012; Peters et al., 2017). Since this is an independence between functions or conditionals rather than between random variables, it is non-trivial to formalize it statistically (Janzing & Schölkopf, 2010; Guo et al., 2022). Hence, various implementations of the principle have been proposed (Daniusis et al., 2010; Janzing et al., 2010; 2012; Shajarisales et al., 2015; Locatello et al., 2018; Besserve et al., 2018; 2021; Janzing, 2021), typically for settings in which both cause and effect are observed. Our notion of independent sub-mechanisms is most closely related to work by Gresele et al. (2021), who also study representation learning and define mechanisms more broadly in terms of the Jacobian $\mathbf{J}\mathbf{f}$: they assume independent latents and formalize mechanism independence as column-orthogonality of the Jacobian. In contrast, our rank condition (Eq. (5)) is inspired by object-centric representation learning with dependent latents.

Identifiable Representation Learning. As this is the first identifiability study of unsupervised object-centric representations, our problem setting differs from existing work both in terms of the assumptions we make on the generative process and the type of identifiability that we aim to achieve.

First, prior work on identifiable representation learning commonly places assumptions on the latent distribution, such as conditional independence given an auxiliary variable (Hyvärinen & Morioka, 2016; 2017; Hyvärinen et al., 2019; Khemakhem et al., 2020a; Hälvä & Hyvärinen, 2020; Hälvä et al., 2021) or access to views arising from pairs of similar latents (Gresele et al., 2019; Klindt et al., 2021; Zimmermann et al., 2021; von Kügelgen et al., 2021), while leaving the generator $\mathbf{f}$ completely unconstrained. In contrast, we place no assumptions on $p_{\mathbf{z}}$ and instead impose structure on (the Jacobian of) the generator $\mathbf{f}$. Recent works have also leveraged assumptions on $\mathbf{Jf}$ such as orthogonality (Gresele et al., 2021; Zheng et al., 2022; Reizinger et al., 2022; Buchholz et al., 2022), unit determinant (Yang et al., 2022), or a *fixed* sparsity structure (Moran et al., 2021; Lachapelle et al., 2021; Lachapelle & Lacoste-Julien, 2022). While the latter relates to our definition of compositionality (Defn. 1), we crucially allow the sparsity pattern on $\mathbf{Jf}$ to vary with $\mathbf{z}$ (in line with the basic notion that objects are not fixed in space), and impose sparsity with respect to slots rather than individual latents. Secondly, existing work typically aims to identify individual latent components z_i up to permutations (or linear transformations). However, this is inappropriate for object-centric representation learning, where we aim to capture and isolate the subsets of latents corresponding to each object in well-defined slots. Identifying such groups of latents is similar to efforts in independent subspace analysis (ISA; Hyvärinen & Hoyer, 2000). However, results for ISA are generally restricted to linear models and independent groups, whereas we allow for nonlinear models and dependence. Our notion of slot identifiability is most closely related to that of block-identifiability introduced by (von Kügelgen et al., 2021) and can be seen as an extension or generalization thereof to a setting with multiple blocks.

5 Experiments

Thm. 2 states that inference models which minimize reconstruction loss $\mathcal{L}_{\text{rec}}$ and compositional contrast C_{comp} achieve *slot identifiability* (Defn. 6). This provides a concrete way to empirically test our main theoretical result. To do so, we perform two main sets of experiments. First, in § 5.1 we generate controlled synthetic data according to the process specified in § 2 and train an inference model on this data which directly optimizes $\mathcal{L}_{\text{rec}}$ and C_{comp} jointly. Second, in § 5.2 we seek to better understand the relationship between $\mathcal{L}_{\text{rec}}$, C_{comp}, and slot identifiability in existing object-centric models. To this end, we analyze a set of models trained on a multi-object sprites dataset.

Quantifying Slot Identifiability. To assess whether a model is slot identifiable in practice, we first establish a metric to measure slot identifiability. Specifically, we want to measure if there exists an invertible function between each ground-truth and exactly one inferred latent slot. To this end, we first fit nonlinear models between inferred and ground-truth slots and measure their quality by the R^2 coefficient of determination. To properly measure this R^2 score, we must first match each ground-truth slot to its corresponding inferred slot as permutations could exist between slots. For our experiments in § 5.1, this permutation will be global i.e. the same for all inferred latents, thus we use the Hungarian Algorithm (Kuhn, 1955) to find the optimal matching based on the R^2 scores for models fit between every pair of slots. For our experiments on image data in § 5.2, however, such a permutation will be local due to the permutation invariance of the generator. To resolve this, we follow a procedure similar to that of Locatello et al. (2020) and Dittadi et al. (2022) using online matching when fitting models between slots. Specifically, at every training iteration, we compute a matching loss for each sample for all possible pairings of ground-truth and inferred slots and use the Hungarian algorithm to find the optimal assignment for minimizing this loss. After resolving permutations, the R^2 scores for the matched slots tell us how much information about each ground-truth slot is contained in one inferred slot. We also need to ensure, however, that inferred slots only contain information about one ground-truth slot and not multiple. To this end, we correct this score by subtracting the maximum R^2 score from models fit between each inferred latent slot and the ground-truth slots that it was not previously matched with. Taking the mean of this score across all slots yields the final score, which we refer to as the *slot identifiability score* (SIS). Further details on the metric are given in Appx. B.4.

5.1 Synthetic Data

Experimental Setup. To generate synthetic data according to § 2, we first sample a KM-dimensional latent vector from a normal distribution $p_{\mathbf{z}} = \mathcal{N}(0, \Sigma)$, where we consider scenarios with both statistically independent latents ($\Sigma = \mathbf{I}$) and dependent latents ($\Sigma \sim \text{Wishart}_{KM}(\mathbf{I}, KM)$). We then partition the latent vector into K slots, each with dimension M, and apply the same multi-layer perceptron (MLP) to each of the K slots separately. The MLP has 2 layers, uses LeakyReLU non-linearities, and is chosen to lead to invertibility almost surely by following the settings used in previous work (Hyvärinen & Morioka, 2016; 2017; Zimmermann et al., 2021). Observations $\mathbf{x}$ are obtained by concatenating the slot-wise MLP outputs such that the generator is compositional according to Defn. 1 as well as invertible.[4] We train models with a number of slots $K \in \{2, 3, 5\}$ and $\lambda \in \{10^{-7}, 10^{-5}, 10^{-2}, 0, 1, 10\}$ (see Thm. 2) each across 10 random seeds (180 models in total). In all cases, we use slot-dimension $M = 3$ and slot-output dimension of 20 such that $\dim(\mathbf{x}) = K \cdot 20$. Further details on this setup may be found in Appx. B.1.

[4] Regarding enforcing irreducibility, see Appx. B.1.

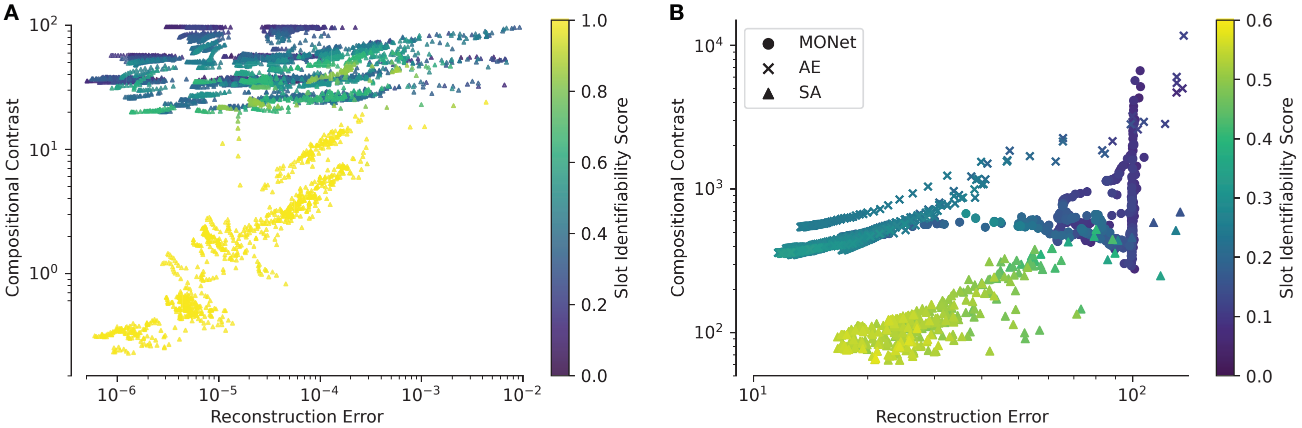

Figure 4. **(A) Experimental validation of Thm. 2.** We trained models on synthetic data generated according to § 2 with 2, 3, 5 independent latent slots (see § 5.1). The color coding indicates the level of identifiability achieved by the model, measured by the Slot Identifiability Score (SIS), where higher values correspond to more identifiable models. As predicted by our theory, if a model sufficiently minimizes both reconstruction error and compositional contrast, then it identifies the ground-truth latent slots. **(B) Application of Thm. 2 to existing object-centric models.** We train 3 existing object-centric architectures—MONet, Slot Attention (SA), and an additive auto-encoder (AE)—on image data and visualize their SIS as a function of both reconstruction error and compositional contrast. We see across models that, in general, SIS increases as reconstruction error and compositional contrast are minimized.

Results. In Fig. 4A, we visualize the SIS as a function of the reconstruction error and compositional contrast for independent latents for all $K \in \{2, 3, 5\}$. We normalize $\mathcal{L}_{\text{rec}}$ and C_{comp} to ensure that their scores are comparable across different K, which we discuss in further detail in Appx. B.3. As predicted by Thm. 2, we can see that all models that minimize both objectives jointly yield high SIS, whereas models that fail to minimize, e.g., the compositional contrast achieve subpar identifiability. Results for dependent latents yield a similar trend which can be seen in Fig. 5.

5.2 Existing Object-Centric Models

Experimental Setup. We now aim to understand the predictions made by our theory in the context of existing object-centric models trained on image data. To this end, we consider image data generated by the Spriteworld renderer (Watters et al., 2019). Specifically, we generate images with 2 to 4 objects, each described by 4 continuous (size, color, x/y position) and 1 discrete (shape) independent latent factors. Samples of this dataset are shown in Fig. 8. We investigate three object-centric approaches on this data: Slot Attention (Locatello et al., 2020), MONet (Burgess et al., 2019), and an additive auto-encoder. We train all models with 4 latent slots, each with dimension 16, leading to an inferred latent dimension larger than the ground-truth. This discrepancy between inferred and ground-truth latent dimensionality is ubiquitous in existing object-centric models. However, it violates our theoretical assumptions which require equal dimensions. See Appx. B.2 for further experimental details.

Results. SIS as a function of reconstruction error and compositional contrast is shown in Fig. 4B. Similar to Fig. 4A, SIS tends to increase as $\mathcal{L}_{\text{rec}}$ and C_{comp} are minimized, highlighting that our theory holds predictive power for slot identifiability in existing object-centric models. Notably, this is in spite of our theoretical assumptions not being exactly met due to the inferred latent dimension exceeding the ground-truth. This mismatch in dimension does seem to have an effect on SIS, however, which can be seen in Fig. 7. Here, we can see that the subtracted R^2 score in the SIS computation is non-zero across models suggesting that these models are using their additional latent capacity to encode information from multiple objects, despite the decoder presumably not using this information during reconstruction.

6 Discussion

Limitations of Experiments. We emphasize that the main goal of this work is to create a theoretical foundation for object-centric learning. Hence, we focus our experiments on validating Thm. 2 (§ 5.1) and exploring our theoretical predictions in existing object-centric models (§ 5.2). While our experiments in § 5.2 provide evidence that existing models which minimize $\mathcal{L}_{\text{rec}}$ and C_{comp} achieve higher SIS, scaling up these experiments to more models and datasets would lead to a more comprehensive understanding of the exact extent to which the performance of existing models can be understood from our theory. We leave such a larger empirical study for future work.

Limitations of Theory. While we believe that our theoretical assumptions capture the essence of important concepts in object-centric learning, they will be violated to various degrees in practical scenarios. For example, the assumption of compositionality (Defn. 1) on the generator $\mathbf{f}$ is broken by translucency/reflection, as a single pixel can then be affected by multiple latent slots. Additionally, occlusions are not yet fully covered by our theory, as pixels at the border of occluding objects would be affected by multiple latent slots. Additionally, it is common to assume in practice that the generator $\mathbf{f}$ is invariant to permutations of the latent slots it acts on. This permutation invariance leads to a lack of invertibility of $\mathbf{f}$, however, as permuted latents will give rise to the same observation. We anticipate that our theoretical results can be adapted to incorporate such a permutation invariant generator but leave this for future work.

Relationship to Existing Definitions of Objects. Under our framework, groups of pixels corresponding to an object have the property that the latent capacity needed to encode partitions of these pixels separately exceeds the latent capacity needed to encode the pixels as a whole (Defn. 5). Intuitively, this implies that there is latent information shared across different parts of an object. By considering the location of objects as one such latent information, our definition relates to the Gestalt law of common fate (Koffka, 1936; Tangemann et al., 2023) and the concept of a Spelke Object (Spelke, 1990; Chen et al., 2022) which posit that pixels belonging to the same object move together. Furthermore, by considering color or texture as shared latent information, our definition relates to the Gestalt law of similarity (Koffka, 1936) that posits that items sharing visual features tend to be grouped together as a single object.

Extensions of Theory. While our theoretical results provide relatively general conditions under which object-centric representations can be identified, there are several potential ways our results could be extended. First, we hypothesize that the reverse implication of our main result may hold as well, i.e., given the generative model in § 3, compositionality and invertibility are not only sufficient but also necessary conditions for slot identifiability. A formal proof of this conjecture would further highlight the importance of these properties. Additionally, it would be interesting to aim to extend our theoretical approach to identifying not just objects but also abstractions such as part-whole hierarchies (Hinton, 2021) or individual object attributes. In this case, our notion of compositionality would need to be adjusted to account for abstractions that interact during generation. Lastly, it would be interesting to extend our results to leverage weakly-supervised information, such as motion, which has been shown empirically to be helpful for object-centric learning (Tangemann et al., 2023; Kipf et al., 2022; Elsayed et al., 2022; Chen et al., 2022).

Optimizing C_{comp} in Object-Centric Models. While creating a new method for object-centric learning is not the focus of this work, one question based on Thm. 2 is whether C_{comp} can be optimized directly in object-centric models on image data to improve slot identifiability. In this setting, explicitly optimizing C_{comp}, as was done in § 5.1, is challenging as the contrast in its current form is based on Jacobians. Thus, naively optimizing it through gradient descent corresponds to second-order optimization, which creates computational challenges for larger models and data dimensionalities. As previously noted, it could also be the case that there exist implicit ways to enforce that C_{comp} is minimized, which could be occurring to some extent through inductive biases in existing object-centric models. We leave finding computationally efficient ways to minimize C_{comp}, whether explicit or implicit, for future work.

Concluding Remarks. Representing scenes in terms of objects is a key aspect of visual intelligence and an important component of generalization in humans. While empirical object-centric learning methods are increasingly successful, we have thus far been lacking a precise theoretical understanding of what properties of the data and model are sufficient to provably learn object-centric representations. To the best of our knowledge, this work is the first to provide such a theoretical understanding. Along with invertibility, two intuitive assumptions on the generator—compositionality and irreducibility–are sufficient to identify the ground-truth object representations. By extending identifiability theory towards object-centric learning, we hope to facilitate a deeper understanding of existing object-centric models as well as provide a solid foundation for the next generation of models to build upon.

Author Contributions. JB developed the theory with technical help from RSZ, insight from YS, and advising from WB and JvK. JB implemented and executed the experiments with help from RSZ and YS, while RSZ implemented the C_{comp} and SIS metrics on image data. JB and JvK led the writing of the manuscript with help from WB, BS, and RSZ. WB and RSZ created all figures in the manuscript.

Acknowledgments. We thank: the anonymous reviewers for helpful suggestions which led to improvements in the manuscript, Andrea Dittadi for helpful discussions regarding experiments, Attila Juhos for pointing out an issue with Thm. 1, Amin Charusaie, Michel Besserve, and Simon Buchholz for helpful technical discussions, and Zac Cranko for theoretical efforts in the early stages of the project.

This work was supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A, 01IS18039B. WB acknowledges financial support via an Emmy Noether Grant funded by the German Research Foundation (DFG) under grant no. BR 6382/1-1 and via the Open Philantropy Foundation funded by the Good Ventures Foundation. WB is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting RSZ and YS.

References

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V. F., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Çaglar Gülçehre, Song, H. F., Ballard, A. J., Gilmer, J., Dahl, G. E., Vaswani, A., Allen, K. R., Nash, C., Langston, V., Dyer, C., Heess, N. M. O., Wierstra, D., Kohli, P., Botvinick, M. M., Vinyals, O., Li, Y., and Pascanu, R. Relational inductive biases, deep learning, and graph networks. *ArXiv*, abs/1806.01261, 2018. [Cited on page 1.]

Besserve, M., Shajarisales, N., Schölkopf, B., and Janzing, D. Group invariance principles for causal generative models. In *AISTATS*, volume 84 of *Proceedings of Machine Learning Research*, pp. 557–565, 2018. [Cited on page 6.]

Besserve, M., Sun, R., Janzing, D., and Schölkopf, B. A theory of independent mechanisms for extrapolation in generative models. In *AAAI*, pp. 6741–6749, 2021. [Cited on page 6.]

Biza, O., van Steenkiste, S., Sajjadi, M. S., Elsayed, G. F., Mahendran, A., and Kipf, T. Invariant slot attention: Object discovery with slot-centric reference frames. *ArXiv preprint*, abs/2302.04973, 2023. [Cited on page 1.]

Buchholz, S., Besserve, M., and Schölkopf, B. Function classes for identifiable nonlinear independent component analysis. In *NeurIPS*, 2022. [Cited on page 7.]

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. Understanding disentangling in β-vae. *ArXiv preprint*, abs/1804.03599, 2018. [Cited on page 22.]

Burgess, C. P., Matthey, L., Watters, N., Kabra, R., Higgins, I., Botvinick, M., and Lerchner, A. Monet: Unsupervised scene decomposition and representation. *ArXiv preprint*, abs/1901.11390, 2019. [Cited on pages 1 and 8.]

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *ICCV*, pp. 9630–9640, 2021. [Cited on page 1.]

Chen, H., Venkatesh, R. M., Friedman, Y., Wu, J., Tenenbaum, J. B., Yamins, D. L. K., and Bear, D. Unsupervised segmentation in real-world images via spelke object inference. In *European Conference on Computer Vision*, 2022. [Cited on page 9.]

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. [Cited on page 1.]

Daniusis, P., Janzing, D., Mooij, J. M., Zscheischler, J., Steudel, B., Zhang, K., and Schölkopf, B. Inferring deterministic causal relations. In *UAI 2010, Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA, July 8-11, 2010*, pp. 143–150, 2010. [Cited on pages 4 and 6.]

Dehaene, S. *How We Learn: Why Brains Learn Better Than Any Machine... for Now*. 2020. [Cited on page 1.]

Dittadi, A., Papa, S. S., Vita, M. D., Schölkopf, B., Winther, O., and Locatello, F. Generalization and robustness implications in object-centric learning. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5221–5285, 2022. [Cited on pages 5, 7, 22, and 23.]

Elsayed, G. F., Mahendran, A., van Steenkiste, S., Greff, K., Mozer, M. C., and Kipf, T. Savi++: Towards end-to-end object-centric learning from real-world videos. In *NeurIPS*, 2022. [Cited on pages 1, 5, and 9.]

Engelcke, M., Jones, O. P., and Posner, I. Reconstruction bottlenecks in object-centric generative models. *ArXiv preprint*, abs/2007.06245, 2020a. [Cited on page 5.]

Engelcke, M., Kosiorek, A. R., Jones, O. P., and Posner, I. GENESIS: generative scene inference and sampling with object-centric latent representations. In *ICLR*, 2020b. [Cited on page 6.]

Engelcke, M., Jones, O. P., and Posner, I. GENESIS-V2: inferring unordered object representations without iterative refinement. In *NeurIPS*, pp. 8085–8094, 2021. [Cited on page 6.]

Fodor, J. A. and Pylyshyn, Z. W. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2): 3–71, 1988. [Cited on page 1.]

Gerstenberg, T. and Tenenbaum, J. B. Intuitive theories. *Oxford handbook of causal reasoning*, pp. 515–548, 2017. [Cited on page 1.]

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., and Tenenbaum, J. B. A counterfactual simulation model of causal judgments for physical events. *Psychological Review*, 128(5):936, 2021. [Cited on page 1.]

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., and Danks, D. A theory of causal learning in children: causal maps and bayes nets. *Psychological review*, 111(1):3, 2004. [Cited on page 1.]

Goyal, A. and Bengio, Y. Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A*, 478(2266):20210068, 2022. [Cited on page 1.]

Green, E. J. A theory of perceptual objects. *Philosophy and Phenomenological Research*, 99(3):663–693, 2019. [Cited on page 2.]

Greff, K., Srivastava, R. K., and Schmidhuber, J. Binding via reconstruction clustering. *ArXiv preprint*, abs/1511.06418, 2015. [Cited on page 6.]

Greff, K., van Steenkiste, S., and Schmidhuber, J. Neural expectation maximization. In *NIPS*, pp. 6691–6701, 2017. [Cited on page 6.]

Greff, K., Kaufman, R. L., Kabra, R., Watters, N., Burgess, C., Zoran, D., Matthey, L., Botvinick, M. M., and Lerchner, A. Multi-object representation learning with iterative variational inference. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2424–2433, 2019. [Cited on pages 1 and 6.]

Greff, K., Van Steenkiste, S., and Schmidhuber, J. On the binding problem in artificial neural networks. *ArXiv preprint*, abs/2012.05208, 2020. [Cited on pages 1 and 2.]

Gresele, L., Rubenstein, P. K., Mehrjou, A., Locatello, F., and Schölkopf, B. The incomplete rosetta stone problem: Identifiability results for multi-view nonlinear ICA. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019*, volume 115 of *Proceedings of Machine Learning Research*, pp. 217–227, 2019. [Cited on page 7.]

Gresele, L., von Kügelgen, J., Stimper, V., Schölkopf, B., and Besserve, M. Independent mechanism analysis, a new concept? In *NeurIPS*, pp. 28233–28248, 2021. [Cited on pages 4, 6, and 7.]

Guo, S., Tóth, V., Schölkopf, B., and Huszár, F. Causal de Finetti: On the identification of invariant causal structure in exchangeable data. *ArXiv preprint*, abs/2203.15756, 2022. [Cited on page 6.]

Hälvä, H. and Hyvärinen, A. Hidden markov nonlinear ICA: unsupervised learning from nonstationary time series. In *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI 2020, virtual online, August 3-6, 2020*, volume 124 of *Proceedings of Machine Learning Research*, pp. 939–948, 2020. [Cited on page 7.]

Hälvä, H., Corff, S. L., Lehéricy, L., So, J., Zhu, Y., Gassiat, E., and Hyvärinen, A. Disentangling identifiable features from noisy data with structured nonlinear ICA. In *NeurIPS*, pp. 1624–1633, 2021. [Cited on page 7.]

He, K., Gkioxari, G., Dollár, P., and Girshick, R. B. Mask R-CNN. In *ICCV*, pp. 2980–2988, 2017. [Cited on page 1.]

Heess, N. M. O. *Learning generative models of mid-level structure in natural images*. PhD thesis, The University of Edinburgh, 2012. [Cited on page 6.]

Hinton, G. E. How to represent part-whole hierarchies in a neural network. *Neural computation*, pp. 1–40, 2021. [Cited on page 9.]

Horn, R. A. and Johnson, C. R. *Matrix analysis*. 2012. [Cited on page 16.]

Hyvärinen, A. and Hoyer, P. O. Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural Comput.*, 12(7):1705–1720, 2000. [Cited on page 7.]

Hyvärinen, A. and Morioka, H. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *NIPS*, pp. 3765–3773, 2016. [Cited on pages 2 and 7.]

Hyvärinen, A. and Morioka, H. Nonlinear ICA of temporally dependent stationary sources. In *AISTATS*, volume 54 of *Proceedings of Machine Learning Research*, pp. 460–469, 2017. [Cited on pages 2 and 7.]

Hyvärinen, A. and Perkiö, J. Learning to segment any random vector. *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, pp. 4167–4172, 2006. [Cited on page 6.]

Hyvärinen, A., Sasaki, H., and Turner, R. E. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *AISTATS*, volume 89 of *Proceedings of Machine Learning Research*, pp. 859–868, 2019. [Cited on pages 2 and 7.]

Hyvärinen, A. and Pajunen, P. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999. ISSN 0893-6080. [Cited on page 2.]

Janzing, D. Causal versions of maximum entropy and principle of insufficient reason. *Journal of Causal Inference*, 9(1):285–301, 2021. [Cited on page 6.]

Janzing, D. and Schölkopf, B. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010. [Cited on pages 4 and 6.]

Janzing, D., Hoyer, P. O., and Schölkopf, B. Telling cause from effect based on high-dimensional observations. In *ICML*, pp. 479–486, 2010. [Cited on page 6.]

Janzing, D., Mooij, J., Zhang, K., Lemeire, J., Zscheischler, J., Daniušis, P., Steudel, B., and Schölkopf, B. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182:1–31, 2012. [Cited on pages 4 and 6.]

Kabra, R., Burgess, C., Matthey, L., Kaufman, R. L., Greff, K., Reynolds, M., and Lerchner, A. Multi-object datasets. https://github.com/deepmind/multi-object-datasets/, 2019. [Cited on page 22.]

Karazija, L., Laina, I., and Rupprecht, C. Clevrtex: A texture-rich benchmark for unsupervised multi-object segmentation. *ArXiv preprint*, abs/2111.10265, 2021. [Cited on page 1.]

Khemakhem, I., Kingma, D. P., Monti, R. P., and Hyvärinen, A. Variational autoencoders and nonlinear ICA: A unifying framework. In *AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2207–2217, 2020a. [Cited on pages 2 and 7.]

Khemakhem, I., Monti, R. P., Kingma, D. P., and Hyvärinen, A. Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ICA. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020b. [Cited on page 2.]

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015. [Cited on page 22.]

Kipf, T., Elsayed, G. F., Mahendran, A., Stone, A., Sabour, S., Heigold, G., Jonschkowski, R., Dosovitskiy, A., and Greff, K. Conditional object-centric learning from video. In *ICLR*, 2022. [Cited on pages 1 and 9.]

Kipf, T. N., van der Pol, E., and Welling, M. Contrastive learning of structured world models. In *ICLR*, 2020. [Cited on page 1.]

Klindt, D. A., Schott, L., Sharma, Y., Ustyuzhaninov, I., Brendel, W., Bethge, M., and Paiton, D. M. Towards nonlinear disentanglement in natural data with temporal sparse coding. In *ICLR*, 2021. [Cited on pages 2 and 7.]

Koffka, K. *Principles Of Gestalt Psychology*. 1936. [Cited on pages 2 and 9.]

Kuhn, H. W. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. [Cited on pages 7 and 23.]

Lachapelle, S. and Lacoste-Julien, S. Partial disentanglement via mechanism sparsity. In *UAI 2022 Workshop on Causal Representation Learning*, 2022. [Cited on page 7.]

Lachapelle, S., Rodriguez, P., Sharma, Y., Everett, K. E., Le Priol, R., Lacoste, A., and Lacoste-Julien, S. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ica. In *First Conference on Causal Learning and Reasoning*, 2021. [Cited on page 7.]

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017. [Cited on page 1.]

Lin, Z., Wu, Y., Peri, S. V., Sun, W., Singh, G., Deng, F., Jiang, J., and Ahn, S. SPACE: unsupervised object-oriented scene representation via spatial attention and decomposition. In *ICLR*, 2020. [Cited on page 1.]

Locatello, F., Vincent, D., Tolstikhin, I., Rätsch, G., Gelly, S., and Schölkopf, B. Competitive training of mixtures of independent deep generative models. *ArXiv preprint*, abs/1804.11130, 2018. [Cited on page 6.]

Locatello, F., Bauer, S., Lucic, M., Rätsch, G., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124, 2019. [Cited on page 2.]

Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., and Kipf, T. Object-centric learning with slot attention. In *NeurIPS*, 2020. [Cited on pages 1, 5, 7, 8, and 22.]

Marcus, G. F. *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. 2001. [Cited on page 1.]

Moran, G. E., Sridhar, D., Wang, Y., and Blei, D. M. Identifiable variational autoencoders via sparse decoding. *ArXiv preprint*, abs/2110.10804, 2021. [Cited on page 7.]

Papa, S., Winther, O., and Dittadi, A. Inductive biases for object-centric representations in the presence of complex textures. In *UAI 2022 Workshop on Causal Representation Learning*, 2022. [Cited on page 1.]

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E. Z., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pp. 8024–8035, 2019. [Cited on page 22.]

Pearl, J. *Causality*. 2 edition, 2009. [Cited on page 6.]

Peters, B. and Kriegeskorte, N. Capturing the objects of vision with neural networks. *Nature Human Behaviour*, 5(9):1127–1144, 2021. [Cited on page 1.]

Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms*. 2017. [Cited on pages 2, 4, and 6.]

Reizinger, P., Gresele, L., Brady, J., von Kügelgen, J., Zietlow, D., Schölkopf, B., Martius, G., Brendel, W., and Besserve, M. Embrace the gap: Vaes perform independent mechanism analysis. In *NeurIPS*, 2022. [Cited on page 7.]

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241, 2015. [Cited on page 1.]

Roux, N. L., Heess, N. M. O., Shotton, J., and Winn, J. M. Learning a generative model of images by factoring appearance and shape. *Neural Computation*, 23:593–650, 2011. [Cited on page 6.]

Sajjadi, M. S. M., Duckworth, D., Mahendran, A., van Steenkiste, S., Pavetic, F., Lucic, M., Guibas, L. J., Greff, K., and Kipf, T. Object scene representation transformer. In *NeurIPS*, 2022. [Cited on pages 1 and 5.]

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. M. On causal and anticausal learning. In *ICML*, 2012. [Cited on page 6.]

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634, 2021. [Cited on page 6.]

Seitzer, M., Horn, M., Zadaianchuk, A., Zietlow, D., Xiao, T., Simon-Gabriel, C.-J., He, T., Zhang, Z., Schölkopf, B., Brox, T., and Locatello, F. Bridging the gap to real-world object-centric learning. In *The Eleventh International Conference on Learning Representations*, 2023. [Cited on pages 1 and 5.]

Shajarisales, N., Janzing, D., Schölkopf, B., and Besserve, M. Telling cause from effect in deterministic linear dynamical systems. In *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 285–294, 2015. [Cited on page 6.]

Singh, G., Deng, F., and Ahn, S. Illiterate DALL-E learns to compose. In *ICLR*, 2022a. [Cited on page 1.]

Singh, G., Wu, Y., and Ahn, S. Simple unsupervised object-centric learning for complex and naturalistic videos. In *NeurIPS*, 2022b. [Cited on pages 1 and 5.]

Spelke, E. S. Principles of object perception. *Cogn. Sci.*, 14: 29–56, 1990. [Cited on pages 2 and 9.]

Spelke, E. S. What makes us smart? core knowledge and natural language. *Language in mind: Advances in the study of language and thought*, pp. 277–311, 2003. [Cited on page 1.]

Spelke, E. S. and Kinzler, K. D. Core knowledge. *Developmental science*, 10(1):89–96, 2007. [Cited on page 1.]

Spirtes, P., Glymour, C., and Scheines, R. *Causation, Prediction, and Search*, volume 1. 2001. [Cited on page 6.]

Tangemann, M., Schneider, S., von Kügelgen, J., Locatello, F., Gehler, P., Brox, T., Kümmerer, M., Bethge, M., and Schölkopf, B. Unsupervised object learning via common fate. In *2nd Conference on Causal Learning and Reasoning (CLeaR)*, 2023. [Cited on pages 6 and 9.]

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331:1279 – 1285, 2011. [Cited on page 1.]

Träuble, F., Creager, E., Kilbertus, N., Locatello, F., Dittadi, A., Goyal, A., Schölkopf, B., and Bauer, S. On disentangled representations learned from correlated data. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10401–10412, 2021. [Cited on page 6.]

van Steenkiste, S., Chang, M., Greff, K., and Schmidhuber, J. Relational neural expectation maximization: Unsupervised discovery of objects and their interactions. In *ICLR (Poster)*, 2018. [Cited on page 6.]

von Kügelgen, J., Ustyuzhaninov, I., Gehler, P., Bethge, M., and Schölkopf, B. Towards causal generative scene models via competition of experts. In *ICLR 2020 Workshop "Causal Learning for Decision Making"*. [Cited on page 6.]

von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M., and Locatello, F. Self-supervised learning with data augmentations provably isolates content from style. In *NeurIPS*, pp. 16451–16467, 2021. [Cited on page 7.]

Watters, N., Matthey, L., Borgeaud, S., Kabra, R., and Lerchner, A. Spriteworld: A flexible, configurable reinforcement learning environment. https://github.com/deepmind/spriteworld/, 2019. [Cited on pages 8 and 22.]

Weis, M. A., Chitta, K., Sharma, Y., Brendel, W., Bethge, M., Geiger, A., and Ecker, A. S. Benchmarking unsupervised object representations for video sequences. *J. Mach. Learn. Res.*, 22:183:1–183:61, 2021. [Cited on page 1.]

Yang, X., Wang, Y., Sun, J., Zhang, X., Zhang, S., Li, Z., and Yan, J. Nonlinear ICA using volume-preserving transformations. In *ICLR*, 2022. [Cited on page 7.]

Yang, Y. and Yang, B. Promising or elusive? unsupervised object segmentation from real-world single images. In *NeurIPS*, 2022. [Cited on page 1.]

Zheng, Y., Ng, I., and Zhang, K. On the identifiability of nonlinear ICA: Sparsity and beyond. In *Advances in Neural Information Processing Systems*, 2022. [Cited on page 7.]

Zimmermann, R. S., Sharma, Y., Schneider, S., Bethge, M., and Brendel, W. Contrastive learning inverts the data generating process. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12979–12990, 2021. [Cited on pages 2 and 7.]

A Proofs

In this section, we present the proofs for the results presented in the main text. First, we recall our notation:

Notation. N will denote the dimensionality of observations $\mathbf{x}$, K the number of latent slots, and M the dimensionality of each latent slot $\mathbf{z_k}$. For $n \in \mathbb{N}$, $[n]$ will denote the set of natural numbers from 1 to n, i.e., $[n] := \{1, \dots, n\}$. If $\mathbf{f}$ is a function with n component functions, then $\mathbf{f}_S$ will denote the restriction of $\mathbf{f}$ to the component functions indexed by $S \subseteq [n]$, i.e. $\mathbf{f}_S := (f_s)_{s \in S}$ where $\mathbf{f}_S$ is ordered according to the natural ordering of the elements of S. Additionally, when restricting $\mathbf{f}$ to the component functions indexed by $I_k(\mathbf{z})$, defined according to Eq. (3), we will drop the dependence on $\mathbf{z}$ for notional convenience i.e. $\mathbf{f}_{I_k}(\mathbf{z}) := \mathbf{f}_{I_k(\mathbf{z})}(\mathbf{z})$. For functions $\mathbf{f}, \hat{\mathbf{f}}$, we will use $I_k(\mathbf{z}), \hat{I_k}(\hat{\mathbf{z}})$, respectively, to distinguish between the indices defined for each function according to Eq. (3). Lastly, we will slightly abuse notation and use $\mathbf{0}$ to denote both the zero vector and a matrix whose entries are all 0.

We begin by proving several lemmata which will be leveraged for our main theoretical result. We start with the intuitive result that sub-mechanisms from different latent slots are independent in the sense of Defn. 4.

Lemma 1 (Sub-Mechanisms of Distinct Mechanisms are Independent)**.** *Let $\mathbf{f}$ be a diffeomorphism that is compositional (Defn. 1), and let $S_1, S_2 \subseteq [N]$ be nonempty. $\forall \mathbf{z} \in \mathcal{Z}, k \in [K]$, if $S_1 \subseteq I_k(\mathbf{z})$, $S_2 \cap I_k(\mathbf{z}) = \varnothing$, then sub-mechanisms $\mathbf{Jf}_{S_1}(\mathbf{z})$, $\mathbf{Jf}_{S_2}(\mathbf{z})$ are independent in the sense of Defn. 4.*

Proof. From the definition of $I_k(\mathbf{z})$ in Eq. (3) it follows that:

$$\forall n \in [N]: \quad \frac{\partial f_n}{\partial \mathbf{z}_k}(\mathbf{z}) \neq \mathbf{0} \implies n \in I_k(\mathbf{z}).$$

Since $S_1 \subseteq I_k(\mathbf{z})$, we know that $\forall n \in S_1 : \frac{\partial f_n}{\partial \mathbf{z}_k}(\mathbf{z}) \neq \mathbf{0}$. Further, since $S_2 \cap I_k(\mathbf{z}) = \varnothing$ it means that $\forall n \in S_2 : \frac{\partial f_n}{\partial \mathbf{z}_k}(\mathbf{z}) = \mathbf{0}$. Put differently, this means that rows of $\mathbf{Jf}_{S_1}(\mathbf{z})$ are non-zero for those rows where $\mathbf{Jf}_{S_2}(\mathbf{z})$ vanishes and vice versa. Therefore, one cannot represent any column of $\mathbf{Jf}_{S_1}(\mathbf{z})$ as a linear combination of those of $\mathbf{Jf}_{S_2}(\mathbf{z})$. Hence,

$$\operatorname{rank}\left(\mathbf{Jf}_{S_1}(\mathbf{z})\right) + \operatorname{rank}\left(\mathbf{Jf}_{S_2}(\mathbf{z})\right) = \operatorname{rank}\left([\mathbf{Jf}_{S_1}(\mathbf{z}); \mathbf{Jf}_{S_2}(\mathbf{z})]\right),$$

where $[\,\cdot\,;\,\cdot\,]$ denotes vertical concatenation. Note that the RHS is equal to $\mathbf{Jf}_{S_1 \cup S_2}(\mathbf{z})$ up to permutations of rows (which do not change the rank). Thus, Eq. (5) holds for S_1, S_2 showing that $\mathbf{Jf}_{S_1}(\mathbf{z})$, $\mathbf{Jf}_{S_2}(\mathbf{z})$ are independent in the sense of Defn. 4. □

We next show that the rank of each sub-mechanism is less than or equal to the latent slot-dimension dimension, M.

Lemma 2. *Let $\mathbf{f} : \mathcal{Z} \to \mathcal{X}$ be a diffeomorphism that is compositional (Defn. 1). $\forall \mathbf{z} \in \mathcal{Z}, k \in [K]$, if $S \subseteq I_k(\mathbf{z})$ is non-empty:*

$$\operatorname{rank}\left(\mathbf{Jf}_S(\mathbf{z})\right) \leq M. \tag{8}$$

Proof. Since $S \subseteq I_k(\mathbf{z})$, then by compositionality of $\mathbf{f}$

$$\forall \mathbf{z} \in \mathcal{Z}, s \in S, j \in [K] \setminus \{k\} : \quad \frac{\partial f_s}{\partial \mathbf{z}_j}(\mathbf{z}) = \mathbf{0}. \tag{9}$$

Thus, $\mathbf{Jf}_S(\mathbf{z})$ has at most M non-zero columns (those corresponding to the non-zero partials w.r.t. $\mathbf{z}_k$) which implies $\operatorname{rank}(\mathbf{Jf}_S(\mathbf{z})) \leq M$. □

We now show that the rank of each mechanism is equal to the latent slot-dimension M.

Lemma 3. *Let $\mathbf{f} : \mathcal{Z} \to \mathcal{X}$ be a diffeomorphism that is compositional (Defn. 1). Then $\forall \mathbf{z} \in \mathcal{Z}, k \in [K]$:*

$$\operatorname{rank}(\mathbf{Jf}_{I_k}(\mathbf{z})) = M.$$

Proof. First note $\mathbf{f}$ is a diffeomorphism and is thus invertible. Therefore, $\mathbf{Jf}$ must be invertible and thus have full column-rank, i.e., $\forall \mathbf{z} \in \mathcal{Z} : \operatorname{rank}(\mathbf{Jf}(\mathbf{z})) = MK$.

Next, $\forall \mathbf{z} \in \mathcal{Z}, k \in [K]$, let $I_k^C := [N] \setminus I_k$ denote the complement of I_k in $[N]$ such that $I_k^C \cap I_k = \varnothing$. Thus, by Lemma 1, the corresponding sub-mechanisms are independent:

$$\forall \mathbf{z} \in \mathcal{Z}, k \in [K] : \quad \operatorname{rank}(\mathbf{Jf}(\mathbf{z})) = \operatorname{rank}(\mathbf{Jf}_{I_k}(\mathbf{z})) + \operatorname{rank}(\mathbf{Jf}_{I_k^C}(\mathbf{z})) = MK. \tag{10}$$

By compositionality of $\mathbf{f}$,

$$\forall \mathbf{z} \in \mathcal{Z}, j \in [K] \setminus \{ k \} : \quad \frac{\partial \mathbf{f}_{I_k}}{\partial \mathbf{z}_j}(\mathbf{z}) = \mathbf{0}. \tag{11}$$

Thus, $\mathbf{Jf}_{I_k}(\mathbf{z})$ has at most M non-zero columns implying that $\operatorname{rank}(\mathbf{Jf}_{I_k}(\mathbf{z})) \leq M$. Furthermore, by definition,

$$\forall \mathbf{z} \in \mathcal{Z} : \quad \frac{\partial \mathbf{f}_{I_k^C}}{\partial \mathbf{z}_k}(\mathbf{z}) = \mathbf{0}, \tag{12}$$

which means that $\mathbf{Jf}_{I_k^C}(\mathbf{z})$ has at most $(K-1)M$ non-zero columns implying $\operatorname{rank}(\mathbf{Jf}_{I_k^C}(\mathbf{z})) \leq (K-1)M$. Inserting this result in Eq. (10) yields

$$\forall \mathbf{z} \in \mathcal{Z}, k \in [K] : \quad M \leq MK - \operatorname{rank}(\mathbf{Jf}_{I_k^C}(\mathbf{z})) = \operatorname{rank}(\mathbf{Jf}_{I_k}(\mathbf{z})) \leq M, \tag{13}$$

which can only be true if $\operatorname{rank}(\mathbf{Jf}_{I_k}(\mathbf{z})) = M$. □

Next, we show that for ground-truth generator $\mathbf{f}$ and inferred generator $\hat{\mathbf{f}}$, the sub-mechanisms at a given point with respect to the same pixel subset S will be have the same rank.

Lemma 4. *Let $\mathbf{f}, \hat{\mathbf{f}} : \mathcal{Z} \to \mathcal{X}$ be diffeomorphisms with inverses $\mathbf{g}, \hat{\mathbf{g}} : \mathcal{X} \to \mathcal{Z}$, respectively. Then $\forall \mathbf{z} \in \mathcal{Z}, S \subseteq [N]$ s.t. $S \neq \varnothing$, $\operatorname{rank}(\mathbf{Jf}_S(\mathbf{z})) = \operatorname{rank}(\mathbf{J}\hat{\mathbf{f}}_S(\hat{\mathbf{z}}))$, where $\hat{\mathbf{z}} := \hat{\mathbf{g}}(\mathbf{f}(\mathbf{z}))$.*

Proof. First, we introduce the function

$$\mathbf{h} := \hat{\mathbf{g}} \circ \mathbf{f} : \mathcal{Z} \to \mathcal{Z} \quad \text{s.t.} \quad \hat{\mathbf{z}} := \hat{\mathbf{g}}(\mathbf{f}(\mathbf{z})) = \mathbf{h}(\mathbf{z}),$$

We can express $\mathbf{f}$ as $\mathbf{f} = \hat{\mathbf{f}} \circ \hat{\mathbf{g}} \circ \mathbf{f} = \hat{\mathbf{f}} \circ \mathbf{h}$. Thus, if $S \subseteq [N], S \neq \varnothing, \ \mathbf{f}_S = \hat{\mathbf{f}}_S \circ \mathbf{h}$. Therefore,

$$\forall \mathbf{z} \in \mathcal{Z}, \ \operatorname{rank}(\mathbf{Jf}_S(\mathbf{z})) = \operatorname{rank}(\mathbf{J}\hat{\mathbf{f}}_S(\hat{\mathbf{z}})\mathbf{Jh}(\mathbf{z})). \tag{14}$$

Because $\mathbf{h}$ is a diffeomorphism, $\mathbf{Jh}(\mathbf{z})$ is invertible. Thus $\operatorname{rank}(\mathbf{AJh}(\mathbf{z})) = \operatorname{rank}(\mathbf{A})$ for any matrix $\mathbf{A}$ s.t. $\mathbf{AJh}(\mathbf{z})$ is defined (Horn & Johnson, 2012, Section 0.4.6). Therefore, by Eq. (14):

$$\forall \mathbf{z} \in \mathcal{Z}, \ \operatorname{rank}(\mathbf{Jf}_S(\mathbf{z})) = \operatorname{rank}(\mathbf{J}\hat{\mathbf{f}}_S(\hat{\mathbf{z}})). \tag{15}$$

□

We now prove several propositions which will be used to build our main result (Thm. 1). Firstly, we show that each inferred latent slot depends on at least one ground-truth slot.

Proposition 1. *Let $\mathcal{Z}$ be a latent space, $\mathcal{X}$ an observation space, and $\mathbf{f} : \mathcal{Z} \to \mathcal{X}$ a diffeomorphism that is compositional (Defn. 1). Let $\hat{\mathbf{g}} : \mathcal{X} \to \mathcal{Z}$ be a diffeomorphism and $\hat{\mathbf{z}} := \hat{\mathbf{g}}(\mathbf{f}(\mathbf{z})), \forall \mathbf{z} \in \mathcal{Z}$. Then, $\forall \mathbf{z} \in \mathcal{Z}, i \in [K], \exists j \in [K] : \frac{\partial \hat{\mathbf{z}}_j}{\partial \mathbf{z}_i}(\mathbf{z}) \neq \mathbf{0}$.*

Proof. We first define the function

$$\mathbf{h} := \hat{\mathbf{g}} \circ \mathbf{f} : \mathcal{Z} \to \mathcal{Z} \quad \text{s.t.} \quad \hat{\mathbf{z}} := \hat{\mathbf{g}}(\mathbf{f}(\mathbf{z})) = \mathbf{h}(\mathbf{z}).$$

As $\hat{\mathbf{g}}$ and $\mathbf{f}$ are both diffeomorphisms, $\mathbf{h}$ is also a diffeomorphism.

Note that $\forall \mathbf{z} \in \mathcal{Z}, \mathbf{Jh}(\mathbf{z})$ is a square matrix. Furthermore, because $\mathbf{h}$ is a diffeomorphism, it follows that $\forall \mathbf{z} \in \mathcal{Z}$, $\mathbf{Jh}(\mathbf{z})$ is full rank. This implies $\mathbf{Jh}(\mathbf{z})$ must have all non-zero columns, which implies

$$\forall \mathbf{z} \in \mathcal{Z}, i \in [K], \exists j \in [K] : \frac{\partial \hat{\mathbf{z}}_j}{\partial \mathbf{z}_i}(\mathbf{z}) \neq \mathbf{0}.$$

□

Next, we show that each inferred latent slot generates the same pixels as at most one ground-truth slot.

Proposition 2. *Let $\mathcal{Z}$ be a latent space and $\mathcal{X}$ an observation space defined as in § 2. Let $\mathbf{f} : \mathcal{Z} \to \mathcal{X}$ be a diffeomorphism that is compositional (Defn. 1) with irreducible mechanisms (Defn. 5). Let $\hat{\mathbf{g}} : \mathcal{X} \to \mathcal{Z}$ be a diffeomorphism with inverse $\hat{\mathbf{f}} : \mathcal{Z} \to \mathcal{X}$ that is compositional (Defn. 1). Then $\forall \mathbf{z} \in \mathcal{Z}, j \in [K]$, there exists exactly one $i \in [K] : \hat{I}_j(\hat{\mathbf{z}}) \cap I_i(\mathbf{z}) \neq \varnothing$, where $\hat{\mathbf{z}} := \hat{\mathbf{g}}(\mathbf{f}(\mathbf{z}))$*

Proof. Our goal is to show that $\hat{\mathbf{f}}$ maps each inferred latent slot $\hat{\mathbf{z}}_j$ to pixels generated by exactly one ground-truth latent slot $\mathbf{z}_i$.

Step 1 We will first show that $\hat{\mathbf{f}}$ maps each inferred latent slot $\hat{\mathbf{z}}_j$ to pixels generated by at least one ground-truth latent slot $\mathbf{z}_i$. More precisely, we aim to show:

$$\forall \mathbf{z} \in \mathcal{Z}, j \in [K], \exists i \in [K] : \hat{I}_j(\hat{\mathbf{z}}) \cap I_i(\mathbf{z}) \neq \varnothing. \tag{16}$$

Suppose for a contradiction to Eq. (16) that:

$$\exists \mathbf{z}^* \in \mathcal{Z}, j \in [K], \nexists i \in [K] : \hat{I}_j(\hat{\mathbf{z}}^*) \cap I_i(\mathbf{z}^*) \neq \varnothing. \tag{17}$$

We will show that this assumption leads to a contradiction and, hence, is false. Let, $\mathbf{z}^*$ denote the value for which Eq. (17) holds. Eq. (17) coupled with the definition of $I_i(\mathbf{z}^*)$ in Eq. (3) imply that there exists pixels which depend on $\hat{\mathbf{z}}^*$ under $\hat{\mathbf{f}}$ but not on $\mathbf{z}^*$ under $\mathbf{f}$. More precisely,

$$\exists i \in \hat{I}_j(\hat{\mathbf{z}}^*) : \mathbf{J}\hat{\mathbf{f}}_i(\hat{\mathbf{z}}^*) \neq \mathbf{0}, \quad \nexists i \in \hat{I}_j(\hat{\mathbf{z}}^*) : \mathbf{J}\mathbf{f}_i(\mathbf{z}^*) \neq \mathbf{0} \tag{18}$$

This then implies that:

$$\operatorname{rank}(\mathbf{J}\hat{\mathbf{f}}_{\hat{I}_j}(\hat{\mathbf{z}}^*)) \neq \mathbf{0}, \quad \operatorname{rank}(\mathbf{J}\mathbf{f}_{\hat{I}_j}(\mathbf{z}^*)) = \mathbf{0} \tag{19}$$

which contradicts the equality of Jacobian ranks between $\mathbf{f}$ and $\hat{\mathbf{f}}$ stated in Lemma 4. Thus, our assumed contradiction in Eq. (17) cannot hold and we conclude that Eq. (16) must hold true.

Step 2 We will now show that $\hat{\mathbf{f}}$ maps each inferred latent slot $\hat{\mathbf{z}}_j$ to pixels generated by at most one ground-truth latent slot $\mathbf{z}_i$. More precisely, for $C := \{ P \subseteq [K] : |P| > 1 \}$ we aim to show:

$$\forall \mathbf{z} \in \mathcal{Z}, j \in [K], \nexists P \in C : \qquad i \in P \implies \hat{I}_j(\hat{\mathbf{z}}) \cap I_i(\mathbf{z}) \neq \varnothing. \tag{20}$$

Suppose for a contradiction to Eq. (20) that:

$$\exists \mathbf{z}^* \in \mathcal{Z}, j \in [K], P \in C : \qquad i \in P \implies \hat{I}_j(\hat{\mathbf{z}}^*) \cap I_i(\mathbf{z}^*) \neq \varnothing. \tag{21}$$

We will let $\mathbf{z}^*$ denote the value for which Eq. (21) holds and without loss of generality let $j = 1$.

Step 2.1 First, $\forall i \in P$ we define the sets:

$$O_{i,1} := \{ j \in I_i(\mathbf{z}^*) \mid j \in \hat{I}_1(\hat{\mathbf{z}}^*) \}, \;\; O_{i,2} := \{ j \in I_i(\mathbf{z}^*) \mid j \notin \hat{I}_1(\hat{\mathbf{z}}^*) \}, \tag{22}$$

Intuitively, the set $O_{i,1}$ represents the pixels which are a function of both ground-truth latent slot $\mathbf{z}_i^*$ and inferred slot $\hat{\mathbf{z}}_1^*$, while $O_{i,2}$ represents the pixels which are a function of $\mathbf{z}_i^*$ but not $\hat{\mathbf{z}}_1^*$. Our aim is now to show that for all $\forall i \in P$, the sets $O_{i,1}, O_{i,2}$ form a partition of $I_i(\mathbf{z}^*)$.

By Eq. (22), $\forall i \in P$, $O_{i,1} \cup O_{i,2} = I_i(\mathbf{z}^*)$, and $O_{i,1} \cap O_{i,2} = \varnothing$. We thus only need to show that $O_{i,1}, O_{i,2} \neq \varnothing$.

We first note that by our assumed contradiction in Eq. (21), there are pixels which are a function of both ground-truth slot $\mathbf{z}_i^*$ and inferred slot $\hat{\mathbf{z}}_1^*$ i.e.:

$$\forall i \in P, \exists j \in I_i(\mathbf{z}^*) : j \in \hat{I}_1(\hat{\mathbf{z}}^*) \implies j \in O_{i,1} \implies O_{i,1} \neq \varnothing. \tag{23}$$

We will now show that $\forall i \in P, \; O_{i,2} \neq \varnothing$. Suppose for a contradiction that

$$\exists i \in P : O_{i,2} = \varnothing, \tag{24}$$

This implies that $I_i(\mathbf{z}^*) = O_{i,1}$ as $I_i(\mathbf{z}^*) = O_{i,1} \cup O_{i,2} = O_{i,1} \cup \varnothing$. Further, Eq. (22) implies that $O_{i,1} \subseteq \hat{I}_1(\hat{\mathbf{z}}^*)$ thus $O_{i,1} = I_i(\mathbf{z}^*) \subseteq \hat{I}_1(\hat{\mathbf{z}}^*)$.

Next, consider another ground-truth slot $\mathbf{z}_k^*$ where $k \neq i \in P$. As previously established, $O_{k,1} \neq \varnothing$. Moreover, by Eq. (22), $O_{k,1} \subseteq \hat{I}_1(\hat{\mathbf{z}}^*)$. Thus, $A := I_i(\mathbf{z}^*) \cup O_{k,1} \subseteq \hat{I}_1(\hat{\mathbf{z}}^*)$. Now, note that because $\hat{\mathbf{f}}$ is compositional, Lemma 2 implies that the rank of the sub-mechanism defined by $A \leq M$. When coupled with the equality of Jacobian ranks between $\mathbf{f}$ and $\hat{\mathbf{f}}$ stated in Lemma 4, we get:

$$\operatorname{rank}(\mathbf{J}\mathbf{f}_A(\mathbf{z}^*)) = \operatorname{rank}(\mathbf{J}\hat{\mathbf{f}}_A(\hat{\mathbf{z}}^*)) \leq M. \tag{25}$$

Moreover, according to Eq. (22), $O_{k,1} \subseteq I_k(\mathbf{z}^*)$. By compositionality of $\mathbf{f}$, it thus follows that $O_{k,1} \cap I_i(\mathbf{z}^*) = \varnothing$ since $i \neq k$. Therefore, by Lemma 1, we know the sub-mechanisms defined by $I_i(\mathbf{z}^*)$ and $O_{k,1}$ are independent such that

$$\mathrm{rank}(\mathbf{Jf}_A(\mathbf{z}^*)) = \mathrm{rank}(\mathbf{Jf}_{I_i}(\mathbf{z}^*)) + \mathrm{rank}(\mathbf{Jf}_{O_{k,1}}(\mathbf{z}^*)). \tag{26}$$

Leveraging Lemma 3 yields $\mathrm{rank}(\mathbf{Jf}_{I_i}(\mathbf{z}^*)) = M$. Inserting this in the previous equation yields

$$\mathrm{rank}(\mathbf{Jf}_A(\mathbf{z}^*)) = M + \mathrm{rank}(\mathbf{Jf}_{O_{k,1}}(\mathbf{z}^*)), \tag{27}$$

which according to Eq. (25) must be $\leq M$ i.e.

$$M \geq \mathrm{rank}(\mathbf{Jf}_A(\mathbf{z}^*)) = M + \mathrm{rank}(\mathbf{Jf}_{O_{k,1}}(\mathbf{z}^*)). \tag{28}$$

Now, note that by the definition of $I_k(\mathbf{z}^*)$ in Eq. (3), $\forall i \in I_k(\mathbf{z}^*)$, $\mathbf{Jf}_i(\mathbf{z}^*) \neq \mathbf{0}$. Because $O_{k,1} \neq \varnothing$ and $O_{k,1} \subseteq I_k(\mathbf{z}^*)$, it follows that $\mathbf{Jf}_{O_{k,1}}(\mathbf{z}^*) \neq \mathbf{0}$. This implies $\mathrm{rank}(\mathbf{Jf}_{O_{k,1}}(\mathbf{z}^*)) > 0$. However, this contradicts Eq. (28) and, hence, also the initial assumption in Eq. (24). Therefore, we conclude that $\forall i \in P,\ O_{i,2} \neq \varnothing$.

Taken together, we have shown that $\forall i \in P$, the sets $O_{i,1}, O_{i,2}$ are nonempty and form a partition of $I_i(\mathbf{z}^*)$.

Step 2.2 Next, we first note that Lemma 3 implies that the rank of the mechanism $\mathbf{Jf}_{I_i}(\mathbf{z}^*)$ is equal to M. Moreover, by assumption, $\mathbf{Jf}_{I_i}(\mathbf{z}^*)$ is irreducible. Because $O_{i,1}$ and $O_{i,2}$ form a partition of $I_i(\mathbf{z}^*)$, irreducibility then implies:

$$\forall i \in P : \mathrm{rank}(\mathbf{Jf}_{O_{i,1}}(\mathbf{z}^*)) + \mathrm{rank}(\mathbf{Jf}_{O_{i,2}}(\mathbf{z}^*)) > M. \tag{29}$$

Due to the equality of Jacobian ranks between $\mathbf{f}$ and $\hat{\mathbf{f}}$ stated in Lemma 4, Eq. (29) implies

$$\forall i \in P : \mathrm{rank}(\mathbf{J}\hat{\mathbf{f}}_{O_{i,1}}(\hat{\mathbf{z}}^*)) + \mathrm{rank}(\mathbf{J}\hat{\mathbf{f}}_{O_{i,2}}(\hat{\mathbf{z}}^*)) > M. \tag{30}$$

By the definition of $O_{i,1}, O_{i,2}$ in Eq. (22), $\forall i \in P : O_{i,1} \subseteq \hat{I}_1(\hat{\mathbf{z}}^*),\ O_{i,2} \cap \hat{I}_1(\hat{\mathbf{z}}^*) = \varnothing$. It thus follows from Lemma 1 that the sub-mechanisms defined by $O_{i,1}$ and $O_{i,2}$ are independent under $\hat{\mathbf{f}}$ in the sense of Defn. 4. Because $O_{i,1}$ and $O_{i,2}$ form a partition of $I_i(\mathbf{z}^*)$, this independence, when coupled with Eq. (30), implies:

$$\forall i \in P : \mathrm{rank}(\mathbf{J}\hat{\mathbf{f}}_{I_i}(\hat{\mathbf{z}}^*)) = \mathrm{rank}(\mathbf{J}\hat{\mathbf{f}}_{O_{i,1}}(\hat{\mathbf{z}}^*)) + \mathrm{rank}(\mathbf{J}\hat{\mathbf{f}}_{O_{i,2}}(\hat{\mathbf{z}}^*)) > M. \tag{31}$$

We know from Lemma 3 that the mechanism defined by $I_i(\mathbf{z}^*)$ has rank M under $\mathbf{f}$. The equality of Jacobian ranks between $\mathbf{f}$ and $\hat{\mathbf{f}}$ stated in Lemma 4 then implies:

$$\mathrm{rank}(\mathbf{J}\hat{\mathbf{f}}_{I_i}(\hat{\mathbf{z}}^*)) = \mathrm{rank}(\mathbf{Jf}_{I_i}(\mathbf{z}^*)) = M, \tag{32}$$

which contradicts Eq. (31), and, hence the initial assumption of this proof by contradiction in Eq. (21) cannot be correct and Eq. (20) must hold true.

We have now shown that $\forall \mathbf{z} \in \mathcal{Z}, j \in [K]$, there exists at least one and at most one $i \in [K] : \hat{I}_j(\hat{\mathbf{z}}) \cap I_i(\mathbf{z}) \neq \varnothing$ implying there exists exactly one, thus completing the proof. □

We now provide a corollary to Prop. 2 stating that the result also holds when the roles of $\hat{I}_j(\hat{\mathbf{z}}), I_i(\mathbf{z})$ are reversed.

Corollary 1. *$\forall \mathbf{z} \in \mathcal{Z}, i \in [K]$, there exists exactly one $j \in [K] : \hat{I}_j(\hat{\mathbf{z}}) \cap I_i(\mathbf{z}) \neq \varnothing$.*

Proof. We will first prove that there exists at least one $j \in [K] : \hat{I}_j(\hat{\mathbf{z}}) \cap I_i(\mathbf{z}) \neq \varnothing$. Assume, for a contradiction that:

$$\exists \mathbf{z}^* \in \mathcal{Z}, i \in [K], \nexists j \in [K] : \hat{I}_j(\hat{\mathbf{z}}^*) \cap I_i(\mathbf{z}^*) \neq \varnothing. \tag{33}$$

This contradiction can be shown not to hold by exactly repeating the procedure in **Step 1** of Prop. 2.

We thus only need to prove that there exists at most one $j \in [K] : \hat{I}_j(\hat{\mathbf{z}}) \cap I_i(\mathbf{z}) \neq \varnothing$. Let $C := \{\, P \subseteq [K] : |P| > 1 \,\}$. Suppose for a contradiction that:

$$\exists \mathbf{z}^* \in \mathcal{Z}, i \in [K], P \in C : \qquad j \in P \implies \hat{I}_j(\hat{\mathbf{z}}^*) \cap I_i(\mathbf{z}^*) \neq \varnothing. \tag{34}$$

Let $A := [K] \setminus P$. We know by Prop. 2 that $\forall j \in A$, there exists exactly one $i \in [K] : \hat{I}_j(\hat{\mathbf{z}}^*) \cap I_i(\mathbf{z}^*) \neq \varnothing$. This implies that at least $|[K]| - |A| = |P|$ ground-truth latent slots generate pixels which do not overlap with the pixels generated by any inferred latent slots in A. In other words, there exists a set $B \subset [K]$ with cardinality $\geq |P| > 1$ s.t.

$$\forall i \in B, \forall j \in A : \hat{I}_j(\hat{\mathbf{z}}^*) \cap I_i(\mathbf{z}^*) = \varnothing \tag{35}$$

Now consider the set P. We know by Eq. (34), that for all $j \in P : \hat{I}_j(\hat{\mathbf{z}}^*) \cap I_i(\mathbf{z}^*) \neq \varnothing$. By Prop. 2, we know that for all $j \in P$, $\hat{I}_j(\hat{\mathbf{z}}^*)$ can intersect only with $I_i(\mathbf{z}^*)$. Given that $|B| > 1$, this then implies

$$\exists i \in B : \forall j \in P : \hat{I}_j(\hat{\mathbf{z}}^*) \cap I_i(\mathbf{z}^*) = \varnothing \tag{36}$$

Now, by construction, $[K] = A \cup P$. Thus, Eq. (35) and Eq. (36) together imply:

$$\exists i \in B \subset [K] : \forall j \in [K] : \hat{I}_j(\hat{\mathbf{z}}^*) \cap I_i(\mathbf{z}^*) = \varnothing \tag{37}$$

We have already shown in the first part of this corollary, however, that Eq. (37) cannot be true by repeating the procedure in **Step 1** of Prop. 2. Thus, our assumed contradiction in Eq. (34) cannot be true.

We have now shown that $\forall \mathbf{z} \in \mathcal{Z}, i \in [K]$, there exists at least one and at most one $j \in [K] : \hat{I}_j(\hat{\mathbf{z}}) \cap I_i(\mathbf{z}) \neq \varnothing$ implying there exists exactly one, thus completing the proof. □

We now build upon Prop. 2 and Cor. 1, to show that all inferred latent slots depend on at most one ground-truth slot.

Proposition 3. *Let $\mathcal{Z}$ be a latent space and $\mathcal{X}$ an observation space. Let $\mathbf{f} : \mathcal{Z} \to \mathcal{X}$ be a diffeomorphism that is compositional (Defn. 1) with irreducible mechanisms (Defn. 5). Let $\hat{\mathbf{g}} : \mathcal{X} \to \mathcal{Z}$ be a diffeomorphism with inverse $\hat{\mathbf{f}} : \mathcal{Z} \to \mathcal{X}$ that is compositional (Defn. 1). Let $\hat{\mathbf{z}} := \hat{\mathbf{g}}(\mathbf{f}(\mathbf{z})), \forall \mathbf{z} \in \mathcal{Z}$. Then, $\forall \mathbf{z} \in \mathcal{Z}, i \in [K]$, there exists at most one $j \in [K] : \frac{\partial \hat{\mathbf{z}}_j}{\partial \mathbf{z}_i}(\mathbf{z}) \neq \mathbf{0}$.*

Proof. Our goal is to show that at most one $\hat{\mathbf{z}}_j$ is a function of a given $\mathbf{z}_i$. More precisely, let $C := \{ P \subseteq [K] : |P| > 1 \}$. We aim to show that:

$$\forall \mathbf{z} \in \mathcal{Z}, i \in [K], \nexists P \in C : \qquad j \in P \implies \frac{\partial \hat{\mathbf{z}}_j}{\partial \mathbf{z}_i}(\mathbf{z}) \neq \mathbf{0}. \tag{38}$$

Suppose for a contradiction to Eq. (38) that:

$$\exists \mathbf{z}^* \in \mathcal{Z}, i \in [K], P \in C : \qquad j \in P \implies \frac{\partial \hat{\mathbf{z}}_j}{\partial \mathbf{z}_i}(\mathbf{z}^*) \neq \mathbf{0}. \tag{39}$$

Let $\mathbf{z}^*$ denote the value for which Eq. (39) holds and without loss of generality let $i = 1$.

We first introduce the function

$$\mathbf{h} := \hat{\mathbf{g}} \circ \mathbf{f} : \mathcal{Z} \to \mathcal{Z} \text{ s.t. } \hat{\mathbf{z}} := \hat{\mathbf{g}}(\mathbf{f}(\mathbf{z})) = \mathbf{h}(\mathbf{z}).$$

Note that $\mathbf{f} = \hat{\mathbf{f}} \circ \hat{\mathbf{g}} \circ \mathbf{f} = \hat{\mathbf{f}} \circ \mathbf{h}$. Thus, $\forall S \subseteq [N], \;\; \mathbf{f}_S = \hat{\mathbf{f}}_S \circ \mathbf{h}$. Therefore,

$$\forall \mathbf{z} \in \mathcal{Z}, j \in [K] : \frac{\partial \mathbf{f}_{\hat{I}_j}}{\partial \mathbf{z}_1}(\mathbf{z}) = \frac{\partial \hat{\mathbf{f}}_{\hat{I}_j}}{\partial \hat{\mathbf{z}}}(\hat{\mathbf{z}}) \frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}_1}(\mathbf{z}) \tag{40}$$

Due to the compositionality of $\hat{\mathbf{f}}$, $\frac{\partial \hat{\mathbf{f}}_{\hat{I}_j}}{\partial \hat{\mathbf{z}}_k}(\hat{\mathbf{z}}) = \mathbf{0}, \forall k \neq j \in [K]$. This implies that these columns can be ignored when taking the product in Eq. (40), s.t.

$$\frac{\partial \hat{\mathbf{f}}_{\hat{I}_j}}{\partial \hat{\mathbf{z}}}(\hat{\mathbf{z}}) \frac{\partial \hat{\mathbf{z}}}{\partial \mathbf{z}_1}(\mathbf{z}) = \frac{\partial \hat{\mathbf{f}}_{\hat{I}_j}}{\partial \hat{\mathbf{z}}_j}(\hat{\mathbf{z}}) \frac{\partial \hat{\mathbf{z}}_j}{\partial \mathbf{z}_1}(\mathbf{z}). \tag{41}$$

Now by Cor. 1, there exists exactly one $j \in P \subseteq [K]$ s.t. $\hat{I}_j(\hat{\mathbf{z}}^*) \cap I_1(\mathbf{z}^*) \neq \varnothing$. By the definition of $I_i(\mathbf{z})$ in Eq. (3), this implies that there exists exactly one $j \in P$ s.t. $\frac{\partial \mathbf{f}_{\hat{I}_j}}{\partial \mathbf{z}_1}(\mathbf{z}^*) \neq \mathbf{0}$. $|P| > 1$, thus there exists a $j \in P$ s.t.

$$\frac{\partial \mathbf{f}_{\hat{I}_j}}{\partial \mathbf{z}_1}(\mathbf{z}^*) = \frac{\partial \hat{\mathbf{f}}_{\hat{I}_j}}{\partial \hat{\mathbf{z}}_j}(\hat{\mathbf{z}}^*) \frac{\partial \hat{\mathbf{z}}_j}{\partial \mathbf{z}_1}(\mathbf{z}^*) = \mathbf{0} \tag{42}$$

where we leveraged Eq. (40), Eq. (41) to get the first equality above. Now, we know by Lemma 3, that $\mathbf{J}\hat{\mathbf{f}}_{\hat{I}_j}(\hat{\mathbf{z}}^*)$ is full column-rank. By compositionality of $\hat{\mathbf{f}}$, we also know that $\mathrm{rank}(\mathbf{J}\hat{\mathbf{f}}_{\hat{I}_j}(\hat{\mathbf{z}}^*)) = \mathrm{rank}(\frac{\partial \hat{\mathbf{f}}_{\hat{I}_j}}{\partial \hat{\mathbf{z}}_j}(\hat{\mathbf{z}}^*))$ as these are the only non-zero columns in $\mathbf{J}\hat{\mathbf{f}}_{\hat{I}_j}(\hat{\mathbf{z}}^*)$. Thus, $\frac{\partial \hat{\mathbf{f}}_{\hat{I}_j}}{\partial \hat{\mathbf{z}}_j}(\hat{\mathbf{z}}^*)$ is also full column-rank. Now, Eq. (42) implies that all columns of $\frac{\partial \hat{\mathbf{z}}_j}{\partial \mathbf{z}_1}(\mathbf{z}^*)$ must be in $\mathrm{null}(\frac{\partial \hat{\mathbf{f}}_{\hat{I}_j}}{\partial \hat{\mathbf{z}}_j}(\hat{\mathbf{z}}^*))$. Because, $\frac{\partial \hat{\mathbf{f}}_{\hat{I}_j}}{\partial \hat{\mathbf{z}}_j}(\hat{\mathbf{z}}^*)$ is full-column rank, $\mathrm{null}(\frac{\partial \hat{\mathbf{f}}_{\hat{I}_j}}{\partial \hat{\mathbf{z}}_j}(\hat{\mathbf{z}}^*)) = \mathbf{0}$. However, by Eq. (39) at least one column of $\frac{\partial \hat{\mathbf{z}}_j}{\partial \mathbf{z}_1}(\mathbf{z}^*)$ is non-zero. Thus, we obtain a contradiction and conclude that Eq. (38) must hold. □

Building on top of the previous propositions, we now prove our main identifiability result:

Theorem 1. *Let $\mathbf{f} : \mathcal{Z} \to \mathcal{X}$ be a diffeomorphism that is compositional (Defn. 1) with irreducible mechanisms (Defn. 5). If an inference model $\hat{\mathbf{g}} : \mathcal{X} \to \mathcal{Z}$ is* (i) *a diffeomorphism with* (ii) *compositional inverse $\hat{\mathbf{f}} = \hat{\mathbf{g}}^{-1}$, then $\hat{\mathbf{g}}$ slot-identifies $\mathbf{z} = \mathbf{g}(\mathbf{x})$ in the sense of Defn. 6.*

Proof. According to Prop. 1 every inferred latent slot $\hat{\mathbf{z}}_j$ depends on *at least* one ground-truth latent slot $\mathbf{z}_i$. At the same time, Prop. 3 states that every inferred latent slot depends on *at most* one ground-truth slot. Hence, every inferred latent slot depends on *exactly* one ground-truth slot.

This implies that the Jacobian $\mathbf{J}\mathbf{h}(\mathbf{z})$ of $\mathbf{h} = \hat{\mathbf{g}} \circ \mathbf{f} : \mathcal{Z} \to \mathcal{Z}$ must be block diagonal up to permutation everywhere:

$$\forall \mathbf{z} \in \mathcal{Z} : \qquad \mathbf{J}\mathbf{h}(\mathbf{z}) = \mathbf{P}(\mathbf{z})\mathbf{B}(\mathbf{z}) \tag{43}$$

where $\mathbf{P}(\mathbf{z})$ is a permutation matrix and $\mathbf{B}(\mathbf{z})$ a block-diagonal matrix.

Next, note that

$$\det(\mathbf{J}\mathbf{h}(\mathbf{z})) = \det(\mathbf{P}(\mathbf{z}))\det(\mathbf{B}(\mathbf{z})) = \det(\mathbf{B}(\mathbf{z})) \neq 0 \tag{44}$$

since $\mathbf{h}$ is diffeomorphic. Hence, $\mathbf{B}(\mathbf{z})$ is invertible with continuous inverse. We conclude that

$$\mathbf{P}(\mathbf{z}) = \mathbf{J}\mathbf{h}(\mathbf{z})\mathbf{B}^{-1}(\mathbf{z}) \tag{45}$$

is continuous. At the same time, $\mathbf{P}(\mathbf{z})$ can only attain a finite set of values since it is a permutation. Hence, $\mathbf{P}(\mathbf{z})$ must be constant in $\mathbf{z}$, that is, the same global permutation is used everywhere.[5]

Thus, for any $j \in K$, there exists a *unique* $i \in K$ such that the function $\mathbf{h}_j = \hat{\mathbf{g}}_j \circ \mathbf{f} : \mathcal{Z} \to \mathcal{Z}_j$ is, in fact, constant in all slots except $\mathcal{Z}_i$, i.e., it can be written as a mapping $\mathbf{h}_j : \mathcal{Z}_i \to \mathcal{Z}_j$.

Finally, all such $\mathbf{h}_j$ are diffeomorphic, since $\mathbf{h}$ is a diffeomorphism.

This concludes the proof that assumptions *(i)* and *(ii)* imply $\hat{\mathbf{g}}$ slot-identifies $\mathbf{z}$. □

We now show that the compositional contrast C_{comp} introduced in Eq. (6) indicates whether a map is compositional:

Lemma 5. *Let $\mathbf{f} : \mathcal{Z} \to \mathcal{X}$ be a differentiable function. $\mathbf{f}$ is compositional in the sense of Defn. 1 if and only if for all $\mathbf{z} \in \mathcal{Z}$:*

$$C_{\text{comp}}(\mathbf{f}, \mathbf{z}) = 0\,.$$

Proof. ($\Rightarrow$) We begin by analyzing $C_{\text{comp}}(\mathbf{f}, \mathbf{z})$:

$$\sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{j=k+1}^{K}\left\|\frac{\partial f_n}{\partial \mathbf{z}_k}(\mathbf{z})\right\|_2\left\|\frac{\partial f_n}{\partial \mathbf{z}_j}(\mathbf{z})\right\|_2 \tag{46}$$

Since all summands are non-negative, the sum can only equal zero if every summand is zero $\forall \mathbf{z} \in \mathcal{Z}$. Since $j \neq k$ in the summand, this means:

$$\forall \mathbf{z} \in \mathcal{Z}, \forall n \in [N], k \neq j \in [K] : \left\|\frac{\partial f_n}{\partial \mathbf{z}_k}(\mathbf{z})\right\|_2\left\|\frac{\partial f_n}{\partial \mathbf{z}_j}(\mathbf{z})\right\|_2 = 0 \tag{47}$$

[5] Suppose for a contradiction that $\mathbf{P}(\mathbf{z})$ attains distinct values at some $\mathbf{z}^A \neq \mathbf{z}^B$ in $\mathcal{Z}$. Since $\mathcal{Z}$ is convex, the line connecting $\mathbf{z}^A$ and $\mathbf{z}^B$ is also in $\mathcal{Z}$ and $\mathbf{P}$ must change value somewhere along this line, leading to a discontinuity and thus a contradiction.

This relation can only be satisfied if one (or both) of the partial derivatives in the summand have a norm of zero, i.e. if they are zero. More precisely,

$$\forall \mathbf{z} \in \mathcal{Z}, \forall n \in [N], k \neq j \in [K] : \frac{\partial f_n}{\partial \mathbf{z}_k}(\mathbf{z}) = \mathbf{0} \vee \frac{\partial f_n}{\partial \mathbf{z}_j}(\mathbf{z}) = \mathbf{0}. \tag{48}$$

According to Defn. 1 a map $\mathbf{f}$ is compositional if

$$\forall \mathbf{z} \in \mathcal{Z} : \qquad k \neq j \implies I_k(\mathbf{z}) \cap I_j(\mathbf{z}) = \varnothing. \tag{49}$$

By the definition of $I_i(\mathbf{z})$ in Eq. (3), we can restate Eq. (49) as:

$$\forall \mathbf{z} \in \mathcal{Z}, k \neq j, \nexists n \in [N] : \frac{\partial f_n}{\partial \mathbf{z}_k}(\mathbf{z}) \neq \mathbf{0} \wedge \frac{\partial f_n}{\partial \mathbf{z}_j}(\mathbf{z}) \neq \mathbf{0} \tag{50}$$

which implies:

$$\forall \mathbf{z} \in \mathcal{Z}, n \in [N], k \neq j : \frac{\partial f_n}{\partial \mathbf{z}_k}(\mathbf{z}) = \mathbf{0} \vee \frac{\partial f_n}{\partial \mathbf{z}_j}(\mathbf{z}) = \mathbf{0} \tag{51}$$

which is equivalent to Eq. (48). Hence, $\forall \mathbf{z} \in \mathcal{Z} : C_{\text{comp}}(\mathbf{f}, \mathbf{z}) = 0$ implies that $\mathbf{f}$ is compositional.

($\Leftarrow$) We now prove the reverse direction i.e. that if $\mathbf{f}$ is compositional, then $\forall \mathbf{z} \in \mathcal{Z} : C_{\text{comp}}(\mathbf{f}, \mathbf{z}) = 0$. Note that the form of compositionality given in Eq. (50) implies that $\forall \mathbf{z} \in \mathcal{Z}$, at least one term in the summand of $C_{\text{comp}}(\mathbf{f}, \mathbf{z})$ in Eq. (51) will be zero. Thus, each summand is equal to zero. This then implies that $\forall \mathbf{z} \in \mathcal{Z} : C_{\text{comp}}(\mathbf{f}, \mathbf{z}) = 0$, completing the proof. □

Finally, by leveraging Lemma 5, we can obtain Thm. 1 in a less abstract form.

Theorem 2. *Let $\mathbf{f} : \mathcal{Z} \to \mathcal{X}$ be a diffeomorphism that is compositional (Defn. 1) with irreducible mechanisms (Defn. 5). If an encoder $\hat{\mathbf{g}} : \mathcal{X} \to \mathcal{Z}$ and decoder $\hat{\mathbf{f}} : \mathcal{Z} \to \mathcal{X}$ are both differentiable and solve the following functional equation*

$$\mathbb{E}_{\mathbf{x} \sim p_{\mathbf{x}}} \left[\left\| \hat{\mathbf{f}}(\hat{\mathbf{g}}(\mathbf{x})) - \mathbf{x} \right\|_2^2 + \lambda C_{\text{comp}} \left(\hat{\mathbf{f}}, \hat{\mathbf{g}}(\mathbf{x}) \right) \right] = 0, \tag{7}$$

for $\lambda > 0$, then $\hat{\mathbf{g}}$ slot-identifies $\mathbf{z}$ in the sense of Defn. 6.

Proof. As both summands of the functional are non-negative, solving the functional equation means solving for each of the summands to be equal to zero. Thus, we can analyze both of them separately. Solving the first sub-functional equation, i.e.,

$$\mathbb{E}_{\mathbf{x} \sim p_{\mathbf{x}}} \left[\left\| \hat{\mathbf{f}}(\hat{\mathbf{g}}(\mathbf{x})) - \mathbf{x} \right\|_2^2 \right] = 0,$$

implies that $\hat{\mathbf{f}}$ is an inverse of $\hat{\mathbf{g}}$ for every $\mathbf{x} \sim p_{\mathbf{x}}$. Because $p_{\mathbf{z}}$ is assumed to have full support over $\mathcal{Z}$, and $p_{\mathbf{x}}$ is defined by applying a diffeomorphism $\mathbf{f} : \mathcal{Z} \to \mathcal{X}$ on $p_{\mathbf{z}}$, this implies that $p_{\mathbf{x}}$ has full support over $\mathcal{X}$. This means that $\hat{\mathbf{f}}$ is an inverse of $\hat{\mathbf{g}}$ over the entire space $\mathcal{X}$ i.e. $\hat{\mathbf{f}} = \hat{\mathbf{g}}^{-1}$. Since per assumption $\hat{\mathbf{g}}$ and $\hat{\mathbf{f}}$ are differentiable it follows that $\hat{\mathbf{g}}$ is a diffeomorphism.

Moreover, per Lemma 5, solving the second sub-functional equation for $\lambda > 0$, i.e.,

$$\mathbb{E}_{\mathbf{x} \sim p_{\mathbf{x}}} \left[\lambda C_{\text{comp}}(\hat{\mathbf{f}}, \hat{\mathbf{g}}(\mathbf{x})) \right] = 0,$$

means that $\hat{\mathbf{f}}$ is compositional as $p_{\mathbf{x}}$ has full support over $\mathcal{X}$ and $\hat{\mathbf{g}}$ is a diffeomorphism between $\mathcal{X}$ and $\mathcal{Z}$. From Thm. 1 it now follows that $\hat{\mathbf{g}}$ slot-identifies $\mathbf{z}$, concluding the proof. □

B Experimental Details

B.1 Synthetic Data § 5.1

Enforcing Irreducibility We choose slot-output dimension, which we will denote $\dim(\mathbf{x}_s)$, to be greater than slot-dimension M as this is required for irreducibility (Defn. 5). To see this, assume the number of rows in each mechanism (Defn. 2), equal in our case to $\dim(\mathbf{x}_s)$, were equal to M. Because mechanisms have $\text{rank} = M$ (Lemma 3) and we have

M rows, this implies that no row is in the span of any others. Hence, the mechanism would be reducible. Beyond enforcing that the slot-output dimension, equal to 20 in this case, is greater than $M = 3$, we do not do anything further to ensure that our ground-truth generator is irreducible. This is because it is extremely unlikely that the generator, as we have constructed it, could be reducible. Specifically, if the generator were reducible, then as $\dim(\mathbf{x}_s)$ becomes larger than M, each new row in the Jacobian would need to lie in the span of some subset of the previous rows. As $\dim(\mathbf{x}_s)$ continues to increase relative to M, however, this becomes increasingly unlikely since the rows in the weight matrices of our MLP generator are randomly sampled i.e. entries are sampled uniformly from $[-10, 10]$.

Inference Model Training and Evaluation For our inference model, we use a 3 layer MLP with 80 hidden units in each layer and LeakyReLU activation functions. We train on 75,000 samples and use 6,000 and 5,000 for validation and test sets, respectively. We train for 100 epochs with the Adam optimizer (Kingma & Ba, 2015) on batches of 64 with an initial learning rate of 10^{-3}, which we decay by factor of 10 after 50 epochs. We use the validation set to find the optimal permutation for the Hungarian matching and then evaluate the SIS on the test set after applying this permutation to the slots. We compute the SIS for models every 4 epochs during training, all of which are plotted in Fig. 4. We trained all models using PyTorch (Paszke et al., 2019).

B.2 Existing Object-Centric Models § 5.2

Data Generation We generate image data using the Spriteworld renderer (Watters et al., 2019). Images consist of 2 to 4 objects, each described by 4 continuous (size, color, x/y position) and 1 discrete (shape) independent latent factors. We sample all factors uniformly where size is sampled from $[.1, .15]$ and x/y position both from $[.1, .8]$. We represent color using HSV and sample hue from $[0, 1]$ while fixing saturation and value to 3 and 1, respectively. The dataset consists of 100,000 images, 90,000 of which are used for training and 10,000 for evaluation.

Inference Model Training and Evaluation We use the same Slot Attention model proposed by Locatello et al. (2020), with the changes being that we use 16 convolutional filters in the decoder opposed to 32 and do not use a learning rate warm-up. For MONet, we follow the setup used by Dittadi et al. (2022) on Multi-dSprites (Kabra et al., 2019). For our additive autoencoder, we use the convolutional encoder/decoder architecture proposed by Burgess et al. (2018). The model decodes each slot separately to get slot-wise reconstructions and mask, applies the normalized mask to each slot-wise reconstruction, and then adds the results together to get the final reconstructed image. For all models, we use 4 slots with a slot-dimension of 16. We train all models for $500,000$ iterations (356 epochs) on batches of 64 with between 5 to 12 random seeds for each model. We train using the Adam optimizer (Kingma & Ba, 2015) with an initial learning rate of 10^{-4}, which we decay throughout training for all models using the same decay scheduler as Locatello et al. (2020). We trained all models using PyTorch (Paszke et al., 2019).

B.3 Compositional Contrast Normalized Variants

When computing C_{comp} in § 5.1 and § 5.2, we use a few different normalized variants of the contrast to overcome potential issues with the definition given in Defn. 7. Firstly, as the number of latent slots K increases, the contrast in Defn. 7 will scale by a factor $K^2 - K$. Thus, when comparing models across different numbers of slots in § 5.1, we divide the contrast by this factor to ensure that comparisons remain meaningful across different values of K. Another issue with the contrast in Defn. 7, is that it is not scale invariant. Specifically, naively minimizing the norm of the gradients for each pixel across slots will also minimize the contrast, despite all slots having similar gradient norms for a given pixel. This scale invariance did not cause issues when optimizing C_{comp} directly in § 5.1. However, when evaluating the C_{comp} of object-centric models in § 5.2, we account for this invariance. Specifically, we divide the gradient norms for each pixel with respect to each slot by the mean gradient norm for this pixel across slots. This gradient normalization creates an additional problem, however: Pixels with a relatively small gradient norm, such as black background pixels, will be weighted equally to pixels with a larger gradient norm such as pixels corresponding to an object. To account for this, we weight each pixel's contribution to the contrast by the pixel's mean gradient across slots.

B.4 Slot Identifiability Score

We are interested in a metric measuring how much information about the ground-truth latent slots is contained in the inferred latent slots without mixing information about different ground-truth slots into the same inferred slot. Let $S_1, S_2 \in [0, 1]$ denote scores that quantify how much information about each ground-truth slot can be extracted from the most and second-most predictive inferred slot, respectively. The aforementioned metric can be computed by just subtracting the two scores, i.e.

$$S = S_1 - S_2. \tag{52}$$

Following previous work, we use the R^2 coefficient of determination as a score for continuous factors of variation (which we restrict to be strictly non-negative) and the accuracy for categorical factors (Dittadi et al., 2022). We compute one S value for each type and take the weighted mean which we then average across all slots to get the final slot identifiability score (SIS).

Computing SIS on Synthetic Data § 5.1 To compute the scores S_1 and S_2 defined in our experiments in § 5.1, we must fit two inference models between ground-truth and inferred slots: one between the best-matching slots and one between the second-best-matching slots. In § 5.1, we fit these models by first fitting a kernel ridge regression model between every pair of inferred and ground-truth slots and computing the R^2 scores for the predictions given by each model. We then use the Hungarian algorithm (Kuhn, 1955) to match each ground-truth slot to its most predictive inferred slot based on these R^2 scores, which gives us S_1. To get S_2, we take the highest R^2 score for each inferred slot with respect to the ground-truth slots that it was not already matched with. For our experiments in Fig. 5 with dependent latent slots, S_2 will inevitably be non-zero even if a model is perfectly identifiable. Thus, for these experiments, we only consider S_1 and refer to this metric as the Slot MCC (Mean Correlation Coefficient).

Computing SIS on Image Data § 5.2 When training models to compute S_1 and S_2 in our experiments on image data in § 5.2, one issue that arises is that the permutation between inferred latent slots and ground-truth slots is not necessarily a global permutation but can also be a local permutation. This is due to the ground-truth generator function being permutation invariant. To resolve this, we take a similar approach to work by Dittadi et al. (2022) and perform an online matching during training of inferred latent slots to ground-truth slots using the training loss. Specifically, we compute the loss for every pairing of the ground-truth and inferred slots and use the Hungarian algorithm to pick the permutation that yields the lowest aggregate loss. As every slot can contain both continuous and categorical variables, we compute the mean squared error for continuous factors and cross-entropy for categorical variables and sum them up to obtain the training loss. In our experiments, we notice that the cross-entropy tends to yield unstable matching results. Therefore, we use the minimum probability margin [6] to compute the categorical loss to solve the matching problem. Before fitting the readout models, we standardized both the ground-truth and inferred latents. We parameterized the readout models as 5-layer MLPs with LeakyReLU nonlinearity and a hidden dimensionality of 256, and trained them for up to 100 epochs using the Lion optimizer with a learning rate of 10^{-4}. To prevent the network from locking in too early on a suboptimal solution, we add a small amount of noise (10 % of the maximum matching loss value) to the losses before determining the optimal matching. Finally, we suggest performing cross-validation and early stopping to prevent overfitting.

For training the model to compute S_2, we proceed as for S_1 but ensure that the model is not using the same permutation used for computing S_1, i.e., it is trained on the second-best matching between ground-truth and inferred slots. Lastly, when computing S_2, we aim to avoid scenarios in which the model finds a spurious permutation yielding a non-zero S_2 despite the model being identifiable. To account for this, we compute S_2 on the ground-truth latent slots, denoted S_2^{gt}, using the same procedure for computing S_2, and use this score to adjust our previous scores. Specifically, by adjusting the value range accordingly, we obtain a score of

$$S = \frac{S_1 - S_2^{\text{gt}}}{1 - S_2^{\text{gt}}} - \frac{S_2 - S_2^{\text{gt}}}{1 - S_2^{\text{gt}}}, \tag{53}$$

To ensure that the subtracting term is not increasing the final score, we restrict it to be positive, yielding the final score:

$$S = \frac{S_1 - S_2^{\text{gt}}}{1 - S_2^{\text{gt}}} - \max\left(\frac{S_2 - S_2^{\text{gt}}}{1 - S_2^{\text{gt}}}, 0\right). \tag{54}$$

We may additionally be interested in considering the two terms on the RHS of Eq. (54) separately. Thus, we define them below as:

$$\hat{S}_1 = \frac{S_1 - S_2^{\text{gt}}}{1 - S_2^{\text{gt}}}, \quad \hat{S}_2 = \frac{S_2 - S_2^{\text{gt}}}{1 - S_2^{\text{gt}}}, \quad S = \hat{S}_1 - \max(\hat{S}_2, 0). \tag{55}$$

[6] i.e., $\max_i p_i - p_y$, where p denotes the predicted probability for different values of the categorical distribution and y the ground-truth value

C Additional Figures and Experiments

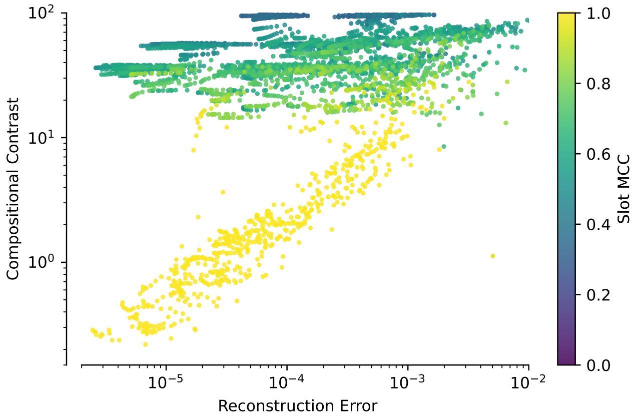

Figure 5. **Experimental validation of Thm. 2 for statistically dependent slots.** We trained models on synthetic data generated according to § 2 with 2, 3, 5 dependent latent slots (see § 5.1). The color coding indicates the level of identifiability achieved by the model, measured by the Slot Mean Correlation Coefficient (MCC), where higher values correspond to more identifiable models. As predicted by our theory, if a model sufficiently minimizes both reconstruction error and compositional contrast, then it identifies the ground-truth latent slots.

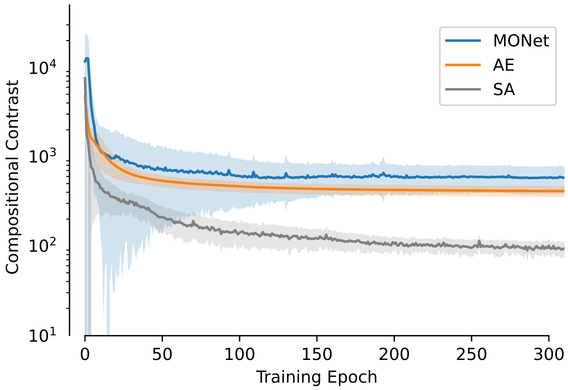

Figure 6. **Compositional Contrast (C_{comp}) throughout training.** Here, we plot the compositional contrast (C_{comp}) over the course of training for MONet, Slot Attention (SA) as well as an additive auto-encoder (AE), on image data. We can see that all models appear to be minimizing C_{comp} to some extent despite it not being explicitly optimized for in any of these models.

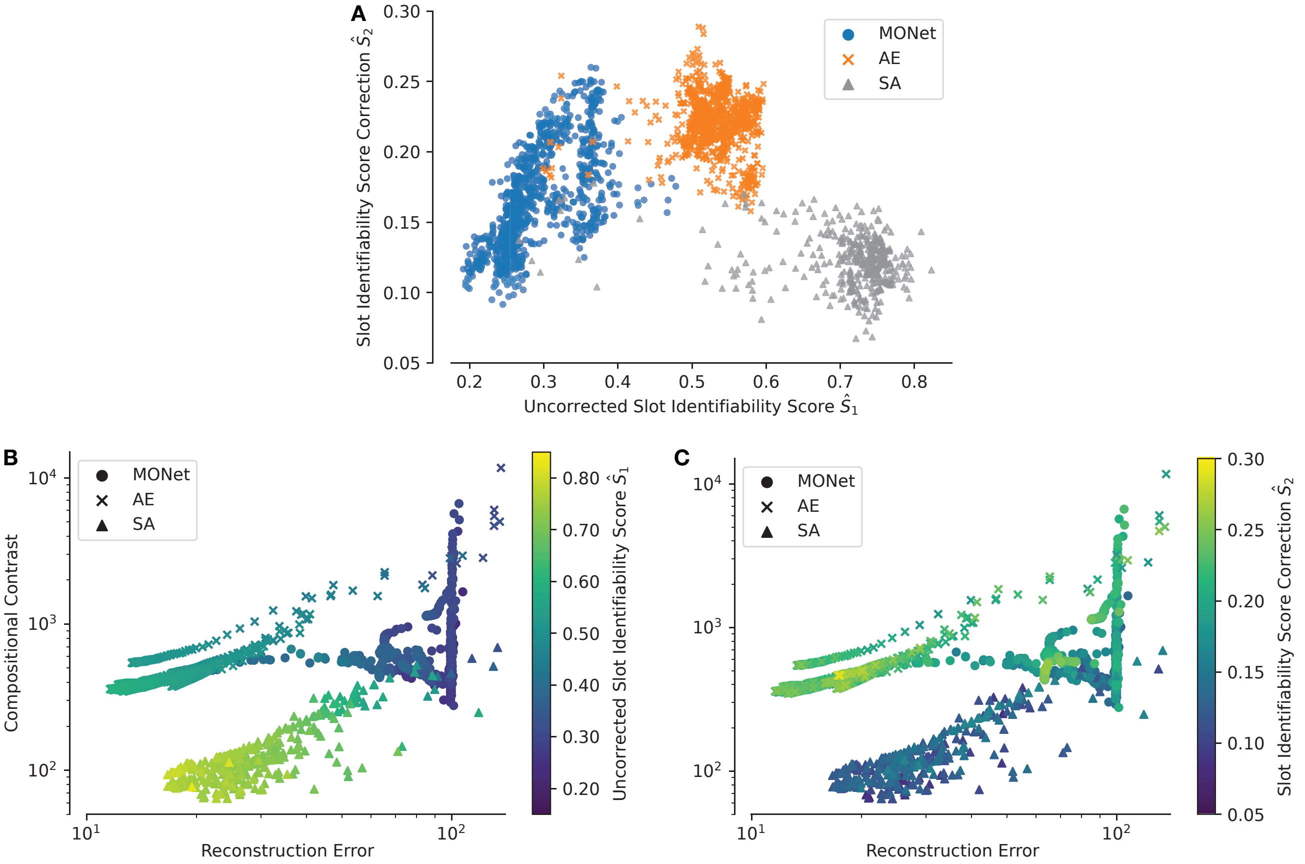

Figure 7. **Analysis of Information Leakage Between Slots from Models Trained in § 5.2. (A) Uncorrected Slot Identifiability Score ($\hat{S}_1$) vs. Correction ($\hat{S}_2$).** We train 3 existing object-centric architectures—MONet, Slot Attention (SA), and an additive auto-encoder (AE)—on image data and investigate whether inferred latent slots encode information from multiple objects when using an inferred latent dimension greater than the ground-truth. To test this, we look at the R^2 score for a model fit between each inferred slot and the second most predictive ground-truth slot for this slot. We refer to this score as the *slot identifiability score correction*, defined as $\hat{S}_2$ in Appx. B.4. We plot this score against the uncorrected slot identifiability score i.e. the most predictive ground-truth slot, defined as $\hat{S}_1$ in Appx. B.4. We can see that for all models, $\hat{S}_2$ is non-zero, even as $\hat{S}_1$ increases, suggesting that models are leveraging their additional latent capacity to encode information about multiple objects in the same latent slot. **(B) and (C) Influence of Reconstruction Error and Compositional Contrast on $\hat{S}_1$ and $\hat{S}_2$.** Here, we further visualize the slot identifiability score correction ($\hat{S}_1$) and the uncorrected score ($\hat{S}_2$) as a function of the reconstruction error and the compositional contrast in panels B and C, respectively. We can see in B that, similar to the SIS in Fig. 4, $\hat{S}_1$ tends to increase as reconstruction loss and compositional contrast decrease. We can additionally see in C that, while $\hat{S}_2$ decreases to some extent with C_{comp}, there is generally less of a correlation between $\hat{S}_2$ and these metrics. This suggests that the latent capacity must also be restricted to minimize $\hat{S}_2$.

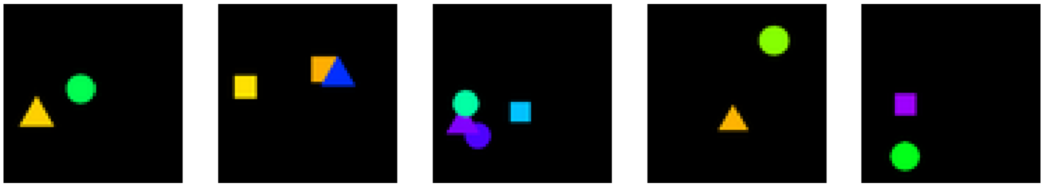

Figure 8. Samples from our multi-sprites dataset used in § 5.2. Objects are described by five latent factors: shape, color, size, and x/y position. Occlusions are present in the dataset, as shown in the samples above (see the second and third images from the left).