Unifying Causal and Object-centric Representation Learning

Avinash Kori [1] **Ben Glocker**[1] **Bernhard Schölkopf** [2] **Francesco Locatello** [3]

[1] Imperial College London
[2] Max Planck Institute for Intelligent Systems, Tübingen, Germany
[3] Institute of Science and Technology Austria
a.kori21@imperial.ac.uk

Abstract

The goal of Object-Centric Learning (OCL) is to enable machine learning systems to decompose complex scenes into discrete, interacting objects, supporting compositional generalization and human-like reasoning. However, existing OCL methods often fail to capture interactions from both attribute-level (*semantic*) and object-level (*spatial*) perspectives. While scene graph methods complement OCL by abstracting scenes as structured graphs, they typically rely on supervision. This position paper argues for a *probabilistic perspective* on Scene Graph Modelling (SGM), grounded in *causal abstraction* as a unifying view on causality, OCL, and scene graphs by considering object interactions as invariant mechanisms within *object-level* graphs, enabling us to generate causally consistent scene compositions. We substantiate our position with thorough conceptual discussion, rigorous definitions, conjectures, and examples, demonstrating how this perspective bridges the gap between unsupervised object discovery and explicit scene graph reasoning.

1 Introduction

To achieve human-level understanding, machine learning systems must be capable of representing the world in terms of distinct, interacting objects Bengio et al. (2013b). The concept of *objectness*—the ability to perceive the environment as a composition of discrete, identifiable entities—has long been recognized as critical for enabling *situation-aware AI systems* that exhibit *human-like reasoning capabilities* (Lake et al., 2017; Schölkopf & von Kügelgen, 2022). This fundamental ability to decompose complex scenes into objects and parts forms the basis of *compositional generalization*, a key feature that allows humans to efficiently learn from limited examples and adapt to new environments (Rock, 1973; Hinton, 1979).

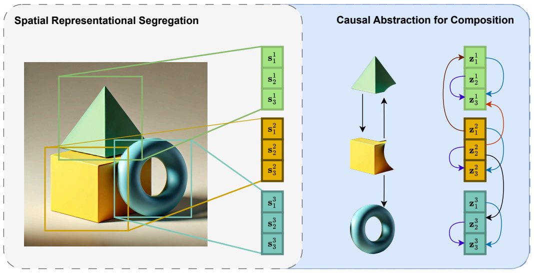
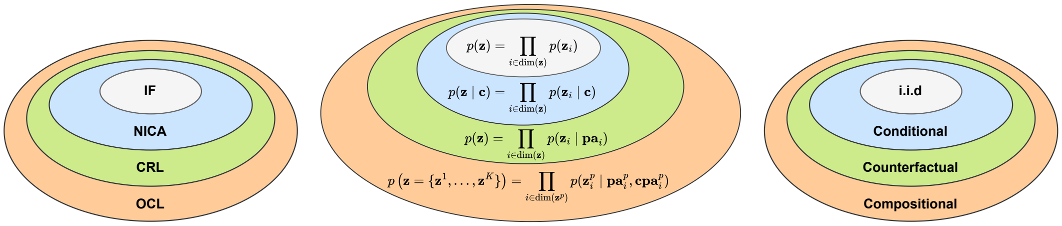


Figure 1: Figure illustrates the process of causal abstraction for a composition resulting in causally related variables in $\mathbf{z}$ using spatially segregated object-centric representations $\mathbf{s}$.

The fundamental requirement of compositional OCL is to ensure attribute and object level disentanglement. Attributes are common features across all objects involved in the scene generation for example features like *light-position, camera position, and so on.* Attribute disentanglement is a key objective in generative modelling such as *Variational Auto Encoders (VAEs)* (Kingma & Welling, 2013). The identifiability of these features has been a common area of research in identifiable representation learning, particularly with the *Non-linear Independent Component Analysis (NICA)* learning paradigm Khemakhem et al. (2020a;b); Willetts & Paige (2021). However, given these attributes are learned and not well defined, learning causal relationships among these attributes is another ill-defined problem and is the core of *Causal Representation Learning (CRL)* Schölkopf et al. (2021); Yao et al. (2024).


IF
NICA
CRL
OCL
$p(\mathbf{z}) = \prod_{i \in \dim(\mathbf{z})} p(\mathbf{z}_i)$
$p(\mathbf{z} \mid \mathbf{c}) = \prod_{i \in \dim(\mathbf{z})} p(\mathbf{z}_i \mid \mathbf{c})$
$p(\mathbf{z}) = \prod_{i \in \dim(\mathbf{z})} p(\mathbf{z}_i \mid \mathbf{pa}_i)$
$p\left(\mathbf{z} = \{\mathbf{z}^1, \ldots, \mathbf{z}^K\}\right) = \prod_{i \in \dim(\mathbf{z}^p)} p(\mathbf{z}_i^p \mid \mathbf{pa}_i^p, \mathbf{cpa}_i^p)$
i.i.d
Conditional
Counterfactual
Compositional

(a) Learning (b) Prior factorisation (c) Generation

Figure 2: Illustration of a hierarchical progression across learning paradigms, latent variable factorization, and sample generation capabilities grouped by color-coded regions. **IF** assumes fully independent latent variables, with a prior factorization $\prod_i p(\mathbf{z}_i)$, enabling *iid sample* generation. **NICA** introduces conditional independencies in the latent space, factorized as $\prod_i p(\mathbf{z}_i \mid \mathbf{c})$, enabling *conditional samples*. **CRL** generalizes NICA by uncovering cause-effect relationships via SCMs, with factorization $p(\mathbf{z}_i \mid \mathbf{pa}_i)$, allowing *conditional and counterfactual sample generation*. **OCL** extends CRL by structuring the latent space as a mixture of objects, capturing both feature-level and object-level causal interactions, with factorization $\prod_i p(\mathbf{z}_i^p \mid \mathbf{pa}_i^p, \mathbf{cpa}_i^p)$, enabling *conditional, counterfactual, and compositional* generation. In all cases, an appropriate mixing function is assumed.

While NICA and CRL have made substantial advances in learning complex latent structures with identifiability guarantees, their applications in OCL remain limited. Most of the works in OCL that perform attribute disentanglement, including Burgess et al. (2019); Chen et al. (2021); Kim & Mnih (2018); Kori et al. (2024); Greff et al. (2019; 2020), rely on *Independent Factorisation (IF)* of attributes similar to VAEs. Some recent advances in OCL have demonstrated the potential to imbue machine learning models with the notion of objectness by learning structured, spatially disentangled representations directly from raw observations (Locatello et al., 2020; Engelcke et al., 2021; Löwe et al., 2024; Kori et al., 2023). While in most of these methods, the main focus is on object discovery, the interaction between and across object features is, in our view, yet to be thoroughly explored. Here, identifiability is crucial because it dictates the conditions under which latent causes can be reliably recovered from observed data (Hyvärinen & Pajunen, 1999; Locatello et al., 2019; Hyvärinen & Oja, 2000). Recent research in OCL has established *theoretical guarantees* with assumptions on *mixing functions or latent distributions* that ensure object-centric representations are identifiable (Brady et al., 2023; Lachapelle et al., 2024; Kori et al., 2024). However, these methods rely on the IF setting, limiting the possibilities of learning relations among and across object features. To model these relations as illustrated in Fig. 1, we need concepts from CRL to capture scene compositions . The problem of composition is non-trivial as the interactions increase combinatorially with the increase in number of objects von Kügelgen et al. (2020); Wiedemer et al. (2023); Tangemann et al. (2021).

Here we argue that an ideal OCL method should aim to capture higher-order *(i) conditional, (ii) causal, and (iii) compositional relationships* between attributes and objects, where the system learns representations that are inherently compositional and capture both feature-level and object-level causal interactions. This view aligns with philosophical perspectives, such as Kant's view that causality is a necessary structure imposed by the mind to organise sensory experience (Kant, 1908), and Fodor's idea of compositional symbolic representations (Fodor, 1975). It also resonates with neuroscientific insights into hierarchical processing in the brain, where dynamic interactions between attributes and objects occur, and causal reasoning, akin to Bayesian inference mechanisms used in the prefrontal cortex (Tenenbaum et al., 2011). By integrating these principles, ideal OCL can bridge low- and high-level semantic understanding, reflecting the complex nature of human perception and the structured nature of real-world environments.

Scene Graph Modelling (SGM) is another line of work which deals with the abstraction of objects/-components in the scene and their complex relationships and provides rich semantic information of an image Gu et al. (2019), in the form of `<subject, object, predicate>`. Scene graphs are mainly considered in scenarios like visual question answering, image captioning, and action recognition Teney et al. (2017); Yao et al. (2018); Li et al. (2022). Scene graphs naturally align with the goals of OCL, representing a scene's latent structure as a graph where nodes correspond to objects, and edges represent semantic relationships, such as spatial, functional, or causal interactions (Krishna et al., 2017; Johnson et al., 2015; Gu et al., 2019); however, most of these methods heavily rely on supervision for identifying objects and their interactions. Recent methods Wei et al. (2024); Gao et al. (2024); Kim et al. (2024); Wang et al. (2024) use large language models (LLMs) o r large vision language models (VLMs) as an implicit scene graph which is used in conditioning the generation of samples. Most of these do not provide explicit control over attributes and object interactions. Gao et al. (2024) uses GPT4 to generate scene graphs from text prompts, providing object-level control with explicit scene graphs. von Kügelgen et al. (2020) provide a graphical abstraction for layer-wise compositional abstraction, Deng et al. (2021) provide a graphical abstraction of part-object relations, capturing the hierarchy involved in scene generation. In a similar spirit, we propose *causal abstraction for composition*, where we model object interactions as interactions between *object-level* graph. Unlike Deng et al. (2021); Yuan et al. (2024), which aims to learn to pose and view equivariant object representations, the *object-level* graph in this setting captures the *invariant mechanisms* across all the objects.

However, the challenge of identifiability of representations and dependencies still persists. Identifiability in modern machine learning models can be considered as a Bayes optimal model that guarantees the test loss to have a unique minimizer, essential to recover the true data-generating process. **In this paper, with probabilistic perspective on SGM, we position compositional OCL as a general learning paradigm that subsumes CRL, NICA, and IF models as special cases and should be an integral part of the research agenda.** With this, we consider the progression of learning paradigms as follows (illustrated in Fig. 2): (i) **IF**: assumes fully independent latent variables and enables the generation of *iid samples*. (ii) **NICA**: introduces *conditional independencies* in latent variables, allowing *conditional sample generation*. (iii) **CRL**: extends NICA by uncovering cause-effect relationships through Structural Causal Models (SCMs), resulting in *counterfactual reasoning*. (iv) **OCL**: builds upon CRL by structuring the latent space as a *mixture of objects*, capturing both feature-and object-level causal interactions, thus enabling the generation of *compositional samples*. In summary, our main contributions are:

1) Highlighting the modelling scenarios in both CRL and OCL fields, unifying them with *causal abstractions*;
2) Provide a probabilistic interpretation of scene graphs through *compositional graphs*, with multiple examples and scenarios demonstrating its implications;
3) Introduce the concept of *latent compositional models with mechanism invariance*, discuss its implications for learning compositional and causal graphs, along with possible modelling approaches and theoretical extensions to validate guarantees.

2 Compositional OCL: A Unifying View

Here, we introduce notations, outline the components of the data generation process and introduce the concept of causal abstraction from a compositional perspective.

2.1 Data Generating Process

Let $\mathcal{K} = \{1, \dots, K\}$ represent all object indices, with a subset $V \subseteq \mathcal{K}$ generating the observation $\mathbf{x} \in \mathcal{X}$. As shown in Fig. 3 (and in Eqn. 1), $\mathbf{x}^V = \{\mathbf{x}^v \in \mathcal{X}^v, \forall v \in V\}$ represents image partitions, where $\mathbf{x}$ is the composition of elements in $\mathbf{x}^V$, and $\mathcal{X}^v \subset \mathcal{X} \subset \mathbb{R}^{d_x}$. Observed variable y groups components ($\mathbf{x}^v$s) across $\mathcal{X}$. Latent variables $\mathbf{z} = \{\mathbf{z}^v \in \mathcal{Z}, \forall v \in V\}$, with $\mathcal{Z} \subset \mathbb{R}^{d_z}$, encode dependencies across features $\mathbf{z}^v_j$. Each partition $\mathbf{x}^v$ depends only on its corresponding $\mathbf{z}^v$ with relation $\mathbf{x}^v = f_d(\mathbf{z}^v)$. Independent counterparts $\mathbf{s} = \{\mathbf{s}^1, \dots, \mathbf{s}^{|V|}\}$ are defined as $\mathbf{s} = \{g_e(\mathbf{x}, y^1), \dots, g_e(\mathbf{x}, y^{|V|})\}$, where $\mathbf{s} \in \mathcal{S} \subseteq \mathcal{S}^1 \times \dots \times \mathcal{S}^K$, $\mathcal{S}^k \subseteq \mathbb{R}^{d_s}$. A volume-preserving transformation $\mathbf{z} = g_z(\mathbf{s})$

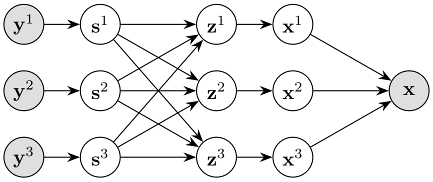


Figure 3: This graphical model illustrates the data generation process, starting with observed grouping variables $y^1, \dots, y^{|V|}$ ($|V| = 3$ shown for illustration), which are used to sample independent object representations $\mathbf{s}^1, \dots, \mathbf{s}^K$. These representations are then combined to form $\mathbf{z}$, capturing both attribute-level and object-level relationships. Finally, $\mathbf{z}$ is used to generate object-level partitions in the observation space, $\mathbf{x}^1, \dots, \mathbf{x}^{|V|}$. All partitions together compose the observed scene, $\mathbf{x}$.

captures relationships among object features. Finally, the compound feature extractor is defined as $\phi(\mathbf{s}) := (g_s \circ (f_c \circ f_d \circ g_z))(\mathbf{s})$, extracting features for a composed image rather than an observed image.

$$\begin{aligned} \mathbf{x} &= f_c([\mathbf{x}^1, \dots, \mathbf{x}^{|V|}]) = f_c(f_d([\mathbf{z}^1, \dots, \mathbf{z}^{|V|}])) = f_c([f_d(\mathbf{z}^1), \dots, f_d(\mathbf{z}^{|V|})]) \\ &= f_c([f_d(g_z(y^1, \mathbf{s})), \dots, f_d(g_z(y^{|V|}, \mathbf{s}))]) \end{aligned} \tag{1}$$

Illustration OCL: Grouping from a compositional perspective refers to spatial disentanglement of and categorisation of objects in V, the total number of distinct groups. Providing control over the generation of scenes with objects of interest given by y^v, this grouping can also be with respect to size, colour, or any of the observable properties which capture the uniqueness of objects in a scene.

Illustration CRL: Note that in the case of CRL methods, we do not need $f_c, \mathbf{x}$ as $\mathbf{x}^v$ are observed, y behaves as the observational grouping variable. This grouping of observed variables is common in many real-world applications, as detailed in Morioka & Hyvärinen (2023), and is used in many different forms, as in the form of multiple views in Yao et al. (2023), as augmentations in Von Kügelgen et al. (2021), and as interventional set in Brehmer et al. (2022).

2.2 Causal Abstraction for Composition

We introduce the notion of compositional graphs, building the formalisation of causal graphs Pearl (2009) and viewing them as probabilistic interpretations of scene graphs.

Remark 2.1. A causal graph $\mathcal{G}$ (definition A.1)factorises the distribution $p(\mathbf{z}) = \prod_{i=1}^{n} p(\mathbf{z}_i \mid \mathbf{pa}_i)$, where $\mathbf{pa}_i \subseteq \{\mathbf{z}_1, \dots, \mathbf{z}_n\} \setminus \mathbf{z}_i$ is a set of parents of $\mathbf{z}_i$ w.r.t. graph $\mathcal{G}$.

Example 1. *(Instantaneous Causal Effects.) As shown in Fig. 1, the observation is modelled as the output of an SCM, with inputs including light source position, object positions, shadows, sizes, and colours. The interactions between these features are depicted in Fig. 4a. Notably, object-specific properties such as pose and orientation are treated as confounders and are excluded from the causal graph, as we focus solely on features that represent invariant mechanisms across all objects in the dataset.*

Definition 2.2. (Compositional Graph: $\mathfrak{G}$) A compositional graph is a *meta-graph* capturing interactions between different causal graphs, which can be denoted with $\mathfrak{G} = (V^C, H) = (\bigcup_{k\in\mathcal{K}} V^k, \bigcup_{k\in\mathcal{K}} E^k \bigcup E^m)$, where:

- $V^C = \{\mathcal{G}^1, \dots, \mathcal{G}^K\}$ represents causal graph for individual *objects*, where $\mathcal{G}^k = (V^k, E^k), \forall k \in [K]$ is a causal graph as defined above.
- $E^m = \{e_{ij}^{pq} \mid e_{ij}^{pq}$ corresponds to *meta* edges capturing interactions between i^{th} variable of $\mathcal{G}^p$ and j^{th} variable of $\mathcal{G}^q\}$, *i.e.,* edge e_{ij}^{pq} captures the causal influence between $\mathbf{z}_j^q \rightarrow \mathbf{z}_i^p$

Remark 2.3. In a compositional setting, we have a set of latent representations $\mathbf{z} = \{\mathbf{z}^1, \dots, \mathbf{z}^{|V|}\}$, where $\mathbf{z}^k$ is sampled from separate environments given by $\mathbf{z}^k \sim p(\mathbf{z}^k \mid \mathbf{cpa}^k, y^k)$, where y^k categorical variables conditioning a variable to a particular environment. Similar to parents in causal setting we consider $\mathbf{cpa}_i^p \subseteq \bigcup_{q\in V, q\neq p} \{\mathbf{z}_j^q \,|\, \forall j \in \dim(\mathbf{z}^q)\}$, describing compositional parents, which is the set of all the variables across environments causing the feature $\mathbf{z}_i^p$, they can be treated as a causal parents but in a different environment. Invariant edges across environments can be considered causal graphs. Additionally, it is essential to note that compositional graphs are scene-dependent, capturing the interaction between invariant causal graphs. The resulting latent distribution can be factorised as:

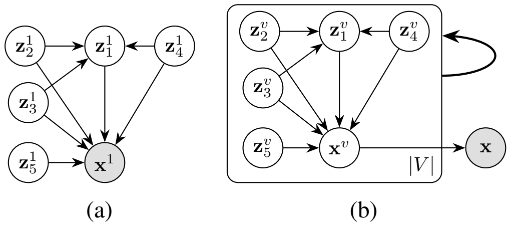
$$p(\mathbf{z}) \propto \prod_{p,q\in\mathcal{K}, p\neq q} p(\mathbf{z}^p \mid \mathbf{z}^q, y^p) = \prod_{p\in\mathcal{K}} \prod_{i} p(\mathbf{z}_i^p \mid \mathbf{cpa}_i^p \cup \mathbf{pa}_i^p, y^p) \tag{2}$$

Example 2. *(Causal Abstraction for Composition.) Building upon Example 1, we extend the discussion to object relations within a scene, using the same instantiation of causal graphs for Torus, Cube, and Pyramid, as shown in Fig. 1. This instantiation remains valid since the underlying mechanisms—such as shadow formation due to object size and light source—are consistent and invariant across different objects. The edges in the graph (ref. Fig. 5) represent compositional* *behaviours; for example, the Cube's position causally influences the Pyramid's position when the Pyramid is placed above it. Similarly, the Torus' position is causally linked to the Cube's position.*

The compositional graph provides a *probabilistic* interpretation of scene graphs, integrating both spatial and semantic relationships. Its edges capture attribute-level (semantic) reasoning, e.g., `<object-size, shadow-length, causes>`, or object-level (spatial) reasoning, e.g., `<torus, cube, leans on>`.

Lemma 2.4. *(Acyclicity of composition) The combined compositional graph remains acyclic when:*

- *L.1: Circular consistency: No circular interactions are introduced between any graph pairs $(\mathcal{G}^p, \mathcal{G}^q)$,* i.e., *there are no loops in E^m.* Example: *If there is an edge $(\mathbf{z}_i^p \rightarrow \mathbf{z}_j^q) \in E^m$, then $(\mathbf{z}_j^q \rightarrow \mathbf{z}_i^p) \notin E^m$.*
- *L.2: Causal consistency: For any two nodes $\mathbf{z}_i^p, \mathbf{z}_j^q$, if there is an edge $(\mathbf{z}_i^p \rightarrow \mathbf{z}_j^p) \in E^p$, then $(\mathbf{z}_j^q \rightarrow \mathbf{z}_i^p) \notin E^m$.*



(a) (b)

Figure 4: **(a) Instantaneous causal graph** representing the causal relationships among variables for a single object, where $\mathbf{z}_1^1$ to $\mathbf{z}_5^1$ correspond to *object-shadow, object-size, light-position, object-position, and object-color*, respectively. **(b) Compositional graph** depicted using plate notation, where the plate labelled $|V|$ represents multiple instances of the invariant causal graph for different objects. Interactions across objects are illustrated with elliptical arrows spanning the plate, and the observed data $\mathbf{x}$ is modelled as the combined result of all $|V|$ objects and their interactions.

Interpretation: *The first constraint is straight forward, ensuring E^m does not contain any direct cycles. The second constraint preserves consistency in the causal direction between features across groups. Together, these constraints ensure $\mathfrak{G}$ retains its acyclic nature. In the context of physical systems, in Fig. 5, acyclicity in the composition of interacting objects is inherently maintained. This means that the cause-and-effect relationships between the objects — such as the influence of the light source on the shadows cast by the cube and torus or the way the position and size of one object can impact the other — follow a directed, non-circular flow. However, some complex biological systems do break these acyclicity properties Sachs et al. (2005), which we do not consider in this work.*

To capture the notion of latent distribution respecting compositional nature, we consider parametrised Markov random fields (MRFs), capturing pairwise interactions between the elements of $\mathbf{s}_j^p$ and $\mathbf{z}_i^q$. Based on the parametrization of MRFs, we consider two different potential functions, g_z and $\bar{g_z}$, where g_z is a pairwise potential function, which captures causal relation across objects modelling compositions. In contrast, $\bar{g_z}$ captures the instantaneous causal interactions for a given component. As illustrated in Morioka & Hyvärinen (2023), the parametric interpretation of $g_z, \bar{g_z}$ is general enough to capture many different classes of SEMs, resulting in a latent distribution (Eqn. 3).

$$p(\mathbf{z}) \propto \prod_{p,q \in \mathcal{K}} \exp\left(\bar{g_z}(\mathbf{z}^p, \mathbf{s}^p; y)\right) \cdot \left[\prod_{(i,j) \in |\mathcal{Z}^p| \times |\mathcal{Z}^q|} \exp\left(\lambda_{ij}^{pq} g_z(\mathbf{z}_i^p, \mathbf{s}_j^q)\right) \right] \tag{3}$$

Terms $g_z(\mathbf{z}_i^p, \mathbf{s}_j^q)$ captures the sample-specific functional relations like the influence of cube size on the shadow of a torus in Example 2, while λ_{ij}^{pq} captures the existence of these compositional edges. Here, the component level interaction $p(\mathbf{z}^p \mid \mathbf{z}^q)$ is modelled with an axillary variable $\mathbf{s}$ which are considered to satisfy ICA assumption resulting in $p(\mathbf{s}) = \prod p(\mathbf{s}^v)$ (ref. Fig. 3). We do this by estimating $\mathbf{z}^p$ with all independent base representations $\{\mathbf{s}^1, \ldots, \mathbf{s}^{|V|}\}$.

Remark 2.5. Note that based on the joint distribution, the potential function $\bar{g_z}$ depends on a component level grouping reflected by $y, \forall k \in [K]$, which in Morioka & Hyvärinen (2023) is assumed to be known in the form of observational groups, while in OCL, this further requires weak supervision (classification labels for objects in the scene), breaking equivariance property, which is implicit in OCL Locatello et al. (2020); Wang et al. (2023); Emami et al. (2022); Kori et al. (2023; 2024). Additionally, the potential function g_z, needs a further set of assumptions to factories Eqn. 3 to reflect the directionality of cause and effect.

Graph Extraction: Considering the invariance of causal graphs across groups, we get the same causal graph $\mathcal{G}$ for all $k \in \mathcal{K}$. It is important to note that causal graph $\mathcal{G}$ is contained in a compositional graph $\mathfrak{G}$ and reflects the inter-group weighted adjacency matrix. Let $\boldsymbol{C} = \{\boldsymbol{C}^{pq}\}_{ij}$ correspond to adjacency matrices of $\mathfrak{G}$ for every node pair $(i, j) \in |\mathcal{Z}^p| \times |\mathcal{Z}^q|$ for a given scene, as detailed in examples 1 & 2. Similar to Reizinger et al. (2023), we link the notion of Jacobian $\boldsymbol{J}_{g_z^{-1}}$ of function g_z to compositional graph. Note we consider $\boldsymbol{C}$ to be concatenated matrix in $\mathbb{R}^{|V||d_z| \times |V||d_z|}$ resulting the combined Jacobian as illustrated in Fig. 6. Finally, we consider two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ to be structurally equivalent ($\boldsymbol{A} \sim_{\text{DAG}} \boldsymbol{B}$) if both matrices follow the same sparsity structure.

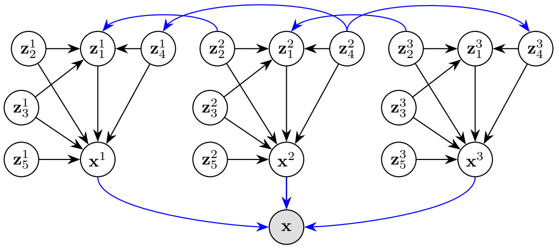
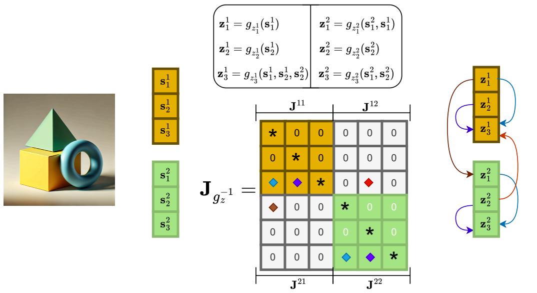


Figure 5: Example of particular instantiation of Fig. 4b (with $V = \{1, 2, 3\}$, resulting in $\mathbf{x}^1, \mathbf{x}^2$ and $\mathbf{x}^3$ corresponding to *torus, cube,* and *pyramid*, respectively), illustrating the scene in Fig. 1. Here, $\mathbf{z}_1^v$ to $\mathbf{z}_5^v$ correspond to *object-shadow, object-size, light-position, object-position, and object-color, respectively*. Blue arrows in the graph correspond to compositional edges (E^m).

Invariant Mechanisms: As detailed in examples 1 & 2, given the mechanisms are consistent across objects in a scene, we need a way to enforce this behaviour when our SCM is modelled in a parametric sense. We propose the following mechanism consistency regularisations (ref. definition 2.6), based on a Jacobian of the considered SCM.

Definition 2.6. (Mechanism consistency) The causal relation among features for a particular component is given by $\mathcal{G}^p$ which is captured by $d_s \times d_s$ Jacobian matrix given $\boldsymbol{J}^{pp}_{g_z^{-1}}$, we define mechanism consistency as: $\mathcal{L}_{\text{MC}} = \mathbb{E}_{\mathbf{z} \in \mathcal{Z}} \left[\sum_{p=1}^{|V|} \| \boldsymbol{J}^{pp}_{g_z^{-1}}(\mathbf{z}) - \boldsymbol{J}^{00}_{g_z^{-1}}(\mathbf{z}) \|_f \right]$.

Interpretation: *The regularisation basically encourages all the block-diagonals with dimension d_z of a Jacobian $\boldsymbol{J}_{g_z^{-1}}$ to follow the same sparsity structure. This basically forces the parametric SCM to use the same set of parameters for all objects.*


$\mathbf{z}_1^1 = g_{z_1^1}(\mathbf{s}_1^1)$
$\mathbf{z}_2^1 = g_{z_2^1}(\mathbf{s}_2^1)$
$\mathbf{z}_3^1 = g_{z_3^1}(\mathbf{s}_1^1, \mathbf{s}_2^1, \mathbf{s}_2^2)$
$\mathbf{z}_1^2 = g_{z_1^2}(\mathbf{s}_1^2, \mathbf{s}_1^1)$
$\mathbf{z}_2^2 = g_{z_2^2}(\mathbf{s}_2^2)$
$\mathbf{z}_3^2 = g_{z_3^2}(\mathbf{s}_1^2, \mathbf{s}_2^2)$
$\mathbf{s}_1^1$
$\mathbf{s}_2^1$
$\mathbf{s}_3^1$
$\mathbf{s}_1^2$
$\mathbf{s}_2^2$
$\mathbf{s}_3^2$
$\mathbf{J}^{11}$
$\mathbf{J}^{12}$
$\mathbf{J}^{21}$
$\mathbf{J}^{22}$
$\mathbf{J}_{g_z^{-1}} =$
$\mathbf{z}_1^1$
$\mathbf{z}_2^1$
$\mathbf{z}_3^1$
$\mathbf{z}_1^2$
$\mathbf{z}_2^2$
$\mathbf{z}_3^2$

*	0	0	0	0	0
0	*	0	0	0	0
◆	◆	*	0	◆	0
◆	0	0	*	0	0
0	0	0	0	*	0
0	0	0	◆	◆	*

Figure 6: The figure demonstrates the Jacobian of g_z, and the resulting compositional graph between attributes $\mathbf{z}$ expressed as a function of independent variables $\mathbf{s}$. Block diagonals reflect the causal relations between attributes, while off-diagonal entries show compositional relations. Coloured diamonds in a Jacobian reflect entries corresponding to their edges in a compositional graph.

DAG regularisation: Lemma 2.4 establishes conditions for compositional interactions to ensure the resulting joint composition graph is a DAG. To enforce DAG structure in SEM g_z, we apply a continuous DAG penalty over the Jacobian $\boldsymbol{J}$ (we later show, $\boldsymbol{J} \sim_{\text{DAG}} \boldsymbol{I} - \boldsymbol{C}$). We leverage the findings of Nazaret et al. (2023), which demonstrate that minimizing the largest eigenvalue $\lambda(\boldsymbol{J})$ ensures DAG soundness. Augmented Lagrangian scheme with hyperparameters α and ρ, which are progressively increased during training, $\mathcal{L}_{\text{DAG}} = \|\boldsymbol{J}\|_f^2 + \rho h(\boldsymbol{J})^2 + \alpha h(\boldsymbol{J})$ where $h(\boldsymbol{J}) = |\lambda(\boldsymbol{J})|$.

Compositional Consistency: Similar to Wiedemer et al. (2023), to ensure compositional generalisability, we use compositional consistency, which basically points towards a fixed point of ϕ, given by $\mathbf{s} = \phi(\mathbf{s})$. Which is achieved by maximising the posterior log probabilities for all possible random compositions.

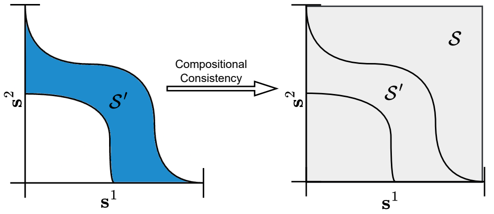


Figure 7: The figure demonstrates the generalisation of latent space as a result of optimising compositional consistency.

Conjecture 2.7. (Compositional consistency) Given the full support over $\mathcal{S}$ by observational data $\mathcal{X}$, maximising the posterior $q(\phi(\mathbf{s})) \forall \mathbf{s} \in \mathcal{S}$, while preserving high likelihood over $p(\mathbf{x})$ results in compositional generalisation of considered LCM. Which means for any random composition $\mathbf{s} \in \mathcal{S}$, $(f_c \circ f_d \circ g_z)(\mathbf{s})$ is optimised to be

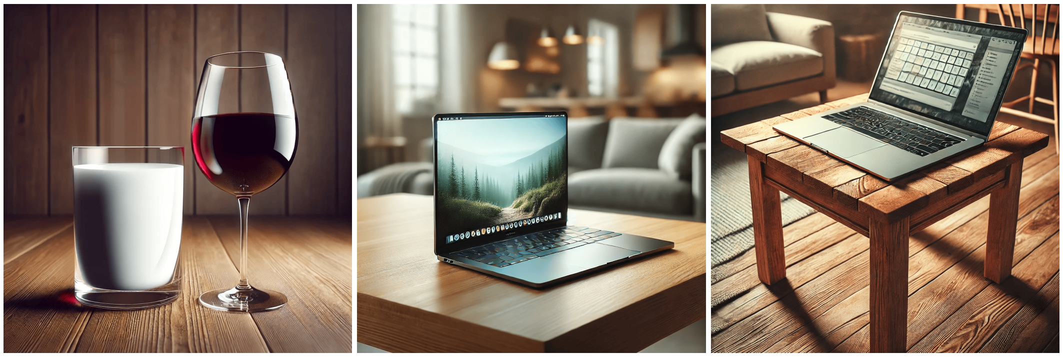


Figure 8: **(a)** “Two glasses, one with milk and the other with wine, both fully filled.", **(b)** “A laptop on top of a table.", and **(c)** “A table placed on top of a laptop." These examples illustrate generated samples from GPT-4o with highlighted predicates, demonstrating issues with VLMs and the need for our unified perspective.

in-distribution of $p(\mathbf{x})$, $\mathcal{L}_{\text{CC}} = -\mathbb{E}_{\mathbf{s} \in \mathcal{S}} [\log q(\phi(\mathbf{s}))]$. **Interpretation:** *As illustrated in Fig. 7, $\mathcal{S}'$ is the inferred latent space of observational data capturing the possible interactions of objects $\mathbf{s}^1$ and $\mathbf{s}^2$ as reflected in observed data, while $\mathcal{S}$ demonstrates the space captured by all possible random compositions while preserving data* fidelity.

Training objective: The model would be trained via likelihood maximisation with additional regularisations, encouraging the representations to have consistent mechanisms that follow DAG structure while ensuring generalisation on compositions. The resulting objective, with hyperparameters β, γ and δ is given by $\mathcal{L}_{\text{total}} := -\mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x},\mathbf{y})} \log(p(\mathbf{x} \mid \mathbf{z}) + \beta \mathcal{L}_{\text{CC}} + \gamma \mathcal{L}_{\text{MC}} + \delta \mathcal{L}_{\text{DAG}}$

3 Why does it matter?

As VLMs and foundation models continue to evolve, the need for structured, interpretable, and compositional representations becomes increasingly evident in achieving human-aligned, safe, fair, robust, and generalisable representations. While these models excel at capturing broad patterns from large datasets, their ability to generalise to novel, compositional, and causal reasoning tasks remains limited. Fig. 8 highlights such challenges. In Fig. 8(a), the mechanism of filling a glass should be consistent, independent of the specific instance (glass of milk vs glass of wine), yet the model fails to learn this invariant mechanism and instead learns instance-specific mechanisms. This demonstrates the necessity of the proposed perspective on invariant mechanisms that maintain causal consistency in a compositional setting.

Figures 8(b) and (c) further illustrate compositional inconsistencies. Both scenarios—*a laptop on a table* and *a table on a laptop*—are physically plausible and causally consistent. However, the model hallucinates in the latter case, failing to generalise in that scenario, highlighting the lack of compositional generalisability. This is crucial for ensuring out-of-distribution (OOD) generalisation capabilities, and accurate modelling of rare yet realistic events Wiedemer et al. (2023). Building on these observations, we outline the theoretical and practical implications of a probabilistic SGM perspective that unifies both CRL and OCL.

Transferability: The proposed causal abstraction provides a direction for developing *System-2* agents by bridging perception and abstract reasoning. This perspective suggests that by grounding dataset-level objects in scenes (Lake et al., 2017; Schölkopf et al., 2021; Kori et al., 2023), models could potentially generalise across unseen scenes through invariant mechanisms. Such grounded representations facilitates task transferability and human-like reasoning (Tenenbaum et al., 2011), providing new avenues for integrating perception-driven neural networks with symbolic reasoning for planning and decision-making (Bengio et al., 2013a; Mao et al., 2019; Proietti & Toni, 2023).

Robustness: This perspective aligns with applying causal inference principles, such as intervention and counterfactual reasoning, to scene modelling, as partially explored by Mansouri et al. (2022; 2023).

Models leveraging invariant mechanisms may capture object features even under partial observability, reducing the risk of hallucinations. Exploring this robustness is particularly promising for critical applications, such as autonomous driving (Schölkopf et al., 2021) and medical diagnosis (Castro et al., 2020).

Generalisation: Enforcing mechanism consistency across image partitions could lead to representations that are both causal and invariant across objects, improving OOD generalisation (Schölkopf et al., 2021). Additionally, compositional consistency enables model to generalise to novel environments in a causally consistent manner.

Counterfactual Reasoning and Controllability: Extending counterfactual reasoning in OCL through causal abstraction holds significant potential. Existing generative models (Goodfellow et al., 2014; Mirza & Osindero, 2014; Karras, 2019; Kingma & Welling, 2013; Song & Ermon, 2019) lack fine-grained control, often resulting in hallucinations or poor OOD generalisation (Reizinger & Krishnan, 2024). By formalising compositional graphs in the latent space, this perspective points towards new methods for controllable and interpretable generation, which extends to applications like image editing and embodied AI systems.

Despite these promising implications, several challenges remain to provide promising future direction:

Identifiability in Complex Scenes. Ensuring the identifiability of object-centric causal variables in complex, the current formalism provides an informal proof under explicitly made assumptions; rigorously establishing identifiability for general setting will be crucial for many use cases.

Mechanism Invariances. The proposed approach enforces invariances across image partitions, but extending these invariances to handle temporal dependencies is an interesting direction to model the physical properties of the system. For instance, considering the scene in Fig. 1, if a torus is observed *pushing the cube* within a temporal sequence, this interaction reveals important properties such as the mass of objects and provides insight into friction within the environment. Such temporal extensions can improve the model's ability to infer *dynamical causal relationships*, making it more suitable for applications in physical reasoning, robotic manipulation, and scene understanding.

Systematic Evaluation. Given all the existing OCL datasets do not provide access to compositional or causal graphs that are used in generating particular scenes, it would be an important direction to define relevant metrics to evaluate these properties additionally, creating a detailed environment to generate synthetic/realistic scenes along with compositional structure would be very beneficial to the field in general. Some of the existing environments Authors (2024); Greff et al. (2022) can be modified for this purpose.

4 Conclusion

In this position paper, we proposed a unifying perspective that integrates OCL and CRL through a probabilistic view SGM with *causal abstraction*. By conceptualizing object interactions as invariant mechanisms embedded within object-level graphs, we outlined a structured approach that captures both semantic and spatial dependencies, extending existing paradigms such as IF, NICA, and CRL.

Here, we provide a perspective substantiated through rigorous conceptual arguments and demonstrative scenarios compelling reasoning for how this approach could mitigate key challenges in *foundation models*, including *hallucinations, lack of structured reasoning, and poor generalization*. Specifically, by embedding causal reasoning within object-centricity, our perspective suggests that future AI systems can: 1) Enhance controllability and reliability in structured scene generation. 2) Improve generalization by aligning learned representations with real-world structured environments. 3) Enable compositional/counterfactual reasoning, allowing robust inference over novel interactions. While our claims remain theoretical, we argue that formalizing composition as a causal abstraction presents a promising direction for developing AI systems that are interpretable, compositional, and robust. This perspective has broad implications for autonomous systems, robotics, and medical AI, where structured and generalizable scene understanding is critical, moving beyond pattern recognition to structured, human-aligned intelligence.

References

Genesis Authors. Genesis: A universal and generative physics engine for robotics and beyond, December 2024. URL https://github.com/Genesis-Embodied-AI/Genesis.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013a.

Yoshua Bengio, Li Yao, Guillaume Alain, and Pascal Vincent. Generalized denoising auto-encoders as generative models. *Advances in neural information processing systems*, 26, 2013b.

Jack Brady, Roland S Zimmermann, Yash Sharma, Bernhard Schölkopf, Julius von Kügelgen, and Wieland Brendel. Provably learning object-centric representations. *arXiv preprint arXiv:2305.14229*, 2023.

Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco S Cohen. Weakly supervised causal representation learning. *Advances in Neural Information Processing Systems*, 35:38319–38331, 2022.

Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. Monet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019.

Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):3673, 2020.

Chang Chen, Fei Deng, and Sungjin Ahn. Roots: Object-centric representation and rendering of 3d scenes. *Journal of Machine Learning Research*, 22(259):1–36, 2021.

Fei Deng, Zhuo Zhi, Donghun Lee, and Sungjin Ahn. Generative scene graph networks. In *International Conference on Learning Representations*, 2021.

Patrick Emami, Pan He, Sanjay Ranka, and Anand Rangarajan. Slot order matters for compositional scene understanding. *arXiv preprint arXiv:2206.01370*, 2022.

Martin Engelcke, Oiwi Parker Jones, and Ingmar Posner. Genesis-v2: Inferring unordered object representations without iterative refinement. *Advances in Neural Information Processing Systems*, 34:8085–8094, 2021.

Jerry Fodor. The language of thought, 1975.

Gege Gao, Weiyang Liu, Anpei Chen, Andreas Geiger, and Bernhard Schölkopf. Graphdreamer: Compositional 3d scene synthesis from scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21295–21304, 2024.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Klaus Greff, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. Neural expectation maximization. *Advances in Neural Information Processing Systems*, 30, 2017.

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

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

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3749–3761, 2022.

Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, and Mingyang Ling. Scene graph generation with external knowledge and image reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1969–1978, 2019.

Geoffrey Hinton. Some demonstrations of the effects of structural descriptions in mental imagery. *Cognitive Science*, 3(3):231–250, 1979.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

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

Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3668–3678, 2015.

Immanuel Kant. Critique of pure reason. 1781. *Modern Classical Philosophers, Cambridge, MA: Houghton Mifflin*, pp. 370–456, 1908.

Tero Karras. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2019.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pp. 2207–2217. PMLR, 2020a.

Ilyes Khemakhem, Ricardo Monti, Diederik Kingma, and Aapo Hyvarinen. Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ica. *Advances in Neural Information Processing Systems*, 33:12768–12778, 2020b.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2018.

Kibum Kim, Kanghoon Yoon, Jaehyeong Jeon, Yeonjun In, Jinyoung Moon, Donghyun Kim, and Chanyoung Park. Llm4sgg: Large language models for weakly supervised scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28306–28316, 2024.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Identifiability of deep generative models without auxiliary information. *Advances in Neural Information Processing Systems*, 35:15687–15701, 2022.

Avinash Kori, Francesco Locatello, Fabio De Sousa Ribeiro, Francesca Toni, and Ben Glocker. Grounded object centric learning. *arXiv preprint arXiv:2307.09437*, 2023.

Avinash Kori, Francesco Locatello, Francesca Toni, Ben Glocker, and Fabio De Sousa Ribeiro. Identifiable object centric representations via probabilistic slot attention. *arXiv preprint arXiv:2307.09437*, 2024.

Markus Krimmel, Jan Achterhold, and Joerg Stueckler. Attention normalization impacts cardinality generalization in slot attention. *arXiv preprint arXiv:2407.04170*, 2024.

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

Sébastien Lachapelle, Divyat Mahajan, Ioannis Mitliagkas, and Simon Lacoste-Julien. Additive decoders for latent variables identification and cartesian-product extrapolation. *arXiv preprint arXiv:2307.02598*, 2023.

Sébastien Lachapelle, Divyat Mahajan, Ioannis Mitliagkas, and Simon Lacoste-Julien. Additive decoders for latent variables identification and cartesian-product extrapolation. *Advances in Neural Information Processing Systems*, 36, 2024.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017.

Rongjie Li, Songyang Zhang, and Xuming He. Sgtr: End-to-end scene graph generation with transformer. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19486–19496, 2022.

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

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in Neural Information Processing Systems*, 33:11525–11538, 2020.

Sindy Löwe, Phillip Lippe, Francesco Locatello, and Max Welling. Rotating features for object discovery. *Advances in Neural Information Processing Systems*, 36, 2024.

Amin Mansouri, Jason Hartford, Kartik Ahuja, and Yoshua Bengio. Object-centric causal representation learning. In *NeurIPS 2022 Workshop on Symmetry and Geometry in Neural Representations*, 2022.

Amin Mansouri, Jason Hartford, Yan Zhang, and Yoshua Bengio. Object-centric architectures enable efficient causal representation learning. *arXiv preprint arXiv:2310.19054*, 2023.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *arXiv preprint arXiv:1904.12584*, 2019.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Hiroshi Morioka and Aapo Hyvärinen. Causal representation learning made identifiable by grouping of observational variables. *arXiv preprint arXiv:2310.15709*, 2023.

Achille Nazaret, Justin Hong, Elham Azizi, and David Blei. Stable differentiable causal discovery. *arXiv preprint arXiv:2311.10263*, 2023.

Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009. doi: 10.1017/CBO9780511803161.

Maurizio Proietti and Francesca Toni. A roadmap for neuro-argumentative learning. In *NeSy*, 2023.

Patrik Reizinger and Rahul Krishnan. Exploring a bayesian view on compositional and counterfactual generalization. In *NeurIPS 2024 Workshop on Compositional Learning: Perspectives, Methods, and Paths Forward*, 2024.

Patrik Reizinger, Yash Sharma, Matthias Bethge, Bernhard Schölkopf, Ferenc Huszár, and Wieland Brendel. Multivariable causal discovery with general nonlinear relationships. In *UAI 2022 Workshop on Causal Representation Learning*, 2023.

Irvin Rock. Orientation and form. *(No Title)*, 1973.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

Bernhard Schölkopf and Julius von Kügelgen. From statistical to causal learning. *Proceedings of the International Congress of Mathematicians*, 2022.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109, 2021.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Matthias Tangemann, Steffen Schneider, Julius Von Kügelgen, Francesco Locatello, Peter Gehler, Thomas Brox, Matthias Kümmerer, Matthias Bethge, and Bernhard Schölkopf. Unsupervised object learning via common fate. *arXiv preprint arXiv:2110.06562*, 2021.

Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285, 2011.

Damien Teney, Lingqiao Liu, and Anton van Den Hengel. Graph-structured representations for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2017.

Julius von Kügelgen, Ivan Ustyuzhaninov, Peter Gehler, Matthias Bethge, and Bernhard Schölkopf. Towards causal generative scene models via competition of experts. *arXiv preprint arXiv:2004.12906*, 2020.

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

Xu Wang, Yifan Li, Qiudan Zhang, Wenhui Wu, Mark Junjie Li, Lin Ma, and Jianmin Jiang. Weakly-supervised 3d scene graph generation via visual-linguistic assisted pseudo-labeling. *IEEE Transactions on Multimedia*, 2024.

Yanbo Wang, Letao Liu, and Justin Dauwels. Slot-vae: Object-centric scene generation with slot attention. *arXiv preprint arXiv:2306.06997*, 2023.

Yao Wei, Martin Renqiang Min, George Vosselman, Li Erran Li, and Michael Ying Yang. Compositional 3d scene synthesis with scene graph guided layout-shape generation. *arXiv preprint arXiv:2403.12848*, 2024.

Thaddäus Wiedemer, Jack Brady, Alexander Panfilov, Attila Juhos, Matthias Bethge, and Wieland Brendel. Provable compositional generalization for object-centric learning. *arXiv preprint arXiv:2310.05327*, 2023.

Matthew Willetts and Brooks Paige. I don't need u: Identifiable non-linear ica without side information. *arXiv preprint arXiv:2106.05238*, 2021.

Dingling Yao, Danru Xu, Sébastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. *arXiv preprint arXiv:2311.04056*, 2023.

Dingling Yao, Dario Rancati, Riccardo Cadei, Marco Fumero, and Francesco Locatello. Unifying causal representation learning with the invariance principle. *arXiv preprint arXiv:2409.02772*, 2024.

Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 684–699, 2018.

Jinyang Yuan, Tonglin Chen, Zhimeng Shen, Bin Li, and Xiangyang Xue. Unsupervised object-centric learning from multiple unspecified viewpoints. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

A Additional definitions and assumptions

Definition A.1. (Causal Graph: $\mathcal{G}$) Let the causal graph $\mathcal{G} = (V, E)$ be a directed acyclic graph (DAG) that represents causal relationships:

- $V = \mathbf{z} \in \mathcal{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\} \in \{\mathcal{Z}_1, \ldots, \mathcal{Z}_n\}$ are the nodes, representing causal variables.
- $E \subseteq V \times V$ are the directed edges encoding pairwise causal dependencies between variables, i.e., $\mathbf{z}_i \rightarrow \mathbf{z}_j$ represents that $\mathbf{z}_i$ causally influences $\mathbf{z}_j$.

Assumption A.2. ($\beta-$disentanglement, Lachapelle et al. (2023)) Let $\mathbf{s} = \{\mathbf{s}_B, \forall\, B \in \mathcal{B}\}$ be a set of features wrt partition set $\mathcal{B}$. The learned mixing function f is said to be $\mathcal{B}$ disentangled wrt true decoder $\tilde{f}$ if there exists a permutation respecting diffeomorphism $v_B = \tilde{f}^{-1} \circ f \,\forall\, B \in \mathcal{B}$ which for a given feature $\mathbf{s}$ can be expressed as $v_B(\mathbf{s}) = v_B(\mathbf{s}_B)$.

Assumption A.3 (*Weak* Injectivity)**.** Let $f : \mathcal{Z} \rightarrow \mathcal{X}$ be a mapping between latent space and image space, where $\dim(\mathcal{Z}) \leq \dim(\mathcal{X})$. The mapping f_d is weakly injective if there exists $\mathbf{x}_0 \in \mathcal{X}$ and $\delta > 0$ such that $|f^{-1}(\{\mathbf{x}\})| = 1, \forall \mathbf{x} \in B(\mathbf{x}_0, \delta) \cap f(\mathcal{Z})$, and $\{\mathbf{x} \in \mathcal{X} : |f^{-1}(\{\mathbf{x}\})| = \infty\} \subseteq f(\mathcal{Z})$ has measure zero w.r.t. to the Lebesgue measure on $f(\mathcal{Z})$ (cf. Kivva et al. (2022)).

Remark A.4. In words, Assumption A.3 says that a mapping f_d is weakly injective if: (i) in a small neighbourhood around a specific point $\mathbf{x}_0 \in \mathcal{X}$ the mapping is injective – meaning each point in this neighbourhood maps to exactly one point in the latent space $\mathcal{Z}$; and (ii) while f_d may not be globally injective, the set of points in $\mathcal{X}$ that map back to an infinite number of points in $\mathcal{Z}$ (non-injective points) is almost non-existent in terms of the Lebesgue measure on the image of $\mathcal{Z}$ under f_d.

Definition A.5. (Additive composition function) A function $f_c : \mathcal{M} \times \mathcal{X} \rightarrow \mathcal{X}$ is said to be additive if, for $\mathbf{x} \sim \mathcal{X}, \mathbf{m} \sim \mathcal{M}$ with $\mathbf{x} = \{\mathbf{x}^1, \ldots, \mathbf{x}^{|V|}\}$ and $\mathbf{m} = \{\mathbf{m}^1, \ldots, \mathbf{m}^{|V|}\}$, it can be expressed as $f_c(\mathbf{m}, \mathbf{x}) = \sum_{v=0}^{|V|} \mathbf{m}^v \odot \mathbf{x}^v = \mathbf{x}$. When $\mathcal{M}$ is constrained, such that $\sum_{v=0}^{|V|} \mathbf{m}^v = 1, \mathbf{m}^v > 0$, the following properties are inherently satisfied:

D.1. when $\mathbf{m}^v \odot \mathbf{x}^v$ satisfy $\beta-$disentanglement assumption A.2, given partition masks, composition function can be approximately inverted with $\mathbf{x}^v = \mathbf{m}^v \odot \mathbf{x}$,

D.2. ordering invariance of elements in $(\mathbf{x}, \mathbf{m})$ together.

Definition A.6. (Latent compositional models (LCM)) $\mathcal{M} = \langle \mathcal{X}, g_z, f_d, f_c, \mathcal{Y} \rangle$, which consists of:

D.1. $\mathcal{X} = \mathcal{X}^1 \times \cdots \times \mathcal{X}^{|V|}$: observational data space,

D.2. $g_z : \mathcal{S} \rightarrow \mathcal{Z}$: Structural Equation Model (SEM) mapping $\mathbf{pa}_i^p \cup \mathbf{cpa}_i^p$ to $\mathbf{z}_i^p$,

D.3. $f_d : \mathcal{Z}^v \rightarrow \mathcal{X}^v$: mixing function mapping slot subspace to observational subspace $\mathcal{X}^v$,

D.4. $f_c : \mathcal{X}^1 \times \cdots \times \mathcal{X}^{|V|} \rightarrow \mathcal{X}$: an additive composition function, given by definition A.5,

D.5. y: observed grouping variable.

Conjecture A.7. (Identifiability of LCM) For any two models $\mathcal{M} = \langle \mathcal{X}, g_z, f_d, f_c, \mathcal{Y} \rangle$ and $\mathcal{M}' = \langle \mathcal{X}, g_z', f_d', f_c', \mathcal{Y} \rangle$, the following properties ensure their identifiability ($\mathcal{M} \sim \mathcal{M}'$), meaning $p_{\mathcal{M}}(\mathbf{x}, y) = p_{\mathcal{M}'}(\mathbf{x}, y)$:

P.1: Identical observational spaces: $\mathcal{X}, \mathcal{Y}$ for both models,

P.2: Latent distribution structure : $p_{\mathcal{M}}(\mathbf{z})$ and $p_{\mathcal{M}'}(\mathbf{z})$ follow the structure defined by Eqn. 2,

P.3: Weak injectivity: f_d maps each $\mathbf{z} \in \mathcal{Z}$ uniquely to a ball of radius $\delta > 0$ in $\mathcal{X}$,

P.4: Scalar latent variables: g_z and g_z' are sets of scalar endogenous $\mathbf{z}_i^v \in \mathbb{R}$ and exogenous $\mathbf{s}_i^v \in \mathbb{R}$ variables,

P.5: Regularizations: Constraints on compositional consistency (Eqn. **??**), DAG regularisation, and mechanism consistency are optimized.

Interpretation: *The conjecture ensures identifiability by constraining LCMs: P.1 and P.2 ensures the same observation $(\mathcal{X}, \mathcal{Y})$ and model classes are considered across LCMs, P.3 ensures that the decoder f_d guarantees that every latent state $\mathbf{z}$ uniquely determines a region in the observational space $\mathcal{X}$, crucial for identifiability Kivva et al. (2022), P.4 ensures that considered variables are scalar in nature, and finally, P.5 ensures all the regularisations are optimised, where compositional consistency and invariant mechanisms are essential for generalisation, making sure that support for object representation is full Wiedemer et al. (2023) and encourages similar latent representations along with functions in both LCMs, DAG consistency enforces a valid latent graph structure by penalizing cycles, ensuring a sound SEM. Together, these properties ensure the LCM results in learning true compositional model.*

Conjecture A.8. ($\boldsymbol{J}_{g_z^{-1}} \sim_{DAG} (\boldsymbol{I} - \boldsymbol{C})$) In case when LCM is identifiable, we can show that Jacobian $\boldsymbol{J}_{g_z^{-1}}$ is structurally equivalent to $(\boldsymbol{I} - \boldsymbol{C})$.

Interpretation: *When an LCM is identifiable, the Jacobian of SEM is faithful to the compositional graph involved in data generation. This means we can use $\boldsymbol{I} - \boldsymbol{C}$ for invariant mechanism and DAG regularisations.*

B Putting Things in Context

In this section, we link our theory to existing OCL and CRL works, highlighting their relations.

B.1 Relation to Spatial Mixtures

Considering a mixing coefficient $\boldsymbol{\pi}$ over a categorical environment variable y, we can restructure Eqn. 3 as a mixture of exponentials:

$$p(\mathbf{z}) \propto \sum_{p=1}^{|V|} \boldsymbol{\pi}^p \exp\left(\bar{\phi}(\mathbf{z}^p, \mathbf{s}^p; y^p)\right) \cdot \left[\prod_{q \in \mathcal{K}} \prod_{(i,j) \in |\mathcal{Z}^p| \times |\mathcal{Z}^q|} \exp\left(\lambda_{ij}^{pq} \phi_1(\mathbf{s}_i^p, \mathbf{z}_j^q)\right)\right] \tag{4}$$

Previous works such as Greff et al. (2017), Kori et al. (2024), Yuan et al. (2024), and Burgess et al. (2019) have shown that, from a probabilistic perspective, the latent distribution in unsupervised object-discovery methods can be interpreted as spatial mixtures. Kori et al. (2024) and Krimmel et al. (2024) demonstrate that more recent methods along slot-attention line of work can reduce the latent distribution to Gaussian Mixture Models (GMMs) or Von Mises-Fisher (vMF) distributions, depending on the chosen distance function. In our case, the model represents a generalized approach to spatial mixtures. When we assume that latent factors are statistically independent both within and across environments (i.e., $\lambda_{ij}^{pq} = 0$ for all (i, j, p, q)), the latent distribution $p(\mathbf{z})$ simplifies to a GMM, as shown in Eqn. 2. This distribution can be further extended to capture compositional graph structures and LCM, making it suitable for modelling more complex dependencies in the latent space.

What does this mean? The latent distribution in Eqn. 4 generalizes traditional *spatial mixture models* by incorporating *pairwise dependencies* between object features. This demonstrates that providing a probabilistic interpretation of scene graphs can be achieved by considering pairwise interactions, where the term $\lambda_{ij}^{pq} \phi_1(\mathbf{s}_i^p, \mathbf{z}_j^q)$ can be interpreted as edges and functions in a scene graph, capturing both object- and attribute-level interactions.

B.2 Relation to Invariance Theory

The proposed formulation of invariant mechanisms can be understood through the lens of invariance principles in CRL, as introduced by Yao et al. (2024). In the context of sample-level invariances (Von Kügelgen et al., 2021; Yao et al., 2023), consider two observed images $(\mathbf{x}^1, \mathbf{x}^2)$ that share underlying structural similarities due to transformations such as augmentations, changes in viewpoint, or pose shifts. By encoding these images into latent representations $\mathbf{z}^1 = g(\mathbf{x}^1)$ and $\mathbf{z}^2 = g(\mathbf{x}^2)$, the invariant CRL frameworks identifies a subset of invariant latent variables $\mathbf{c}^1, \mathbf{c}^2 \subseteq \mathcal{Z}$ that remain stable across both samples, such that for any $z \notin \mathbf{c}_1, \mathbf{c}_2$, we have $\frac{\partial h_1(\mathbf{c}_1)}{\partial z} = \frac{\partial h_2(\mathbf{c}_2)}{\partial z} = 0$. However, similar to observational grouping in Morioka & Hyvärinen (2023), we consider grouping of image $\mathbf{x}$ components by grouping variable y. Each partition corresponds to an object or component in the image and by minimising mechanism consistency, ensures learning invariant object- and attribute-level interactions across scenes.

What does this mean? The proposed formalisation can be reduced to invariant representation learning as in Yao et al. (2024) by appropriately grouping observed data via categorical conditioning variable y. In essence, the approach unifies CRL and scene graph modelling by grounding scene understanding in invariant causal structures, by capturing object- and attribute-level interactions.