IDENTIFIABLE OBJECT REPRESENTATIONS UNDER SPATIAL AMBIGUITIES

Avinash Kori **Francesca Toni** **Ben Glocker**

Imperial College London
a.kori21@imperial.ac.uk

ABSTRACT

Modular object-centric representations are essential for *human-like reasoning* but are challenging to obtain under spatial ambiguities, *e.g. due to occlusions and view ambiguities*. However, addressing challenges presents both theoretical and practical difficulties. We introduce a novel multi-view probabilistic approach that aggregates view-specific slots to capture *invariant content* information while simultaneously learning disentangled global *viewpoint-level* information. Unlike prior single-view methods, our approach resolves spatial ambiguities, provides theoretical guarantees for identifiability, and requires *no viewpoint annotations*. Extensive experiments on standard benchmarks and novel complex datasets validate our method's robustness and scalability.

1 INTRODUCTION

The ability to capture the notion of *objectness* in learned representations is considered to be a critical aspect for developing situation-aware AI systems with human-like reasoning capabilities (Schölkopf & von Kügelgen, 2022; Lake et al., 2017). Objectness can be characterised as understanding the environment from the perspective of its building blocks. These can further be divided into object-part composition Hinton (1979; 2022), which might be a potential reason why humans generalise across environments with few examples to learn from Tenenbaum et al. (2011). Recent advances in object-centric representation learning (OCL) have shown great potential in segregating objects in observed scenes (Locatello et al., 2020b; Kori et al., 2023; Löwe et al., 2024). Indeed, the goal of OCL is to enable agents to learn representations of *objects* in an observed scene in the context of their environment, as opposed to learning global representations as in the case of traditional generative models such as variational auto-encoders (Kingma & Welling, 2013). OCL approaches enable agents to learn spatially disentangled representations, which is an important step in compositional scene generation (Bengio et al., 2013; Lake et al., 2017; Battaglia et al., 2018; Greff et al., 2020) and understanding of causal (and physical) interactions between the objects (Marcus, 2003; Gerstenberg et al., 2021; Gopnik et al., 2004).

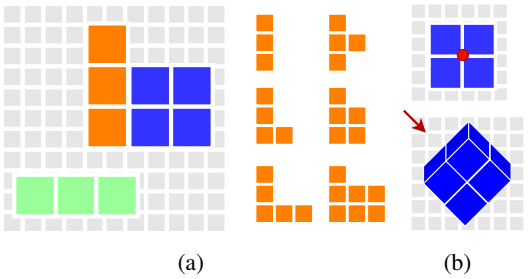


Figure 1: **(a) Occlusion Ambiguity:** the orange object, which is occluded by the blue object, could be any of the six plausible objects shown on the right. **(b) View Ambiguity:** the blue object is observed from two different viewpoints (represented with a red arrow and a dot), leading to a change in its overall shape. In general, identifiable representations resolve ambiguities by determining the most plausible object under occlusion and correct object properties in case of view transformation by leveraging information from multiple viewpoints.

Recent progress in OCL has been limited to learning scene representations from single-viewpoints (Locatello et al., 2020b; Engelcke et al., 2021; Singh et al., 2021; Kori et al., 2023; Chang et al., 2022; Seitzer et al., 2022; Löwe et al., 2024). Although these approaches can learn meaningful object-

specific representations, they encounter significant challenges stemming from spatial ambiguities such as occlusion and view ambiguities (see Fig. 1 for examples). Additionally, while it has been hypothesised that these models learn un-occluded object representations even in the case of occlusions. Learning from a single viewpoint fails to capture effective object representations, due to the presence of multiple plausibilities of partially or fully occluded objects and the effects of view transformations, as demonstrated in Fig. 1 and highlighted by the results in Fig. 2 (we will revisit these results later in section 4). Another example of spatial ambiguities can be observed in Fig. 3, where object $\mathcal{O}_4$ in $\mathbf{x}^1$ and $\mathbf{x}^2$ can be interpreted as a cube, but only after considering $\mathbf{x}^3$ we can conclude that being a pyramid.

A handful of approaches, including MULMON(Li et al., 2020), DYMON(Li et al., 2021), OCLOC(Yuan et al., 2024), have considered multiple viewpoints for object representations. Additionally, methods such as (Liu et al., 2025; Chen et al., 2021; Luo et al., 2024) effectively use NERF(Mildenhall et al., 2021) for constructing a 3D environment from multi-viewpoint images, where the occlusions are addressed by construction. Among these methods, MULMON, DYMON, and all NERF based approaches assume that the viewpoint annotations are known, which simplifies the problem of learning to disentangle object representations conditioned on viewpoint information.

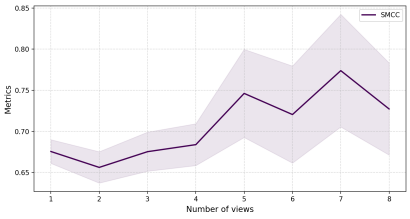


Figure 2: Identifiability across a number of views measured with Slot Mean Correlation Coefficient (SMCC).

The problem setting in this work aligns with OCLOC, in that, our aim is to learn invariant object representations while simultaneously learning global view information with respect to an *implicit global coordinate frame*. This eliminates the requirement for paired viewpoint-image data. While OCLOC introduces an innovative approach for learning global view information independently of the scene, its primary focus is on achieving *object-consistency* unconditional to views rather than explicitly learning view-invariant object representations. Additionally, learning global unconditional view representations does not guarantee learning identifiable view/object representations, which was not studied for OCLOC. In this work, we provide a novel model, where object representations satisfy view-invariance and view representations satisfy *approximate equivariance* properties, allowing us to exploit objects' inherent geometry and semantics to establish correspondences across views.

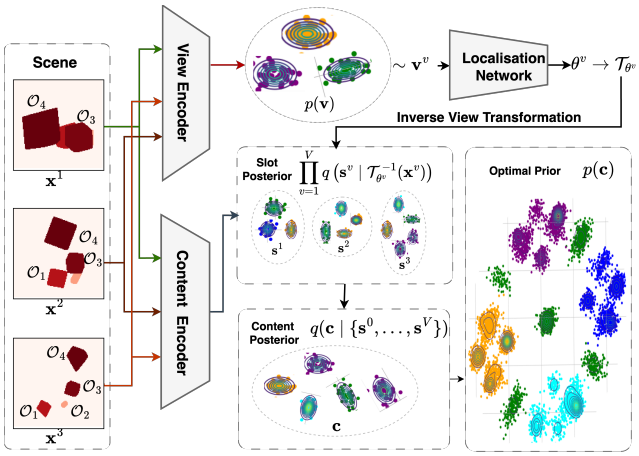


Figure 3: This illustrates a scene with four objects $\mathcal{O}^s = \{\mathcal{O}_1, \mathcal{O}_2, \mathcal{O}_3, \mathcal{O}_4\}$, observed from three different viewpoints, each described with a set of clearly visible objects: $\mathcal{O}^1 = \{\mathcal{O}_3, \mathcal{O}_4\}, \mathcal{O}^2 = \{\mathcal{O}_1, \mathcal{O}_3, \mathcal{O}_4\}, \mathcal{O}^3 = \{\mathcal{O}_1, \mathcal{O}_2, \mathcal{O}_3, \mathcal{O}_4\}$. The corresponding images are passed through view and content encoders, and sampled global view vector $\mathbf{v}$ is used to estimate transformation function $\mathcal{T}_{\theta^v}$ given by parameters θ^v predicted using a localisation network. We apply a view-specific inverse $\mathcal{T}_{\theta^v}^{-1}$ on respective images projecting them to an *implicit* space, which is used to learn view conditioned slot posterior corresponding to GMMs represented by $q(\mathbf{s}^v \mid \mathcal{T}_{\theta^v}^{-1}(\mathbf{x}^v))$, which are further aggregated to marginalize viewpoint information, resulting in a content posterior, also a GMM $q(\mathbf{c} \mid \{\mathbf{s}^1, \ldots, \mathbf{s}^V\})$, which is further accumulated across all samples resulting in optimal prior $p(\mathbf{c})$.

In single-view OCL, Kori et al. (2024); Brady et al. (2023); Lachapelle et al. (2023) make an effort in rigorously formalising the underpinning, explicit and implicit assumptions and provide conditions under which models result in learning identifiable slot representations, leaving out ambiguous scenarios. Unlike them, our approach resolves spatial ambiguities, provides theoretical guarantees for identifiability, and requires no viewpoint annotations.

To the best of our knowledge, this is the first work addressing explicit formalisations of assumptions and theory required for achieving *identifiable* object representations under occlusions with multi-view observational data. To this end, we make use of the spatial Gaussian mixture models(GMM) in latent distribution across viewpoints to encourage identifiability without additional auxiliary data. Our main contributions in this work can be summarised as follows:

(i) We propose a probabilistic slot attention variant, *View-Invariant Slot Attention (VISA)* for learning identifiable object-centric representations from multiple viewpoints, resolving spacial ambiguities such as occlusions and view ambiguities (Section 2).

(ii) We prove that our object-centric representations are identifiable in the case of partial or full occlusions without additional view information up to an equivalence relation with a mixture model specification (Section 3).

(iii) We provide conclusive evidence of our identifiability results, including visual verification on synthetic datasets; we also demonstrate the scalability of the proposed method on two new, carefully designed complex datasets MVMOVI-C and MVMOVI-D (Section 4).

2 VISA FORMALISM

Let $\mathbf{x}^{1:V} = \{\mathbf{x}^1, \dots \mathbf{x}^V\} \in \mathcal{X} = \mathcal{X}^1 \times \cdots \times \mathcal{X}^V$, V views of the same scene observed from different viewpoints with an observational space $\mathcal{X} \subseteq \mathbb{R}^{V \times H \times W \times C}$. We consider $[V]$ as a shorthand notation for $\{1, \dots, V\}$. Let $\mathcal{O}^e = \mathcal{O}^1 \cup \cdots \cup \mathcal{O}^V$ correspond to an abstract notion of object sets of an environment, while $\mathcal{O}^v, \forall v \in [V]$ is a set of objects present in a considered viewpoint v. Importantly, we consider that the number of objects per viewpoint can vary, *i.e.,* $|\mathcal{O}^1 \cup \cdots \cup \mathcal{O}^V| \geq |\mathcal{O}^v| \; \forall \; v \in [V]$, allowing for partial or full occlusion in some viewpoints. Let $\mathbf{v}^{1:V} \in \mathcal{V} = \mathcal{V}^1 \times \cdots \times \mathcal{V}^V \subseteq \mathbb{R}^{V \times d_v}$ be inferred viewpoint-specific information[1], while $\mathbf{s}^{1:V}_{1:K} \in \mathcal{S} = \mathcal{S}^1 \times \cdots \times \mathcal{S}^V \subseteq \mathbb{R}^{V \times K \times d_s}$ correspond to a viewpoint-specific slot representation. Let $\mathbf{c}_{1:K} \in \mathcal{C} \subseteq \mathbb{R}^{K \times d_c}$ capture the notion of an *aggregate*, effectively accumulating the object knowledge across viewpoints. For any subset A of $[V]$, we represent scene observations as $\mathbf{x}^A = \{\mathbf{x}^i : \forall i \in A\} \in \times_{i \in A} \mathcal{X}^i$. The inferred viewpoints and the view specific slots are denoted as $\mathbf{v}^A = \{\mathbf{v}^i : \forall i \in A\} \in \times_{i \in A} \mathcal{V}^i$, and $\mathbf{s}^A_{1:K} = \{\mathbf{s}^i_{1:K} : \forall i \in A\} \in \times_{i \in A} \mathcal{S}^i$, respectively. We define $p_A(\mathbf{c})$ as the distribution of $\mathbf{c}$ over A. A more comprehensive summary of notations and terminologies is provided in App. A.

In modelling, `w.l.o.g`, we consider access to a certain subset $A \subseteq [V]$, ensuring the model's applicability across different scenarios. Furthermore, to simplify notation, we sometimes do not include the superscript denoting the full set of views, thereby using $\mathbf{x} = \mathbf{x}^A$, $\mathbf{s}_{1:K} = \mathbf{s}^A_{1:K}$, and $\mathbf{v} = \mathbf{v}^A$ interchangeably. Likewise, if we do not specify the subscripts for $\mathbf{c}$ and $\mathbf{s}$, it implies they represent the entire collection of objects, specifically as $\mathbf{s} = \mathbf{s}^A_{1:K}$ and $\mathbf{c} = \mathbf{c}_{1:K}$. Lastly, for any function f that operates on two distinct inputs $\mathbf{x} = f(\mathbf{z}, \mathbf{v})$, its inverse is denoted by $\mathbf{z} = f^{-1}(\mathbf{x}; \mathbf{v})$, which signifies the reversal of f conditioned on a variable $\mathbf{v}$. In the rest of this section, we introduce all the components involved in our model. We also introduce assumptions, examples, and intuition wherever necessary. Considering the generative model Eqn. 1, which is overviewed in graphical model Fig. 4, any scene $\mathbf{x}$ is generated using view $\mathbf{v}$ and content $\mathbf{c}$. Here, both $\mathbf{c}$ and $\mathbf{v}$ are latent variables learned with variational inference(Kingma & Welling, 2013).

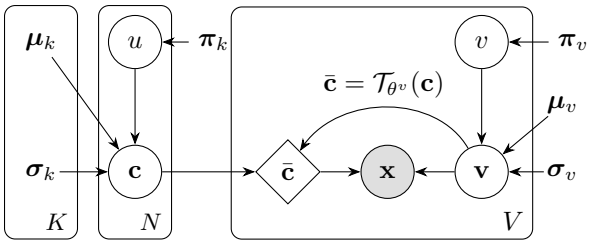


Figure 4: **Graphical model for multi-view probabilistic slot attention:** For every image in a dataset a view $\mathbf{v} \in \mathbb{R}^{d_v} \sim p(\mathbf{v})$, this view is used to compute transformation $\mathcal{T}_{\theta^v}$. Similarly, desired number ($< K$) of content representations $\mathbf{c} \in \mathbb{R}^{N \times d_s}$ are sampled content distribution $p(\mathbf{c})$. Finally, the image $\mathbf{x}$ is generated using the transformed content $\mathcal{T}_{\theta^v}(\mathbf{c})$ and view $\mathbf{v}$.

$$p(\mathbf{x}) = \iint p(\mathbf{x}^v \mid \mathcal{T}_{\theta^v}(\mathbf{c}), \mathbf{v}^v)\; p(\mathbf{c})\; p(\mathbf{v}^v)\; d\mathbf{v}^v\; d\mathbf{c} \tag{1}$$

[1]We abuse the terminology by considering viewpoint, lighting, object dimension, to be encoded in a representation $\mathbf{v}$. Note that the $\mathbf{v}$ is inferred by the model.

View model. Given that the view property remains consistent across all objects, we treat the view as a global, image-level property as opposed to Yuan et al. (2024), where view is treated as an object-level property. Assuming access to a discrete set of viewpoints denoted by A, we consider prior over a view distribution to be a GMM represented by $p(\mathbf{v}) = \sum_{v=1}^{|A|} \boldsymbol{\pi}^v \mathcal{N}(\mathbf{v}; \boldsymbol{\mu}_v, \boldsymbol{\sigma}_v^2)$. To learn the parameters of this GMM, we consider the posterior of the form $q_\phi(\mathbf{v} \mid \mathbf{x}^v)\ \forall\ v \in A$[2]. In both prior and posterior, we consider the covariance to be diagonal, implicitly making an ICA assumption (Khemakhem et al., 2020a). The sampled variable $\mathbf{v} \sim q_\theta(\mathbf{v} \mid \mathbf{x}^v)$ is used to estimate transformation parameters $\theta^v \in \mathbb{R}^{3\times 2}$ as in Jaderberg et al. (2015) which makes an affine transformation map $\mathcal{T}_{\theta^v}$, which is later applied on content $\mathbf{c}$ and on view-specific slots $\mathbf{s}$. It is important to note that we use the same set of parameters ϕ across all viewpoints in A for inferring view information $\mathbf{v}$.

Viewpoint specific slots. As illustrated in Fig. 3 the inference of $\mathbf{c}$ depends on the view-specific slots $\mathbf{s}$. For a considered image $\mathbf{x}^v, v \in A$, we first apply an inverse view transformation $\mathcal{T}_{\theta^v}^{-1}$ and model the slot distribution as a spatial mixture model represented by $q(\mathbf{s}_{1:K}^A \mid \mathcal{T}_{\theta^v}^{-1}(\mathbf{x}^A))$. The inverse transformation makes sure that the estimated object representations across all view in A are in a common implicit representation space. As this is an intermediate variable which does not show up in our generative model in Eqn. 1, we update the corresponding parameters with closed-form equations via expectation maximisation algorithm as in Kori et al. (2024). The resulting slot posterior is a conditional GMM as described in Eqn. 2, where $\bar{\mathbf{x}}^v = \mathcal{T}_{\theta^v}^{-1}(\mathbf{x}^v)$ is a transformed inputs, $(\mu_k(\bar{\mathbf{x}}^v), \sigma_k^2(\bar{\mathbf{x}}^v), \pi_k(\bar{\mathbf{x}}^v))$ are mean, diagonal covariance, and mixing coefficients for the considered a view and object.

$$q(\mathbf{s}^v \mid \bar{\mathbf{x}}^v) = \sum_{k=1}^{K} \boldsymbol{\pi}_k(\mathbf{x}^v) \mathcal{N}\Big(\mathbf{s}_k^v; \boldsymbol{\mu}_k(\bar{\mathbf{x}}^v), \boldsymbol{\sigma}_k^2(\bar{\mathbf{x}}^v)\Big) \tag{2}$$

Representation matching. Given the permutation equivariance property of slot representations, we use a matching function with a permutation matrix $\boldsymbol{P}^v$, $m_s : \mathcal{S}^A \to \mathcal{S}^A$ such that $m_s(\mathbf{s}_{1:K}^v) = \boldsymbol{P}^v \mathbf{s}_{1:K}^v$ mapping representation axis `w.r.t` $\boldsymbol{P}^v$. The permutation matrix $\boldsymbol{P}^v$ is estimated by considering the slots of the first viewpoint $\mathbf{s}^1$ as a base representation, and other representations $\mathbf{s}^v\ \forall v \in A$ are matched to align with it. We utilise Hungarian matching, as illustrated in Locatello et al. (2020b); Wang et al. (2023), to estimate this permutation matrix $\boldsymbol{P}^v$, to control the noise in the matching algorithm, we introduce view-warmup strategy, which we detail in App. G.5.

Content aggregator. We consider $g : \mathcal{S} \to \mathcal{C}$ as a content aggregator function, which marginalises the effect of view conditioning. To achieve this, we consider a convex combination of all the aligned slot representations (aligned to a base representation), considering mixing coefficients $\boldsymbol{\pi}_k(\mathbf{x}^v)$ (we use $\boldsymbol{\pi}_k^v$ for simplicity) in Eqn. 2 as a combination weight. The convex combination accounts for potential object occlusions, which may cause objects to be absent in particular views ensuring only active representations are combined (refer to an intuition below), resulting in a content posterior $(q(\mathbf{c} \mid \mathbf{s}))$, which is a GMM with mixing coefficients $\tilde{\boldsymbol{\pi}}_k = \left(\sum_{v=1}^{|A|} \boldsymbol{\pi}_k^v\right) / |A|$ and the parameters described in Eqn. 4 (`w.l.o.g` we consider $\mathbf{s}, \boldsymbol{\pi}$ to represent aligned representations), refer to Lemma F.3, with $w_i = 1/|A|\ \forall i \in A$. Additionally, algorithm 1 details the entire forward process.

Intuition: Content aggregation

Based on illustrated example in Fig. 3, for images $\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3$, the resulting matched slots and mixing coefficients correspond to $\mathbf{s}^1 = \{\mathbf{s}_r^1, \mathbf{s}_r^1, \mathbf{s}_{\mathcal{O}_3}^1, \mathbf{s}_{\mathcal{O}_4}^1, \mathbf{s}_b^1\}, \mathbf{s}^2 = \{\mathbf{s}_{\mathcal{O}_1}^2, \mathbf{s}_r^2, \mathbf{s}_{\mathcal{O}_3}^2, \mathbf{s}_{\mathcal{O}_4}^2, \mathbf{s}_b^2\}, \mathbf{s}^3 = \{\mathbf{s}_{\mathcal{O}_1}^3, \mathbf{s}_{\mathcal{O}_2}^3, \mathbf{s}_{\mathcal{O}_3}^3, \mathbf{s}_{\mathcal{O}_4}^3, \mathbf{s}_b^3\}$, where $\mathbf{s}_{\mathcal{O}_i}^v, \mathbf{s}_r^v$, and $\mathbf{s}_b^v$ correspond to slot representation for object $\mathcal{O}_i$, random slot representation and background information, respectively, with mixing coefficients $\boldsymbol{\pi}^1 = \{0, 0, 1, 1, 1\}, \boldsymbol{\pi}^2 = \{1, 0, 1, 1, 1\}$, and $\boldsymbol{\pi}^1 = \{1, 1, 1, 1, 1\}$. Proposed aggregation merges the slots ignoring the random slots $\mathbf{s}_r^v$, resulting in $\mathbf{c}_{\mathcal{O}_1} = (\mathbf{s}_{\mathcal{O}_1}^2 + \mathbf{s}_{\mathcal{O}_1}^3)/2, \mathbf{c}_{\mathcal{O}_2} = \mathbf{s}_{\mathcal{O}_2}^3$ and so on.

[2]We consider the parametric form of q to be Gaussian.

$$g(\mathbf{s}_{1:K}^{1:V}, \boldsymbol{\pi}^{1:V}) = \sum_{v=1}^{|A|} \frac{\boldsymbol{\pi}_{1:k}^{v}}{|A|\tilde{\boldsymbol{\pi}}_{k}^{v}} \mathbf{s}_{1:K}^{v}; \tag{3}$$

$$\tilde{\boldsymbol{\mu}}_k(\mathbf{x}^A) = \sum_{v=1}^{|A|} \frac{\boldsymbol{\pi}_k^v}{|A|\tilde{\boldsymbol{\pi}}_k^v} \boldsymbol{\mu}_k(\mathbf{x}^v);\ \tilde{\boldsymbol{\sigma}}^2(\mathbf{x}^A) = \sum_{v=1}^{|A|} \left(\frac{\boldsymbol{\pi}_k^v}{|A|\tilde{\boldsymbol{\pi}}_k^v}\right)^2 \boldsymbol{\sigma}_k^2(\mathbf{x}^v); \tag{4}$$

Mixing function and training objective. We consider both *additive and non-additive* (ref. definition E.1) mixing functions $f_d : \mathcal{C} \times \mathcal{V}^v \to \mathcal{X}^v$. For additive decoders, we use a spatial-broadcasting (Greff et al., 2019) and MLP decoders, and for non-additive mixing function, we use auto-regressive transformers (Vaswani et al., 2017). We use the shared decoder f_d for all views and objects, modelling the conditional distribution $p(\mathbf{x}^v \mid \mathcal{T}_{\theta^v}(\mathbf{c}), \mathbf{v}^v)$. To train our model in an end-to-end fashion, we maximise the log-likelihood of the joint $p(\mathbf{x}^A)$, which results in the evidence lower bound (ELBO), Eqn. 5, check Lemma F.1. Here, we consider the distribution form of $p(\mathbf{x}^v \mid \mathbf{c}, \mathbf{v}^v)$ to be Gaussian with learnable mean with isotropic covariance.

$$\mathbb{E} \log p(\mathbf{x} \mid \mathcal{T}_{\theta^v}(\mathbf{c}), \mathbf{v}) - \mathrm{KL}\left(q(\mathbf{v} \mid \mathbf{x}) \parallel p(\mathbf{v})\right) \tag{5}$$

3 Theoretical Analysis

In this section, we leverage the properties of the proposed model to theoretically demonstrate the learning of identifiable representations under challenging spatial ambiguities. In this work, we consider our data-generating process to satisfy a viewpoint sufficiency assumption (refer to 3.1).

Assumption 3.1. (View-point sufficiency) For any set $A \subseteq [V]$, we consider set A to be view-point sufficient iff $|\mathcal{O}^A| = |\mathcal{O}^e|$. This basically means that all the objects are visible across all the considered views A, even when an individual view may not contain all the objects.

Example 1. *Based on illustrated example in Figure 3, the scene is composition of four objects $\mathcal{O}^e = \{\mathcal{O}_1, \mathcal{O}_2, \mathcal{O}_3, \mathcal{O}_4\}$, view point subset $A = [V] = \{1, 2, 3\}$ is considered to be view point sufficient since $\bigcup_{v \in A} \mathcal{O}^v = \{\mathcal{O}_3, \mathcal{O}_4\} \cup \{\mathcal{O}_1, \mathcal{O}_3, \mathcal{O}_4\} \cup \{\mathcal{O}_1, \mathcal{O}_2, \mathcal{O}_3, \mathcal{O}_4\} = \mathcal{O}^e$.*

Given that we learn the parameters of our view-specific spatial GMM with closed-form updates, we do not use an explicit prior minimising KL divergence. Instead, we rely on the fact that marginalising the effect of data points from posterior (*aggregate posterior*) is an optimal prior (Hoffman & Johnson, 2016; Kori et al., 2024), resulting in $p(\mathbf{c}) = \iint q(\mathbf{c}|\mathbf{s}^A, \mathbf{x}^A) d\mathbf{s}^A d\mathbf{x}^A$. Given that GMMs are universal density approximates given enough components (even GMMs with diagonal covariances), the resulting aggregate posterior $q(\mathbf{c}) = p(\mathbf{c})$ is highly flexible and multi-modal. It often suffices to approximate it using a sufficiently large subset of the dataset if marginalising out the entire dataset becomes computationally restrictive.

Lemma 3.2 (Optimal Prior)**.** *For $A \in [V]$, given the a local content distribution $q(\mathbf{c}_{1:K} \mid \mathbf{s}_{1:K}^A, \mathbf{x}^A)$ (per-scene $\mathbf{x}^A \in \{\mathbf{x}_i^A\}_{i=1}^M$), which can be expressed as a GMM with K components, the aggregate posterior $q(\mathbf{c})$ is obtained by marginalizing out $\mathbf{x}, \mathbf{s}$ is a non-degenerate global Gaussian mixture with MK components:*

$$p(\mathbf{c}) = q(\mathbf{c}) = \frac{1}{M} \sum_{i=1}^{M} \sum_{k=1}^{K} \widehat{\boldsymbol{\pi}}_{ik} \mathcal{N}\left(\mathbf{c}; \widehat{\boldsymbol{\mu}}_{ik}, \widehat{\boldsymbol{\sigma}}_{ik}^2\right). \tag{6}$$

Proof Sketch. The result is obtained by integrating the product of involved posterior densities $q(\mathbf{c} \mid \mathbf{s}) q(\mathbf{s} \mid \mathbf{x}) p(\mathbf{x})$. Further, we verify if the mixing coefficients sum to one in the new mixture, proving the aggregate to be well-defined. □

With this, we show three main results: firstly, we show that aggregate content representations ($\mathbf{c}$) are identifiable without supervision (up to $\sim_s$). Secondly, we show that these representations are invariant to the choice of viewpoints under assumption 3.1. Finally, we show that the model exhibits in an approximate view equivariance.

Theorem 3.3. *(Affine Equivalence) For any subset $A \subseteq [V]$, such that $|A| > 0$, given a set of images $\mathbf{x}^A \in \mathcal{X}^A$ and a corresponding aggregate content $\mathbf{c} \in \mathcal{C}$ and a non-degenerate content posterior $q(\mathbf{c} \mid \mathbf{s}^A)$, considering two mixing function $f_d, \tilde{f}_d$ satisfying assumption F.4, with a shared image, then $\mathbf{c}$ are identifiable up to $\sim_s$ equivalence.*

Proof Sketch. To prove the following result, we follow multiple steps as described below: (i). We demonstrate the distribution $p(\mathbf{c})$ obtained as a result of lemma 3.2 is non-degenerate and a valid distribution, (ii). With the above results, we demonstrate invertibility restrictions on mixing functions, (iii). Finally, we constrain the subspace to affine, demonstrating $\sim_s$ of aggregate content $\mathbf{c}$. □

Intuition: Affine equivalence Considering an example 1, with two perfectly trained models f_d and $\tilde{f}_d$. Resulting aggregate contents are described as $\mathbf{c} = f_d^{-1}(\mathbf{x}^A; \mathbf{v}^A) = \{\mathbf{c}_{\mathcal{O}_1}, \mathbf{c}_{\mathcal{O}_2}, \mathbf{c}_{\mathcal{O}_3}, \mathbf{c}_{\mathcal{O}_4}, \mathbf{c}_{\mathcal{O}_b}\}$ and $\tilde{\mathbf{c}} = \tilde{f}_d^{-1}(\mathbf{x}^A; \mathbf{v}^A) = \{\tilde{\mathbf{c}}_{\mathcal{O}_2}, \tilde{\mathbf{c}}_{\mathcal{O}_4}, \tilde{\mathbf{c}}_{\mathcal{O}_3}, \tilde{\mathbf{c}}_{\mathcal{O}_1}, \tilde{\mathbf{c}}_{\mathcal{O}_b}\}$ for $A = [V] = \{1, 2, 3\}$. $\sim_s$ equivalence states that there exists a permutation matrix $\boldsymbol{P}$ which aligns the object order in $\tilde{\mathbf{c}}$ to match with $\mathbf{c}$ and there exists and invertible affine mapping $\boldsymbol{A}$ such that $\tilde{\mathbf{c}}_{\mathcal{O}_k} = \boldsymbol{A}\mathbf{c}_{\mathcal{O}_k} \forall k \in \{1, 2, 3, 4\}$.

Theorem 3.4. *(Invariance of aggregate content) For any subset $A, B \subseteq [V]$, such that $|A| > 0, |B| > 0$ and both A, B satisfy an assumption 3.1, we consider aggregate content to be invariant if $f_A \sim_s f_B$ for data $\mathcal{X}^A \times \mathcal{X}^B$.*

Proof Sketch. To prove this, we extend the proof of Thm. 3.3, and establish that there exist two inevitable affine functions h_A, h_B for mixing functions $f_A, f_B : \mathcal{C} \times \mathcal{V} \to \mathcal{X}$ to map representations $\mathbf{c}$ with a given view set $\mathbf{v}^A$ to observations $\mathbf{x}^A$. Later, we show that, in the case of invariance, an affine mapping exists from h_A to h_B. □

Intuition: Invariant slots. Considering an example 1, with $A = \{1, 3\}, B = \{2, 3\}$, such that sets A, B are viewpoint sufficient. Let f_A and f_B, be trained models on $\mathcal{X}^A$ and $\mathcal{X}^B$ respectively. Resulting in $\mathbf{c} = f_A^{-1}(\mathbf{x}^A; \mathbf{v}^A) = \{\mathbf{c}_{\mathcal{O}_1}, \mathbf{c}_{\mathcal{O}_2}, \mathbf{c}_{\mathcal{O}_3}, \mathbf{c}_{\mathcal{O}_4}, \mathbf{c}_{\mathcal{O}_b}\}$ and $\tilde{\mathbf{c}} = \tilde{f}_B^{-1}(\mathbf{x}^B; \mathbf{v}^B) = \{\tilde{\mathbf{c}}_{\mathcal{O}_2}, \tilde{\mathbf{c}}_{\mathcal{O}_4}, \tilde{\mathbf{c}}_{\mathcal{O}_3}, \tilde{\mathbf{c}}_{\mathcal{O}_1}, \tilde{\mathbf{c}}_{\mathcal{O}_b}\}$. Thm. 3.4 states that the representations $\mathcal{T}_{\theta^B}^{-1}(\tilde{\mathbf{c}}_{\mathcal{O}_k})$ can be mapped to $\mathcal{T}_{\theta^A}^{-1}(\mathbf{c}_{\mathcal{O}_k})$ by permuting object indices and an affine transformation.

Theorem 3.5. *(Approximate representational equivariance) For a given aggregate content $\mathbf{c}$, for any two views $\mathbf{v}, \tilde{\mathbf{v}} \sim p_A(\mathbf{v})$, resulting in respective scenes $\mathbf{x} \sim p_A(\mathbf{x} \mid \mathbf{v}, \mathbf{c})$ and $\tilde{\mathbf{x}} \sim p_A(\mathbf{x} \mid \tilde{\mathbf{v}}, \mathbf{c})$, for any homeomorphic transformation $h_x \in \mathcal{H}_x$ such that $h_x(\mathbf{x}) = \tilde{\mathbf{x}}$, their exists another homeomorphic transformation $h_v \in \mathcal{H}_v$ such that $\mathcal{H}_v \subseteq \mathcal{H}_x \subseteq \mathbb{R}^{\dim(\mathbf{x})}$ and $\mathbf{v} = h_v^{-1}\left(f_d^{-1}(h_x(\mathbf{x}); \mathbf{c})\right)$.*

Remark 3.6. Note that the theorem only says that the transformation function transforming the view representations $\mathbf{v}$ as an effect of the homeomorphic transformation of $\mathbf{x}$ lies in the same subspace of input transformations.

Proof Sketch. We prove the following result by following the steps in Thm. 3.4, over a view distribution $p(\mathbf{v})$ but for a fixed content vector $\mathbf{c}$. □

Intuition: Approximate equivariance In the scenario when the cameras are positioned such that they have overlapping fields of view, and their relative pose (rotation and translation) must avoid degeneracies like aligning on the same plane or mapping points to infinity. This results in the transformation between views being smooth, invertible, and consistent. If the scene is planar or depth variations are minimal, the homography can capture the transformation accurately without the need for inverse rendering. Notably, the cameras should have non-zero rotation and translation to avoid collapsing the scene, and their intrinsic parameters must be known or identical to prevent distortions. When the scenario satisfies all the above properties, the 2D homography transformation $\boldsymbol{H}$ between two camera views can be learned as a homeomorphic transformation (Hartley & Zisserman, 2003).

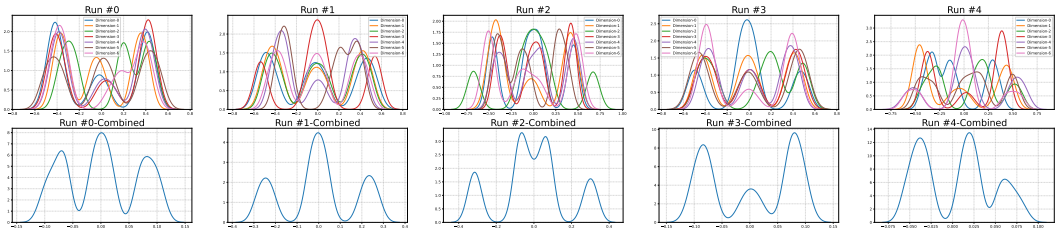


Figure 5: **Identifiability of** $q(\mathbf{c})$. The top row indicates individual feature distribution across five different runs. The bottom row reflects the feature distribution, which we use as a proxy for multi-dimensional features given Lemma F.2. As observed, mean feature distribution across runs is either scaled, shifted, or split (increase in number of modes); this provides strong evidence of recovery of the latent space up to affine transformations, empirically verifying our claims in Thm. 3.3.

Table 1: Comparing identifiability of $q(\mathbf{s})$, $q(\mathbf{c})$, and $p(\mathbf{v})$ scores wrt existing OCL methods.

METHOD	CLEVR-MV			GQN			GSO		
	SMCC ↑	INV-SMCC ↑	MCC ↑	SMCC ↑	INV-SMCC ↑	MCC ↑	SMCC ↑	INV-SMCC ↑	MCC ↑
AE	$0.32 \pm .02$	-	-	$0.29 \pm .02$	-	-	0.24 ± 0.08	-	-
SA	$0.47 \pm .03$	-	-	$0.38 \pm .02$	-	-	0.28 ± 0.06	-	-
PSA	$0.49 \pm .02$	-	-	$0.38 \pm .02$	-	-	0.30 ± 0.04	-	-
MulMON	$0.61 \pm .03$	$0.62 \pm .02$	-	$0.59 \pm .06$	$0.61 \pm .02$	-	0.56 ± 0.04	0.48 ± 0.06	-
OCLOC	$0.63 \pm .02$	$0.64 \pm .01$	$0.48 \pm .04$	$0.60 \pm .03$	$0.60 \pm .01$	$0.42 \pm .08$	0.58 ± 0.04	0.54 ± 0.03	0.46 ± 0.04
VISA	$0.67 \pm .01$	$0.66 \pm .01$	$0.60 \pm .04$	$0.59 \pm .01$	$0.63 \pm .01$	$0.52 \pm .03$	$0.60 \pm .03$	$0.61 \pm .02$	$0.58 \pm .03$

4 EMPIRICAL EVALUATION

Given the work's theoretical focus, experimentally, we aim to provide strong empirical evidence of our identifiability, invariance, and equivariance claims in a multiview setting. We also extend our experiments to standard imaging benchmarks, including CLEVR-MV, CLEVR-AUG, GQN (Li et al., 2020); we additionally demonstrate the framework's scalability to highly diverse setting with GSO (Downs et al., 2022) and proposed datasets MV-MOVIC, MV-MOVID which are multiview versions of MoViC dataset with fixed and varying scene-specific cameras (Greff et al., 2022).

Experimental setup. To verify our claims on (i) identifiability claim, we train our model on a given view subset $A \subseteq [V]$ and compare view averaged slot mean correlation coefficient (SMCC) measure Kori et al. (2024), (ii) invariance claim, we train multiple models on different subsets of viewpoints $A, B \subseteq [V]$ and compare the aggregate content representations across models, quantifying the similarities with SMCC, we consider this measure to be invariant SMCC (INV-SMCC), and finally, (iii) for subspace equivariance, we consider a trained model with a view subset $A \subseteq [V]$ and compute MCC of view information $\mathbf{v}$ by applying random homeomorphic transformations on samples $\mathbf{x}^A \sim \mathcal{X}^A$ (which can also be done by considering samples $\mathbf{x}^B \sim \mathcal{X}^B$, where cameras relative position satisfy the required constraints 3.5, and analyse $p(\mathbf{v}^A)$ and $p(\mathbf{v}^B)$).

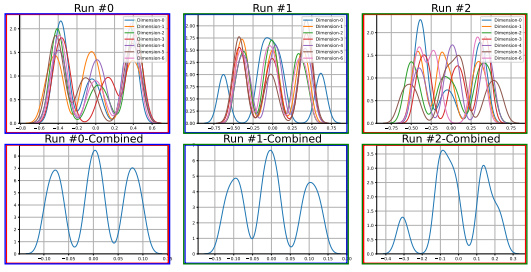


Figure 6: **Viewpoint invariance for** $q(\mathbf{c})$. The top and bottom row indicates individual feature levels and mean feature distributions, respectively. Each columns reflect marginalised aggregate content distribution $q(\mathbf{c})$ when trained with different view pairs {(blue, red), (green, blue), and (green, red)}, respectively. As the resulting distributions with different datasets only vary by an affine transformation, providing strong evidence for Thm. 3.4.

Models & baselines. We consider two ablations with two types of decoders: (i) additive with MLPs and spatial broadcasting CNNs and (ii) non-additive decoders, which include transformer models. In all cases, we use LeakyReLU activations to satisfy the weak injectivity conditions (Assumption F.4). In terms of object-centric learning baselines, we compare with standard additive autoencoder setups following (Brady et al., 2023), slot-attention (SA) (Locatello et al., 2020b), probabilistic slot-attention (PSA) (Kori et al., 2024), MulMON (Li et al., 2020), and OCLOC (Yuan et al., 2024).

Table 2: Identifiability and generalisability analysis on MV-MoVIC dataset.

METHOD	IN-DOMAIN RESULTS				OUT-OF-DOMAIN RESULTS			
	mBO ↑	SMCC ↑	INV-SMCC ↑	MCC ↑	mBO ↑	SMCC ↑	INV-SMCC ↑	MCC ↑
SA-MLP	0.28 ± 0.091	0.36 ± 0.004	-	-	0.26 ± 0.08	0.38 ± 0.006	-	-
PSA-MLP	0.30 ± 0.022	0.38 ± 0.002	-	-	0.30 ± 0.03	0.40 ± 0.005	-	-
VISA-MLP	0.28 ± 0.021	0.52 ± 0.021	0.61 ± 0.023	0.54 ± 0.026	0.27 ± 0.02	0.51 ± 0.029	0.58 ± 0.031	0.52 ± 0.021
SA-TRANSFORMER	0.34 ± 0.014	0.36 ± 0.016	-	-	0.33 ± 0.041	0.36 ± 0.043	-	-
PSA-TRANSFORMER	0.37 ± 0.021	0.38 ± 0.007	-	-	0.37 ± 0.033	0.39 ± 0.016	-	-
VISA-TRANSFORMER	0.38 ± 0.008	0.44 ± 0.003	0.46 ± 0.001	0.53 ± 0.011	0.36 ± 0.017	0.46 ± 0.033	0.46 ± 0.018	0.55 ± 0.082

CASE STUDY 1: ILLUSTRATION OF IDENTIFIABILITY. To definitively show the validity of our claims about identifiability (Thm 3.3, Thm 3.4, and Thm 3.5), we created a synthetic unconfounded scenario for modelling. This provides us with two data modalities, Fig, 7 (i) projected point cloud data, and (ii) corresponding imagery data, we detail point cloud illustrations in appendix G.1. Additionally, this dataset also provides us with the ground truth object and viewpoint features for evaluation. To visualise the aggregate mixture, following Lemma F.2, we use the projected GMM to interpret the distribution of random variables in $\mathbb{R}^d$.

The data-generating process is thoroughly explained in the App. D.1. In Fig. 5, we display the distributions of marginalized aggregate content distribution $q(\mathbf{c})$, comparing individual features and a mean feature across different runs that are either scaled, shifted, or split (increase in number of modes), which is reflective of affine transformation of features across runs. To quantitatively measure the same, we computed SMCC and observed it to be $\mathbf{0.72} \pm 0.04$, empirically verifying our Thm. 3.3. Furthermore, to illustrate the invariance of distribution $q(\mathbf{c})$ across viewpoints (Thm. 3.4), we consider three different viewpoints. We use all possible pairs to learn $q(\mathbf{c})$ distributions as illustrated in Fig. 6, where the distributions are described `w.r.t` viewpoints described by {g, r}, {r, b}, and {g, b}, respectively. These distributions were also found to have similar properties as before, with an observed SMCC of $\mathbf{0.71} \pm 0.11$, further confirming the claims in Thm. 3.4. Additionally, Fig. 2 demonstrates the improvement in identifiability as the number of viewpoints increases.

CASE STUDY 2: IMAGING APPLICATIONS. We first evaluate the framework on standard benchmarks, specifically focusing on CLEVR-MV, CLEVR-AUG, GQN, and GSO with simple objects. Given the *true generative factors* are unobserved, we derive our quantitative assessments from multiple runs. The results are shown in Table 1, confirming the validity of our theory on imaging datasets. Regarding the baseline comparisons that utilize a single viewpoint, the INV-SMCC mirrors the SMCC due to its inherent design (*i.e.,* aggregation of a set with a single element is the same element). Moreover, in the case of AE, SA, PSA, and MULMON, the models do not estimate view information but either treat them independently or use the observed view conditioning, rendering the MCC metric inapplicable. Fig. 12 showcases how the number of viewpoints impacts the identifiability of the $\mathbf{s}, \mathbf{v}$, and $\mathbf{c}$ variables; the involved experiments reflect the increase in performance with an increase in the number of views, across all benchmark datasets.

Additionally, we demonstrate our methodology on proposed complex datasets, MV-MoVIC and MV-MoVID, the latter dataset enables us to examine the model performs when the assumption 3.1 is not satisfied. To evaluate model behaviour in an environment with consistent objects but with different viewpoints, we conducted in-domain and out-of-domain (OOD) evaluations. For in-domain analysis, the model is trained and assessed on the same viewpoint group $A = [1, 2, 3]$. Conversely, for OOD evaluation, we consider the previously trained model but test it against a new set of viewpoints $B = [3, 4, 5]$. The findings presented in Table 2 regarding the MV-MoVIC dataset reveal that the SMCC, INV-SMCC, and MCC metrics show similar performance across both domains. This indicates that the distributional characteristics remain unchanged when both the training and testing environments contain the same objects. The MV-MoVID dataset analysis can be found in App. G.

5 CONCLUSION & DISCUSSION

Understanding when object-centric representations are both unambiguous and identifiable is essential for developing large-scale models with provable correctness guarantees. Unlike most existing work on identifiability, which largely focuses on single-view setups, we offer identifiability guarantees in multi-view scenarios. We use distributional assumptions for latent slot and view representations, drawing inspiration from mixture model-based structures. To achieve this, we propose a model that

is viewpoint-agnostic and does not require additional view-conditioning information. Our model specifically guarantees the identifiability of view-specific slot representations, viewpoint-invariant content representations, and view representations, all without the need for additional supervision (up to an equivalence relation). We visually validate our theoretical claims with unconfounded synthetic dataset with illustrative 2D data plots. We then empirically demonstrate the model's identifiability properties on multiple object-centric benchmarks, highlighting its ability to resolve view ambiguities in imaging applications. Furthermore, we showcase the scalability of our approach on large-scale datasets and more complex decoders using realistic datasets and transformer decoders, respectively, demonstrating its capacity to scale effectively with both data volume and decoder complexity.

REFERENCES

Kartik Ahuja, Yixin Wang, Divyat Mahajan, and Yoshua Bengio. Interventional causal representation learning. *arXiv preprint arXiv:2209.11924*, 2022.

Amir Arsalan Soltani, Haibin Huang, Jiajun Wu, Tejas D Kulkarni, and Joshua B Tenenbaum. Synthesizing 3d shapes via modeling multi-view depth maps and silhouettes with deep generative networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1511–1519, 2017.

Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

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

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.

Simon Buchholz, Michel Besserve, and Bernhard Schölkopf. Function classes for identifiable nonlinear independent component analysis. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 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.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Michael Chang, Thomas L Griffiths, and Sergey Levine. Object representations as fixed points: Training iterative refinement algorithms with implicit differentiation. *arXiv preprint arXiv:2207.00787*, 2022.

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.

Eric Crawford and Joelle Pineau. Exploiting spatial invariance for scalable unsupervised object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3684–3692, 2020.

Imant Daunhawer, Alice Bizeul, Emanuele Palumbo, Alexander Marx, and Julia E Vogt. Identifiability results for multimodal contrastive learning. *arXiv preprint arXiv:2303.09166*, 2023.

Nat Dilokthanakul, Pedro AM Mediano, Marta Garnelo, Matthew CH Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016.

Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2553–2560. IEEE, 2022.

Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.

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

Martin Engelcke, Adam R Kosiorek, Oiwi Parker Jones, and Ingmar Posner. Genesis: Generative scene inference and sampling with object-centric latent representations. *arXiv preprint arXiv:1907.13052*, 2019.

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.

SM Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S Morcos, Marta Garnelo, Avraham Ruderman, Andrei A Rusu, Ivo Danihelka, Karol Gregor, et al. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, 2018.

Tobias Gerstenberg, Noah D Goodman, David A Lagnado, and Joshua B Tenenbaum. A counterfactual simulation model of causal judgments for physical events. *Psychological review*, 128(5):936, 2021.

Alison Gopnik, Clark Glymour, David M Sobel, Laura E Schulz, Tamar Kushnir, and David Danks. A theory of causal learning in children: causal maps and bayes nets. *Psychological review*, 111(1): 3, 2004.

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.

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

Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

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

Geoffrey Hinton. How to represent part-whole hierarchies in a neural network. *Neural Computation*, pp. 1–40, 2022.

Matthew D Hoffman and Matthew J Johnson. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference, NIPS*, volume 1, 2016.

Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li F Fei-Fei, and Juan Carlos Niebles. Learning to decompose and disentangle representations for video prediction. *Advances in neural information processing systems*, 31, 2018.

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

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89, pp. 859–868. PMLR, 2019.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015.

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. In *Advances in Neural Information Processing Systems*, volume 33, 2020b. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/962e56a8a0b0420d87272a682bfd1e53-Paper.pdf.

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

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

Thomas Kipf, Gamaleldin F Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional object-centric learning from video. *arXiv preprint arXiv:2111.12594*, 2021.

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.

Adam Kosiorek, Hyunjik Kim, Yee Whye Teh, and Ingmar Posner. Sequential attend, infer, repeat: Generative modelling of moving objects. *Advances in Neural Information Processing Systems*, 31, 2018.

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.

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.

Nanbo Li, Cian Eastwood, and Robert Fisher. Learning object-centric representations of multi-object scenes from multiple views. *Advances in Neural Information Processing Systems*, 33:5656–5666, 2020.

Nanbo Li, Muhammad Ahmed Raza, Wenbin Hu, Zhaole Sun, and Robert Fisher. Object-centric representation learning with generative spatial-temporal factorization. *Advances in neural information processing systems*, 34:10772–10783, 2021.

Zhixuan Lin, Yi-Fu Wu, Skand Vishwanath Peri, Weihao Sun, Gautam Singh, Fei Deng, Jindong Jiang, and Sungjin Ahn. Space: Unsupervised object-oriented scene representation via spatial attention and decomposition. *arXiv preprint arXiv:2001.02407*, 2020.

Yu Liu, Baoxiong Jia, Yixin Chen, and Siyuan Huang. Slotlifter: Slot-guided feature lifting for learning object-centric radiance fields. In *European Conference on Computer Vision*, pp. 270–288. Springer, 2025.

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

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

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

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

Rundong Luo, Hong-Xing Yu, and Jiajun Wu. Unsupervised discovery of object-centric neural fields. *arXiv preprint arXiv:2402.07376*, 2024.

Gary F Marcus. *The algebraic mind: Integrating connectionism and cognitive science*. MIT press, 2003.

Joe Marino, Yisong Yue, and Stephan Mandt. Iterative amortized inference. In *International Conference on Machine Learning*, pp. 3403–3412. PMLR, 2018.

Emile Mathieu, Tom Rainforth, N Siddharth, and Yee Whye Teh. Disentangling disentanglement in variational autoencoders. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

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

Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, et al. Bridging the gap to real-world object-centric learning. *arXiv preprint arXiv:2209.14860*, 2022.

Gautam Singh, Fei Deng, and Sungjin Ahn. Illiterate dall-e learns to compose. *arXiv preprint arXiv:2110.11405*, 2021.

Gautam Singh, Yeongbin Kim, and Sungjin Ahn. Neural block-slot representations. *arXiv preprint arXiv:2211.01177*, 2022.

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.

Joshua Tobin, Wojciech Zaremba, and Pieter Abbeel. Geometry-aware neural rendering. *Advances in Neural Information Processing Systems*, 32, 2019.

Sjoerd Van Steenkiste, Michael Chang, Klaus Greff, and Jürgen Schmidhuber. Relational neural expectation maximization: Unsupervised discovery of objects and their interactions. *arXiv preprint arXiv:1802.10353*, 2018.

Sjoerd Van Steenkiste, Karol Kurach, Jürgen Schmidhuber, and Sylvain Gelly. Investigating object compositionality in generative adversarial networks. *Neural Networks*, 130:309–325, 2020.

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

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.

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

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

Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in neural information processing systems*, 29, 2016.

Xiaojiang Yang, Yi Wang, Jiacheng Sun, Xing Zhang, Shifeng Zhang, Zhenguo Li, and Junchi Yan. Nonlinear ICA using volume-preserving transformations. In *International Conference on Learning Representations*, 2022.

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.

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

A NOTATIONS

$\mathcal{O}^v$	: Abstract object set as observed from viewpoint v.
$[V] = \{1, \ldots, V\}$	: Exhaustive set of viewpoints, representing all possible views.
$A, B \subset [V]$	: Subsets of viewpoints, used for training.
$\mathcal{X} = \times_{v \in A} \mathcal{X}^v$	: Data space, formed by the Cartesian product of data spaces for each view in subset A.
$\mathbf{x}^A = \{\mathbf{x}^v : \forall v \in A\} \in \mathcal{X}$	: Data sample, where $\mathbf{x}^v$ is the data from view v, and $\mathbf{x}^A$ represents the set of data across all views in A.
f_e	: Encoder model maps input data to a latent space or feature representation.
$\mathbf{z}$	: Spatial latent features, representing inferred spatial properties from the data.
$\mathcal{S}$	: View-specific slot space, a space for features that are tied to particular viewpoints.
$\mathcal{C}$	: View-invariant content space, representing features that are constant across different viewpoints.
$\mathbf{s} \in \mathcal{S}$	: Samples from the view-specific slot space, representing view-dependent latent features.
$\mathbf{c} \in \mathcal{C}$	: Samples from the view-invariant content space, representing features that remain consistent across views.
$f_s, \tilde{f}_s$	: Slot attention module, responsible for attending to and disentangling different parts of the input related to different views.
$f_d, \tilde{f}_d$	: Mixing function, which combines view-specific and view-invariant features into a unified representation.
$\mathcal{V}$	: View information space, a space that encodes information specific to each viewpoint (e.g., angle, position).
$\mathbf{v} \in \mathcal{V}$	: A sample from the view information space representing a specific view or camera configuration.
$f_v, \tilde{f}_v$	: View extractor function, which extracts viewpoint-related information from the data.
$\boldsymbol{\mu}_c, \boldsymbol{\mu}_s, \boldsymbol{\mu}_v$	: Mean of invariant content, view-specific slots, and view distributions.
$\boldsymbol{\sigma}_c, \boldsymbol{\sigma}_s, \boldsymbol{\sigma}_v$	: Standard deviation of invariant content, view-specific slots, and view distributions.
$\boldsymbol{\pi}_c, \boldsymbol{\pi}_s, \boldsymbol{\pi}_v$	: Mixing coefficients of invariant content, view-specific slots, and view distributions.
A_{nk}	: Assignment confidence of a slot k getting mapped to token n.
$\boldsymbol{P} \in \mathcal{P} \subseteq \{0, 1\}^{K \times K}$	: Permutation matrix.
m_s	: Matching function, used to align object representations across views.
Δ^K	: Simplex in the space of dimension K.
$\mathcal{H}_x, \mathcal{H}_v$	: Space of homeomorphic transformation.

B Related Works

Identifiable representation learning. Learning meaningful representations from unlabeled data has long been a primary objective of deep learning (Bengio et al., 2013). Several approaches, such as those proposed by (Higgins et al., 2017; Kim & Mnih, 2018; Eastwood & Williams, 2018; Mathieu et al., 2019), relied on independence assumptions between latent variables to learn disentangled representations. However, Hyvärinen & Pajunen (1999); Locatello et al. (2019) demonstrated the provable impossibility of unsupervised methods for learning independent latent representations from i.i.d. data. Which is tackled by restricting mixing functions to conformal maps (Buchholz et al., 2022) or volume-preserving transformations (Yang et al., 2022), or with additional data assumptions (Zimmermann et al., 2021; Locatello et al., 2020a; Brehmer et al., 2022; Ahuja et al., 2022; Von Kügelgen et al., 2021), or by imposing structure in the latent space as in nonlinear Independent Component Analysis (ICA) (Hyvarinen et al., 2019; Khemakhem et al., 2020a;b), resulting in identifiable models. In the context of nonlinear ICA, Dilokthanakul et al. (2016) introduced a VAE model with a GMM prior, and Willetts & Paige (2021) empirically demonstrated the effectiveness of the GMM prior, which was later rigorously proven by Kivva et al. (2022). Kori et al. (2024) use this notion of latent GMM in the context of OCL, achieving identifiability guarantees for object-centric representations. Here, we use this notion in the context of multi-view object-centric representations, tackling the issues with spatial ambiguities and uncertainties in bindings.

Identifiable Object-centric learning. Extending nonlinear Independent Component Analysis (ICA) from representation learning to object-specific representational learning has been heavily explored before (Burgess et al., 2019; Engelcke et al., 2019; Greff et al., 2019) by employing an iterative variational inference approach (Marino et al., 2018), whereas Van Steenkiste et al. (2020); Lin et al. (2020) adopt more of a generative perspective, studied the effect of object binding and scene composition empirically. Recently, the use of iterative attention mechanisms has gained a significant interest (Locatello et al., 2020b; Engelcke et al., 2021; Singh et al., 2021; Wang et al., 2023; Singh et al., 2022; Emami et al., 2022). Most of these works operate in a single-view setting, which causes fundamental issues of viewpoint ambiguities in terms of occlusions and uncertainties in binding. Recent methods, including Eslami et al. (2018); Arsalan Soltani et al. (2017); Tobin et al. (2019); Wu et al. (2016) consider a single object from multiple views to tackle this particular problem. Additionally, Kosiorek et al. (2018); Hsieh et al. (2018); Li et al. (2020) explore multi-object binding in videos and multiple views, tackling object binding issues across frames. Despite their empirical effectiveness, most of these works lack formal identifiability guarantees. In line with recent efforts analysing theoretical guarantees in object-centric representations (Lachapelle et al., 2023; Brady et al., 2023; Kori et al., 2024), we formally investigate the modelling assumptions and their implications for achieving identifiability guarantees in the context of multi-object, multiview object-centric representation learning settings.

Multiview nonlinear ICA. It has been noted that addressing the challenge of nonlinear ICA can involve incorporating a learnable clustering task within the latent representations, thereby imposing asymmetry in the latent distribution (Willetts & Paige, 2021; Kivva et al., 2022). Moreover, Gresele et al. (2020) delve into multiview nonlinear ICA, particularly in scenarios involving corrupted observations, where they aim to recover invariant representations while accounting for certain ambiguities. Along similar lines, Daunhawer et al. (2023); Von Kügelgen et al. (2021) explore the concept of style-content identification using contrastive learning, focusing on addressing the multiview nonlinear ICA problem. Here, we work along similar lines by emphasising the learning of invariant content and identifiable object-centric representations. We achieve this by formulating a reconstruction objective where the enforced invariance and equivariance stem from the underlying probabilistic graphical model rather than relying on a contrastive learning objective. Similar to the noiseless setting in Gresele et al. (2020), we demonstrate the recovery of invariant content representations using different subsets of viewpoints.

Multi-view Object-centric learning. Recent progress in multi-view object-centric learning has seen notable contributions from methods like MulMON (Li et al., 2020), ROOTS (Chen et al., 2021), SlotLifter(Liu et al., 2025), and uOCF(Luo et al., 2024), each offering distinct approaches to compositional representation learning. However, these methods rely heavily on viewpoint annotations, which limit their applicability in fully unsupervised settings. MulMON refines object representations

iteratively using annotated viewpoint-image pairs, while ROOTS, SLOTLIFTER, UOCF estimates 3D object positions performing an inverse rendering operation within a grid and projects them into image space via viewpoint transformations. In contrast, we deal with fully unsupervised framework without the need of viewpoint annotations while providing approximate viewpoint equivariance for object representations.

Temporal Object-centric learning. An alternative approach to bypass the need for viewpoint annotations leverages temporal information. Methods for learning from single-viewpoint video sequences, such as Relational N-EM (Van Steenkiste et al., 2018), SQAIR (Kosiorek et al., 2018), SILOT (Crawford & Pineau, 2020), and SAVI (Kipf et al., 2021), focus on modeling object motion, interactions, and identity tracking across frames, even under occlusion. However, these methods assume fixed viewpoints, making them unsuitable for multi-view scenarios where objects appear in different spatial configurations. Additionally, object motion affects individual objects independently, unlike viewpoint changes, which influence the entire scene. Recent advances such as DYMON (Li et al., 2021) extend multi-view approaches like MULMON (Li et al., 2020) to dynamic scenes by disentangling object motion and viewpoint changes, assuming one dominates in adjacent frames. However, DYMON relies on viewpoint annotations, limiting its utility in unsupervised settings. Temporal methods such as SIMONE (Luo et al., 2024) address this by leveraging temporal coherence across multi-view videos, using spatial and temporal positional embeddings to disentangle object and viewpoint representations. Yet, SIMONe's reliance on temporal continuity restricts its generalizability to scenarios where such coherence is absent. In contrast, our framework does not assume temporal dependencies.

C ALGORITHM

Here we illustrate all the steps involved in the of proposed method VISA, refer 1.

Algorithm 1 View Invariant Slot Attention VISA

Input: $A \in [V]$, $\mathbf{z}^A = \{f_e(\mathbf{x}^v)\ \forall v \in A\} \in \mathbb{R}^{|A|\times N\times d}$ ▷ input representations
View: $\mathbf{v}^A = \{\mathbf{v}^v \sim \mathcal{N}(\mathbf{v}^v; \boldsymbol{\mu}(\mathbf{z}^v), \boldsymbol{\sigma}^2(\mathbf{z}^v))\ \forall v \in A\} \in \mathbb{R}^{|A|\times d}$ ▷ view representations
View Transformation: $\theta^A = \{\theta^v = \mathrm{STN}(\mathbf{v}^v)\ \forall v \in A\} \in \mathbb{R}^{|A|\times 2\times 3}$ ▷ transformation parameters
$\mathrm{key}^A \leftarrow \boldsymbol{W}_k \mathcal{T}_{\theta^v}^{-1}(\mathbf{z}^A) \in \mathbb{R}^{|A|\times N\times d}$, $\mathrm{value}^A \leftarrow \boldsymbol{W}_v \mathcal{T}_{\theta^v}^{-1}(\mathbf{z}^A) \in \mathbb{R}^{|A|\times N\times d}$ ▷ optional value := key
$\mathbf{s} \leftarrow \emptyset$; $\hat{\boldsymbol{\pi}} \leftarrow \emptyset$
for $v \in A$ **do**
  $\forall k, \boldsymbol{\pi}(0)_k \leftarrow 1/K$, $\boldsymbol{\mu}(0)_k \sim \mathcal{N}(0, \mathbf{I}_d)$, $\boldsymbol{\sigma}(0)_k^2 \leftarrow \mathbf{1}_d$
  for $t = 0 \rightarrow T-1$ **do**
    $A_{nk} \leftarrow \frac{\boldsymbol{\pi}(t)_k \mathcal{N}\left(\mathrm{key}_n; \boldsymbol{W}_q \boldsymbol{\mu}(t)_k, \boldsymbol{\sigma}(t)_k^2\right)}{\sum_{j=1}^{K} \boldsymbol{\pi}(t)_j \mathcal{N}\left(\mathrm{key}_n; \boldsymbol{W}_q \boldsymbol{\mu}(t)_j, \boldsymbol{\sigma}(t)_j^2\right)}$ ▷ compute attention
    $\hat{A}_{nk} \leftarrow \frac{A_{nk}}{\sum_{l=1}^{N} A_{lk}}$ ▷ normalize attention
    $\boldsymbol{\mu}(t+1)_k \leftarrow \sum_{n=1}^{N} \hat{A}_{nk} \cdot \mathrm{value}_n$ ▷ update slot mean
    $\boldsymbol{\sigma}(t+1)_k^2 \leftarrow \sum_{n=1}^{N} \hat{A}_{nk} \cdot (\mathrm{value}_n - \boldsymbol{\mu}(t+1)_k)^2$ ▷ update slot variance
    $\boldsymbol{\pi}(t+1)_k \leftarrow \frac{1}{N} \sum_{n=1}^{N} A_{nk}$ ▷ update mixing coefficient
  end for
  $\mathbf{s} \leftarrow \mathbf{s} \cup \{(\boldsymbol{\mu}_{1:K}(T), \boldsymbol{\sigma}_{1:K}^2(T))\}$; $\hat{\boldsymbol{\pi}} \leftarrow \hat{\boldsymbol{\pi}} \cup \{\boldsymbol{\pi}_{1:K}(T)\}$ ▷ slot collection
end for
return $\mathrm{ConvexCombination}(\mathbf{s}, \hat{\boldsymbol{\pi}})$ ▷ K view invariant content =0

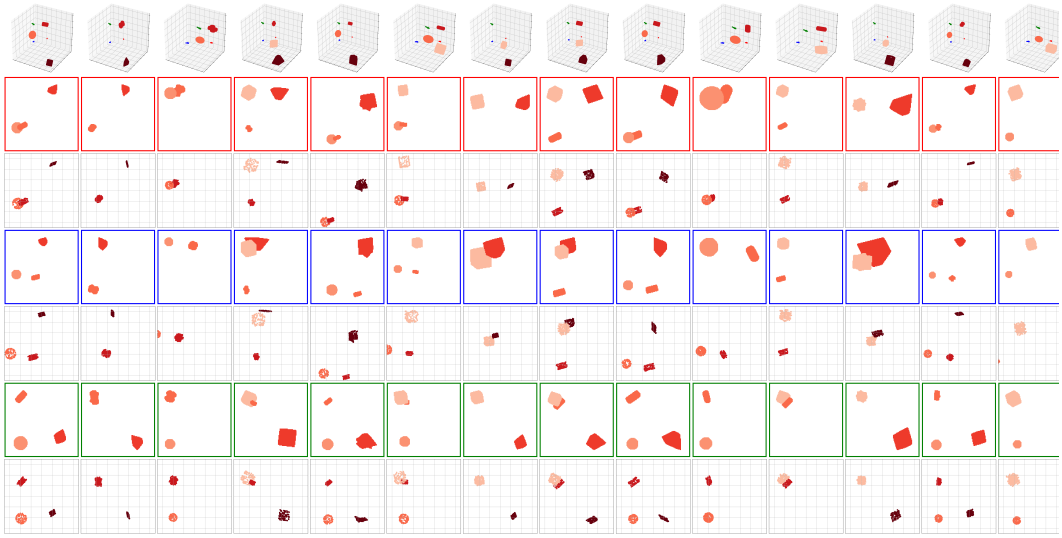


Figure 7: **Data generating process:** The figure illustrates 3D point cloud data in the first row, with camera location highlighted in red, blue, and green arrow. Following rows indicates projected images and point cloud as observed from red, blue, and green cameras, respectively.

D DATASETS

D.1 ILLUSTRATIVE DATASET

To visually illustrate the effectiveness of our theory we experiment with two dimensional illustrative dataset. For this, similar to Kori et al. (2024), we defined a $K = 5$ component GMM, with differing mean parameters $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_5\}$, and shared isotropic covariances, which we use to sample locations for an object. For a given location, we randomly select one object from `{'cube', 'cylinder', 'pyramid', 'sphere'}` and generate 1000 random points on the surface of the selected shape uniformly covering it. To create a single data point, we randomly select three of the five locations and place a randomly selected object at the location. To include multiple viewpoints, we consider V camera location and project the objects, creating V different scenes. We then fill this by considering convex hull operation resulting in projected images as illustrated in Fig. 7. To maintain uniformity, we only use imaging modality in the main paper while also demonstrating point cloud illustrations here in the appendix. We use different colours representing different objects in Fig. 8, **??** and used $10,000$ data points in total to train our toy models. Unlike existing benchmark datasets, here we remove all the confounding effects caused by lighting and depth. This provides an ideal test bed to validate all our theoretical claims.

D.2 PROPOSED DATASET

In this work, we introduce the MV-MOVI datasets, created using Kubric Greff et al. (2022), which feature multi-view scenes with segmentation annotations. We propose two variants of the dataset: MV-MOVIC, where the camera locations for every viewpoint remain fixed across all scenes, and MV-MOVID, where the camera locations dynamically change for each scene.

Both MV-MOVIC and MV-MOVID primarily consist of scenes generated by randomly selecting a background from a set of **458** available options and choosing K objects, where $3 \leq K \leq 6$, from a pool of **930** objects. In total, a significantly high number of images can be generated in general. In contrast, for this work, we generate **72,000** scenes, each captured from 5 different viewpoints, with object segmentation masks for every view to facilitate the evaluation of model performance. In the case of MV-MOVIC, the locations of all five cameras are fixed across the $72,000$ scenes, while in MV-MOVID, the camera positions are dynamically sampled and vary across scenes.

E MASK GENERATION

In the case of additive decoders, the decoder outputs K three channelled tensors along with K single channelled mask. We consider normalising these masks with `softmax` transformation along slot dimension, ensuring that each pixel only contributes to a single slot. The resulting softmaxed masks are used in composing ($\text{image} = \sum_k \text{mask}_k \cdot \text{image}_k$) the slots to reconstruct an image for training. During inference, we normalise masks with `sigmoid` transformation, allowing us to estimate occluded objects visually, resolving the spatial ambiguities with occluded objects. In a later section, we illustrate the results with both `softmax` and `sigmoid` transformations.

E.1 ADDITIVITY IMPLICATIONS

Definition E.1. (Additive models) Function f is considered to be an additive decoder if, for any object decoders f_{obj} and masking mechanism m_{obj}, if they can be expressed as:

$$f(\mathbf{z}) = \sum_{k \in [K]} m_{\text{obj}}(\mathbf{z}_k) \odot f_{\text{obj}}(\mathbf{z}_k) \tag{7}$$

As pointed out in Lachapelle et al. (2023), `softmax`-based masks do not truly fall under the category of additive decoders due to the competition between masks for groups of pixels. This implies that the additive decoders studied in Lachapelle et al. (2023) are not expressive enough to represent the "masked decoders" typically employed in object-centric representation learning. The issue arises from the normalization of alpha masks, and care must be taken when extrapolating the findings from Lachapelle et al. (2023) to the models used in practice.

Although `sigmoid`-based masks satisfy the condition of additivity during inference, it is important to note that the model is still trained using `softmax` normalization in our setting. The effect of using `sigmoid` masks during inference can be visually observed in App. G.

F PROOFS

Lemma F.1 (ELBO). *With prior distributions $p(\mathbf{v})$ and $p(\mathbf{c})$ for view and content latent random variables, the likelihood $p(\mathbf{x})$ can be maximised by maximising the following expression:*

$$\log p(\mathbf{x}) \geq \mathbb{E} \log p(\mathbf{x} \mid \mathcal{T}_{\theta^v}(\mathbf{c}), \mathbf{v}) - \text{KL}\left(q(\mathbf{v} \mid \mathbf{x}) \parallel p(\mathbf{v})\right) := \text{ELBO}(\mathbf{x}) \tag{8}$$

Proof. Considering the generative model in Eqn. 1 respecting the graphical model in Fig. 4, we get:

$$p(\mathbf{x}) = \iint p(\mathbf{x}^A \mid \mathcal{T}_{\theta^v}(\mathbf{c}), \mathbf{v}^A)\, p(\mathbf{c})\, p(\mathbf{v}^A)\, d\mathbf{v}\, d\mathbf{c} \tag{9}$$

$$\log p(\mathbf{x}) = \log \iint p(\mathbf{x}^A \mid \mathbf{c}_{1:K}, \mathbf{v}^A)\, p(\mathbf{c}_{1:K})\, p(\mathbf{v}^A) \frac{q(\mathbf{v}, \mathbf{c} \mid \mathbf{x}^A)}{q(\mathbf{v}, \mathbf{c} \mid \mathbf{x}^A)}\, d\mathbf{v}\, d\mathbf{c}_{1:K} \tag{10}$$

$$\geq \iint q(\mathbf{v}^A \mid \mathbf{x}^A) q(\mathbf{c}_{1:K} \mid \mathcal{T}_{\theta^v}^{-1}(\mathbf{x}^A)) \log p(\mathbf{x}^A \mid \mathcal{T}_{\theta^v}(\mathbf{c})_{1:K}, \mathbf{v}^A) \frac{p(\mathbf{v}^A_{1:K})}{q(\mathbf{v}^A \mid \mathbf{x}^A)} \frac{p(\mathbf{c}^A_{1:K})}{q(\mathbf{c}_{1:K} \mid \mathbf{x}^A)}\, d\mathbf{v}^A\, d\mathbf{c}_{1:K} \tag{11}$$

$$= \sum_{v \in A} \iint q(\mathbf{v}^v \mid \mathbf{x}^v) q(\mathbf{c}_{1:K} \mid \mathcal{T}_{\theta^v}^{-1}(\mathbf{x}^v))) \log p(\mathbf{x}^v \mid \mathcal{T}_{\theta^v}(\mathbf{c})_{1:K}, \mathbf{v}^v) \frac{p(\mathbf{v}^A_{1:K})}{q(\mathbf{v}^A \mid \mathbf{x}^v)} \frac{p(\mathbf{c}^A_{1:K})}{q(\mathbf{c}_{1:K} \mid \mathbf{x}^v)}\, d\mathbf{v}^v\, d\mathbf{c}_{1:K} \tag{12}$$

Given the iterative update for $\mathbf{c}$ with EM algorithm, ideally we expect posterior to converge to prior, which results in:

$$\log p(\mathbf{x}) = \sum_{v \in A} \iint q(\mathbf{v}^v \mid \mathbf{x}^v) q(\mathbf{c}_{1:K} \mid \mathcal{T}_{\theta^v}^{-1}(\mathbf{x}^v))) \log p(\mathbf{x}^v \mid \mathcal{T}_{\theta^v}(\mathbf{c})_{1:K}, \mathbf{v}^v) \frac{p(\mathbf{v}_{1:K}^A)}{q(\mathbf{v}^A \mid \mathbf{x}^v)} \; d\mathbf{v}^v \; d\mathbf{c}_{1:K} \tag{13}$$

$$= \sum_{v \in A} \mathbb{E}_{\mathbf{c},\mathbf{v}} \log p(\mathbf{x}^v \mid \mathcal{T}_{\theta^v}(\mathbf{c}), \mathbf{v}) - \mathrm{KL}\left(q(\mathbf{v} \mid \mathbf{x}^v) \parallel p(\mathbf{v})\right) \tag{14}$$

Given the subscript notation, the above expression can also be expressed as:

$$\mathbb{E}_{\mathbf{c},\mathbf{v}} \log p(\mathbf{x} \mid \mathcal{T}_{\theta^v}(\mathbf{c}), \mathbf{v}) - \mathrm{KL}\left(q(\mathbf{v} \mid \mathbf{x}) \parallel p(\mathbf{v})\right) := \mathrm{ELBO}(\mathbf{x}) \tag{15}$$

□

Lemma F.2 (Mean GMM)**.** *Let $\mathbf{z} \in \mathbb{R}^{N \times d}$ be a random variable drawn from a GMM with K components:*

$$\mathbf{z} \sim \sum_{k=1}^{K} \boldsymbol{\pi}_k \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{16}$$

where $\boldsymbol{\pi}_k$ are the mixture weights, $\boldsymbol{\mu}_k \in \mathbb{R}^d$ are the mean vectors, and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{d \times d}$ are the covariance matrices. Assuming the mixture satisfies the ICA assumption, such that the components of $\mathbf{z}$ are statistically independent. A projected random variable $\bar{z}$ as the average over the dimensions of $\mathbf{z}$:

$$\bar{z} = \frac{1}{d} \sum_{j=1}^{d} \mathbf{z}_j, \tag{17}$$

is also distributed according to a GMM with K components, with appropriately transformed means and variances.

Proof. Given the random variable $\mathbf{z}$ follows a GMM, so its density can be expressed as:

$$p(\mathbf{z}) = \sum_{k=1}^{K} \boldsymbol{\pi}_k \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{18}$$

where:

$$\boldsymbol{\mu}_k = [\mu_{k,1}, \mu_{k,2}, \ldots, \mu_{k,d}]^\top; \qquad \boldsymbol{\Sigma}_k = \mathrm{diag}([\sigma_{k,1}^2, \sigma_{k,2}^2, \ldots, \sigma_{k,d}^2]). \tag{19}$$

Considering, the projection of $\mathbf{z}$ onto $\bar{z}$ is defined as:

$$\bar{z} = \frac{1}{d} \sum_{j=1}^{d} \mathbf{z}_j. \tag{20}$$

Given the ICA assumption, the components $\mathbf{z}_{:,j}$ are independent. For a fixed component k, the projected mean and variance of $\bar{z}$ can be derived as:

$$\mathbb{E}[\bar{z}] = \frac{1}{d} \sum_{j=1}^{d} \mu_{k,j}; \qquad \mathrm{Var}(\bar{z}) = \frac{1}{d^2} \sum_{j=1}^{d} \sigma_{k,j}^2. \tag{21}$$

Since the projection $\bar{z}$ is a linear combination of independent Gaussian variables, $\bar{z}$ remains Gaussian for each component k. Thus, the overall distribution of $\bar{z}$ is also a GMM:

$$p(\bar{z}) = \sum_{k=1}^{K} \boldsymbol{\pi}_k \mathcal{N}(\bar{z}; \mu_{\bar{z},k}, \sigma_{\bar{z},k}^2), \tag{22}$$

where:

$$\mu_{\bar{z},k} = \frac{1}{d} \sum_{j=1}^{d} \mu_{k,j}; \qquad \sigma_{\bar{z},k}^2 = \frac{1}{d^2} \sum_{j=1}^{d} \sigma_{k,j}^2. \tag{23}$$

This concludes the proof.

□

Lemma F.3 (Convex Combination of GMMs). *Let* $\mathbf{s}^1 = \{\mathbf{s}_1^1, \ldots, \mathbf{s}_K^1\}$ *and* $\mathbf{s}^2 = \{\mathbf{s}_1^2, \ldots, \mathbf{s}_K^2\}$ *be two sets of* K *random vectors in* $\mathbb{R}^d$*, each distributed according to GMMs:*

$$\mathbf{s}^1 \sim \sum_{k=1}^{K} \boldsymbol{\pi}_{1,k} \mathcal{N}(\boldsymbol{\mu}_{1,k}, \boldsymbol{\Sigma}_{1,k}); \quad \mathbf{s}^2 \sim \sum_{k=1}^{K} \boldsymbol{\pi}_{2,k} \mathcal{N}(\boldsymbol{\mu}_{2,k}, \boldsymbol{\Sigma}_{2,k}) \tag{24}$$

where $\boldsymbol{\mu}_{i,k} \in \mathbb{R}^d$*,* $\boldsymbol{\Sigma}_{i,k} \in \mathbb{R}^{d \times d}$*, and* $\boldsymbol{\pi}_{i,k}$ *are the means, covariances, and mixing coefficients respectively.*

Then for any weights $w_1, w_2 \in \mathbb{R}$ *such that* $w_1 + w_2 = 1$*, the convex combination* $\mathbf{s} = w_1 \mathbf{s}^1 + w_2 \mathbf{s}^2$ *is also distributed according to a GMM with* K *components.*

Proof. Without loss of generality, assume the components of both GMMs are aligned. For each component k, we derive the parameters of the resulting mixture:

The mixing coefficients of the resulting GMM are weighted combinations of the original coefficients:

$$\tilde{\boldsymbol{\pi}}_k = w_1 \boldsymbol{\pi}_{1,k} + w_2 \boldsymbol{\pi}_{2,k} \tag{25}$$

For each component k, the convex combination of Gaussians results in a Gaussian distribution. The mean of the resulting Gaussian is:

$$\tilde{\boldsymbol{\mu}}_k = \frac{w_1 \boldsymbol{\pi}_{1,k} \boldsymbol{\mu}_{1,k} + w_2 \boldsymbol{\pi}_{2,k} \boldsymbol{\mu}_{2,k}}{\tilde{\boldsymbol{\pi}}_k} \tag{26}$$

The covariance of the resulting Gaussian for each component k can be derived as follows. Firstly, recall that for a random variable X, the covariance is:

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top] = \mathbb{E}[XX^\top] - \mathbb{E}[X]\mathbb{E}[X]^\top \tag{27}$$

First lets compute $\mathbb{E}[\mathbf{s}_k \mathbf{s}_k^\top]$:

$$\mathbb{E}[\mathbf{s}_k \mathbf{s}_k^\top] = \mathbb{E}\left[\left(\frac{w_1 \boldsymbol{\pi}_{1,k} \mathbf{s}_k^1 + w_2 \boldsymbol{\pi}_{2,k} \mathbf{s}_k^2}{\tilde{\boldsymbol{\pi}}_k} \right) \left(\frac{w_1 \boldsymbol{\pi}_{1,k} \mathbf{s}_k^1 + w_2 \boldsymbol{\pi}_{2,k} \mathbf{s}_k^2}{\tilde{\boldsymbol{\pi}}_k} \right)^\top \right] \tag{28}$$

$$= \frac{w_1^2 \boldsymbol{\pi}_{1,k}^2 \mathbb{E}[\mathbf{s}_k^1 (\mathbf{s}_k^1)^\top] + w_2^2 \boldsymbol{\pi}_{2,k}^2 \mathbb{E}[\mathbf{s}_k^2 (\mathbf{s}_k^2)^\top]}{(\tilde{\boldsymbol{\pi}}_k)^2} \tag{29}$$

$$+ \frac{w_1 w_2 \boldsymbol{\pi}_{1,k} \boldsymbol{\pi}_{2,k}}{(\tilde{\boldsymbol{\pi}}_k)^2} \mathbb{E}[\mathbf{s}_k^1 (\mathbf{s}_k^2)^\top + \mathbf{s}_k^2 (\mathbf{s}_k^1)^\top] \tag{30}$$

Then, substitute known expectations:

$$\mathbb{E}[\mathbf{s}_k^i (\mathbf{s}_k^i)^\top] = \boldsymbol{\Sigma}_{i,k} + \boldsymbol{\mu}_{i,k} \boldsymbol{\mu}_{i,k}^\top \tag{31}$$

$$\mathbb{E}[\mathbf{s}_k^1 (\mathbf{s}_k^2)^\top] = \boldsymbol{\mu}_{1,k} \boldsymbol{\mu}_{2,k}^\top \tag{32}$$

Finally, by subtract $\mathbb{E}[\mathbf{s}_k^i]\mathbb{E}[\mathbf{s}_k^i]^T = \tilde{\boldsymbol{\mu}}_k \tilde{\boldsymbol{\mu}}_k^\top$ we get the covariance:

$$\tilde{\boldsymbol{\Sigma}}_k = \frac{w_1^2 \boldsymbol{\pi}_{1,k}^2 \boldsymbol{\Sigma}_{1,k} + w_2^2 \boldsymbol{\pi}_{2,k}^2 \boldsymbol{\Sigma}_{2,k}}{(\tilde{\boldsymbol{\pi}}_k)^2} \tag{33}$$

To verify this forms a non degenerate GMM, we show the mixing coefficients sum to 1:

$$\sum_{k=1}^{K} \tilde{\boldsymbol{\pi}}_k = \sum_{k=1}^{K} (w_1 \boldsymbol{\pi}_{1,k} + w_2 \boldsymbol{\pi}_{2,k}) \tag{34}$$

$$= w_1 \sum_{k=1}^{K} \boldsymbol{\pi}_{1,k} + w_2 \sum_{k=1}^{K} \boldsymbol{\pi}_{2,k} \tag{35}$$

$$= w_1 \cdot 1 + w_2 \cdot 1 = 1 \tag{36}$$

Therefore, the convex combination results in a valid Gaussian mixture model with K components, where each component has mean $\tilde{\boldsymbol{\mu}}_k$, covariance $\tilde{\boldsymbol{\Sigma}}_k$, and mixing coefficient $\tilde{\boldsymbol{\pi}}_k$. □

Lemma 3.2. (Optimal Content Mixture) For $A \in [V]$, given the a local content distribution $q(\mathbf{c}_{1:K} \mid \mathbf{s}^A_{1:K}, \mathbf{x}^A)$ (per-scene $\mathbf{x}^A \in \{\mathbf{x}^A_i\}_{i=1}^M$), which can be expressed as a GMM with K components, the aggregate posterior $q(\mathbf{c})$ is obtained by marginalizing out $\mathbf{x}, \mathbf{s}$ is a non-degenerate global Gaussian mixture with MK components:

$$p(\mathbf{c}) = q(\mathbf{c}) = \frac{1}{M} \sum_{i=1}^{M} \sum_{k=1}^{K} \widehat{\boldsymbol{\pi}}_{ik} \mathcal{N}\left(\mathbf{c}; \widehat{\boldsymbol{\mu}}_{ik}, \widehat{\boldsymbol{\sigma}}^2_{ik}\right). \tag{37}$$

Proof. We extend the proof in Kori et al. (2024), by incorporating hierarchical slot to aggregate content formalisation. For which, we begin by noting that the aggregate posterior $q(\mathbf{c})$ is the optimal prior $p(\mathbf{c})$ so long as our posterior approximation $q(\mathbf{c} \mid \mathbf{s}^A, \mathbf{x}^A)$ is close enough to the true posterior $p(\mathbf{c} \mid \mathbf{s}^A, \mathbf{x}^A)$, since for a dataset $\mathbf{x}^A \in \{\mathbf{x}^A_i\}_{i=1}^M$, for which we start with $q(\mathbf{s}^A \mid \mathbf{x}^A)$, wlog, given view point transformation is deterministic, we consider $\mathbf{x}^A = \mathcal{T}_{\theta^A}(\mathbf{x}^A)$ we have that:

$$\begin{aligned}
p(\mathbf{s}^A) &= \int p(\mathbf{s}^A \mid \mathbf{x}^A) p(\mathbf{x}^A) d\mathbf{x}^A && (38)\\
&= \mathbb{E}_{\mathbf{x}^A \sim p(\mathbf{x}^A)} \left[p(\mathbf{s}^A \mid \mathbf{x}^A) \right] && (39)\\
&\approx \frac{1}{M} \sum_{i=1}^{M} p(\mathbf{s}^A \mid \mathbf{x}^A_i) \quad \text{(empirical approximation)} && (40)\\
&\approx \frac{1}{M} \sum_{i=1}^{M} q(\mathbf{s}^A \mid \mathbf{x}^A_i) \quad \text{(posterior approximation)} && (41)\\
&=: q(\mathbf{s}^A), && (42)
\end{aligned}$$

We further extend this to $q(\mathbf{c})$, with the result from Lemma F.3, we know that the $q(\mathbf{c} \mid \mathbf{s}^A)$ is a GMM with same number of components as $q(\mathbf{s}^v \mid \mathbf{s}^v)$ for any $v \in [V]$ as follows

$$\begin{aligned}
p(\mathbf{c}) &= \int p(\mathbf{c} \mid \mathbf{s}^A) p(\mathbf{s}^A) d\mathbf{s}^A && (43)\\
&= \mathbb{E}_{\mathbf{s}^A \sim p(\mathbf{s}^A)} \left[p(\mathbf{c} \mid \mathbf{s}^A) \right] && (44)\\
&\approx \frac{1}{M} \sum_{i=1}^{M} p(\mathbf{c} \mid \mathbf{s}^A_i) && (45)\\
&\approx \frac{1}{M} \sum_{i=1}^{M} q(\mathbf{c} \mid \mathbf{s}^A_i) && (46)\\
&=: q(\mathbf{c}), && (47)
\end{aligned}$$

where we approximated $p(\mathbf{x})$ using the empirical distribution, then substituted in the approximate posterior, marginalizing $\mathbf{x}$ to get $p(\mathbf{s})$, we later consider the distributional form of $p(\mathbf{s})$ and marginalise $\mathbf{s}^A$ to get $p(\mathbf{c})$. This observation was first made by Hoffman & Johnson (2016) and was used in Kori et al. (2024) we use it to motivate our setup. Given our model fits a local GMM to each latent variable sampled from the approximate posterior: $\mathbf{z}^A \sim q(\mathbf{z}^A \mid \mathbf{x}^A_i)$, for $i = 1, \ldots, M$. Let $f_s(\mathbf{z}^A)$ denote

the (local) the product of GMM density, its expectation is given by:

$$\mathbb{E}_{p(\mathbf{x}^A),q(\mathbf{z}^A|\mathbf{x}^A)}\left[f_s(\mathbf{z}^A)\right] = \iint p(\mathbf{x}^A)q(\mathbf{z}^A \mid \mathbf{x}^A)f_s(\mathbf{z}^A)d\mathbf{x}^A d\mathbf{z}^A \tag{48}$$

$$\approx \iint \frac{1}{M}\sum_{i=1}^{M}\delta(\mathbf{x}^A - \mathbf{x}_i^A)q(\mathbf{z}^A \mid \mathbf{x}^A)f(\mathbf{z}^A)d\mathbf{x}^A d\mathbf{z}^A \tag{49}$$

$$= \int \frac{1}{M}\sum_{i=1}^{M} q(\mathbf{z}^A \mid \mathbf{x}_i^A)f(\mathbf{z}^A)d\mathbf{z}^A \tag{50}$$

$$= \int \frac{1}{M}\sum_{i=1}^{M}\mathcal{N}\big(\mathbf{z}^A;\boldsymbol{\mu}(\mathbf{x}_i^A),\boldsymbol{\sigma}^2(\mathbf{x}_i^A)\big)\cdot\sum_{k=1}^{K}\boldsymbol{\pi}_k(\mathbf{x}_i^A)\mathcal{N}\left(\mathbf{z}^A;\boldsymbol{\mu}_k(\mathbf{x}_i^A),\boldsymbol{\sigma}_k^2(\mathbf{x}_i^A)\right)d\mathbf{z}^A$$

$$\approx \int \frac{1}{M}\sum_{i=1}^{M}\delta(\mathbf{z}^A - \boldsymbol{\mu}(\mathbf{x}_i^A))\cdot\sum_{k=1}^{K}\boldsymbol{\pi}_k(\mathbf{x}_i^A)\mathcal{N}\left(\mathbf{z}^A;\boldsymbol{\mu}_k(\mathbf{x}_i^A),\boldsymbol{\sigma}_k^2(\mathbf{x}_i^A)\right)d\mathbf{z}^A \tag{51}$$

$$= \frac{1}{M}\sum_{i=1}^{M}\sum_{k=1}^{K}\boldsymbol{\pi}_k(\mathbf{x}_i^A)\mathcal{N}\left(\mathbf{z}^A;\boldsymbol{\mu}_k(\mathbf{x}_i^A),\boldsymbol{\sigma}_k^2(\mathbf{x}_i^A)\right) \tag{52}$$

$$=: q(\mathbf{z}^A), \tag{53}$$

where we again used the empirical distribution approximation of $p(\mathbf{x})$, and the following basic identity of the Dirac delta to simplify: $\int \delta(\mathbf{x}-\mathbf{x}')f_e(\mathbf{x})d\mathbf{x} = f_e(\mathbf{x}')$.

For the general case, however, we must instead compute the product of $q(\mathbf{z}^A \mid \mathbf{x}^A)$ and $f_s(\mathbf{z}^A)$ rather than use a Dirac delta approximation as in Eqn. 51. To that end we may proceed as follows w.r.t. to each datapoint $\mathbf{x}_i^A$:

$$q(\mathbf{z}^A \mid \mathbf{x}_i^A)\cdot f_s(\mathbf{z}^A) = \mathcal{N}\big(\mathbf{z}^A;\boldsymbol{\mu}(\mathbf{x}_i^A),\boldsymbol{\sigma}^2(\mathbf{x}_i^A)\big)\cdot\sum_{k=1}^{K}\boldsymbol{\pi}_k(\mathbf{x}_i^A)\mathcal{N}\left(\mathbf{z}^A;\boldsymbol{\mu}_k(\mathbf{x}_i^A),\boldsymbol{\sigma}_k^2(\mathbf{x}_i^A)\right) \tag{54}$$

$$= \sum_{k=1}^{K}\boldsymbol{\pi}_k(\mathbf{x}_i^A)\left[\mathcal{N}\big(\mathbf{z}^A;\boldsymbol{\mu}(\mathbf{x}_i^A),\boldsymbol{\sigma}^2(\mathbf{x}_i^A)\big)\cdot\mathcal{N}\left(\mathbf{z}^A;\boldsymbol{\mu}_k(\mathbf{x}_i^A),\boldsymbol{\sigma}_k^2(\mathbf{x}_i^A)\right)\right] \tag{55}$$

Given that means across all views are aligned, similar to Lemma F.3, we know the resulting combined GMM has same number of components:

$$q(\mathbf{z}^A \mid \mathbf{x}_i^A)\cdot f_s(\mathbf{z}^A) = \prod_{v=1}^{|A|}\sum_{k=1}^{K}\bar{\boldsymbol{\pi}}_{ik}^v\mathcal{N}\left(\mathbf{z};\bar{\boldsymbol{\mu}}_{ivk},\bar{\boldsymbol{\sigma}}_{ivk}^2\right), \tag{56}$$

Given the product of GMM is a GMM with the number of components equal to the product of a number of components in individual GMM, however in our setting we consider all the components in individual GMM across viewpoints to be aligned resulting in GMM with a number of components equal to the sum of individual components which in our case correspond to K. The posterior parameters of the resulting mixture are given in closed form by:

$$\bar{\boldsymbol{\sigma}}_{ivk}^2 = \left(\frac{1}{\boldsymbol{\sigma}_k^2(\mathbf{x}_i^v)} + \frac{1}{\boldsymbol{\sigma}^2(\mathbf{x}_i^v)}\right)^{-1}, \qquad \bar{\boldsymbol{\mu}}_{ivk} = \bar{\boldsymbol{\sigma}}_{ivk}^2\left(\frac{\boldsymbol{\mu}(\mathbf{x}_i^v)}{\boldsymbol{\sigma}^2(\mathbf{x}_i^v)} + \frac{\boldsymbol{\mu}_k(\mathbf{x}_i^v)}{\boldsymbol{\sigma}_k^2(\mathbf{x}_i^v)}\right), \tag{57}$$

The resulting GMM is still on the view-specific slots, the aggregation of these slots to obtain content vectors marginalises the viewpoint-level information with convex combination of parameters across all the viewpoints considered as described in cf. F.3, results in:

$$\prod_{v=1}^{|A|}\sum_{k=1}^{K}\bar{\boldsymbol{\pi}}_{ik}^v\mathcal{N}\left(\mathbf{z};\bar{\boldsymbol{\mu}}_{ivk},\bar{\boldsymbol{\sigma}}_{ivk}^2\right) = \sum_{k=1}^{K}\widehat{\boldsymbol{\pi}}_{ik}\mathcal{N}\left(\mathbf{z};\widehat{\boldsymbol{\mu}}_{ik},\widehat{\boldsymbol{\sigma}}_{ik}^2\right), \tag{58}$$

$$\widehat{\boldsymbol{\sigma}}_{ik}^2 = g(\bar{\boldsymbol{\sigma}}_{ik}, \bar{\boldsymbol{\pi}}_{ik}) = \sum_{v=1}^{|A|} \left(\frac{\bar{\boldsymbol{\pi}}_{ik}^v}{\sum_{v=1}^{|A|} \bar{\boldsymbol{\pi}}_{ik}^v} \right)^2 \bar{\boldsymbol{\sigma}}_{ik}^2, \tag{59}$$

$$\widehat{\boldsymbol{\mu}}_{ivk} = g(\bar{\boldsymbol{\mu}}_{ik}, \bar{\boldsymbol{\pi}}_{ik}) = \sum_{v=1}^{|A|} \frac{\bar{\boldsymbol{\pi}}_{ik}^v}{\sum_{v=1}^{|A|} \bar{\boldsymbol{\pi}}_{ik}^v} \bar{\boldsymbol{\mu}}_{ik}, \tag{60}$$

Now to show that the resulting GMM is non-degenerate we need to show $\sum_{k=1}^{K} \widehat{\boldsymbol{\pi}}_{ik} = 1$, for $i = 1, 2, \ldots, M$. Based on Eqn. 52:

$$\implies \frac{1}{M} \sum_{i=1}^{M} \sum_{k=1}^{K} \widehat{\boldsymbol{\pi}}_{ik} = \frac{1}{M|A|} \sum_{i=1}^{M} \sum_{k=1}^{K} \sum_{v=1}^{|A|} \bar{\boldsymbol{\pi}}_{ik}^v = \frac{1}{M|A|} \sum_{i=1}^{M} |A| = \frac{1}{M|A|} \cdot M|A| = 1, \tag{61}$$

$$\implies \frac{1}{M} \sum_{i=1}^{M} \sum_{k=1}^{K} \widehat{\boldsymbol{\pi}}_{ik} = 1. \tag{62}$$

based on the above equation we can say that the scaled sum of the mixing proportions of all K components in all M GMMs when the components are aligned must equal 1, show that the resulting aggregate posterior is non-degenerate and a valid probability distribution. □

Assumption F.4 (*Weak* Injectivity)**.** Let $f : \mathcal{Z} \to \mathcal{X}$ 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 F.5. In words, Assumption F.4 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.

Theorem F.6 (Mixture of Concatenated Slots)**.** *Let f_s denote a permutation equivariant probabilistic slot attention function such that $f_s(\mathbf{z}^v, P\mathbf{s}^v) = P f_s(\mathbf{z}^v, \mathbf{s}^v)$, where $P \in \{0,1\}^{K \times K}$ is an arbitrary permutation matrix. Let $\mathbf{c} = (g(\mathbf{s}_1^A, .), \ldots, g(\mathbf{s}_K^A, .)) \in \mathbb{R}^{Kd}$ be the concatenation of K individual content vectors, where each vector is an aggregate of all the slots obtained from considered viewpoints in a viewpoint-set $A \subseteq [V]$ (cf. Kori et al. (2024)). Due to the nature of the aggregator function, the individual content vector is Gaussian distributed within a K-component mixture: $\mathbf{c}_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \in \mathbb{R}^d, \forall k \in \{1, \ldots K\}$. Then, $\mathbf{c}$ is also GMM distributed with $K!$ mixture components:*

$$p(\mathbf{c}) = \sum_{p=1}^{K!} \boldsymbol{\pi}_p \mathcal{N}(\mathbf{c}; \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p), \text{ where } \boldsymbol{\pi} \in \Delta^{K!-1}, \boldsymbol{\mu}_p \in \mathbb{R}^{Kd}, \boldsymbol{\Sigma}_p \in \mathbb{R}^{Kd \times Kd}. \tag{63}$$

We additionally borrow some theorems and definitions from Kivva et al. (2022) which are essential for our proofs. First, we restate the definition of a *generic point* as outlined by Kivva et al. (2022) below.

Definition F.7. (Generic point) A point $\mathbf{x} \in f_d(\mathbb{R}^m) \subseteq \mathbb{R}^n$ is generic if there exists $\delta > 0$, such that $f_d : B(\mathbf{s}, \delta) \to \mathbb{R}^n$ is affine for every $\mathbf{s} \in f_d^{-1}(\{\mathbf{x}\})$

Theorem F.8 (Kivva et al. Kivva et al. (2022))**.** *Given $f_d : \mathbb{R}^m \to \mathbb{R}^n$ is a piecewise affine function such that $\{\mathbf{x} \in \mathbb{R}^n : |f_d^{-1}(\{\mathbf{x}\})| = \infty\} \subseteq f_d(\mathbb{R}^m)$ has measure zero with respect to the Lebesgue measure on $f_d(\mathbb{R}^m)$, this implies $\dim(f_d(\mathbb{R}^m)) = m$ and almost every point in $f_d(\mathbb{R}^m)$ (with respect to the Lebesgue measure on $f_d(\mathbb{R}^m)$) is generic with respect to f_d.*

Theorem F.9 (Kivva et al. Kivva et al. (2022))**.** *Consider a pair of finite GMMs in $\mathbb{R}^m$:*

$$\mathbf{y} = \sum_{j=1}^{J} \boldsymbol{\pi}_j \mathcal{N}(\mathbf{y}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), \qquad \text{and} \qquad \mathbf{y}' = \sum_{j=1}^{J'} \boldsymbol{\pi}'_j \mathcal{N}(\mathbf{y}'; \boldsymbol{\mu}'_j, \boldsymbol{\Sigma}'_j). \tag{64}$$

Assume that there exists a ball $B(\mathbf{x}, \delta)$ such that $\mathbf{y}$ and $\mathbf{y}'$ induce the same measure on $B(\mathbf{x}, \delta)$. Then $\mathbf{y} \equiv \mathbf{y}'$, and for some permutation τ we have that $\boldsymbol{\pi}_i = \boldsymbol{\pi}'_{\tau(i)}$ and $(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = (\boldsymbol{\mu}'_{\tau(i)}, \boldsymbol{\Sigma}'_{\tau(i)})$.

Theorem F.10 (Kivva et al. Kivva et al. (2022)). *Given* $\mathbf{z} \sim \sum_{i=1}^{J} \boldsymbol{\pi}_i \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ *and* $\mathbf{z}' \sim \sum_{j=1}^{J'} \boldsymbol{\pi}'_j \mathcal{N}(\mathbf{z}'; \boldsymbol{\mu}'_j, \boldsymbol{\Sigma}'_j)$ *and* $f_d(\mathbf{z})$ *and* $\tilde{f}_d(\mathbf{z}')$ *are equally distributed. We can assume for* $\mathbf{x} \in \mathbb{R}^n$ *and* $\delta > 0$, f_d *is invertible on* $B(\mathbf{x}, 2\delta) \cap f_d(\mathbb{R}^m)$. *This implies that there exists* $\mathbf{x}_1 \in B(\mathbf{x}, \delta)$ *and* $\delta_1 > 0$ *such that both* f_d *and* $\tilde{f}_d$ *are invertible on* $B(\mathbf{x}_1, \delta_1) \cap f_d(\mathbb{R}^m)$.

Theorem 3.3 (Affine Equivalence of aggregate content) For any subset $A \subseteq [V]$, such that $|A| > 0$, given a set of images $\mathbf{x}^A \in \mathcal{X}^A$ and a corresponding aggregate content $\mathbf{c} \in \mathcal{C}$ and a non-degenerate content posterior $q(\mathbf{c} \mid \mathbf{s}^A)$, considering two mixing function $f_d, \tilde{f}_d$ satisfying assumption F.4, with a shared image, then $\mathbf{c}$ are identifiable up to $\sim_s$ equivalence.

Proof. Based on the results of Kori et al. (2024) we know that when $p(\mathbf{s})$ is aggregate posterior of $q(\mathbf{s} \mid \mathbf{x})$, $p(\mathbf{s})$ is identifiable up to $\sim_s$ equivalence. Additionally, based on lemma 3.2 we know that both $q(\mathbf{s} \mid \mathbf{x})$ and $q(\mathbf{c} \mid \mathbf{s})$ are a non-degenerate GMM with valid probability distribution. Using similar arguments in Kori et al. (2024); Kivva et al. (2022) we show that $p(\mathbf{c})$ and $p(\mathbf{s})$ are identifiable up to $\sim_s$ equivalence. `W.l.o.g`, given view point transformation is deterministic, we consider $\mathbf{x}^A = \mathcal{T}_{\theta^A}(\mathbf{x}^A)$.

We know that

$$
\begin{aligned}
p(\mathbf{s}^A) &= \int q(\mathbf{s}^A_{1:K} \mid \mathbf{x}^A) p(\mathbf{x}^A) d\mathbf{x}^A && (65)\\
&= \int \prod_{v \in A} q(\mathbf{s}^v \mid \mathbf{x}^v) p(\mathbf{x}^v) d\mathbf{x}^A && (66)\\
&= \int \prod_{v \in A} \left(\sum_{k=1}^{K} \boldsymbol{\pi}_k^v \mathcal{N}\left(\mathbf{s}^v; \boldsymbol{\mu}_k(\mathbf{x}^v), \boldsymbol{\sigma}_k^2(\mathbf{x}^v)\right) \right) p(\mathbf{x}^v) d\mathbf{x}^A && (67)\\
&= \prod_{v \in A} \frac{1}{|\mathcal{X}|} \int \left(\sum_{k=1}^{K} \boldsymbol{\pi}_k^v \mathcal{N}\left(\mathbf{c}^v; \boldsymbol{\mu}_k(\mathbf{x}^v), \boldsymbol{\sigma}_k^2(\mathbf{x}^v)\right) \right) \delta(\mathbf{x}^v - \mathbf{x}_i^v) d\mathbf{x}^A && (68)\\
&= \prod_{v \in A} \left(\sum_{k=1}^{|\mathcal{X}|K} \frac{1}{|\mathcal{X}|} \hat{\boldsymbol{\pi}}_{ik}^v \mathcal{N}\left(\mathbf{s}^v; \hat{\boldsymbol{\mu}}_{ivk}, \tilde{\boldsymbol{\sigma}}_{ivk}^2\right) \right) && (69)
\end{aligned}
$$

Change of variables from $\mathbf{s}$ to $\mathbf{c}$ to get prior over random variable $\mathbf{c}$, with matching function g, results in:

$$
p(\mathbf{c}_{1:K}) = \int p(\mathbf{s}^A_{1:K}) \delta\left(\mathbf{s}^A_{1:K} - g(\mathbf{s}^A_{1:K}, \boldsymbol{\pi}^A_{1:K})\right) d\mathbf{c}^A_{1:K} \tag{70}
$$

Given the transformation g is linear, resulting us with the distribution with mean given by:

$$
\begin{aligned}
\mathbb{E}_{\mathbf{c}}\left(\mathbf{c}_{1:K}\right) &= \mathbb{E}_{\mathbf{s}}\left(g(\mathbf{s}^A_{1:K}, \boldsymbol{\pi}_{A,1:K})\right) && (71)\\
&= g\left(\mathbb{E}_{\mathbf{s}}(\mathbf{s}^A_{1:K}), \boldsymbol{\pi}^A_{1:K}\right) && (72)\\
&= \sum_{v \in A} \frac{\boldsymbol{\pi}^v_{1:K}}{\sum_{v \in A} \boldsymbol{\pi}^v_{1:K}} \mathbb{E}_{\mathbf{s}}(\mathbf{s}^A_{1:K}) && (73)
\end{aligned}
$$

and the covariance follows the diagonal structure as in $p(\mathbf{c})$, which can be described as follows:

$$
\mathbb{V}\mathrm{ar}(\mathbf{c}_{1:K}) = \sum_{v \in A} \left(\frac{\boldsymbol{\pi}^v_{1:K}}{\sum_{v \in A} \boldsymbol{\pi}^v_{1:K}} \right)^2 \mathbb{V}\mathrm{ar}_{\mathbf{c}}(\mathbf{c}^A_{1:K}) \tag{74}
$$

Finally, the mixture components can be expressed as:

$$
\tilde{\boldsymbol{\pi}}_{1:K} = \frac{\sum_{v \in A} \boldsymbol{\pi}^v_{1:K}}{|A|} \tag{75}
$$

With distribution parameters described in equations 73, 74, and 75, we define the aggregate content distribution as GMM expressed as follows:

$$p(\mathbf{c}) = \sum_{k=1}^{|\mathcal{X}|K} \frac{1}{|\mathcal{X}|} \tilde{\boldsymbol{\pi}}_k^v \mathcal{N}\left(\mathbf{v}; \mathbb{E}(\mathbf{c})_k), \mathbb{V}\mathrm{ar}(\mathbf{c})_k\right) \tag{76}$$

Validity of $p(\mathbf{c})$: The outer summation in equation 76 can be split into two one for image samples and other for original mixing coefficients, which results in the equation:

$$p(\mathbf{c}) = \sum_{i=1}^{|\mathcal{X}|} \sum_{k=1}^{K} \frac{1}{|\mathcal{X}|} \tilde{\boldsymbol{\pi}}_{ik}^v \mathcal{N}\left(\mathbf{v}; \mathbb{E}(\mathbf{c})_{ik}), \mathbb{V}\mathrm{ar}(\mathbf{c})_{ik}\right) \tag{77}$$

Based on this we can observe the each component in our GMM corresponds to particular slots for a given image in a given viewpoint, triple describing each component is:

$$\left\{\tilde{\boldsymbol{\pi}}_{ik}^v, \tilde{\boldsymbol{\mu}}_{vik}, \tilde{\boldsymbol{\sigma}}_{vik}^2\right\}, \quad \text{for} \quad v = 1, \ldots, |A| \quad i = 1, 2, \ldots, |\mathcal{X}|, \quad \text{and} \quad k = 1, 2, \ldots, K. \tag{78}$$

To verify that $p(\mathbf{c})$ is a non-degenerate mixture, we observe the following implication:

$$\sum_{i=1}^{|\mathcal{X}|} \sum_{k=1}^{K} \frac{1}{|\mathcal{X}|} \frac{\sum_{v \in A} \tilde{\boldsymbol{\pi}}_{ik}^v}{|A|} = 1, \tag{79}$$

$$\implies \frac{1}{|\mathcal{X}|} \frac{1}{|A|} \sum_{i=1}^{|\mathcal{X}|} \sum_{v \in A} \sum_{k=1}^{K} \boldsymbol{\pi}_{ik}^v = \frac{1}{|\mathcal{X}|} \frac{1}{|A|} |\mathcal{X}| \cdot |A| \cdot 1 = 1 \tag{80}$$

similar to lemma 3.2, this says that the scaled sum of the mixing proportions of all K components in all $|\mathcal{X}|$ GMMs must equal 1, proving that the associated aggregate posterior mixture $p(\mathbf{c})$ is a well-defined and non degenerate probability distribution.

Invertibility restrictions: Given two piece-wise affine compositional functions $f_d, \tilde{f}_d : \mathcal{C} \times \mathcal{V} \to \mathcal{X}$, for a given set of views $\mathbf{v}^A$, let $\mathbf{c} = (\mathbf{c}_1, \ldots, \mathbf{c}_K), \ni \mathbf{c}_k \sim \mathcal{N}(\mathbf{c}_k; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ and $\mathbf{c}' = (\mathbf{c}'_1, \ldots, \mathbf{c}'_K), \ni \mathbf{c}'_k \sim \mathcal{N}(\mathbf{c}'_k; \boldsymbol{\mu}'_k, \boldsymbol{\Sigma}'_k)$ be a pair of aggregate content representations, result of sampling a concatenated higher dimensional GMM distribution in $\mathbb{R}^{Kd}$, as shown in Theorem F.6, Kori et al. (2024). In the case when, $f_{d\sharp}(\mathcal{C}, \{\mathbf{v}^A\})$ and $\tilde{f}_{d\sharp}(\mathcal{C}', \{\mathbf{v}^A\})$[3] are equally distributed. Now assume that there exists $\mathbf{x}^A \in \mathcal{X}$ and $\delta > 0$ such that f_d and $\tilde{f}_d$ are invertible and piecewise affine on $B(\mathbf{x}^A, \delta) \cap f_d(\mathcal{S})$, for a given set of views $\mathbf{v}^A$, which implies $\dim f_d(\mathcal{C}, \{\mathbf{v}^A\}) = |\mathcal{C}|$.

Affine subspace: We now restrict the space $B(\mathbf{x}^A, \delta)$ to a subspace $B(\mathbf{x}'^A, \delta')$ where $\mathbf{x}^A \in B(\mathbf{x}'^A, \delta')$ such that f_d and $\tilde{f}_d$ are now invertible and affine on $B(\mathbf{x}'^A, \delta') \cap f_d(\mathcal{C} \times \{\mathbf{v}^A\})$. With $L \subseteq \mathcal{X}^A$ be an $|\mathcal{C}|$-dimensional affine subspace (assuming $|\mathcal{X}^A| \geq |\mathcal{C}|$), such that $B(\mathbf{x}'^A, \delta') \cap f_{d\sharp}(\mathcal{C}, \{\mathbf{v}^A\}) = B(\mathbf{x}'^A, \delta') \cap L$. We also define $h_f, h_{\tilde{f}} : \mathcal{C} \to L$ to be a pair of invertible affine functions where $h_{f\sharp}^{-1}(B(\mathbf{x}'^A, \delta') \cap L) = f_{d\sharp}^{-1}(B(\mathbf{x}'^A, \delta') \cap L; \mathbf{v}^A)$ and $h_{\tilde{f}\sharp}^{-1}(B(\mathbf{x}'^A, \delta') \cap L) = \tilde{f}_{d\sharp}^{-1}(B(\mathbf{x}'^A, \delta') \cap L; \mathbf{v}^A)$. Implying $h_f(\mathbf{c})$ and $h_{\tilde{f}}(\mathbf{c}')$ are finite GMMs that coincide with $B(\mathbf{x}'^A, \delta') \cap L$ and $h_f(\mathbf{c}) \equiv h_{\tilde{f}}(\mathbf{c}')$, theorem F.9, Kivva et al. (2022). Given, $h = h_{\tilde{f}}^{-1} \circ h_f$ and $h_f(\mathbf{c})$ and $h_{\tilde{f}}(\mathbf{c}')$ then h is an affine transformation such that $h(\mathbf{c}) = \mathbf{c}'$.

$\sim_s$ equivalence: Given Theorems F.8 and F.10, there exists a point $\mathbf{x} \in f_{d\sharp}(\mathcal{C}, \{\mathbf{v}^A\})$ that is generic with respect f_d and $\tilde{f}_d$ and invertible on $B(\mathbf{x}, \delta) \cap f_{d\sharp}(\mathcal{C}, \{\mathbf{v}^A\})$. Having established that there is an affine transformation $h(\mathbf{c}) = \mathbf{c}'$ and invertiblility of piece-wise affine functions f_d and $\tilde{f}_d$ on $B(\mathbf{x}^A, \delta) \cap f_{d\sharp}(\mathcal{C}, \{\mathbf{v}^A\})$, this implies that $\mathbf{c}$ is identifiable up to an affine transformation and permutation of $\mathbf{c}_k \in \mathbf{c}$, concluding our proof.

[3] $f_{d\sharp}$ correspond to push forward operation, applying function f_d on all the elements of the given set.

Remark: Given Theorem F.9, we know that for each higher dimensional mixture component in $p(\mathbf{c})$ induces the same measure on $B(\mathbf{x}^A, \delta)$ and hence for some permutation τ we have that $(\boldsymbol{\mu}_{\pi(i)}, \boldsymbol{\Sigma}_{\pi(i)}) = (\boldsymbol{\mu}'_{\tau(\pi(i))}, \boldsymbol{\Sigma}'_{\tau(\pi(i))})$. Therefore, each mixture component $\mathbf{c}_{\pi(i)}$ is identifiable up to affine transformation, and permutation of aggregate content representations in $\mathbf{c}$. Now, given sampling $\mathbf{c}_k$ is equivalent to obtaining K samples from the GMM, $q(\mathbf{z})$ and concatenating, this makes $q(\mathbf{z})$ identifiable up to affine transformation, h and permutation of slot representations in $\mathbf{c}$. It now trivially follows that each of the aggregate content representation $\mathbf{c}_k \sim \mathcal{N}(\mathbf{c}_k; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \in \mathbb{R}^d, \forall\, k \in \{1, \ldots, K\}$ is identifiable up to affine transformation, h based on the following observed property of GMMs:

$$\sum_{k=1}^{K} \boldsymbol{\pi}_k h_\sharp \left(\mathcal{N}(\mathbf{s}_k; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) \sim h_\sharp\Big(\sum_{k=1}^{K} \boldsymbol{\pi}_k \mathcal{N}(\mathbf{s}'_k; \boldsymbol{\mu}'_k, \boldsymbol{\Sigma}'_k)\Big), \tag{81}$$

□

Theorem 3.4 (Invariance of aggregate content) For any subset $A, B \subseteq [V]$, such that $|A| > 0, |B| > 0$ and both A, B satisfy an assumption 3.1, we consider aggregate content to be invariant to viewpoints if $f_A \sim_s f_B$ for data $\mathcal{X}^A \times \mathcal{X}^B$.

Proof. Based on equation 76, $p_A(\mathbf{s})$ and $p_B(\mathbf{s})$ can be expressed as follows:

$$p_A(\mathbf{c}) = \sum_{k=1}^{|\mathcal{X}|K} \frac{1}{|\mathcal{X}|} \frac{\sum_{v \in A} \boldsymbol{\pi}_k^v}{|A|} \mathcal{N}\left(\mathbf{c}; \sum_{v \in A} \frac{\boldsymbol{\pi}_k^v}{\sum_{v \in A} \boldsymbol{\pi}_k^v} \boldsymbol{\mu}_{vk}, \sum_{v \in A} \left(\frac{\boldsymbol{\pi}_{vk}}{\sum_{v \in A} \boldsymbol{\pi}_k^v}\right)^2 \boldsymbol{\sigma}_{vk}^2\right) \tag{82}$$

$$p_B(\mathbf{c}) = \sum_{k=1}^{|\mathcal{X}|K} \frac{1}{|\mathcal{X}|} \frac{\sum_{u \in B} \boldsymbol{\pi}_k^u}{|B|} \mathcal{N}\left(\mathbf{c}; \sum_{u \in B} \frac{\boldsymbol{\pi}_{uk}}{\sum_{u \in B} \boldsymbol{\pi}_k^u} \boldsymbol{\mu}_k^u, \sum_{u \in B} \left(\frac{\boldsymbol{\pi}_k^u}{\sum_{u \in B} \boldsymbol{\pi}_k^u}\right)^2 \boldsymbol{\sigma}_{uk}^2\right) \tag{83}$$

Given the assumption of viewpoint sufficiency 3.1 we know the objects observed in viewpoint set A are same as the object observed in set B. Following the results of Theorem 3.3, we know that both $p_A(\mathbf{s})$ and $p_B(\mathbf{s})$ are independently identifiable up to $\sim_s$ equivalence, which means f_A and f_B are invertible for a given views $\mathbf{v}^A$ and $\mathbf{v}^B$ respectively.

Affine mapping. Without loss of generality, for a given set of views $\mathbf{v}^A$, there exists some $L \subseteq \mathcal{X}^A$ be an $|\mathcal{S}|$-dimensional affine subspace, such that $B(\mathbf{x}'^A, \delta) \cap f_{A\sharp}(\mathcal{C}, \{\mathbf{v}^A\}) \cap f_{B\sharp}(\mathcal{C}, \{\mathbf{v}^A\}) = B(\mathbf{x}'^A, \delta) \cap L$. This implies their exists an affine map between $\mathbf{c} = f_A^{-1}(\mathbf{x}^A; \mathbf{v}^A)$ and $\tilde{\mathbf{c}} = f_B^{-1}(\mathbf{x}^B; \mathbf{v}^A)$. Let $h_A : \mathcal{C} \to L$ to be an invertible affine functions where $h_{A\sharp}^{-1}(B(\mathbf{x}'^A, \delta') \cap L) = f_{A\sharp}^{-1}(B(\mathbf{x}'^A, \delta') \cap L; \mathbf{v}^A) = f_{B\sharp}^{-1}(B(\mathbf{x}'^B, \delta') \cap L; \mathbf{v}^A)$ resulting in $h_A(\mathbf{c}) = \mathbf{c}'$. Similarly, we can show their exists an affine map between $\tilde{\mathbf{c}} = f_A^{-1}(\mathbf{x}^A; \mathbf{v}^B)$ and $\tilde{\mathbf{c}}' = f_B^{-1}(\mathbf{x}^B; \mathbf{v}^B)$, such that $h_B(\tilde{\mathbf{c}}) = \tilde{\mathbf{c}}'$.

Invariance setup. In the case when representations are invariant, $p_A(\mathbf{c})$ and $p_B(\mathbf{c})$ are equally distributed, which means aggregate content domain in both cases are same or similar $\mathcal{C}_A = \mathcal{C}_B$.

$$\mathbf{c}' = h(\tilde{\mathbf{c}}') \tag{84}$$

$$\implies h_A(\mathbf{c}) = (h \circ h_B)(\tilde{\mathbf{c}}) \tag{85}$$

$$\implies \mathbf{c} = (h_A^{-1} \circ h \circ h_B)(\tilde{\mathbf{c}}) \tag{86}$$

Given composition of affine maps is affine, we can consider the mapping $(h_A^{-1} \circ h \circ h_B)$ to be an affine, resulting in an $\sim_s$ equivalence between f_A and f_B.

□

Theorem 3.5 (Approximate representational equivariance) For a given aggregate content $\mathbf{c}$, for any two views $\mathbf{v}, \tilde{\mathbf{v}} \sim p_A(\mathbf{v})$, resulting in respective scenes $\mathbf{x} \sim p_A(\mathbf{x} \mid \mathbf{v}, \mathbf{c})$ and $\tilde{\mathbf{x}} \sim p_A(\mathbf{x} \mid \tilde{\mathbf{v}}, \mathbf{c})$, for any homeomorphic, monotonic transformation $h_x \in \mathcal{H}_x$ such that $h_x(\mathbf{x}) = \tilde{\mathbf{x}}$, their exists another homeomorphic and monotonic transformation $h_v \in \mathcal{H}_v$ such that $\mathcal{H}_v \subseteq \mathcal{H}_x \subseteq \mathbb{R}^{\dim(\mathbf{x})}$ and $\mathbf{v} = h_v^{-1}\left(f_d^{-1}(h_x(\mathbf{x}); \mathbf{c})\right)$.

Proof. For a given view $\mathbf{v}$ and a mixing function f_d that satisfy assumptions F.4 and is piecewise affine, from theorem 3.3 we know the latent view representations are identifiable up to $\sim_s$ equivalence for a given aggregate content vector. We know that $p(\mathbf{v})$ is expressed as GMM with a considered set of viewpoints, ideally learning each component for each viewpoint.

$$p(\mathbf{v}) = \sum_{v=1}^{|A|} \boldsymbol{\pi}^v \mathcal{N}(\mathbf{v}; \boldsymbol{\mu}_v, \boldsymbol{\sigma}_v)$$

Following similar arguments in Theorem 3.3 and Kivva et al. (2022), we can show that for a given content representation $\mathbf{c}$ the view distribution $p(\mathbf{v})$ is identifiable up to affine transformation. This means, for any two considered models $f_d, \tilde{f}_d$, such that $f_{d\sharp}(\mathcal{V}; \{\mathbf{c}\})$ and $\tilde{f}_{d\sharp}(\mathcal{V}; \{\mathbf{c}\})$ are equally distributed, then for any $\mathbf{x}^A \sim \mathcal{X}$ with the corresponding content representations given by $\mathbf{c}$ the views $\mathbf{v} = f_d^{-1}(\mathbf{x}^v; \mathbf{c}),\ \mathbf{v}' = \tilde{f}_d^{-1}(\mathbf{x}^v; \mathbf{c})$ are related in by an affine transformation $h(\mathbf{v}) = \mathbf{v}'$, results in:

$$\sum_{v=1}^{|A|} \boldsymbol{\pi}^v h_\sharp\left(\mathcal{N}(\mathbf{v}; \boldsymbol{\mu}_v, \boldsymbol{\sigma}_v^2)\right) \sim h_\sharp\left(\sum_{v=1}^{|A|} \boldsymbol{\pi}^v \mathcal{N}(\mathbf{v}; \boldsymbol{\mu}_v, \boldsymbol{\sigma}_v^2)\right), \tag{87}$$

Without loss of generality we can consider any function $f : \mathcal{C} \times \mathcal{V} \to \mathcal{X}$ is identifiable up to affine transformation, with this for given views $\mathbf{v}, \tilde{\mathbf{v}} \sim p(\mathbf{v})$ and for any object representations $\mathbf{c}$, the resulting scenes are sampled by distributions learned with mixing function f is given by $\mathbf{x} \sim p_f(\mathbf{x} \mid \mathbf{c}, \mathbf{v}), \tilde{\mathbf{x}} \sim p_f(\mathbf{x} \mid \mathbf{c}, \tilde{\mathbf{v}})$. As previously established for some affine transformation h,

$$h(\mathbf{v}) = f^{-1}(\tilde{\mathbf{x}}; \mathbf{c}) \implies \mathbf{v} = h^{-1}\left(f^{-1}(\tilde{\mathbf{x}}; \mathbf{c})\right) \tag{88}$$

Given $h_x(\mathbf{x}) = \tilde{\mathbf{x}}$, when combined with above equation we know $\mathbf{v} = h^{-1}\left(f^{-1}(\mathbf{x}; \mathbf{c})\right), \tilde{\mathbf{v}} = h'^{-1}\left(f^{-1}(h_x(\mathbf{x}); \mathbf{c})\right)$, for some invertible affine transformations h and h'.

Given h_x is homeomorphic and monotonic, and f is piecewise linear, the inverse can be transferred resulting in $\tilde{\mathbf{v}} = h'^{-1}\left(\bar{h}_v(f^{-1}(\mathbf{x}; \mathbf{c}))\right)$, similarly we can also swap h'^{-1} with $\bar{h}_v$, resulting in $\tilde{\mathbf{v}} = \bar{h}_v\left(h'^{-1}\left(f^{-1}(\mathbf{x}; \mathbf{c})\right)\right)$. Additionally combining the results from theorem 3.3 and Kivva et al. (2022), we know that $h'^{-1} \circ h$ is an affine transformation $\bar{h}$. This results in:

$$\bar{h} = h'^{-1} \circ h \tag{89}$$

$$\implies \tilde{\mathbf{v}} = (\bar{h}_v \circ h \circ \bar{h})\left(f^{-1}(\mathbf{x}; \mathbf{c})\right) \tag{90}$$

$$\implies \tilde{\mathbf{v}} = h_v(\mathbf{v}) \tag{91}$$

Given affine transformation preserves monotonicity and homeomophism, the resulting transformation $h_v \in \mathcal{H}_v$ and $h_v \in \mathcal{H}_x$, concluding the proof.

□

G Experiments

G.1 Toy Setting

Here, we repeat the experiments in CASE STUDY 1 with point cloud giving us two dimensional distributions, which can analysed visually. In Fig. 8, we display the distributions of marginalized aggregate content distribution $q(\mathbf{c})$, comparing different runs that are either rotated, skewed, or mirrored with respect to each other, indicating identifiability up to affine transformation. To quantitatively measure

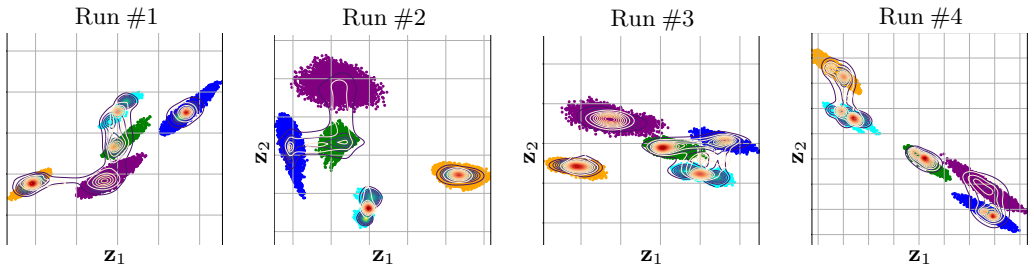


Figure 8: **Identifiability of $q(\mathbf{c})$ and $q(\mathbf{s})$.** Estimated marginalised slot distribution ($q(\mathbf{s})$–blue contours) and marginalised content distribution ($q(\mathbf{c})$–orange contours, across 4 runs of VISA. This provides strong evidence of recovery of the latent space up to affine transformations, empirically verifying our claims in Thm. 3.3.

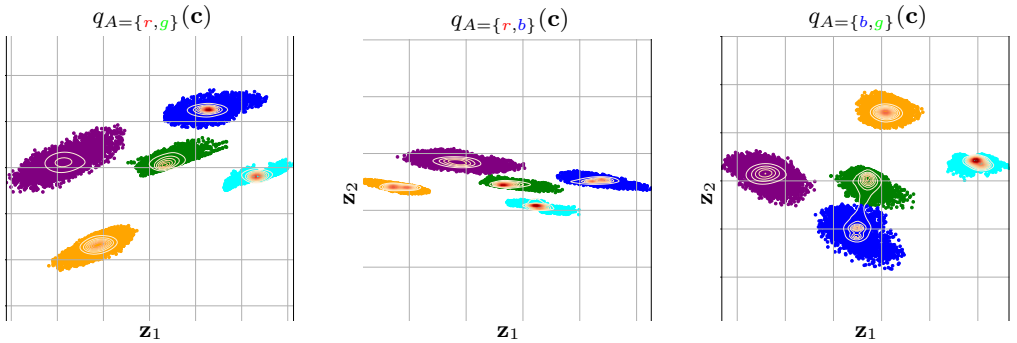


Figure 9: **Viewpoint invariance for $q(\mathbf{c})$.** Estimated marginalised aggregate content distribution $q(\mathbf{c})$ when trained with different view pairs {(green, red), (red, blue), (green, blue)} are illustrated in later figures. As the resulting distributions with different datasets only vary by an affine transformation, providing strong evidence for Thm. 3.4.

the same, we computed SMCC and observed it to be $\mathbf{0.95} \pm 0.01$, empirically verifying our Thm. 3.3. Furthermore, to illustrate the invariance of distribution $q(\mathbf{c})$ across viewpoints (Thm. 3.4), we consider three different views. We use all possible pairs to learn $q(\mathbf{c})$ distributions as illustrated in Fig. 9, where the distributions from second to last sub-figures are learned wrt viewpoints described by {g, r}, {r, b}, and {g, b}, respectively. Similar to our previous findings, these distributions were also found to be rotated, skewed, or mirrored relative to each other, with an observed SMCC of $\mathbf{0.87} \pm 0.11$, further confirming the claims in Thm. 3.4.

G.2 Synthetic dataset results

Here, we illustrate visual results reflecting object binding in the case of view ambiguities. Table 3, demonstrates identifiability results on CLEVR-AUG datasets. In Fig. 10, we demonstrate the results of VISA across three different views. We additionally highlight some of the occluded regions which seem to be better captured by our proposed model, which can be attributed to the multi-view setting and the `sigmoid` mask.

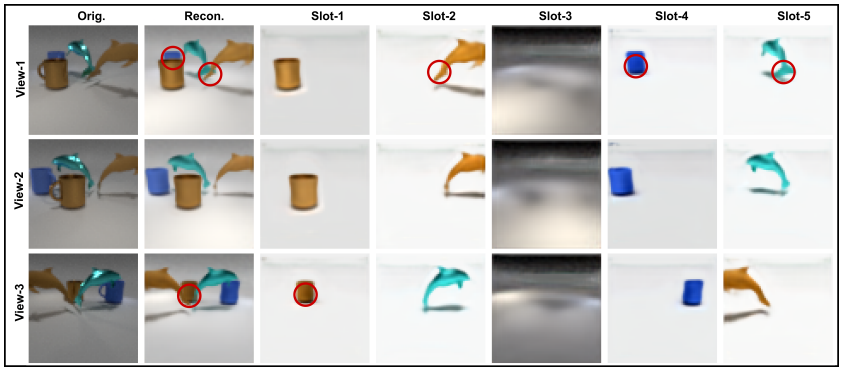
Additionally, we also illustrate the results from CLEVR-MV dataset in figure 11.

G.3 Influence of Number of Views

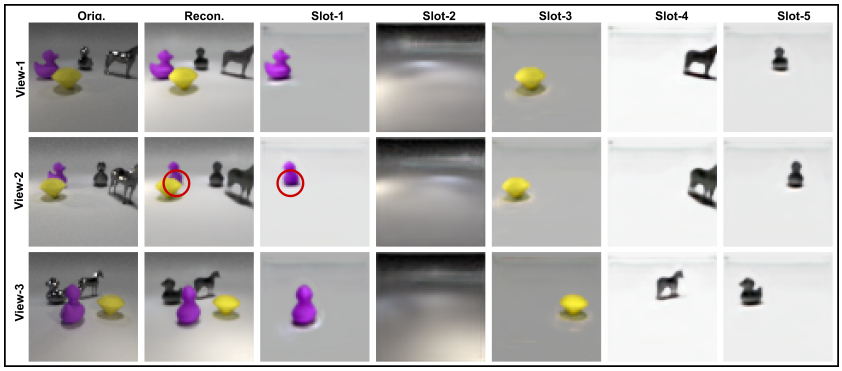
Here, we demonstrate the influence of the number of views on the overall identifiability of object-centric representations. Similar to Fig. 2, in Fig. 12, we observe an increasing number of views increase overall results.

Table 3: Comparing identifiability of $q(\mathbf{s})$, $q(\mathbf{c})$, and $p(\mathbf{v})$ scores wrt existing OCL methods on CLEVR-AUG dataset.

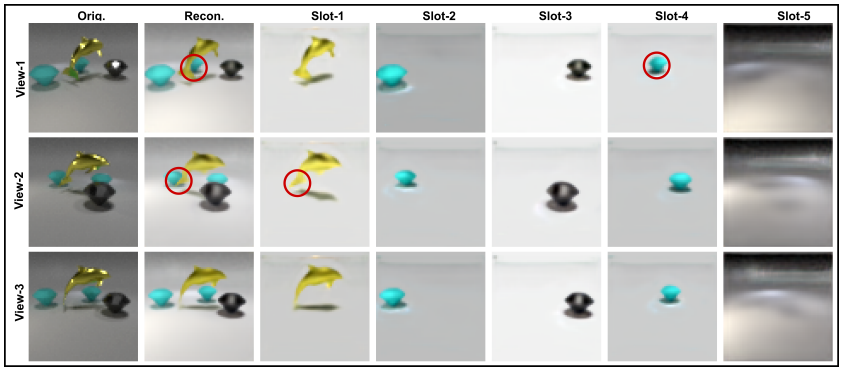
METHOD	SMCC ↑	INV-SMCC ↑	MCC ↑
AE	$0.26 \pm .01$	-	-
SA	$0.45 \pm .05$	-	-
PSA	$0.48 \pm .03$	-	-
MulMON	$0.56 \pm .04$	$0.57 \pm .01$	-
OCLOC	$0.58 \pm .02$	$0.60 \pm .01$	$0.48 \pm .04$
VISA	$0.64 \pm .01$	$0.66 \pm .01$	$0.57 \pm .04$



(a)

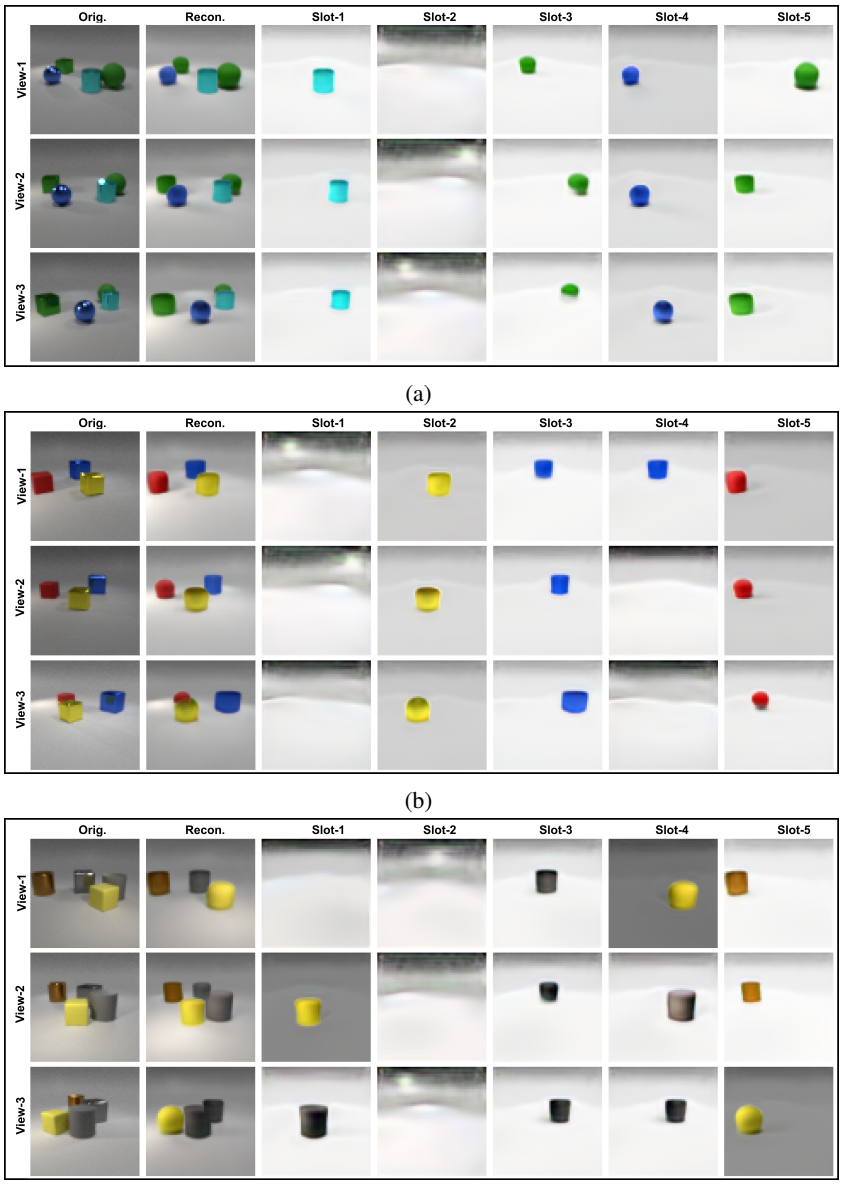


(b)



(c)

Figure 10: Visual illustrations of benchmark results on CLEVR-AUG dataset.



(a)

(b)

(c)

Figure 11: Visual illustrations of benchmark results on CLEVR-MV dataset.

G.4 MVMOVI RESULTS

Here, we discuss the results obtained from the proposed dataset. To reiterate, MVMOVI-C is a variant where fixed camera positions are maintained for all viewpoints across all scenes in the dataset. This setup helps assign a fixed type of viewpoint conditioning for all images captured from a particular camera.

The detection and binding quality of different models are illustrated in Table 2. From these results, we can clearly observe that while the model demonstrates similar binding capabilities, the identifiability of object representations is improved in our proposed model. This suggests that the use of fixed camera positions in MVMOVI-C enhances the consistency and quality of object representation learning, leading to better detection performance across different viewpoints.

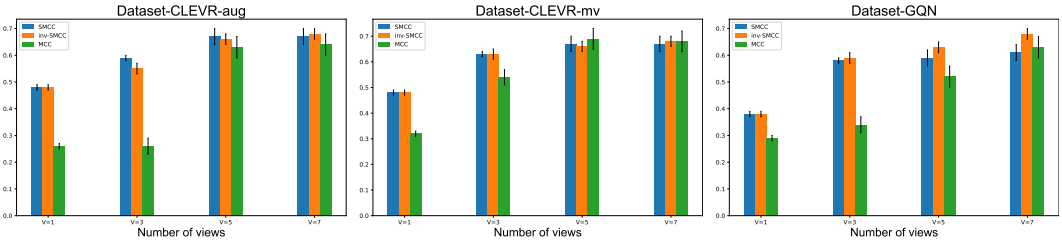
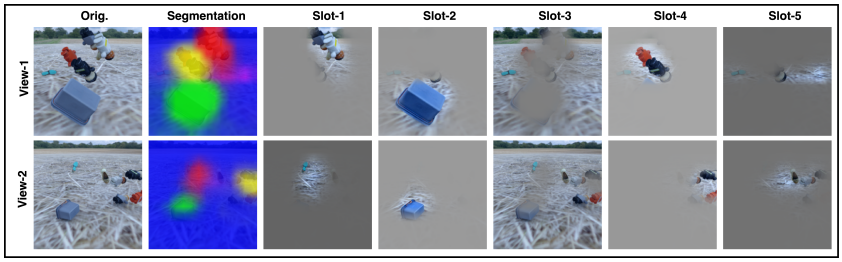
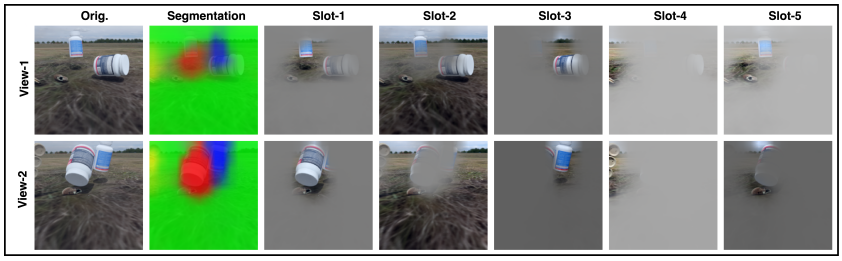


Figure 12: Influence of Number of viewpoints on identifiability for synthetic datasets.

Figure 13 & 14 showcases the object discovery capabilities of the VISA. In the iteration of the MVMoVi-D dataset, we vary the camera position for each scene, making the dataset more dynamic and allowing for the potential violation of assumption 3.1 in certain cases. Table 4 presents the binding and identifiability results for both in-domain and out-of-domain data, following a similar analysis as in Table 2. We observe consistent trends and behaviours, suggesting that the impact of the assumption is minimal. A more detailed analysis of the assumption's effects will be left for future work.



(a)

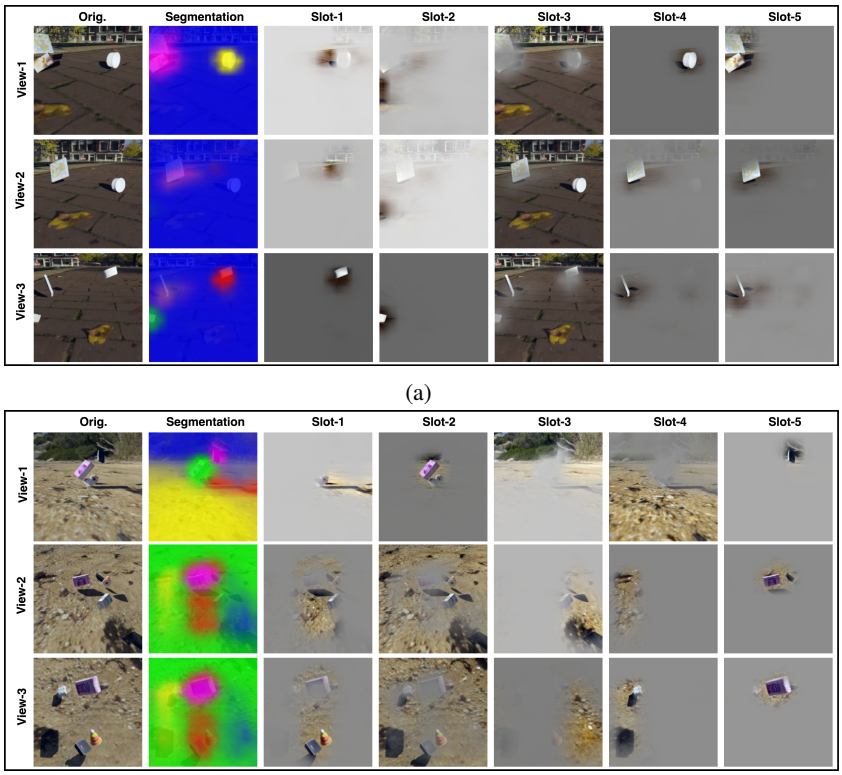


(b)

Figure 13: Visual illustrations of benchmark results on MVMoVi-C dataset with 2 views.

Table 4: Identifiability and generalisability analysis on MV-MoViD dataset.

Method	Indomain Analysis				Out of domain			
	mBO ↑	SMCC ↑	Inv-SMCC ↑	MCC ↑	mBO ↑	SMCC ↑	Inv-SMCC ↑	MCC ↑
SA-MLP	0.24 ± 0.031	0.44 ± 0.005	-	-	0.24 ± 0.097	0.45 ± 0.008	-	-
PSA-MLP	0.26 ± 0.022	0.44 ± 0.006	-	-	0.25 ± 0.012	0.42 ± 0.006	-	-
VISA-MLP	0.24 ± 0.099	0.48 ± 0.009	0.46 ± 0.054	0.57 ± 0.021	0.25 ± 0.011	0.48 ± 0.006	0.51 ± 0.021	0.55 ± 0.021
SA-Transformer	0.34 ± 0.017	0.40 ± 0.041	-	-	0.34 ± 0.066	0.38 ± 0.031	-	-
PSA-Transformer	0.37 ± 0.021	0.38 ± 0.007	-	-	0.36 ± 0.024	0.36 ± 0.016	-	-
VISA-Transformer	0.39 ± 0.016	0.46 ± 0.001	0.48 ± 0.001	0.54 ± 0.032	0.37 ± 0.051	0.46 ± 0.022	0.45 ± 0.010	0.54 ± 0.029



(a)

(b)

Figure 14: Visual illustrations of benchmark results on MVMOVI-C dataset with 3 views.

G.5 VIEW WARM-UP

Given the stochasticity during the initial phase of training, to facilitate meaningful representation in content aggregator function, we consider view warm-up strategy. For the initial 100,000 iterations, we randomly use the view-specific slots for reconstruction instead of invariant content with a probability of 0.5.

This primary makes sure the feature extractor extracts meaningful representations before aggregation, which helps to stabilize the training process and allows the model to effectively bind and integrate information from different perspectives in later stages of training.

G.6 HYPERPARAMETERS

In Table 5, we detail all the hyper-parameters used in our experiments. In the case of benchmark experiments, we use trainable CNN encoder as used in Locatello et al. (2020b); Kori et al. (2023), while in the case of proposed MVMOVI datasets we use DINO (Caron et al., 2021) encoder to extract image features and change our objective to reconstruct these features rather than the original image as proposed in Seitzer et al. (2022). For most of hyperparameters we use the values suggested by Locatello et al. (2020b); Seitzer et al. (2022), based on their ablation results.

G.7 COMPUTATIONAL RESOURCES

We run all our experiments on a cluster with a Nvidia NVIDIA L40 48GB GPU cards. Our training usually takes between eight hours to a couple of days, depending on the model and the dataset. It is to be noted that speed might differ slightly with respect to the considered system and the background processes. All experimental scripts will be made available on GitHub at a later stage.

Table 5: Experimental details w.r.t datasets

DATASETS(↓)	PARAMETERS	VALUES
CLEVR, GSO	No. Layers	4
	No. Views	10 (GSO: 8)
	No. Slots	7
	Training Epochs	5000
	Batch Size	32
	Optimizer	ADAM
	Learning Rate	0.0002
	Initial Slot $\boldsymbol{\mu}$	$\mathcal{N}(0, 1)$
	Initial Slot $\boldsymbol{\sigma}$	$\mathbb{I}$
	Warmup Steps	10000
	Decoder	SPATIAL BROADCASTING-CNN
	$\mathbf{x}-$ likelihood	$\mathcal{N}(\boldsymbol{\mu}_x, \sigma_x^2\mathbb{I})$
GQN	No. Layers	4
	No. Views	10
	No. Slots	4
	Training Epochs	5000
	Batch Size	64
	Optimizer	ADAM
	Learning Rate	0.0002
	Initial Slot $\boldsymbol{\mu}$	$\mathcal{N}(0, 1)$
	Initial Slot $\boldsymbol{\sigma}$	$\mathbb{I}$
	Warmup Steps	10000
	Decoder	SPATIAL BROADCASTING-CNN
	$\mathbf{x}-$ likelihood	$\mathcal{N}(\boldsymbol{\mu}_x, \sigma_x^2\mathbb{I})$
MVMOVI-C, MVMOVI-D	No. Layers	4
	No. Views	5
	No. Slots	7
	Training Epochs	560
	Batch Size	64
	Optimizer	ADAMW
	Learning Rate	0.0002
	Initial Slot $\boldsymbol{\mu}$	$\mathcal{N}(0, 1)$
	Initial Slot $\boldsymbol{\sigma}$	$\mathbb{I}$
	Warmup Steps	10000
	Pretrained Encoder	DINO_VITB16
	Decoder	MLP, TRANSFORMER
	$\mathbf{x}-$ likelihood	$\mathcal{N}(\boldsymbol{\mu}_x, \mathbb{I})$

G.8 LIMITATIONS & FUTURE WORK.

We recognize that our assumptions, particularly regarding the *viewpoint sufficiency*, are strong and may not always hold in practice. However, we did not observe limiting effects of this assumption on the proposed MV-MOVID dataset. A more extensive analysis of this assumption and its implications in real-world applications is left for future work. We would also highlight that the *weak injectivity* of the mixing function may not always hold for different types of architectures. While generally applicable, the piecewise-affine functions we use may not always capture valid assumptions for real-world problems, *e.g.*, when the model is misspecified. Nevertheless, to the best of our knowledge, our theoretical results on multi-object, multi-view identifiability are unique and capture key concepts in object-centric representation learning, opening various new avenues for future research along the lines of generalisability, world-modelling, and planning.