Anchor–Stabilizer Theory for Identifiable Representation Learning: Generalized Anchors as Symmetry Breakers

Yuan-Hao Wei *

Posted Date: 3 June 2026

doi: 10.20944/preprints202606.0188.v1

Keywords: identifiable representation learning; nonlinear ICA; causal representation learning; generalized anchors; symmetry breaking; anchor stabilizer; latent reparameterization

 doi:10.20944/preprints202606.0188.v1

Article

Anchor–Stabilizer Theory for Identifiable Representation Learning: Generalized Anchors as Symmetry Breakers

Yuan-Hao Wei

The Hong Kong polytechnic university, Hong Kong; Yuan-Hao.Wei@outlook.com or yuan-hao.wei@connect.polyu.hk or yuanhao.wei1993@gmail.com

Abstract

Identifiable representation learning asks when a learned latent representation corresponds to generative, causal, or task-relevant factors rather than an arbitrary nonlinear reparameterization of the same observed distribution. This article develops anchor–stabilizer theory as a symmetry-breaking framework for this problem. The central idea is that positive identifiability results rely, explicitly or implicitly, on generalized anchors: conditional, temporal, structural, mechanistic, relational, semantic, objective-induced, decoder-level, or query-level resources that restrict the latent transformations left unresolved by the observed marginal distribution. An anchor is formalized as an augmented observation or constraint operator that refines baseline observational equivalence. The residual ambiguity left by the anchor is its stabilizer, namely the set of latent reparameterizations that preserve both the original observations and the anchored structure. Identifiability is obtained when this stabilizer is contained in an accepted ambiguity class, such as permutation, component-wise transformations, block transformations, graph-preserving transformations, or query-preserving transformations. Composite anchors reduce ambiguity by intersecting stabilizers, explaining why individually insufficient resources may become jointly identifying. The framework unifies auxiliary-variable nonlinear ICA, identifiable VAEs, structured nonlinear ICA, sparse and geometric decoders, causal representation learning, grouping, contrastive learning, augmentation invariance, supervised labels, and no-auxiliary approaches as instances of the same symmetry-breaking principle.

Keywords: identifiable representation learning; nonlinear ICA; causal representation learning; generalized anchors; symmetry breaking; anchor stabilizer; latent reparameterization

1. Introduction

Modern representation learning is often motivated by the assumption that high-dimensional observations are generated from lower-dimensional latent factors. In images, such factors may correspond to position, lighting, pose, shape, or identity; in scientific data, they may correspond to hidden physical, biological, or causal states. Deep latent-variable models, including variational autoencoders, normalizing flows, energy-based models, and nonlinear causal generative models, provide flexible tools for modeling such observations. Yet fitting the observed distribution is not the same as recovering the latent variables that generated the data. A model may reproduce the same marginal law of observations while representing the underlying factors in a transformed, entangled, or causally misleading coordinate system. (Hyvärinen et al., 2024; Kingma & Welling, 2014; Rezende et al., 2014; Schölkopf et al., 2021)

This gap is the identifiability problem. If observations are generated from latent variables through a nonlinear mechanism, then many invertible transformations of the latent space can be absorbed by changing the decoder or mechanism. The observed marginal law alone therefore cannot usually determine which latent coordinate system is the intended one. In nonlinear ICA this obstruction is

 doi:10.20944/preprints202606.0188.v1

fundamental: independent sources under an unrestricted nonlinear mixing map are not identifiable without additional structure. A parallel lesson appears in disentangled representation learning, where purely unsupervised disentanglement is impossible in general without inductive biases on both model and data. (Hyvärinen & Pajunen, 1999; Locatello et al., 2019)

The consequence is practical as well as conceptual. Without identifiability, learned latent variables may be unstable across retraining, incomparable across models, and unreliable for transfer, intervention, explanation, or downstream scientific discovery. The relevant question is therefore not merely whether a flexible model can fit the observed distribution, but what additional resources make some latent representation recoverable up to an acceptable residual ambiguity.

The literature on identifiable representation learning can be read as a progressive search for such symmetry-breaking resources. Classical linear ICA becomes identifiable under independence and non-Gaussianity assumptions, up to familiar permutation and scaling ambiguities. Nonlinear ICA is generally not identifiable from independent sources alone, so positive results introduce additional information: auxiliary variables, temporal dependence, nonstationarity, hidden regimes, interventions, decoder structure, grouping, multi-view relations, contrastive pairs, labels, or downstream causal queries. These resources differ in surface form, but they play the same role: they rule out latent reparameterizations that would otherwise preserve the observed marginal law. (Comon, 1994; Hälvä et al., 2021; Hyvärinen et al., 2019; Khemakhem, Kingma, et al., 2020; Lippe et al., 2022; Moran et al., 2022; von Kügelgen et al., 2023)

This article calls such resources *generalized anchors*. An anchor is any observed, relational, temporal, structural, mechanistic, semantic, objective-induced, decoder-level, or query-level resource that reduces latent reparameterization freedom. Auxiliary variables, conditional priors, temporal histories, nonstationary regimes, interventions, sparse decoders, grouping maps, contrastive pairs, augmentation invariances, labels, and downstream causal queries are all anchor candidates. They differ in where they live and what they preserve, but they share a common role: they break symmetries that would otherwise make latent representations arbitrary.

The main contribution of this article is to develop *anchor–stabilizer theory*. The baseline unidentifiability of a latent-variable model is represented by an observational equivalence relation or a latent reparameterization group. A generalized anchor induces an augmented observation or constraint operator that refines this equivalence relation. The transformations that survive the anchor form an anchor stabilizer. Identifiability up to a target ambiguity class is obtained precisely when the anchor stabilizer is contained in that target class, modulo unavoidable isotropy of the parameterization. Composite anchors act by stabilizer intersection, which provides a formal language for anchor synergy, redundancy, and incompatibility.

This perspective changes how identifiable representation learning is organized. Instead of treating nonlinear ICA, iVAE, structured nonlinear ICA, sparse decoding, causal representation learning, grouping, contrastive learning, and supervised semantic learning as unrelated technical routes, anchor–stabilizer theory asks the same questions for each result: what object is anchored, which latent transformations are removed, which transformations remain, how much of the latent system is covered, and whether the remaining ambiguity preserves the relevant downstream query.

The article makes four theoretical contributions. First, it formalizes generalized anchors as augmented observation or constraint operators that refine the baseline observational equivalence relation. Second, it defines the residual ambiguity of an anchor as a stabilizer subgroup of the baseline latent reparameterization group, yielding a common criterion for identifiability up to a target ambiguity class. Third, it proves that composite anchors compose by stabilizer intersection, which explains why individually insufficient resources may become jointly identifying. Fourth, it extends exact asymptotic identifiability with local anchor rank, robust anchor margin, anchor strength preorder, partial coverage, soft anchors, and query-preserving identifiability.

 doi:10.20944/preprints202606.0188.v1

2. Existing Routes to Identifiability as Anchor Families

The preceding section framed identifiable representation learning as a problem of reducing latent ambiguity. This section reviews the main technical routes through which the literature has attempted to achieve this reduction. The goal is not to provide a chronological catalogue of a few influential methods, but to reinterpret many methods as different ways of introducing symmetry-breaking resources into otherwise underdetermined nonlinear latent-variable models. Each route supplies a different kind of structure: an observed auxiliary variable, temporal or spatial organization, sparse generative geometry, interventions, multiple views, grouping information, contrastive relations, or environment-level variation. These resources differ in form, but they all constrain the set of latent reparameterizations that remain compatible with the observed data.

2.1. Observed Auxiliary and Conditional-Distribution Anchors

The observed-auxiliary route is one of the cleanest starting points for modern identifiable representation learning. In nonlinear ICA, source independence alone is insufficient under unrestricted nonlinear mixing, but an auxiliary variable can modulate the latent distribution in a way that rules out many spurious reparameterizations. Generalized contrastive learning uses an auxiliary variable u, such as a time index, history, class label, or other side information, to discriminate true (x, u) pairs from mismatched pairs, thereby turning conditional variation into an identifying signal. (Hyvärinen et al., 2019)

Deep generative versions of the same idea condition the latent prior or energy function on observed side information. Identifiable VAEs show that, under suitable conditional factorization and sufficient variation, a broad class of latent-variable models can identify the joint distribution over observations and latents up to simple transformations. Related flow and energy-based extensions show that the same anchor logic can be expressed through likelihood-based, flow-based, or conditional energy-based objectives, not only through one VAE architecture. (Khemakhem, Kingma, et al., 2020; Khemakhem, Monti, et al., 2020; Sorrenson et al., 2020)

The important conceptual point is not that there is one privileged auxiliary-variable method. It is that an observed u can act as an explicit anchor by forcing the representation to preserve a family of conditional distributions rather than only a marginal distribution. The strength of this anchor depends on the richness of modulation, the coverage of latent factors, and whether the factorization assumptions match the true data-generating process. (Hyvärinen et al., 2024, 2023)

The limitation is equally important. Many datasets do not provide clean, observed, correctly specified side variables. Some labels are coarse, noisy, or affect only part of the latent state; some covariates are themselves effects rather than causes; some domains are latent; and some variables useful for identifiability are never measured. This motivates routes that replace explicit auxiliary variables with process structure, decoder structure, interventions, relations among samples, or downstream queries.

2.2. Process Anchors: Time, Nonstationarity, Hidden Regimes, and Non-Invertible Observations

A second route uses structure internal to the data-generating process. Temporal dependence, spatial organization, hidden states, nonstationarity, and autocorrelation can play the role of anchors because they restrict which latent transformations preserve the joint distribution of sequences or structured observations. The identifying object is no longer simply a side variable u; it may be a transition law, a dependence pattern, or a latent regime. (Hälvä & Hyvärinen, 2020; Hälvä et al., 2021; Hyvärinen & Morioka, 2017)

Structured nonlinear ICA is a representative framework in this family. Its contribution is not only a particular model, but the broader observation that identifiability can arise from temporal structures, spatial dependencies, hidden states, and noisy observations. In anchor language, the process itself supplies constraints that the latent representation must respect. (Hälvä et al., 2021)

 doi:10.20944/preprints202606.0188.v1

Recent temporal causal representation work pushes this further. Temporally Disentangled Representation Learning studies time-delayed causally related latent processes with nonparametric transitions and asks when latent temporal variables and their relations can be identified from nonlinear mixtures. Unknown-nonstationarity methods treat domain or phase indices as latent anchors to be inferred, while sparse-transition methods identify distribution shifts through sufficiently distinct and sparse transition changes. (Song et al., 2024, 2023; Yao et al., 2022)

This family also exposes a major fragility: many results assume that the observation map is invertible. Non-invertible temporal generation, such as occlusion or projection from richer states to lower-dimensional observations, can break standard assumptions. CaRiNG addresses this by using temporal context to recover information lost at a single time point, showing that the anchor may be a window of observations rather than the current observation alone. (Chen et al., 2024)

Process anchors are valuable because they often exist without manual labels. Their difficulty is diagnosis: temporal order or nonstationarity is not automatically an anchor. It becomes one only when the process variation is rich enough to rule out the relevant reparameterizations.

2.3. Decoder, Sparsity, Geometric, and No-Auxiliary Anchors

A third route asks whether identifiability can be obtained without observed auxiliary variables. These works replace external side information with constraints on the mixing function, decoder, latent prior, or architecture. From the generalized-anchor perspective, this route is essential: it shows that an anchor need not be attached to each sample. It can be a structural property of the generative map itself. (Kivva et al., 2022; Moran et al., 2022; Zheng et al., 2022)

Sparse decoding gives the most literal example. In sparse deep generative models, each observed feature depends only on a small subset of latent factors. If each latent factor has enough anchor features, observed coordinates depending on that factor alone, then the sparse generative model can be identifiable. These anchor features pin down latent factors by breaking rotations and nonlinear entanglements that would otherwise preserve $p(x)$. (Moran et al., 2022)

Structural sparsity generalizes the same idea from explicit anchor features to the support pattern of the Jacobian of the mixing function. If only certain latent variables can influence certain observed variables, the admissible transformations are restricted. Later work relaxes all-or-nothing assumptions by considering undercomplete settings, partial sparsity, partial source dependence, and flexible grouping structures. This progression is important because real anchors may cover only part of the latent system. (Zheng et al., 2022; Zheng & Zhang, 2023)

Another no-auxiliary family constrains the function class geometrically. Conformal maps, orthogonal coordinate transformations, and related differential-geometric restrictions limit how the decoder may warp latent space. These anchors should be separated from sparsity: sparsity constrains the support of influence, whereas geometric anchors constrain local metric, angle, or derivative structure. Both reduce latent symmetry, but through different invariants. (Buchholz et al., 2022)

A further family constrains decoder algebra and composition. Additive decoders assume observations decompose into sums of block-specific contributions; under suitable conditions, latent blocks can be identified up to block-wise transformations. Interaction asymmetry extends this idea to higher-order generator interactions, proposing that within-concept interactions are more complex than across-concept interactions and using derivative block-diagonality to support disentanglement and compositional generalization. (Brady et al., 2025; Lachapelle et al., 2023)

Finally, no-auxiliary identifiability can come from prior-family and architecture constraints. Mixture priors combined with piecewise-affine decoders yield a hierarchy of identifiability strengths, including affine recovery under weak conditions and stronger recovery under additional assumptions. This route is important because it prevents the theory from implying that observed side information is always necessary. Anchor theory instead says that some resource must break the symmetry; in this case the resource is model structure. (Kivva et al., 2022)

 doi:10.20944/preprints202606.0188.v1

2.4. Intervention and Causal-Mechanism Anchors

Causal representation learning strengthens the target of identification. The goal is not only to recover statistically meaningful factors, but to learn latent variables that support interventions, mechanism modeling, graph recovery, counterfactual reasoning, and transfer. This is harder than ordinary nonlinear ICA because representation learning and causal discovery are combined: both the hidden variables and their causal structure may be unknown. (Hyvärinen et al., 2024; Schölkopf et al., 2021)

Known-target temporal intervention methods sit at the strong end of the supervision ladder. CITRIS uses temporal sequences with observed intervention targets to identify scalar and multi-dimensional causal factors. The intervention target label anchors the factor directly affected by an action, while temporal ordering helps separate causal dynamics from instantaneous reparameterization freedom. (Lippe et al., 2022)

A weaker setting uses paired observations before and after unknown interventions. Weakly supervised causal representation learning shows that pre/post pairs can identify causal variables and causal mechanisms under assumptions such as stochastic perfect interventions and sufficient coverage, even without target labels. Here the anchor is relational and mechanistic: the pair says that one mechanism has changed while other aspects of the system are preserved. (Brehmer et al., 2022; Locatello et al., 2020)

A further step removes paired counterfactual views and uses multiple interventional environments with unknown targets. Nonparametric identifiability results show that, with sufficient environments and genericity assumptions, latent causal variables and their graph can be recovered up to unavoidable ambiguities. Recent work on general environments pushes this direction further by seeking less restrictive environment changes under nonparametric mixing. (Jiang & Aragam, 2023; Ng et al., 2025; von Kügelgen et al., 2023)

Causal Component Analysis occupies an intermediate point between nonlinear ICA and full causal representation learning. It assumes the latent causal graph is known and studies recovery of the unmixing function and causal mechanisms from interventional datasets. This separates coordinate recovery from graph discovery; impossibility results under known graphs transfer to harder CRL settings, while possibility results provide stepping stones. (Liang et al., 2023)

Other work studies Gaussian or linear causal structure under arbitrary nonlinear mixing. Unknown single-node interventions can identify linear causal representations and latent causal structure by exploiting the geometry of latent precision changes after nonlinear density transformation. This bridges classical SEM thinking and modern deep representation learning: the observation map may be highly nonlinear, while the intervention family reveals a linear causal representation in latent space. (Buchholz et al., 2023)

Mechanistic anchors can also be evaluated by downstream causal semantics. Intervention extrapolation shows that affine identifiability of a latent representation can be sufficient when the goal is predicting unseen interventions on actions. Counterfactual identifiability of bijective causal models similarly shifts the target from unique coordinate recovery to equivalence classes sufficient for counterfactual queries. (Nasr-Esfahany et al., 2023; Saengkyongam et al., 2024)

Overall, interventional anchors differ from ordinary distributional anchors because they often identify modular mechanisms rather than only latent coordinates. Their residual ambiguity should be judged relative to causal tasks such as graph recovery, intervention prediction, and counterfactual estimation.

2.5. Relational, Grouping, Contrastive, Multi-View, and Semantic Anchors

A fifth route uses relations among observations. Multi-view, grouping, contrastive, augmentation-based, supervised, and concept-based methods all exploit the fact that samples can be organized by what they share and what they change. The anchor is therefore relational: it specifies an invariance,

 doi:10.20944/preprints202606.0188.v1

difference, overlap, or semantic alignment across observations. (Locatello et al., 2020; von Kügelgen et al., 2021; Zimmermann et al., 2021)

Weakly supervised disentanglement illustrates this idea in a minimal form. Observations may be paired so that some factors change while others remain fixed, and even coarse knowledge about the number of changed factors can reduce the set of admissible latent transformations. This is closely related to paired interventions in causal representation learning, but it can be studied even when the causal graph is trivial or absent. (Brehmer et al., 2022; Locatello et al., 2020)

Contrastive learning provides a broader relational mechanism. Positive pairs specify information that should be shared; negative pairs specify information that should be separated. Identifiability results for contrastive and multimodal contrastive learning show that objectives in this family can recover latent factors or shared multimodal blocks under suitable assumptions. CEBRA is an application-facing example in neuroscience: behavioral variables, time, or hybrid labels define positive and negative samples so that neural embeddings become consistent and behaviorally informative. (Daunhawer et al., 2023; Schneider et al., 2023; Zimmermann et al., 2021)

Data augmentation supplies an invariance anchor. In augmentation-based self-supervised learning, two views are modeled as sharing content while style may change, and the invariant content partition can be identified even when latent variables are statistically or causally dependent. This result is important because it weakens the assumption that useful disentanglement always requires mutually independent factors. (von Kügelgen et al., 2021)

Grouping assumptions offer an observation-space relational anchor. If observed variables are known to be grouped by sensor, modality, location, time point, or measurement design, that grouping can restrict how latent causal variables generate observations. Recent CRL results show that suitable grouping can make causal representations identifiable without temporal structure, interventions, or weak supervision. This is especially relevant for scientific settings where measurement design is often known even when interventions are unavailable. (Morioka & Hyvärinen, 2024)

Supervised labels and concept labels are semantic anchors. Cross-entropy classification can recover ground-truth factors up to a linear transformation under a cluster-centric data-generating process, while concept-based models such as GlanceNets define interpretability as alignment between learned concepts and a partially interpretable data-generating process, with attention to concept leakage. These examples broaden the anchor lens beyond unsupervised representation learning: labels can be anchors, but their residual ambiguity and leakage risks must be analyzed. (Marconato et al., 2022; Reizinger et al., 2025)

Relational anchors are often weaker than clean auxiliary-variable anchors, but they are flexible and realistic. They can identify shared factors, invariant content, blocks, concept slots, or task-aligned subspaces rather than all latent coordinates. Anchor theory therefore treats relational structure as a first-class source of identifiability, not as an informal heuristic.

2.6. Environment, Invariance, and Query-Oriented Anchors

Environment and domain variation provide another major anchor family. Data are often not i.i.d. samples from one fixed distribution; they are collected across regimes, domains, tasks, interventions, nonstationary phases, or policy-induced distributions. An environment becomes an anchor only when its variation constrains latent reparameterizations by revealing what is stable, what changes, and which mechanisms are responsible for those changes. (Song et al., 2023; von Kügelgen et al., 2023; Yao et al., 2022)

Nonstationary temporal representation learning is the clearest example. History can serve as side information, but nonstationarity adds another source of variation. TDRL exploits fixed causal dynamics and distribution shifts to recover nonparametric latent temporal processes, while unknown-nonstationarity methods infer hidden domain or phase indices and use them as latent anchors. Sparse-transition work further shows that sufficiently distinct transition clusters can reveal distribution shifts without directly observing domain variables or assuming a simple Markov prior over them. (Song et al., 2024, 2023; Yao et al., 2022)

 doi:10.20944/preprints202606.0188.v1

Environment anchors also appear in causal representation learning. Unknown interventional environments, general environments, and mechanism-change assumptions all attempt to extract latent causal variables from families of distributions rather than from one distribution. The anchor is the pattern of variation across environments, not merely the environment label itself. (Jiang & Aragam, 2023; Ng et al., 2025; von Kügelgen et al., 2023)

Downstream tasks can define what kind of identifiability is sufficient. Intervention extrapolation, counterfactual estimation, robust reinforcement learning, and concept-based prediction may not require exact component recovery. They require a representation whose residual ambiguity preserves the relevant query. This suggests a query-oriented version of anchor theory: the accepted stabilizer should be judged by whether it preserves the causal, predictive, or interpretive quantity of interest. (Dunion et al., 2023; Marconato et al., 2022; Nasr-Esfahany et al., 2023; Saengkyongam et al., 2024)

This route connects the taxonomy to the open problems. Environment variation can be powerful but hard to diagnose. Latent anchors can be learned but may introduce new nonidentifiabilities. Downstream tasks can tolerate weaker ambiguities, but only if the equivalence class preserves the query. These qualifications motivate the later sections on anchor strength, coverage, composition, robustness, and usable identifiability.

2.7. Cross-Route Comparison and Transition

Across these routes, the same technical motif reappears: add enough structure so that latent transformations preserving the observational law must also preserve a richer conditional, temporal, geometric, relational, causal, or environmental object. The routes therefore differ less in purpose than in where the extra structure lives. Observed-auxiliary approaches place it in conditional latent distributions; process approaches place it in temporal, spatial, hidden-state, or nonstationary organization; sparse and geometric approaches place it in the decoder, prior family, function class, or Jacobian structure; causal approaches place it in interventions and mechanism changes; relational approaches place it in paired, grouped, contrastive, multimodal, or semantic relations among observations; environment approaches place it in cross-distribution patterns of stability and change. (Hälvä et al., 2021; Hyvärinen et al., 2019; Lippe et al., 2022; Morioka & Hyvärinen, 2024; von Kügelgen et al., 2023; Zheng et al., 2022)

This comparison clarifies why no single route should be treated as the canonical solution to identifiability. Observed side information is mathematically transparent, but requires informative and correctly specified variables. Process anchors are natural for time series and spatial data, but depend on correctly modeled dependence. Decoder and function-class anchors reduce the need for supervision, but can be brittle when structural assumptions fail. Interventional anchors are semantically strong, but often demand data that are expensive or impossible to obtain. Relational and grouping anchors are flexible, but may identify only shared factors, blocks, or subspaces. Environment anchors are broad, but their sufficient-change assumptions are difficult to diagnose empirically. (Daunhawer et al., 2023; Kivva et al., 2022; Lachapelle et al., 2024; Locatello et al., 2019; Song et al., 2024)

Taken together, these routes show that identifiable representation learning has moved beyond a single template. The main lesson is not that the field has discovered unrelated tricks, but that it has repeatedly found resources that shrink latent reparameterization classes in systematically different ways. The next sections abstract from these routes to their shared mechanism: each turns a passive modeling assumption into an active anchor by breaking latent reparameterization freedom.

3. Observational Equivalence and Generalized Anchors

Let Θ denote a latent-variable model class. A parameter $\theta \in \Theta$ may include a decoder or mixing map, a latent law, a transition law, an intervention family, a structural causal mechanism, a sparsity pattern, a grouping map, or nuisance parameters such as observation noise. The baseline object learned from unanchored observations is the observed marginal distribution. We write this as an observation operator

$$\Phi_0 : \Theta \to \mathcal{D}_0, \qquad \Phi_0(\theta) = P_\theta(X).$$

 doi:10.20944/preprints202606.0188.v1

Two parameters are observationally equivalent without anchors when

$$\theta \sim_0 \theta' \quad \Longleftrightarrow \quad \Phi_0(\theta) = \Phi_0(\theta').$$

The baseline observational equivalence class is

$$[\theta]_0 = \{\theta' \in \Theta : \Phi_0(\theta') = \Phi_0(\theta)\}.$$

Identifiability fails when this class contains nontrivially different latent coordinate systems.

This equivalence class is large in nonlinear latent-variable models. If $X = f(Z)$ and $h : \mathcal{Z} \to \mathcal{Z}$ is an admissible invertible latent transformation, define

$$\widetilde{Z} = h(Z), \qquad \widetilde{f} = f \circ h^{-1}.$$

Then

$$\widetilde{f}(\widetilde{Z}) = (f \circ h^{-1})(h(Z)) = f(Z) = X.$$

Thus (f, P_Z) and $(\widetilde{f}, h_{\#}P_Z)$ induce the same law of X, where $h_{\#}P_Z$ denotes the pushforward of P_Z under h. This is the basic latent reparameterization symmetry behind nonlinear ICA and unsupervised disentanglement impossibility results. (Hyvärinen & Pajunen, 1999; Locatello et al., 2019)

Definition 1 (generalized anchor).

A generalized anchor is an additional object A that induces an augmented observation or constraint operator

$$\Phi_A : \Theta \to \mathcal{D}_A.$$

The anchor may be an observed auxiliary variable, a conditional distribution, a time or space index, a transition law, a hidden regime structure, an intervention family, a grouping map, a multi-view relation, a contrastive sampling rule, a sparsity pattern, an additive decoder, a semantic label, or a downstream query. The anchor-induced equivalence relation is

$$\theta \sim_A \theta' \quad \Longleftrightarrow \quad \Phi_A(\theta) = \Phi_A(\theta'),$$

with equivalence class

$$[\theta]_A = \{\theta' \in \Theta : \Phi_A(\theta') = \Phi_A(\theta)\}.$$

The anchor is informative at θ when $[\theta]_A$ is a strict refinement of $[\theta]_0$.

Definition 2 (valid, effective, and identifying anchor).

Let $\Phi_0 : \Theta \to \mathcal{D}_0$ denote the baseline observation operator. An anchor A with operator

$$\Phi_A : \Theta \to \mathcal{D}_A$$

is valid relative to Φ_0 if there exists a forgetting map

$$\pi_A : \mathcal{D}_A \to \mathcal{D}_0$$

such that

$$\Phi_0 = \pi_A \circ \Phi_A.$$

It is effective at θ if its anchored equivalence class is a proper refinement of the baseline equivalence class,

$$[\theta]_A \subsetneq [\theta]_0.$$

 doi:10.20944/preprints202606.0188.v1

It is identifying up to a target ambiguity class T if every anchor-compatible alternative belongs to the T-orbit of θ.

For purely structural or decoder-level anchors, the anchor operator may be written as a product operator

$$\Phi_A(\theta) = \big(\Phi_0(\theta), C_A(\theta)\big),$$

where C_A records the structural certificate, such as a sparsity pattern, grouping map, decoder algebra, Jacobian support, intervention family, augmentation relation, semantic partition, or objective-induced invariance. In this case the forgetting map is projection onto the first coordinate. Thus a structural anchor is valid not because the structure alone determines the observed distribution, but because the anchored object combines the observed distribution with an additional constraint that refines the baseline equivalence class.

Lemma 1 (equivalence refinement).

If $\Phi_0 = \pi_A \circ \Phi_A$, then

$$[\theta]_A \subseteq [\theta]_0.$$

Proof. If $\theta' \in [\theta]_A$, then $\Phi_A(\theta') = \Phi_A(\theta)$. Applying π_A to both sides gives

$$\pi_A(\Phi_A(\theta')) = \pi_A(\Phi_A(\theta)).$$

Since $\Phi_0 = \pi_A \circ \Phi_A$, this implies

$$\Phi_0(\theta') = \Phi_0(\theta),$$

hence $\theta' \in [\theta]_0$. Therefore

$$[\theta]_A \subseteq [\theta]_0.$$

□

Proposition 1 (constraint anchors as product refinements).

Let $C_A : \Theta \to \mathcal{C}_A$ be a structural, relational, semantic, mechanistic, or objective-induced certificate. Define

$$\Phi_A(\theta) = \big(\Phi_0(\theta), C_A(\theta)\big).$$

Then A is valid relative to Φ_0, and

$$[\theta]_A = [\theta]_0 \cap \{\theta' \in \Theta : C_A(\theta') = C_A(\theta)\}.$$

Consequently, A is effective at θ exactly when the certificate removes at least one baseline-observationally equivalent alternative.

Proof. The forgetting map is the projection

$$\pi_A(d_0, c) = d_0.$$

Therefore

$$\pi_A(\Phi_A(\theta)) = \Phi_0(\theta),$$

so A is valid. Moreover, $\theta' \in [\theta]_A$ holds if and only if

$$\Phi_0(\theta') = \Phi_0(\theta) \quad \text{and} \quad C_A(\theta') = C_A(\theta).$$

This gives the stated intersection representation. The final claim follows from whether the intersection is strict. □

This proposition is important because it prevents the theory from reducing every assumption to an informal anchor. A structural assumption becomes an anchor only when it can be represented as a

 doi:10.20944/preprints202606.0188.v1

certificate that refines the baseline observational equivalence class. If a regularizer has no associated operator, no certificate, and no stabilizer reduction, it remains an inductive bias rather than a rigorous anchor.

4. Anchor–Stabilizer Theory

The previous section defined anchors as refinements of observational equivalence. We now express the same idea through group actions. Let G_0 be a baseline latent reparameterization group acting on Θ. In a noiseless nonlinear ICA model $X = f(Z)$, an element $g \in G_0$ may be represented by an invertible transformation $h : \mathcal{Z} \to \mathcal{Z}$ acting as

$$h \cdot (f, P_Z) = (f \circ h^{-1}, h_{\#} P_Z).$$

By construction,

$$\Phi_0(h \cdot (f, P_Z)) = \Phi_0(f, P_Z).$$

More generally, G_0 denotes transformations that preserve the baseline observed object.

Identifiability is never absolute. It is always relative to an accepted residual ambiguity. Let

$$T \subseteq G_0$$

denote the target ambiguity class. In linear ICA, T may contain permutations, scalings, and sign flips. In nonlinear ICA, a common target is permutation and component-wise invertible transformations,

$$T_{\text{comp}} = \mathfrak{S}_d \ltimes \prod_{j=1}^{d} \text{Diff}(\mathcal{Z}_j).$$

In causal or task-oriented representation learning, T may be graph-preserving, mechanism-preserving, content-preserving, or query-preserving.

Definition 3 (anchor stabilizer).

Given a baseline symmetry group G_0 and an anchor A with augmented operator Φ_A, the stabilizer of A at θ is

$$G_A(\theta) = \{g \in G_0 : \Phi_A(g \cdot \theta) = \Phi_A(\theta)\}.$$

Thus $G_A(\theta)$ is the set of latent transformations that survive after the anchor has been imposed. It is the residual ambiguity left by A.

The three relevant objects are therefore as follows. The group G_0 denotes the symmetries preserving the baseline observations. The stabilizer $G_A(\theta)$ denotes the symmetries preserving both the baseline observations and the anchor. The class T denotes the symmetries regarded as acceptable residual ambiguity. An anchor is informative when

$$G_A(\theta) \subsetneq G_0,$$

and identifying when its stabilizer is contained in the accepted ambiguity class, modulo isotropy.

Definition 4 (isotropy group).

The isotropy group of θ is

$$\text{Iso}(\theta) = \{g \in G_0 : g \cdot \theta = \theta\}.$$

It captures parameter redundancies that act trivially on θ. Such redundancies are common in overparameterized neural latent-variable models.

 doi:10.20944/preprints202606.0188.v1

Theorem 1 (anchor–stabilizer identifiability, modulo isotropy).

Let G_0 be a baseline latent reparameterization group acting on Θ, and suppose the baseline observational alternatives to θ inside the model class are contained in the orbit $G_0 \cdot \theta$. Let

$$G_A(\theta) = \{g \in G_0 : \Phi_A(g \cdot \theta) = \Phi_A(\theta)\}$$

be the stabilizer of anchor A. Let $T \subseteq G_0$ be the accepted target ambiguity class. Then the anchored model is identifiable up to T at θ if and only if

$$G_A(\theta) \subseteq T \operatorname{Iso}(\theta).$$

If the action is free at θ, or if $\operatorname{Iso}(\theta) \subseteq T$, the condition reduces to

$$G_A(\theta) \subseteq T.$$

Proof. Identifiability up to T means that for every $g \in G_A(\theta)$, the anchored alternative $g \cdot \theta$ lies in the trivial orbit $T \cdot \theta$. Hence there exists $t \in T$ such that

$$g \cdot \theta = t \cdot \theta.$$

Equivalently,

$$t^{-1} g \cdot \theta = \theta,$$

so $t^{-1} g \in \operatorname{Iso}(\theta)$, and therefore

$$g \in T \operatorname{Iso}(\theta).$$

This proves

$$G_A(\theta) \subseteq T \operatorname{Iso}(\theta).$$

Conversely, if $g \in T \operatorname{Iso}(\theta)$, then $g = ts$ for some $t \in T$ and $s \in \operatorname{Iso}(\theta)$. Thus

$$g \cdot \theta = ts \cdot \theta = t \cdot \theta,$$

so the alternative is trivial up to T. □

This theorem is the formal statement behind the phrase "anchors break symmetries." A positive identifiability theorem proves that a specific anchor has a small stabilizer. Different literatures choose different anchors, but their mathematical form is the same.

5. Anchor Strength, Composition, and Robustness

Definition 5 (anchor strength preorder).

For two anchors A and B, say that A is at least as strong as B at θ, written

$$A \succeq_\theta B,$$

if

$$G_A(\theta) \subseteq G_B(\theta).$$

Say that A and B are stabilizer-equivalent at θ, written

$$A \equiv_\theta B,$$

if

$$G_A(\theta) = G_B(\theta).$$

 doi:10.20944/preprints202606.0188.v1

Say that A strictly dominates B at θ if

$$G_A(\theta) \subsetneq G_B(\theta).$$

Proposition 2 (strength order and equivalence refinement).

If $A \succeq_\theta B$, then every ambiguity that survives A also survives B. If A strictly dominates B, then A removes at least one nontrivial latent reparameterization that B fails to remove.

Proof. The statement follows directly from the stabilizer inclusion

$$G_A(\theta) \subseteq G_B(\theta).$$

□

Many practical settings contain several imperfect anchors: weak domain labels, partial temporal structure, approximate sparsity, noisy intervention information, incomplete views, or coarse grouping. The stabilizer formalism gives an exact rule for combining them.

Definition 6 (composite anchor).

Given anchors $A_1, \dots, A_m$, define their composite anchor by

$$A_{1:m} = A_1 \oplus \cdots \oplus A_m$$

with augmented operator

$$\Phi_{A_{1:m}}(\theta) = \big(\Phi_{A_1}(\theta), \dots, \Phi_{A_m}(\theta)\big).$$

Theorem 2 (stabilizer intersection).

The stabilizer of a composite anchor is

$$G_{A_{1:m}}(\theta) = \bigcap_{r=1}^{m} G_{A_r}(\theta).$$

Consequently, the model is identifiable up to T under the composite anchor whenever

$$\bigcap_{r=1}^{m} G_{A_r}(\theta) \subseteq T \ \mathrm{Iso}(\theta).$$

Proof. A transformation $g \in G_0$ lies in $G_{A_{1:m}}(\theta)$ if and only if

$$\Phi_{A_{1:m}}(g \cdot \theta) = \Phi_{A_{1:m}}(\theta).$$

By definition of $\Phi_{A_{1:m}}$, this is equivalent to

$$\Phi_{A_r}(g \cdot \theta) = \Phi_{A_r}(\theta) \quad \text{for every } r.$$

Thus

$$g \in G_{A_r}(\theta) \quad \text{for every } r,$$

which is equivalent to

$$g \in \bigcap_{r=1}^{m} G_{A_r}(\theta).$$

The identifiability statement follows from Theorem 1. □

 doi:10.20944/preprints202606.0188.v1

Definition 7 (anchor synergy, redundancy, and incompatibility).

Anchors $A_1, \ldots, A_m$ are synergistic for target ambiguity T at θ if no single anchor is identifying up to T, but their composite anchor is identifying:

$$G_{A_r}(\theta) \not\subseteq T\ \mathrm{Iso}(\theta) \quad \text{for every } r,$$

while

$$\bigcap_{r=1}^{m} G_{A_r}(\theta) \subseteq T\ \mathrm{Iso}(\theta).$$

They are redundant at θ if their stabilizers are equal or nested. They are incompatible if their joint constraint excludes the true data-generating parameter from the anchored model class.

This definition explains why identifiability can arise from combinations of individually weak assumptions. A distributional anchor may remove transformations invisible to a sparsity anchor, while the sparsity anchor may remove transformations invisible to the distributional anchor. Their intersection may therefore fall inside the accepted ambiguity class even though neither stabilizer does so alone. (Li et al., 2025; Zheng et al., 2022; Zheng & Zhang, 2023)

The composition rule also separates three common empirical situations. Complementarity means that two anchors remove different symmetry directions. Redundancy means that two anchors have the same stabilizer, or one stabilizer is contained in the other. Incompatibility means that the true generative process is not contained in the model class satisfying both anchors. Without these distinctions, the field risks accumulating isolated sufficient conditions rather than developing a calculus of anchor composition.

5.1. Local Anchor Rank

The preceding results are global. To describe anchor strength locally, suppose G_0 is a Lie group with Lie algebra $\mathfrak{g}_0$. Let

$$\rho_\theta : \mathfrak{g}_0 \to T_\theta \Theta$$

be the infinitesimal action,

$$\rho_\theta(\xi) = \left.\frac{d}{dt}\right|_{t=0} \exp(t\xi) \cdot \theta.$$

The differential of the anchor operator is

$$D\Phi_A(\theta) : T_\theta \Theta \to T_{\Phi_A(\theta)} \mathcal{D}_A.$$

Definition 8 (infinitesimal anchor stabilizer).

The infinitesimal stabilizer of anchor A is

$$\mathfrak{g}_A(\theta) = \{\xi \in \mathfrak{g}_0 : D\Phi_A(\theta)\rho_\theta(\xi) = 0\}.$$

It contains the infinitesimal latent transformations that the anchor cannot detect.

Definition 9 (local anchor rank).

The local rank of anchor A at θ is

$$r_A(\theta) = \mathrm{rank}\big(D\Phi_A(\theta) \circ \rho_\theta\big).$$

Equivalently,

$$r_A(\theta) = \dim \mathfrak{g}_0 - \dim \mathfrak{g}_A(\theta).$$

Thus r_A measures how many infinitesimal symmetry directions are removed by the anchor.

 doi:10.20944/preprints202606.0188.v1

Theorem 3 (local anchor identifiability).

Let $\mathfrak{t}$ be the Lie algebra of the accepted residual group T. If

$$\mathfrak{g}_A(\theta) \subseteq \mathfrak{t},$$

then the model is locally identifiable up to T at θ.

Proof. A local non-identifiability direction is an infinitesimal symmetry $\xi \in \mathfrak{g}_0$ such that the path

$$\exp(t\xi) \cdot \theta$$

leaves the anchored object unchanged to first order:

$$\left.\frac{d}{dt}\right|_{t=0} \Phi_A(\exp(t\xi) \cdot \theta) = 0.$$

By the chain rule this is

$$D\Phi_A(\theta)\rho_\theta(\xi) = 0,$$

so $\xi \in \mathfrak{g}_A(\theta)$. If

$$\mathfrak{g}_A(\theta) \subseteq \mathfrak{t},$$

then every undetected local direction is trivial. Therefore the model is locally identifiable up to T. □

5.2. Robust Margins and Approximate Identifiability

Exact stabilizer inclusion is an asymptotic population-level condition. In practice, anchors may be noisy, partial, weak, estimated from finite data, or misspecified. A robust version can be stated using an anchor discrepancy d_A on $\mathcal{D}_A$. Define the anchor violation of a transformation $g \in G_0$ as

$$\Delta_A(g;\theta) = d_A\big(\Phi_A(g \cdot \theta), \Phi_A(\theta)\big).$$

For $\delta > 0$, define the anchor margin outside T by

$$\gamma_A(\theta;\delta) = \inf_{g \in G_0 : \mathrm{dist}(g,T) \geq \delta} \Delta_A(g;\theta).$$

Theorem 4 (robust anchor recovery).

Suppose an empirical anchored object $\widehat{\Phi}_A$ satisfies

$$d_A(\widehat{\Phi}_A, \Phi_A(\theta)) \leq \epsilon.$$

If

$$2\epsilon < \gamma_A(\theta;\delta),$$

then no empirical solution of the form $g \cdot \theta$ with

$$\mathrm{dist}(g, T) \geq \delta$$

can match the anchored data within error ϵ. Hence all such empirical solutions must lie within δ of the accepted ambiguity class T.

Proof. Assume, for contradiction, that there exists $g \in G_0$ with

$$\mathrm{dist}(g, T) \geq \delta$$

and

$$d_A(\Phi_A(g \cdot \theta), \widehat{\Phi}_A) \leq \epsilon.$$

 doi:10.20944/preprints202606.0188.v1

By the triangle inequality,

$$\begin{aligned} & d_A(\Phi_A(g\cdot\theta),\Phi_A(\theta)) \\ & \quad \le d_A\Big(\Phi_A(g\cdot\theta),\widehat{\Phi}_A\Big) \\ & \qquad + d_A\Big(\widehat{\Phi}_A,\Phi_A(\theta)\Big) \\ & \quad \le 2\epsilon. \end{aligned}$$

But by definition of $\gamma_A(\theta;\delta)$, every such g must satisfy

$$d_A(\Phi_A(g\cdot\theta),\Phi_A(\theta)) \ge \gamma_A(\theta;\delta).$$

Thus $\gamma_A(\theta;\delta) \le 2\epsilon$, contradicting the assumption. □

A useful diagnostic reading of the margin is that finite-sample or noisy estimates of the anchored object can exclude nontrivial symmetries only when the population anchor separates them by a positive amount. Anchor strength asks how much of G_0 is removed, locally measured by the rank of $D\Phi_A(\theta)\circ\rho_\theta$. Anchor coverage asks which latent components are constrained. Anchor observability asks whether the anchor is measured, inferred, or assumed. Anchor reliability asks whether it is noisy or misspecified. These quantities are not yet standardized, but they are needed if anchor theory is to become a practical framework rather than only a vocabulary.

5.3. Query-Preserving Identifiability

For causal and task-oriented representation learning, the target is often not full latent recovery but preservation of a query. Let

$$q:\Theta\to\mathcal{Q}$$

be a target query, such as a causal graph, an intervention distribution, a counterfactual functional, a mechanism class, or a downstream prediction functional.

Definition 10 (query-preserving identifiability).

The query-preserving ambiguity class at θ is

$$T_q(\theta) = \{g\in G_0 : q(g\cdot\theta) = q(\theta)\}.$$

An anchor A identifies the query q at θ if

$$G_A(\theta) \subseteq T_q(\theta)\,\mathrm{Iso}(\theta).$$

This definition separates full latent identifiability from query identifiability. A representation may fail to identify all latent coordinates while still identifying the graph, intervention distribution, counterfactual, content variable, or downstream scientific quantity of interest. (Dunion et al., 2023; Nasr-Esfahany et al., 2023; Saengkyongam et al., 2024)

6. Existing Results as Anchor–Stabilizer Instances

The anchor–stabilizer formalism is useful only if it recovers existing positive identifiability results. The following principle states how this works.

Theorem 5 (anchor representation principle).

Suppose an identifiability theorem has the form

$$\Phi_A(\theta) = \Phi_A(\theta') \quad\Longrightarrow\quad \theta' \in T\cdot\theta,$$

 doi:10.20944/preprints202606.0188.v1

where Φ_A is an augmented object used by the theorem, and suppose that $\Phi_0 = \pi_A \circ \Phi_A$ for some forgetting map π_A. Then the theorem can be written as the anchor–stabilizer statement

$$G_A(\theta) \subseteq T \operatorname{Iso}(\theta).$$

Proof. If $g \in G_A(\theta)$, then

$$\Phi_A(g \cdot \theta) = \Phi_A(\theta).$$

Applying the assumed identifiability theorem with $\theta' = g \cdot \theta$ yields

$$g \cdot \theta \in T \cdot \theta.$$

By Theorem 1, this is equivalent to

$$g \in T \operatorname{Iso}(\theta).$$

Thus

$$G_A(\theta) \subseteq T \operatorname{Iso}(\theta).$$

□

Auxiliary-variable nonlinear ICA and iVAE.

Auxiliary-variable nonlinear ICA and iVAE provide the cleanest observed-anchor case. The anchor is an observed variable U, such as a time index, history, label, or other side information:

$$A = U, \qquad \Phi_U(\theta) = P_\theta(X \mid U).$$

The forgetting map marginalizes over U:

$$\pi_U(P(X \mid U)) = \int P(X \mid U = u)\, dP(u).$$

The sufficient variation of the conditional latent law restricts the transformations that preserve $P(X \mid U)$, shrinking the stabilizer from arbitrary nonlinear reparameterizations to a much smaller ambiguity class. In iVAE, the conditional factorized exponential-family prior supplies precisely this anchor. In anchor language,

$$G_U(\theta) \subseteq T_A,$$

where T_A is the residual ambiguity class allowed by the iVAE identifiability theorem, and under stronger assumptions this can reduce to block-permutation or component-wise ambiguity. (Hyvärinen et al., 2019; Khemakhem, Kingma, et al., 2020; Khemakhem, Monti, et al., 2020; Sorrenson et al., 2020)

Structured nonlinear ICA.

For structured nonlinear ICA, the anchor is not necessarily an observed auxiliary variable but the structured joint law of a process indexed by time, space, or another structured set:

$$A = \text{process structure},$$
$$\Phi_A(\theta) = \left\{ P_\theta(X_{t_1}, \ldots, X_{t_m}) \;:\; m \geq 2 \right\}.$$

Temporal dependence, spatial structure, nonstationarity, hidden regimes, and noise structure constrain which latent transformations preserve the joint process law. In stabilizer terms,

$$G_{\text{SNICA}}(\theta) \subseteq T_{\text{translation}} \circ T_{\text{comp}},$$

 doi:10.20944/preprints202606.0188.v1

and after centering or ignoring unavoidable noise translation,

$$G_{\mathrm{SNICA}}(\theta) \subseteq T_{\mathrm{comp}}.$$

Thus process structure functions as an anchor even when no clean observed side variable is supplied. (Hälvä & Hyvärinen, 2020; Hälvä et al., 2021; Song et al., 2024, 2023; Yao et al., 2022)

Sparse decoding and structural sparsity.

Sparse decoding and structural sparsity move the anchor into the decoder. For sparse decoding, the anchor may be an observation–factor dependency graph with pure or anchor features:

$$\Phi_{\mathrm{feat}}(\theta) = \mathcal{H}_{X,Z}(\theta).$$

For structural sparsity, the anchor may be the Jacobian support of the mixing function:

$$\Phi_{\mathrm{sparse}}(\theta) = \mathrm{supp}(J_{f_\theta}).$$

The stabilizer is

$$G_{\mathrm{sparse}}(\theta) = \{h \in G_0 : \mathrm{supp}\, J_{f\circ h^{-1}} = \mathrm{supp}\, J_f\}.$$

Since

$$J_{f\circ h^{-1}}(z) = J_f(h^{-1}(z))J_{h^{-1}}(z),$$

a generic dense transformation h destroys sparsity. Under appropriate graph or sparsity assumptions, only permutation, component-wise, block-wise, or otherwise restricted transformations remain. (Moran et al., 2022; Zheng et al., 2022; Zheng & Zhang, 2023)

Geometric and decoder-algebra anchors.

Function-class restrictions such as conformality, orthogonality, additive decoders, mixture priors with piecewise-affine decoders, and interaction-asymmetry constraints also act as structural anchors. Their common form is a certificate

$$C_A(\theta) = \text{decoder geometry or algebra},$$

so that

$$\Phi_A(\theta) = \big(\Phi_0(\theta), C_A(\theta)\big).$$

The stabilizer contains transformations that preserve the specified decoder geometry or algebra. When the certificate is sufficiently rigid, the residual ambiguity reduces to affine, block, component-wise, or compositional transformations. (Brady et al., 2025; Buchholz et al., 2022; Kivva et al., 2022; Lachapelle et al., 2023)

Intervention and causal-mechanism anchors.

For intervention-based causal representation learning, the anchor is an environment or intervention family:

$$\Phi_{\mathrm{int}}(\theta) = \{P^e_\theta(X)\}_{e\in E}.$$

The stabilizer contains transformations preserving the pattern of mechanism changes:

$$G_{\mathrm{int}}(\theta) = \{g \in G_0 : \{P^e_{g\cdot\theta}(X)\}_{e\in E} = \{P^e_\theta(X)\}_{e\in E}\}.$$

If a transformation mixes intervened and non-intervened factors in a way that destroys modularity, it is removed. Depending on the assumptions, the residual group may preserve causal variables, causal graphs, mechanisms, or only a class of causal queries. (Brehmer et al., 2022; Buchholz et al., 2023; Jiang

 doi:10.20944/preprints202606.0188.v1

& Aragam, 2023; Lachapelle et al., 2024; Liang et al., 2023; Lippe et al., 2022; Locatello et al., 2020; Ng et al., 2025; von Kügelgen et al., 2023)

Grouping, multi-view, contrastive, augmentation, and semantic anchors.

Relational anchors specify what is shared, changed, grouped, invariant, or semantically aligned across observations. Pairing and contrastive sampling may define

$$\Phi_A(\theta) = P_\theta(X, X^+),$$

grouping may define

$$\Phi_A(\theta) = \Pi_X,$$

augmentation may define a content-invariant view law, and supervision may define

$$\Phi_A(\theta) = P_\theta(Y \mid X) \quad \text{or} \quad P_\theta(X \mid Y).$$

These anchors often do not identify all latent coordinates. Instead, their stabilizers preserve shared factors, content variables, observed groups, class partitions, semantic concepts, or query-relevant subspaces. (Daunhawer et al., 2023; Locatello et al., 2020; Marconato et al., 2022; Morioka & Hyvärinen, 2024; Reizinger et al., 2025; Schneider et al., 2023; von Kügelgen et al., 2021; Zimmermann et al., 2021)

Table 1. Anchor families as constraint refinements of the baseline observational equivalence class.

Anchor family	Anchored object	Stabilizer condition	Typical residual class
Auxiliary variable	conditional law	preserves modulation by U	component-wise / linear
Process structure	structured joint law	preserves temporal or spatial dependence	component-wise + translation
Sparse decoder	Jacobian support or decoder graph	preserves sparsity certificate	component-wise / block
Sparse features	feature–factor graph	preserves pure-feature structure	permutation / scaling
Geometric decoder	metric or differential structure	preserves decoder geometry	affine / orthogonal / component-wise
Interventions	environment family	preserves mechanism changes	causal / graph-preserving
Grouping	observed partition	preserves group constraints	block / causal
Contrastive pairs	pair relation	preserves shared-vs-changing factors	content / shared subspace
Augmentations	invariant view law	preserves content invariance	content-preserving
Labels	class-conditional structure	preserves semantic partition	linear / semantic
Queries	target functional	preserves $q(\theta)$	query-preserving

7. Residual Ambiguity and Soft Anchors

7.1. A Residual-Ambiguity Ladder

The stabilizer view implies that identifiability should always be reported with its residual ambiguity. A useful ladder is

$$T_{\text{perm}} \subseteq T_{\text{comp}} \subseteq T_{\text{block}} \subseteq T_{\text{subspace}} \subseteq G_0.$$

If

$$G_A(\theta) \subseteq T_{\text{comp}},$$

we have component-wise identifiability. If

$$G_A(\theta) \subseteq T_{\text{block}},$$

we have block identifiability. If

$$G_A(\theta) \subseteq T_{\text{subspace}},$$

we have subspace or partition recovery. For causal or task-oriented applications, the relevant target may be a query-preserving class:

$$T_q(\theta) = \{g \in G_0 : q(g \cdot \theta) = q(\theta)\}.$$

 doi:10.20944/preprints202606.0188.v1

If

$$G_A(\theta) \subseteq T_q(\theta)\,\mathrm{Iso}(\theta),$$

then the representation is identifiable for the query even if it is not component-wise identifiable.

This ladder prevents identifiability from being treated as a binary property. Different anchors break different symmetries and leave different residual groups. A mature theory should therefore state not only whether a representation is identifiable, but also whether the remaining ambiguity is permutation, component-wise transformation, affine map, block transformation, subspace rotation, graph-preserving map, content-style decomposition, or query-preserving equivalence.

7.2. Soft Anchors and Strict Distinctness

Not every symmetry-breaking resource delivers full identifiability. Some anchors are soft: they bias the learning procedure against collapse, redundancy, or branch homogeneity without eliminating all nontrivial latent reparameterizations at the population level. Structured latent architectures are examples of such soft anchors. They can break exact permutation symmetry among branches, encourage different dimensions to serve different roles, and create pressure toward non-redundant representations, but they should not be claimed to yield full identifiability without additional rigidity assumptions.

Let

$$Z = g_\phi(X) = (Z_1, \ldots, Z_d),$$

and suppose each branch Z_j is associated with a structural parameter α_j, such as a kernel parameter, transition parameter, mixture parameter, prior family parameter, or branch-specific decoder parameter. Let $\mathcal{A}$ denote the structural-parameter space, and let

$$d_{\mathcal{A}} : \mathcal{A} \times \mathcal{A} \to \mathbb{R}_{\geq 0}$$

be a distance or discrepancy on that space.

Definition 11 ($(\nu, \epsilon, \eta, \delta)$-strictly distinct representation).

A representation $Z = (Z_1, \ldots, Z_d)$ is $(\nu, \epsilon, \eta, \delta)$-strictly distinct if it satisfies the following conditions:

$$\mathrm{Var}(Z_j) \geq \nu \quad \text{for all } j,$$

$$\mathrm{Dep}(Z_i, Z_j) \leq \epsilon \quad \text{for all } i \neq j,$$

$$I(X; Z_j \mid Z_{-j}) \geq \eta \quad \text{for all } j,$$

and

$$d_{\mathcal{A}}(\alpha_i, \alpha_j) \geq \delta \quad \text{for all } i \neq j.$$

Here Dep may be covariance, mutual information, HSIC, distance correlation, or another dependence measure.

These conditions rule out collapsed, redundant, and structurally homogeneous branches. They do not by themselves imply full identifiability, because a representation can be non-collapsed and non-redundant without recovering the true latent factors. Strict distinctness is therefore weaker than identifiability but stronger than ordinary anti-collapse.

The hierarchy is

$$\begin{aligned} \text{anti-collapse} &< \text{strict distinctness} \\ &< \text{disentanglement} \\ &< \text{identifiability}. \end{aligned}$$

 doi:10.20944/preprints202606.0188.v1

A soft structural anchor may justify the implication

$$G_A(\theta) \subseteq G_{\text{distinct}},$$

where G_{distinct} denotes transformations preserving strict distinctness. If an additional rigidity condition implies

$$G_{\text{distinct}} \subseteq T_{\text{block}},$$

then strict distinctness can be upgraded to block-level identifiability. If a stronger rigidity condition implies

$$G_{\text{distinct}} \subseteq T,$$

then it can be upgraded to full identifiability up to T.

A training objective for this purpose may combine reconstruction, source-wise prior matching, anti-collapse, redundancy reduction, branch usage, and structural separation:

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{prior}} + \lambda_{\text{var}}\mathcal{R}_{\text{var}} + \lambda_{\text{dep}}\mathcal{R}_{\text{dep}} + \lambda_{\text{info}}\mathcal{R}_{\text{info}} \\ + \lambda_{\text{sep}}\mathcal{R}_{\text{sep}}.\end{aligned}$$

The variance term prevents collapse, the dependence term reduces redundancy, the information term encourages each branch to contribute unique information, and the separation term discourages structurally identical branches. Such a construction should be presented as a soft-anchor route to strict distinctness, not as an automatic proof of full identifiability. (D'Amour et al., 2022; Locatello et al., 2019)

8. Conclusion

Identifiable representation learning is best understood as the study of how additional resources break latent reparameterization symmetries. The observed marginal distribution usually leaves a large equivalence class of latent explanations. Generalized anchors refine this equivalence class by requiring the representation to preserve conditional, temporal, structural, mechanistic, relational, semantic, objective-induced, decoder-level, or query-level objects beyond the marginal law.

Anchor–stabilizer theory gives this idea a formal structure. The residual ambiguity under an anchor is the stabilizer

$$G_A(\theta) = \{g \in G_0 : \Phi_A(g \cdot \theta) = \Phi_A(\theta)\}.$$

Identifiability up to a target ambiguity class T is obtained when this stabilizer lies inside T, modulo isotropy. Composite anchors compose by stabilizer intersection, local anchor rank measures infinitesimal symmetry breaking, robust anchor margins quantify approximate identifiability, and query-preserving ambiguity classes distinguish full latent recovery from task-relevant recovery.

This framework unifies auxiliary-variable nonlinear ICA, identifiable VAEs, structured nonlinear ICA, sparse and geometric decoders, intervention-based causal representation learning, grouping, contrastive learning, augmentation invariance, supervised labels, and no-auxiliary approaches as instances of the same symmetry-breaking template. Their differences are not superficial: each anchors a different object, removes a different subset of latent transformations, and leaves a different residual ambiguity.

The purpose of anchor theory is therefore not to claim that every assumption is an anchor. It is to provide criteria for when a resource genuinely refines observational equivalence, how much symmetry it removes, how it composes with other resources, and whether its residual ambiguity is acceptable for the scientific or causal query at hand. In this sense, identifiable representation learning should move from a list of isolated sufficient assumptions toward a comparative theory of anchor strength, coverage, composition, robustness, and query-preserving identifiability.

 doi:10.20944/preprints202606.0188.v1

Several open directions follow naturally from this view. First, many practical anchors are weak or partial: they may cover only some latent variables, mechanisms, blocks, views, or environments. Future theory should therefore make anchor coverage explicit and characterize when partial anchors yield component-wise, block-level, subspace-level, mechanism-level, or query-level identifiability. (Lachapelle et al., 2024; Zheng & Zhang, 2023)

Second, anchors may be noisy or misspecified. Side information can be unreliable, intervention labels can be incorrect, grouping assumptions can be approximate, and sparsity patterns can be violated by weak cross-factor effects. A weak but valid anchor may still provide useful partial recovery, whereas a strong but invalid anchor may stabilize the wrong representation. This distinction suggests that anchor validity, anchor strength, and anchor robustness should be analyzed separately. (D'Amour et al., 2022; Hälvä et al., 2021)

Third, many important anchors are latent rather than observed. Hidden regimes, unknown interventions, latent environments, sparse transition clusters, and implicit groupings may need to be inferred jointly with the representation. This creates a circularity: the anchor helps identify the representation, but the representation may be needed to discover the anchor. Understanding when such latent-anchor discovery is identifiable remains an important frontier. (Hälvä & Hyvärinen, 2020; Song et al., 2024, 2023; von Kügelgen et al., 2023)

Finally, identifiability in principle should be distinguished from recovery in practice. Most theorems assume correct specification, infinite data, exact optimization, and valid anchors, while deep representation learning is affected by finite samples, model misspecification, posterior collapse, nonconvex optimization, and proxy evaluation metrics. (Locatello et al., 2019) A usable theory of identifiability should therefore provide diagnostics for anchor validity, residual ambiguity, sensitivity to misspecification, finite-sample reliability, and optimization failure. Anchor–stabilizer theory provides the language for this agenda, but a complete theory will require quantitative tools for measuring how observed, structural, relational, mechanistic, semantic, environmental, and query-defined resources break latent symmetries in realistic settings.

References

Brady, J., von Kügelgen, J., Lachapelle, S., Buchholz, S., Kipf, T., & Brendel, W. (2025). Interaction Asymmetry: A General Principle for Learning Composable Abstractions. In *International conference on learning representations*. Available online: https://openreview.net/forum?id=cCl10IU836 (accessed on).

Brehmer, J., de Haan, P., Lippe, P., & Cohen, T. S. (2022). Weakly Supervised Causal Representation Learning. In *Advances in neural information processing systems* (Vol. 35, pp. 38319–38331). Available online: https://proceedings.neurips.cc/paper_files/paper/2022/hash/fa567e2b2c870f8f09a87b6e73370869-Abstract-Conference.html (accessed on).

Buchholz, S., Besserve, M., & Schölkopf, B. (2022). Function Classes for Identifiable Nonlinear Independent Component Analysis. In *Advances in neural information processing systems* (Vol. 35, pp. 16946–16961). Available online: https://proceedings.neurips.cc/paper_files/paper/2022/hash/6c5da478b9d13f541993d67897a0bb30-Abstract-Conference.html (accessed on).

Buchholz, S., Rajendran, G., Rosenfeld, E., Aragam, B., Schölkopf, B., & Ravikumar, P. (2023). Learning Linear Causal Representations from Interventions under General Nonlinear Mixing. In *Advances in neural information processing systems* (Vol. 36). Available online: https://openreview.net/forum?id=q131tA7HCT (accessed on).

Chen, G., Shen, Y., Chen, Z., Song, X., Sun, Y., Yao, W., Liu, X., & Zhang, K. (2024). CaRiNG: Learning Temporal Causal Representation under Non-Invertible Generation Process. In *Proceedings of the 41st international conference on machine learning* (Vol. 235, pp. 7236–7259). PMLR. Available online: https://proceedings.mlr.press/v235/chen24ai.html (accessed on).

Comon, P. (1994). Independent Component Analysis, a New Concept? *Signal Processing*, *36*(3), 287–314. Available online: https://doi.org/10.1016/0165-1684(94)90029-9 (accessed on). https://doi.org/10.1016/0165-1684(94)90029-9.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., et al. (2022). Underspecification Presents Challenges for Credibility in Modern Machine Learning. *Journal of Machine Learning Research*, *23*(226), 1–61. Available online: https://www.jmlr.org/papers/v23/20-1335.html (accessed on).

 doi:10.20944/preprints202606.0188.v1

Daunhawer, I., Bizeul, A., Palumbo, E., Marx, A., & Vogt, J. E. (2023). Identifiability Results for Multimodal Contrastive Learning. In *International conference on learning representations*. Available online: https://openreview.net/forum?id=U_2kuqoTcB (accessed on).

Dunion, M., McInroe, T., Luck, K. S., Hanna, J. P., & Albrecht, S. V. (2023). Conditional Mutual Information for Disentangled Representations in Reinforcement Learning. In *Advances in neural information processing systems* (Vol. 36, pp. 80111–80129). Available online: https://openreview.net/forum?id=EmYWJsyad4 (accessed on).

Hälvä, H., & Hyvärinen, A. (2020). Hidden Markov Nonlinear ICA: Unsupervised Learning from Nonstationary Time Series. In *Proceedings of the 36th conference on uncertainty in artificial intelligence* (Vol. 124, pp. 939–948). PMLR. Available online: https://proceedings.mlr.press/v124/halva20a.html (accessed on).

Hälvä, H., Le Corff, S., Lehéricy, L., So, J., Zhu, Y., Gassiat, E., & Hyvärinen, A. (2021). Disentangling Identifiable Features from Noisy Data with Structured Nonlinear ICA. In *Advances in neural information processing systems* (Vol. 34). Available online: https://proceedings.neurips.cc/paper/2021/hash/0cdbb4e65815fbaf79689b15482e7575-Abstract.html (accessed on).

Hyvärinen, A., Khemakhem, I., & Monti, R. (2024). Identifiability of Latent-Variable and Structural-Equation Models: From Linear to Nonlinear. *Annals of the Institute of Statistical Mathematics*, *76*(1), 1–33. Available online: https://doi.org/10.1007/s10463-023-00884-4 (accessed on). https://doi.org/10.1007/s10463-023-00884-4.

Hyvärinen, A., Khemakhem, I., & Morioka, H. (2023). Nonlinear independent component analysis for principled disentanglement in unsupervised deep learning. *Patterns*, *4*(10), 100844. Available online: https://doi.org/10.1016/j.patter.2023.100844 (accessed on). https://doi.org/10.1016/j.patter.2023.100844.

Hyvärinen, A., & Morioka, H. (2017). Nonlinear ICA of Temporally Dependent Stationary Sources. In *Proceedings of the 20th international conference on artificial intelligence and statistics* (Vol. 54, pp. 460–469). PMLR. Available online: https://proceedings.mlr.press/v54/hyvarinen17a.html (accessed on).

Hyvärinen, A., & Pajunen, P. (1999). Nonlinear Independent Component Analysis: Existence and Uniqueness Results. *Neural Networks*, *12*(3), 429–439. Available online: https://doi.org/10.1016/S0893-6080(98)00140-3 (accessed on). https://doi.org/10.1016/S0893-6080(98)00140-3.

Hyvärinen, A., Sasaki, H., & Turner, R. E. (2019). Nonlinear ICA Using Auxiliary Variables and Generalized Contrastive Learning. In *Proceedings of the twenty-second international conference on artificial intelligence and statistics* (Vol. 89, pp. 859–868). PMLR. Available online: https://proceedings.mlr.press/v89/hyvarinen19a.html (accessed on).

Jiang, Y., & Aragam, B. (2023). Learning Nonparametric Latent Causal Graphs with Unknown Interventions. In *Advances in neural information processing systems* (Vol. 36). Available online: https://proceedings.neurips.cc/paper_files/paper/2023/hash/bdeab378efe6eb289714e2a5abc6ed42-Abstract-Conference.html (accessed on).

Khemakhem, I., Kingma, D. P., Monti, R. P., & Hyvärinen, A. (2020). Variational Autoencoders and Nonlinear ICA: A Unifying Framework. In *Proceedings of the twenty third international conference on artificial intelligence and statistics* (Vol. 108, pp. 2207–2217). PMLR. Available online: https://proceedings.mlr.press/v108/khemakhem20a.html (accessed on).

Khemakhem, I., Monti, R. P., Kingma, D. P., & Hyvärinen, A. (2020). ICE-BeeM: Identifiable Conditional Energy-Based Deep Models Based on Nonlinear ICA. In *Advances in neural information processing systems* (Vol. 33). Available online: https://proceedings.neurips.cc/paper/2020/hash/962e56a8a0b0420d87272a682bfd1e53-Abstract.html (accessed on).

Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. In *International conference on learning representations*. Available online: https://arxiv.org/abs/1312.6114 (accessed on).

Kivva, B., Rajendran, G., Ravikumar, P., & Aragam, B. (2022). Identifiability of Deep Generative Models Without Auxiliary Information. In *Advances in neural information processing systems* (Vol. 35, pp. 15687–15701). Available online: https://proceedings.neurips.cc/paper_files/paper/2022/hash/649f080d8891ab4d4b262cb9cd52e69a-Abstract-Conference.html (accessed on).

Lachapelle, S., Mahajan, D., Mitliagkas, I., & Lacoste-Julien, S. (2023). Additive Decoders for Latent Variables Identification and Cartesian-Product Extrapolation. In *Advances in neural information processing systems* (Vol. 36). Available online: https://openreview.net/forum?id=R6KJN1AUAR (accessed on).

Lachapelle, S., Rodríguez López, P., Sharma, Y., Everett, K., Le Priol, R., Lacoste, A., & Lacoste-Julien, S. (2024). *Nonparametric partial disentanglement via mechanism sparsity: Sparse actions, interventions and sparse temporal dependencies*. Available online: https://arxiv.org/abs/2401.04890 (accessed on).

 doi:10.20944/preprints202606.0188.v1

Li, Z., Fan, S., Zheng, Y., Ng, I., Xie, S., Chen, G., Dong, X., Cai, R., & Zhang, K. (2025). Synergy Between Sufficient Changes and Sparse Mixing Procedure for Disentangled Representation Learning. In *International conference on learning representations*. Available online: https://openreview.net/forum?id=G1r2rBkUdu (accessed on).

Liang, W., Kekić, A., von Kügelgen, J., Buchholz, S., Besserve, M., Gresele, L., & Schölkopf, B. (2023). Causal Component Analysis. In *Advances in neural information processing systems* (Vol. 36). Available online: https://proceedings.neurips.cc/paper_files/paper/2023/hash/67089958e98b243d5cc1881ad60418b8-Abstract-Conference.html (accessed on).

Lippe, P., Magliacane, S., Löwe, S., Asano, Y. M., Cohen, T., & Gavves, S. (2022). CITRIS: Causal Identifiability from Temporal Intervened Sequences. In *Proceedings of the 39th international conference on machine learning* (Vol. 162, pp. 13557–13603). PMLR. Available online: https://proceedings.mlr.press/v162/lippe22a.html (accessed on).

Locatello, F., Bauer, S., Lucic, M., Rätsch, G., Gelly, S., Schölkopf, B., & Bachem, O. (2019). Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. In *Proceedings of the 36th international conference on machine learning* (Vol. 97, pp. 4114–4124). PMLR. Available online: https://proceedings.mlr.press/v97/locatello19a.html (accessed on).

Locatello, F., Poole, B., Rätsch, G., Schölkopf, B., Bachem, O., & Tschannen, M. (2020). Weakly-Supervised Disentanglement Without Compromises. In *Proceedings of the 37th international conference on machine learning* (Vol. 119, pp. 6348–6359). PMLR. Available online: https://proceedings.mlr.press/v119/locatello20a.html (accessed on).

Marconato, E., Passerini, A., & Teso, S. (2022). GlanceNets: Interpretable, Leak-Proof Concept-Based Models. In *Advances in neural information processing systems* (Vol. 35, pp. 21212–21227). Available online: https://openreview.net/forum?id=J7zY9j75GoG (accessed on).

Moran, G. E., Sridhar, D., Wang, Y., & Blei, D. M. (2022). Identifiable Deep Generative Models via Sparse Decoding. *Transactions on Machine Learning Research*. Available online: https://openreview.net/forum?id=vd0onGWZbE (accessed on).

Morioka, H., & Hyvärinen, A. (2024). Causal Representation Learning Made Identifiable by Grouping of Observational Variables. In *Proceedings of the 41st international conference on machine learning* (Vol. 235, pp. 36249–36293). PMLR. Available online: https://proceedings.mlr.press/v235/morioka24a.html (accessed on).

Nasr-Esfahany, A., Alizadeh, M., & Shah, D. (2023). Counterfactual Identifiability of Bijective Causal Models. In *Proceedings of the 40th international conference on machine learning* (Vol. 202, pp. 25733–25754). PMLR. Available online: https://proceedings.mlr.press/v202/nasr-esfahany23a.html (accessed on).

Ng, I., Xie, S., Dong, X., Spirtes, P., & Zhang, K. (2025). Causal Representation Learning from General Environments under Nonparametric Mixing. In *Proceedings of the 28th international conference on artificial intelligence and statistics* (Vol. 258, pp. 3700–3708). PMLR. Available online: https://proceedings.mlr.press/v258/ng25a.html (accessed on).

Reizinger, P., Bizeul, A., Juhos, A., Vogt, J. E., Balestriero, R., Brendel, W., & Klindt, D. (2025). Cross-Entropy Is All You Need To Invert the Data Generating Process. In *International conference on learning representations*. Available online: https://openreview.net/forum?id=hrqNOxpItr (accessed on).

Rezende, D. J., Mohamed, S., & Wierstra, D. (2014). Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *Proceedings of the 31st international conference on machine learning* (Vol. 32, pp. 1278–1286). PMLR. Available online: https://proceedings.mlr.press/v32/rezende14.html (accessed on).

Saengkyongam, S., Rosenfeld, E., Ravikumar, P., Pfister, N., & Peters, J. (2024). Identifying Representations for Intervention Extrapolation. In *International conference on learning representations*. Available online: https://openreview.net/forum?id=3cuJwmPxXj (accessed on).

Schneider, S., Lee, J. H., & Mathis, M. W. (2023). Learnable Latent Embeddings for Joint Behavioural and Neural Analysis. *Nature*, *617*, 360–368. Available online: https://doi.org/10.1038/s41586-023-06031-6 (accessed on). https://doi.org/10.1038/s41586-023-06031-6.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward Causal Representation Learning. *Proceedings of the IEEE*, *109*(5), 612–634. Available online: https://doi.org/10.1109/JPROC.2021.3058954 (accessed on). https://doi.org/10.1109/JPROC.2021.3058954.

Song, X., Li, Z., Chen, G., Zheng, Y., Fan, Y., Dong, X., & Zhang, K. (2024). Causal Temporal Representation Learning with Nonstationary Sparse Transition. In *Advances in neural information processing systems* (Vol. 37, pp. 77098–77131). Available online: https://proceedings.neurips.cc/paper_files/paper/2024/hash/8cef4e4bcb85f7d4a1005a9db018d6b6-Abstract-Conference.html (accessed on).

 doi:10.20944/preprints202606.0188.v1

Song, X., Yao, W., Fan, Y., Dong, X., Chen, G., Niebles, J. C., Xing, E., & Zhang, K. (2023). Temporally Disentangled Representation Learning under Unknown Nonstationarity. In *Advances in neural information processing systems* (Vol. 36). Available online: https://openreview.net/forum?id=V8GHCGYLkf (accessed on).

Sorrenson, P., Rother, C., & Köthe, U. (2020). Disentanglement by Nonlinear ICA with General Incompressible-flow Networks (GIN). In *International conference on learning representations*. Available online: https://openreview.net/forum?id=rygeHgSFDH (accessed on).

von Kügelgen, J., Besserve, M., Liang, W., Gresele, L., Kekić, A., Bareinboim, E., Blei, D. M., & Schölkopf, B. (2023). Nonparametric Identifiability of Causal Representations from Unknown Interventions. In *Advances in neural information processing systems* (Vol. 36). Available online: https://proceedings.neurips.cc/paper_files/paper/2023/hash/97fe251c25b6f99a2a23b330a75b11d4-Abstract-Conference.html (accessed on).

von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M., & Locatello, F. (2021). Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style. In *Advances in neural information processing systems* (Vol. 34). Available online: https://proceedings.neurips.cc/paper/2021/hash/8929c70f8d710e412d38da624b21c3c8-Abstract.html (accessed on).

Yao, W., Chen, G., & Zhang, K. (2022). Temporally Disentangled Representation Learning. In *Advances in neural information processing systems* (Vol. 35, pp. 26492–26503). Available online: https://proceedings.neurips.cc/paper_files/paper/2022/hash/a938292feb86b94ebe3e6200ff7786ef-Abstract-Conference.html (accessed on).

Zheng, Y., Ng, I., & Zhang, K. (2022). On the Identifiability of Nonlinear ICA: Sparsity and Beyond. In *Advances in neural information processing systems* (Vol. 35). Available online: https://openreview.net/forum?id=Wo1HF2wWNZb (accessed on).

Zheng, Y., & Zhang, K. (2023). Generalizing Nonlinear ICA Beyond Structural Sparsity. In *Advances in neural information processing systems* (Vol. 36). Available online: https://openreview.net/forum?id=gI1SOgW3kw (accessed on).

Zimmermann, R. S., Sharma, Y., Schneider, S., Bethge, M., & Brendel, W. (2021). Contrastive Learning Inverts the Data Generating Process. In *Proceedings of the 38th international conference on machine learning* (Vol. 139, pp. 12979–12990). PMLR. Available online: https://proceedings.mlr.press/v139/zimmermann21a.html (accessed on).