Factored Generative Models through Mechanism Diversity

Anonymous CVPR submission

Paper ID 35

Abstract

A generative model is factored *when each latent dimension independently controls one factor of variation: changing a single latent predictably changes one semantic attribute while leaving the rest unchanged, enabling controllable generation, compositional generalization, and reproducible representations. Existing approaches either constrain the latent distribution,* e.g.*, requiring it to shift with an auxiliary variable, or regularize the model,* e.g.*, sparsity, quantization, or Hessian penalties, neither of which matches how modern conditional generative models actually work: a conditioning signal* $\mathbf{u}$ *reshapes the generator* $g(\cdot, \mathbf{u})$ *while the latent prior stays fixed. We prove an identifiability theorem showing that* generating mechanism diversity*, the natural variation that arises when* $\mathbf{u}$ *sufficiently reshapes* g*, is sufficient for the model to be provably factored, with no parametric assumption on the latent distribution. To actively enforce this condition, we propose* Mechanistic Contrastive Learning *(MCL), a model-agnostic contrastive objective over generator Jacobians. Empirically, MCL achieves state-of-the-art latent concept disentanglement on three benchmarks equipped with a latent diffusion model, and improves prediction quality and zero-shot cross-task transfer in latent action world models with a 1.4B video generative model as the backbone.*

1. Introduction

What makes a generative model *factored*? Intuitively, a factored model has a latent space where each coordinate controls one factor of variation: one latent for color, another for shape, a third for viewpoint, and so on. Such representations underpin controllable generation [50], compositional generalization [37], and reproducible science [49]. Yet whether the latent spaces of modern conditional generative models are provably factored remains an open question.

Modern conditional generative models share a common architecture: a conditioning signal $\mathbf{u}$ (text prompts, labels, actions) modulates a generator $g(\cdot, \mathbf{u})$ mapping latents $\mathbf{z}$ to observations $\mathbf{x}$ [38, 50, 53]. Prior identifiability theory [27] only covers the opposite regime, requiring $\mathbf{u}$ to shift the latent distribution $p(\mathbf{z} \mid \mathbf{u})$ while the generator stays fixed, which is a regime we call *Latent Representation Diversity*, LRD. However, modern generative models keep $p(\mathbf{z})$ fixed (typically Gaussian) and let $\mathbf{u}$ reshape the generator instead, *e.g.*, text-to-image diffusion models [67] inject text into the denoising network, and world models [2] let actions modulate the visual dynamics. In this paper, we refer to this regime as *Generating Mechanism Diversity* (GMD), and our contributions are mainly threefold: ***(i) Theory.*** We prove that sufficient GMD in $g(\cdot, \mathbf{u})$ guarantees a factored latent space (Theorem 1), without parametric assumptions on $p(\mathbf{z})$; this strictly generalizes the classical LRD framework (Proposition 1) and extends to self-supervised and multi-view settings [55, 62]. ***(ii) Method.*** We introduce MCL, which contrasts generator mechanism gradients with conditioning signals and uses shuffled conditions as negatives, making mechanism diversity explicit. ***(iii) Empirics.*** We validate the theory on synthetic data with known factors and instantiate it in image concept disentanglement through MD-VAE and MDDiff, respectively, where MCL improves FactorVAE and DCI score across `Cars3D`, `Shapes3D`, and `MPI3D` while preserving generation quality; we also formulate the temporal world-model setting, where actions play the role of the conditioning mechanism. MCL generally improves the world model quality, action-following, and task-transferability. Notably, MCL is model-agnostic: the same lightweight head plugs into a VAE, diffusion model, or a 1.4B-parameter pretrained video generative backbone, so it does not trade off the scalability of large-scale generative models. Instead, scalable generative models require not only more pooled data, but also factored structure that supports controllable generation, causal consistency, and related capabilities.

2. Background and Related Work

From Nonlinear ICA to Factored Generative Model. *Nonlinear ICA* asks when a generative model $\mathbf{x} = g(\mathbf{z})$ with independent latent factors $\mathbf{z}$ can be inverted to recover those factors from observations alone, and *identifiability theory* formalizes when such inversion is unique up to trivial ambiguities. The two are central to factored generative modeling because a model that fits $p(\mathbf{x})$ but is not identifiable

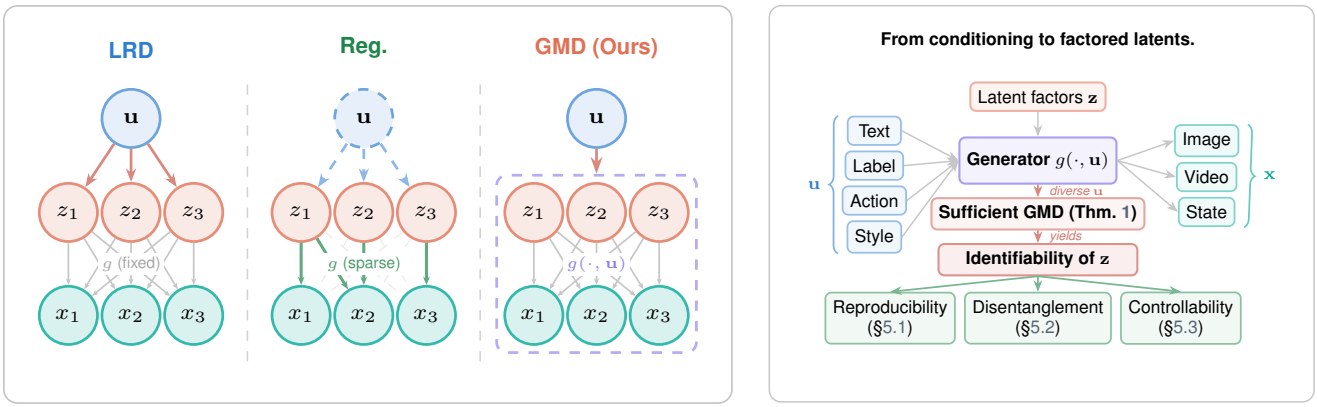


Figure 1. **Factored generative models through mechanism diversity. Left:** LRD varies $p(\mathbf{z}|\mathbf{u})$, Reg. constrains g, and GMD (ours) varies $g(\cdot, \mathbf{u})$. **Right:** sufficient GMD links conditioning signals, identifiability guarantees, and the practical benefits.

can place semantically meaningful structure on arbitrary, non-reproducible directions of $\mathbf{z}$ [26]. This is the gap that generative models face: fitting $p(\mathbf{x})$ alone says nothing about whether the latent code corresponds to disentangled factors of variation. Resolving it has driven a decade of work, and we organize the literature below and in Figure 2.

Three principles for factored representations. Existing answers to nonlinear ICA and generative-model identifiability largely follow these two routes: **LRD**, which varies the latent distribution sufficiently, and **regularization**, which constrains the model's degree of freedom. ***Beyond these***, we identify **GMD** as a third principle, varying the generator itself through conditioning.

(1) Latent distribution diversity (LRD). The auxiliary $\mathbf{u}$ creates diversity in the latent distribution $p(\mathbf{z} \mid \mathbf{u})$ while the generator g stays fixed. Hyvarinen and Morioka [25] established this for temporal nonstationarity, Hyvarinen et al. [27] extended it to arbitrary auxiliaries within an exponential family, and Khemakhem et al. [29] unified the framework with VAEs by showing that the ELBO with an auxiliary-conditioned prior recovers identifiable representations. Subsequent work removed the exponential-family assumption via higher-order log-density derivatives [32, 63] and extended LRD identifiability to multi-view [55], partially observed [62], and additive noisy [17] settings. Apart from theorems, several practical methods exploit LRD implicitly: contrastive predictive coding leverages temporal structure [40], and Zimmermann et al. [70] showed that it can invert the data-generating process to yield identifiable features. Reizinger et al. [46] proved that cross-entropy classification yields identifiable representations, and Reizinger et al. [47] demonstrated that policy diversity produces identifiable state representations in reinforcement learning. The shared technical formal condition underlying these results is summarized below, and our GMD condition (Theorem 1) replaces the requirement on the latent log-density q_i with a condition on the generator Jacobian.

Assumption 1 (Sufficient LRD [27])

For any $\mathbf{z} \in \mathbb{R}^{d_z}$, *there exist* $2d_z + 1$ *values* $\mathbf{u}^{(0)}, \ldots, \mathbf{u}^{(2d_z)}$ *such that the* $2d_z$ *vectors* $\mathbf{w}(\mathbf{z}, \mathbf{u}^{(j)}) - \mathbf{w}(\mathbf{z}, \mathbf{u}^{(0)})$, $j = 1, \ldots, 2d_z$, *are linearly independent, where* $\mathbf{w}(\mathbf{z}, \mathbf{u}) = \left(\frac{\partial q_i(z_i, \mathbf{u})}{\partial z_i}, \frac{\partial^2 q_i(z_i, \mathbf{u})}{\partial z_i^2}\right)_{i=1}^{d_z}$ *and* q_i *is the log-conditional density of* z_i.

(2) Regularization. Instead of requiring diversity in the data, this family constrains the model architecture or objective to favor factored solutions. Independence-based methods penalize statistical dependence among latents: β-VAE [21] upweights the KL term, while FactorVAE [30] and β-TCVAE [5] directly minimize total correlation. Sparsity-based methods constrain the generator: Lachapelle et al. [33] penalize off-diagonal Jacobian entries (mechanism sparsity), Zheng et al. [68] prove identifiability under sparsity on g without auxiliary variables, and Zhang et al. [66] assume a causal graph among latents across distributions. Compression-based methods restrict the latent bottleneck: Tripod [22] combines finite scalar quantization, kernel-based multiinformation minimization, and a normalized Hessian penalty, showing that three complementary regularizers together achieve strong disentanglement in autoencoders. Locatello et al. [37] showed that without such inductive biases, unsupervised disentanglement is impossible.

(3) GMD (ours). Rather than diversifying the latent distribution or constraining the model, GMD induces diversity in the generator itself: a conditioning signal $\mathbf{u}$ (text prompt, class label, control action) reshapes $g(\cdot, \mathbf{u})$ while $p(\mathbf{z})$ stays fixed, and sufficient variation across $\mathbf{u}$ suffices for a factored latent space. Modern conditional generators already operate in this regime ($\mathbf{x} = g(\mathbf{z}, \mathbf{u})$ with $\mathbf{z} \perp \mathbf{u}$): conditional GANs [38, 65], VAEs [20, 53], and diffusion models [9, 41, 50] condition on text or class labels, while

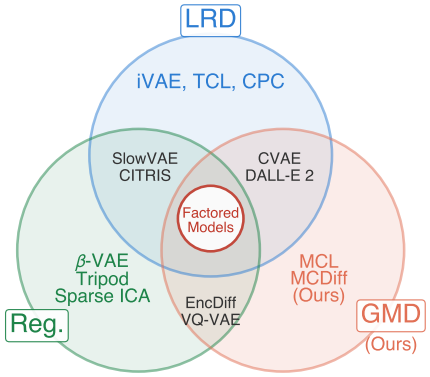


Figure 2. **Three principles for factored models.** **LRD**: diversify the latent distribution. **Regularization**: constrain the model. **GMD** (ours): diversify the generator. GMD $\supset$ LRD (Prop. 1).

action-conditioned video world models [2, 4, 16] condition on actions. Existing identifiability theory targets only LRD or regularization, leaving ***conditional generative models as the missing principle of factored representations***. We close this gap with a sufficient GMD condition (Theorem 1) that strictly subsumes LRD in representational capacity (Proposition 1), and extend it to two practical regimes: multi-view generation (Proposition 3), where three or more conditioning values play the role of distinct views, and self-supervised temporal data (Proposition 2), where a historical latent serves as $\mathbf{u}$. We actively encourage this condition via MCL (Section 4), rather than passively learning it as previous conditional generative models.

3. When Is a Generative Model Factored?

In this section, we formalize the conditions under which a conditional generative model is provably factored. The key insight is that, when $g(\cdot, \mathbf{u})$ varies sufficiently across conditioning signals, the model is factored, hence identifiable. We develop this in three steps. We first clarify our motivation in Section 3.1: GMD strictly contains LRD in representational capacity, so the GMD regime is genuinely the larger family worth analyzing. Building on this, Section 3.2 states the sufficient GMD condition and the resulting identifiability theorem. Section 3.3 then relaxes the two main restrictions of the theorem, observed conditioning and pointwise invertibility, to cover the self-supervised and multi-view settings that contemporary generators operate in.

3.1. GMD and LRD Are Not Equivalent in Generating Data

Consider two data-generation families over $\mathbf{u} \in \mathcal{U}$, $\mathbf{z} \in \mathcal{Z}$, $\mathbf{x} \in \mathcal{X}$, with source noise $\boldsymbol{\epsilon} \sim P_{\boldsymbol{\mathcal{E}}}$:

$$\underbrace{\mathbf{z} = f_L(\mathbf{u}, \boldsymbol{\epsilon}), \quad \mathbf{x} = g_L(\mathbf{z})}_{\text{LRD: } \mathbf{u} \text{ shifts } p(\mathbf{z}|\mathbf{u}),\ g \text{ fixed}}, \tag{175}$$

$$\underbrace{\mathbf{z} = f_G(\boldsymbol{\epsilon}), \quad \mathbf{x} = g_G(\mathbf{u}, \mathbf{z})}_{\text{GMD: } \mathbf{u} \text{ reshapes } g,\ p(\mathbf{z}) \text{ fixed}}. \tag{176}$$

inducing conditional families $\mathcal{P}_{\text{LRD}} := \{\mathcal{L}(g_L(f_L(\mathbf{u}, \boldsymbol{\epsilon})) \mid \mathbf{u}) : \boldsymbol{\epsilon} \sim P_{\boldsymbol{\mathcal{E}}}\}$ and $\mathcal{P}_{\text{GMD}} := \{\mathcal{L}(g_G(\mathbf{u}, f_G(\boldsymbol{\epsilon})) \mid \mathbf{u}) : \boldsymbol{\epsilon} \sim P_{\boldsymbol{\mathcal{E}}}\}$. We show that these two forms are not equivalent below.

Proposition 1 (GMD has strictly greater representation capacity than LRD)**.** *Let $\Theta_{\text{LRD}} \subset \mathbb{R}^{d_L}$, $\Theta_{\text{GMD}} \subset \mathbb{R}^{d_G}$ be open parameter spaces with all mappings analytic in their parameters, $P_{\boldsymbol{\mathcal{E}}}$ Lebesgue, and suppose $g_G(\mathbf{u}, \cdot)$ contains a conditioning direction not realisable by any LRD shift. Then $\mathcal{P}_{\text{LRD}} \subsetneq \mathcal{P}_{\text{GMD}}$, and the LRD-realisable subset $\Theta_{\text{LRD-in-GMD}} := \{\theta_G : \exists\, \theta_L,\ g_G^{\theta_G}(\mathbf{u}, f_G^{\theta_G}(\boldsymbol{\epsilon})) = g_L^{\theta_L}(f_L^{\theta_L}(\mathbf{u}, \boldsymbol{\epsilon}))\ \forall (\mathbf{u}, \boldsymbol{\epsilon})\}$ has Lebesgue measure zero in Θ_{GMD}.*

Equivalently, ***GMD strictly contains LRD as a generative blueprint in data coverage***, and almost all GMD parameter configurations cannot be reproduced by any fixed-decoder LRD model. The GMD region is the standard regime of modern conditional generators: a fixed distribution prior $p(\mathbf{z})$ paired with a $\mathbf{u}$-conditioned generator $g(\cdot, \mathbf{u})$.

3.2. Mechanism Diversity Guarantees Factored Latent Spaces

We first formalize what it means for a generative model to have a factored latent space.

Definition 1 (Factored Generative Model)**.** *A model $\mathbf{x} = g(\mathbf{z}, \mathbf{u})$ is* factored *if for every observationally equivalent model $\mathbf{x} = \hat{g}(\hat{\mathbf{z}}, \mathbf{u})$, there exist a permutation π and invertible functions $\{h_i\}_{i=1}^{d_z}$ such that $\hat{z}_i = h_i(z_{\pi(i)})$.*

Definition 1 is precisely the standard *identifiability* notion of nonlinear ICA [27, 29], recast at the level of the full generative model: a factored model is an identifiable model whose latents are recovered up to permutation and componentwise transformations. Throughout the paper we use *factored* and *identifiable up to permutation–componentwise* interchangeably. Practically, this property produces disentangled latents, enables controllable generation, generalization, and yields reproducible representations across different training runs [49].

To formalize the GMD condition, let $m_i'(\mathbf{z}, \mathbf{u}) := \partial_{z_i} \log |J_g(\mathbf{z}, \mathbf{u})|$ and $m_{ii}''(\mathbf{z}, \mathbf{u}) := \partial_{z_i}^2 \log |J_g(\mathbf{z}, \mathbf{u})|$ denote the first and diagonal-second derivatives of the log-Jacobian, with analogous $\hat{m}_i'$, $\hat{m}_{ii}''$ for the learned model.

Stacking these derivatives gives the *mechanism-derivative vector*

$$V(\mathbf{z},\mathbf{u}) := \left(m'_1,\dots,m'_{d_z},\ m''_{11},\dots,m''_{d_z d_z}\right) \in \mathbb{R}^{2d_z},$$

whose contrasts across distinct $\mathbf{u}$ drive the GMD condition and identifiability theorem below.

Assumption 2 (Sufficient GMD)

For any $\mathbf{z} \in \mathbb{R}^{d_z}$, there exist $2d_z + 1$ values $\mathbf{u}^{(0)},\dots,\mathbf{u}^{(2d_z)}$ such that the $2d_z$ vectors $V(\mathbf{z},\mathbf{u}^{(n)}) - V(\mathbf{z},\mathbf{u}^{(0)})$, $n = 1,\dots,2d_z$, are linearly independent.

Theorem 1 (Identifiability under GMD)

Let $\mathbf{x} = g(\mathbf{z},\mathbf{u})$ with $\mathbf{z} \perp \mathbf{u}$, and let $(\hat{g},\hat{f})$ achieve observational equivalence. Under Assumption 2 and the technical condition $\mathrm{span}\{(m'_i, m''_i)\} \cap \mathrm{span}\{(\hat{m}'_i, \hat{m}''_i)\} = \{0\}$, there exist a permutation π and invertible functions $\{h_i\}_{i=1}^{d_z}$ such that $\hat{z}_i = h_i(z_{\pi(i)})$, i.e. the model is factored in the sense of Definition 1.

Assumption 2 in practice. Assumption 2 is a statement about the auxiliary variable $\mathbf{u}$: it asks that the conditioning signal (a text prompt, action, viewpoint, or regime index) reshape the generator's mechanism in $2d_z + 1$ linearly independent ways. In practice, the auxiliary alphabet must be diverse enough that distinct conditioning values produce genuinely distinct generative mechanisms, a property that modern conditional diffusion and temporal dynamics models are already trained to satisfy.

Proof sketch. Because $\mathbf{z} \perp \mathbf{u}$, the latent log-density carries no $\mathbf{u}$-dependence, so all $\mathbf{u}$-dependent derivative terms in the change-of-variables identity must come from the generator. Change-of-variables relates the learned and true models through $h_{\mathbf{u}} = \hat{g}_{\mathbf{u}}^{-1} \circ g_{\mathbf{u}}$; taking mixed second derivatives and subtracting across conditioning values cancels every $\mathbf{u}$-independent term, leaving a homogeneous system in the off-diagonal entries of $J_{h_{\mathbf{u}}}$. The linearly independent contrasts of Assumption 2 then force each row of $J_{h_{\mathbf{u}}}$ to have at most one nonzero entry, and invertibility upgrades this to a permutation, yielding the componentwise structure.

3.3. Extensions to a Broader Class of Models

Theorem 1 as stated requires (i) an explicit conditioning signal $\mathbf{u}$ and (ii) pointwise invertibility of the generator $g(\cdot,\mathbf{u})$. Both assumptions can fail in practice: in self-supervised settings $\mathbf{u}$ may not be directly observed, and modern image- or video-generators are typically not pointwise invertible (each output discards information). We therefore extend the main theorem along these two axes: Proposition 2 replaces the observed $\mathbf{u}$ with a learned latent context, and Proposition 3 relaxes pointwise invertibility to distribution-level invertibility once enough conditioning values are available. Together, these two informal extensions cover the regimes that contemporary GMD generators actually operate in: self-supervised temporal data and many-view generation. Formal statements are collected in Appendix B, and the proofs are given in Appendix C.

Proposition 2 (Self-supervised GMD, temporal case; *informal*)**.** *When $\mathbf{u}$ is not directly observed, it can be replaced by $\hat{\mathbf{z}}_t$ (the* estimated context representation *at different time steps). If the resulting model satisfies the conditions of Theorem 1, identifiability holds with the same guarantees.*

Proposition 3 (GMD as a multi-view generative model; *informal*)**.** *If the number of distinct conditioning values satisfies $|\{\mathbf{u}^{(n)}\}| \geq 3$, Theorem 1 holds under* distribution-level invertibility *between different $\mathbf{u}$, without requiring pointwise invertibility.*

Proposition 2 extends our framework to temporal cases (Figure 3a), which is justified by findings in the identifiable temporal representation learning [25]. Proposition 3 admits a clean multi-view reading (Figure 3b): each $\mathbf{u}^{(n)}$ yields a different "view" $\mathbf{x}^{(n)} = g(\mathbf{z},\mathbf{u}^{(n)})$ of the shared latent $\mathbf{z}$, and three or more such views play the role of repeated measurements in the nonparametric identification literature [12, 24]. Notably, the threshold of $|\{\mathbf{u}^{(n)}\}| = 3$ smoothly matches classical 3D generation.

4. How to Actively Build Factored Generative Models

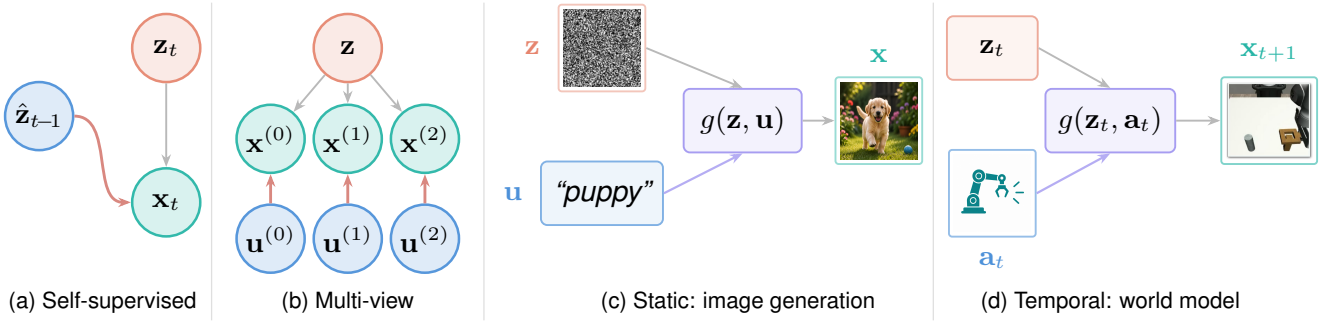
When the conditioning signal lacks sufficient variation, a generator in the GMD pattern may fail to diversify its mechanism across conditions, so Assumption 2 is not satisfied even though the architecture is in the GMD regime. We first describe two complementary instantiations of GMD, then present MCL, a training objective that contrasts mechanism-critic gradients across conditioning signals to enforce Assumption 2 explicitly.

4.1. GMD Instantiations

Static case: conditional generative models. For controlled synthetic validation (Section 5.1) we construct a *Mechanism-Diverse VAE* (**MD-VAE**): a conditional VAE with encoder $q_\phi(\mathbf{z} \mid \mathbf{x},\mathbf{u})$ and decoder $p_\theta(\mathbf{x} \mid \mathbf{z},\mathbf{u})$, fixed prior $p(\mathbf{z}) = \mathcal{N}(0,\mathbf{I})$, trained by maximizing the conditional ELBO

$$\log p_\theta(\mathbf{x} \mid \mathbf{u}) \geq \mathbb{E}_{q_\phi(\mathbf{z}\mid\mathbf{x},\mathbf{u})}[\log p_\theta(\mathbf{x} \mid \mathbf{z},\mathbf{u})] - \mathrm{KL}(q_\phi(\mathbf{z} \mid \mathbf{x},\mathbf{u}) \,\|\, p(\mathbf{z})), \quad (1)$$

denoted $\mathcal{L}_{\text{MD-VAE}}$. Maximizing it reshapes the decoder $g(\cdot,\mathbf{u})$ with $\mathbf{u}$ at fixed prior, instantiating GMD by construction. In the spirit of latent diffusion [50], where a standard



(a) Self-supervised (b) Multi-view (c) Static: image generation (d) Temporal: world model

Figure 3. **Extensions and instantiations of GMD.** (a) Self-supervised GMD: a past learned latent $\hat{\mathbf{z}}_{t-1}$ acts as $\mathbf{u}$. (b) Multi-view GMD: three or more conditioning values $\mathbf{u}^{(n)}$ play the role of repeated measurements of a shared latent $\mathbf{z}$. (c) Static GMD: a text prompt $\mathbf{u}$ reshapes the denoiser. (d) Temporal GMD: an action $\mathbf{a}_t$ reshapes the dynamics *at the pixel level*, in contrast to the latent-state-level dynamics of traditional action-conditioned world models.

VAE encoder feeds a diffusion model in latent space, our **MDDiff** (Section 5.2) plugs MCL into the same role and runs $\mathbf{u}$-conditioned diffusion on the resulting mechanism-diverse latents.

Temporal case: state-space models. The conditioning signal is naturally provided by a history:

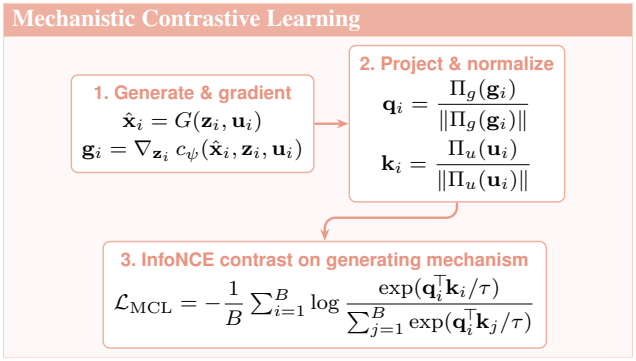
$$\mathbf{x}_t = g(\mathbf{x}_{t-H:t-1},\, \mathbf{u}_{t-H:t-1})\,,$$

where $\mathbf{u}_{t-H:t-1}$ can be actions in action-conditioned world models [16, 18], learned latent states in the self-supervised learning, *e.g.*, time-series data, or any domain index. Sufficient diversity across the temporal dynamics then factors the latent state.

4.2. Mechanistic Contrastive Learning

Beyond a controlled instantiation tied to the variational ELBO or traditional generative models, to make GMD usable across realistic conditional generators in an active way, we now derive a model-agnostic training objective whose minimizer drives the generator Jacobian $\partial_{\mathbf{z}} G(\mathbf{z}, \mathbf{u})$ to vary with $\mathbf{u}$, which is precisely the linear-independence content of Assumption 2.

MCL augments any differentiable conditional generator G with three lightweight modules trained jointly with it: a *mechanism critic* $c_\psi : \mathcal{X} \times \mathcal{Z} \times \mathcal{U} \to \mathbb{R}$ that scores how strongly the realized mechanism $G(\cdot, \mathbf{u})$ depends on $\mathbf{z}$, and two L_2-normalized projection heads Π_g, Π_u mapping mechanism-gradient features and conditioning features into a shared embedding space. For each in-batch sample $(\mathbf{z}_i, \mathbf{u}_i)$ we generate $\hat{\mathbf{x}}_i = G(\mathbf{z}_i, \mathbf{u}_i)$, score it with the critic, and read off the *mechanism gradient* $\mathbf{g}_i = \nabla_{\mathbf{z}_i} c_\psi(\hat{\mathbf{x}}_i, \mathbf{z}_i, \mathbf{u}_i)$, i.e. the gradient of the critic through the realized generator. The gradient and conditioning embeddings are then contrasted with InfoNCE. The three stages are shown below.



The full training loss is $\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_{\text{MCL}}\, \mathcal{L}_{\text{MCL}}$, where $\mathcal{L}_{\text{rec}}$ is the host generator's native reconstruction objective and $\lambda_{\text{MCL}} > 0$ trades off reconstruction quality against mechanism contrast. Intuitively, the InfoNCE optimum aligns each mechanism gradient $\mathbf{g}_i$ more strongly with the embedding of its own $\mathbf{u}_i$ than with any other $\mathbf{u}_j$ in the batch, which is precisely the linear-independence requirement of Assumption 2.

Intuitively, the InfoNCE optimum aligns each mechanism gradient $\mathbf{g}_i$ more strongly with the embedding of its own $\mathbf{u}_i$ than with any other $\mathbf{u}_j$ in the batch. Under standard alignment-and-spread conditions on the projection heads [58] and a $2d_z + 1$ discriminative conditioning support, minimizers of $\mathcal{L}_{\text{MCL}}$ provably satisfy the linear-independence content of Assumption 2, hence drive the model into the GMD regime of Theorem 1; the formal statement appears in Theorem 2 (Appendix B.2).

Comparison with contrastive learning and L_1/L_2 regularization. MCL generalizes two existing methodologies (Figure 4). **Contrastive learning** on latent codes is the practical instantiation of LRD [27, 40, 46, 70]; **L_1/L_2 regularization** (sparsity [33, 68], quantization [44, 54], combined penalties [22]) tightens a single latent code without injecting diversity. **MCL** takes a third stance: $p(\mathbf{z})$ stays fixed and $\mathbf{u}$

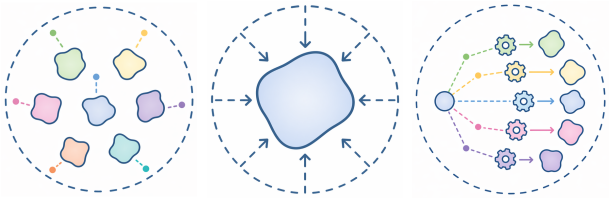


Figure 4. **What each family of objectives does in latent space.** (Left) **Contrastive learning**, LRD; (Center) L_1/L_2 **regularization**; (Right) **MCL**.

reshapes the *generator*, with the contrastive signal computed on Jacobian gradients — the natural identifiability objective for architectures built with $\mathbf{z} \perp \mathbf{u}$.

5. Experiments

We validate the theory on synthetic data with known ground truth (§5.1), then evaluate concept disentanglement (static case, §5.2) and world modeling (temporal case, §5.3).

Implementation. MCL is implemented as a drop-in regularizer on top of any conditional generator. The mechanism critic c_ψ is a 3-layer MLP ($d_z + d_{\hat{\mathbf{x}}} + d_u \to 256 \to 256 \to 1$, GELU activations); the projection heads Π_g, Π_u are 2-layer MLPs into a $d = 128$-dimensional unit sphere. We use temperature $\tau = 0.07$ following SimCLR [6], in-batch negatives only ($B \in \{128, 256\}$ depending on the host model), and a single Adam optimizer for $(G_\theta, c_\psi, \Pi_g, \Pi_u)$ at learning rate 10^{-4}. The MCL weight is $\lambda_{\text{MCL}} \in [0.01, 0.1]$ tuned per host model on a held-out validation split. The critic gradient is detached on the InfoNCE path to avoid the trivial solution where c_ψ collapses to a constant; mechanism-gradient norms are clipped at 10. All experiments use mixed-precision training on a single H100 GPU; world-model experiments scale to four H100s with data-parallel.

5.1. Experiments on Synthetic Data for Factor Verification

Setup and metric. We generate data from a known GMD process (full data generation pipeline in Appendix D.3) with $d_z \in \{5, 10, 20\}$ and a default of $|\mathcal{U}| = 8$ conditioning values. We compare three baselines drawn from the three principles of Section 2: (i) **β-VAE** [21] (regularization-based), (ii) **iVAE** [29] (LRD-based), and (iii) **MD-VAE** (ours, GMD-based), trained with the conditional ELBO of Equation (1). Identifiability is measured by the Mean Correlation Coefficient (MCC) between true and recovered latent variables.

Results. Across $d_z \in \{5, 10, 20\}$ MD-VAE attains median MCC $\{0.96, 0.87, 0.81\}$, exceeding iVAE, *i.e.*, 0.80 to 0.65, and β-VAE, *i.e.*, 0.53 to 0.39, at every d_z, as shown in Figure 5. The ordering matches Theorem 1, the MD-VAE exploits mechanism diversity directly, whereas iVAE relies on Assumption 1 which weakens as d_z grows against a fixed $|\mathcal{U}|$, and β-VAE has no identifiability guarantee. Notably, increasing $|\mathcal{U}|$ can induce more distinct mechanisms (see Appendix F.1).

5.2. Experiments on Concept Disentanglement

As a representative static GMD task, we evaluate whether MCL improves concept disentanglement in conditional diffusion models (concept tokens $= \mathbf{u}$, denoiser $= g(\cdot, \mathbf{u})$). Our **MDDiff** builds MCL and mechanism diversity on top of a latent diffusion generative model with cross-attention, following Yang et al. [61]; full architecture and training details are in Appendix D.

Datasets. We use three standard disentanglement benchmarks: `Shapes3D` [30], `MPI3D` [13], and `Cars3D` [45], which together cover synthetic factors, 3D object factors, and viewpoint variation. All experiments use 64×64 images, following [5, 30, 48, 60].

Baselines and metrics. We compare VAE baselines (FactorVAE [30], β-TCVAE [5]), GAN baselines (InfoGAN-CR [35], GANSpace [19], LatentDisco [56], DisCo [48]), and diffusion baselines (DisDiff [60], EncDiff [61]) against MDDiff; full descriptions are in Appendix D.2. All methods use $N = 20$ scalar representations [60]; we report mean $\pm$ std over 15 runs using the FactorVAE score and DCI disentanglement (as in Appendix D.1).

Results and analysis. Quantitatively (Table 1), even before MCL, the diffusion-based methods already outperform the baselines by a clear margin, consistent with our theory. MDDiff is best or tied-best in nearly all columns; the only exception is DCI on `MPI3D`. The largest gains appear on `Cars3D` (the hardest benchmark) DCI, *i.e.*, 0.493 vs. 0.279 for EncDiff, ours obtain $+0.214$. The training dynamics on `Cars3D` (Figure 5) trace this gain to the two-stage MDDiff pipeline: MDDiff pretraining first learns the conditional generator, then MDDiff + MCL fine-tuning drives both DCI and FV upward relative to EncDiff. `MPI3D`'s complex 3D interactions break LRD-based methods but not MDDiff, supporting our claim that mechanism diversity is a stronger inductive bias. Qualitatively (Figure 6), swapping a single latent dimension between source and target images transfers exactly one attribute while preserving all others , the operational signature of an axis-aligned factored representation, in the sense of Theorem 1. Generation quality (FID, LPIPS) is preserved competitively with EncDiff (Appendix F.2, Table 6); latent traversals and cross-attention maps further confirming localised mechanism diversity in the generator are reported in Appendix F.

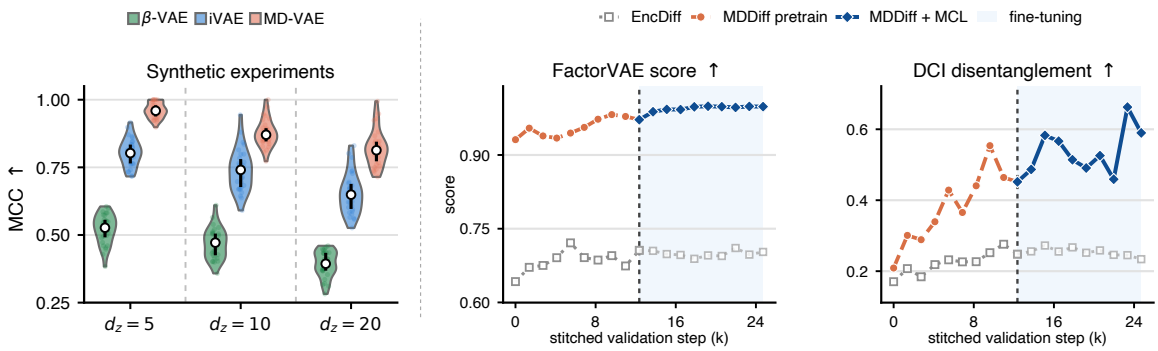


Figure 5. **Synthetic experiments and `Cars3D` metrics curves during training.** Left: MCC violin plots over 30 seeds per cell. Right: stitched FactorVAE and DCI trajectories on `Cars3D`.

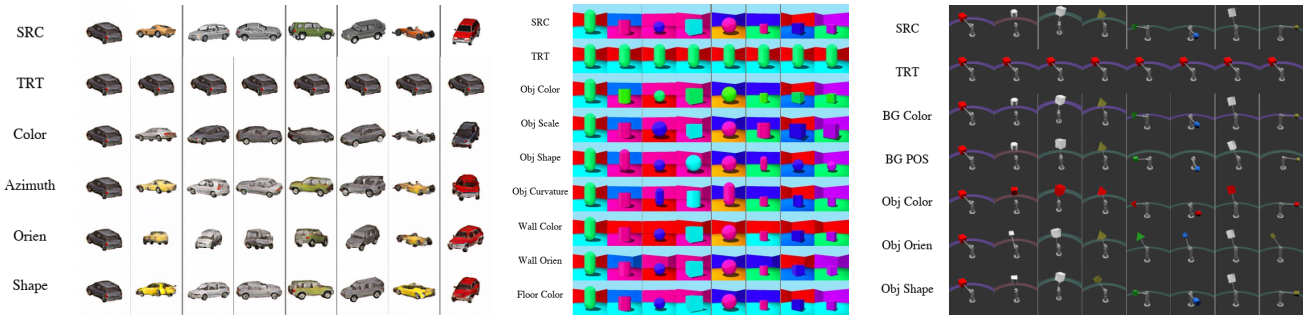


Figure 6. **Factor swapping across all three benchmarks.** Row 1: source images (SRC); row 2: target images (TRT); subsequent rows: SRC with one latent dimension replaced by the corresponding TRT dimension, we match the concept through their consistent visual variations across samples.

5.3. Experiments on World Modeling

We show that MCL can improve world modeling of temporal robotics data. We first relate our theorem to practical scenarios in principle, and then present our methodology and results.

Forward/inverse dynamic models. Recent VA/WA models [34, 64] cast the world as an SEM on $(\mathbf{o}_{1:T}, \mathbf{a}_{1:T})$ with two halves: a forward action-conditioned generative model $\mathbf{o}_{t+1} = g_{\text{fwd}}(\mathbf{z}_t, \mathbf{a}_t)$, and an inverse-dynamics model $\mathbf{a}_t = g_{\text{inv}}(\mathbf{o}_t, \mathbf{o}_{t+1})$. State-of-the-art pipelines train the inverse half by extracting actions from pixel dynamics; we target the *forward* half, where actions reshape the latent dynamics as the conditioning signal $\mathbf{u}_t = \mathbf{a}_t$. The two halves form an inverse pair of the same SEM,

$$\underbrace{\mathbf{o}_{t+1} = g_{\text{fwd}}(\mathbf{z}_t, \mathbf{a}_t)}_{\text{our action-conditioned generation (forward)}}$$

$$\iff \underbrace{\mathbf{a}_t = g_{\text{inv}}(\mathbf{o}_t, \mathbf{o}_{t+1})}_{\text{VA/WA action extraction (inverse)}},$$

and the GMD identifiability theory we develop here directly applies to the forward half: when the action sequence sufficiently reshapes g_{fwd}, the latent state is provably factored (Theorem 1). This is the mirror image of how the conditioning signal diversifies temporal dynamics, with both action alignment and visual fidelity reflecting the controllability or identifiability of $\mathbf{u}$.

Architecture. Our action-conditioned generative world model (ACG) instantiates the forward SEM on top of **Wan 1.4B** [57], a pretrained video generation backbone. Two regimes differ in whether the action conditioning is observed or inferred: **Wan-ACG** feeds the recorded action $\mathbf{a}_t$ directly to the forward model; **Wan-LACG** (the L denotes *latent* action) replaces it with an IDM-inferred latent action. Both regimes use an inverse dynamics model, an AdaLN action interface, and a flow-matching forward model initialized from Wan 1.4B.

MCL on the action conditioning. MCL attaches to the action channel: a mechanism critic produces $\mathbf{g}_t = \nabla_{\mathbf{z}_t} c_\psi(\hat{\mathbf{x}}_{t+1}, \mathbf{z}_t, \mathbf{u}_t)$, contrasted via InfoNCE against the projected action embedding. The IDM, action interface, forward model, and MCL heads train jointly with the causal VAE frozen. Full hyperparameters are in Appendix D.

Table 1. **Disentanglement comparison** (mean ± std; best results are in **Bold**).

Type	Method	Cars3D		Shapes3D		MPI3D	
		FV score ↑	DCI ↑	FV score ↑	DCI ↑	FV score ↑	DCI ↑
VAE	FactorVAE	0.906± 0.052	0.161± 0.019	0.840± 0.066	0.611± 0.082	0.152± 0.025	0.240± 0.051
	β-TCVAE	0.855± 0.082	0.140± 0.019	0.873± 0.074	0.613± 0.114	0.179± 0.017	0.237± 0.056
GAN	InfoGAN-CR	0.411± 0.013	0.020± 0.011	0.587± 0.058	0.478± 0.055	0.439± 0.061	0.241± 0.075
	LatentDisco	0.852± 0.039	0.216± 0.072	0.805± 0.064	0.380± 0.062	0.391± 0.039	0.196± 0.038
	GANSpace	0.932± 0.018	0.209± 0.031	0.788± 0.091	0.284± 0.034	0.465± 0.036	0.229± 0.042
	DisCo	0.855± 0.074	0.271± 0.037	0.877± 0.031	0.708± 0.048	0.371± 0.030	0.292± 0.024
Diffusion	DisDiff	0.976± 0.018	0.232± 0.019	0.902± 0.043	0.723± 0.013	0.617± 0.070	0.337± 0.057
	EncDiff	0.773± 0.060	0.279± 0.022	0.999± 0.000	0.969± 0.030	0.872± 0.049	**0.685**± 0.044
	MDDiff (Ours)	**0.985**± 0.048	**0.493**± 0.025	**1.000**± 0.000	**0.993**± 0.030	**0.930**± 0.049	0.679± 0.010

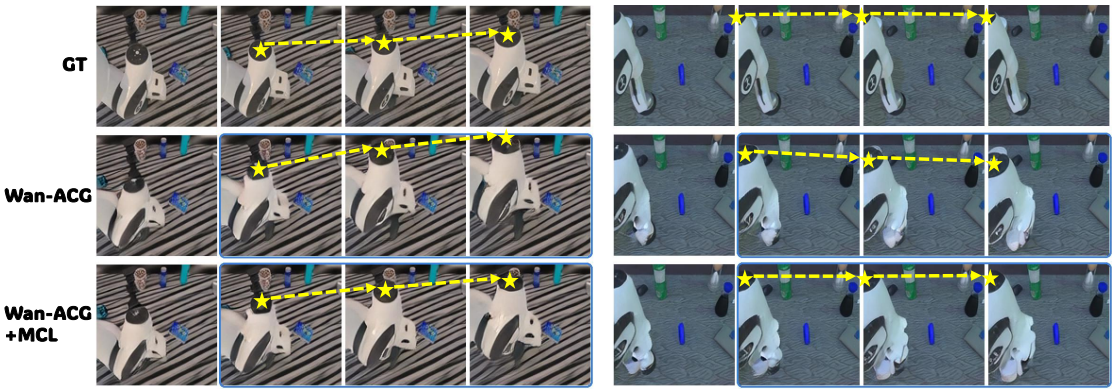


Figure 7. **World-model qualitative ablation.** Two manipulation rollouts driven by the same initial observation and ground-truth action sequence. **Top**: ground-truth frames (**GT**). **Middle**: Wan-ACG alone. **Bottom**: Full method **Wan-ACG+MCL**. The dashed yellow trajectory traces the gripper end-effector across frames; Wan-ACG+MCL follows the ground truth more faithfully than ACG alone, confirming that MCL in the action conditioning sharpens action-following accuracy.

Table 2. **World-model evaluation on Robotwin under the low-data target-embodiment setting. Target task**: held-out evaluation on `place_a2b_left` (same task as training). **Transfer task**: zero-shot evaluation on `place_a2b_right`. Higher is better for SSIM, PSNR, and SSIM-L; lower is better for MSE. **Bold** marks the best result per column.

Method	Target task				Transfer task			
	SSIM ↑	PSNR ↑	MSE ↓	SSIM-L ↑	SSIM ↑	PSNR ↑	MSE ↓	SSIM-L ↑
Wan-ACG	0.763	18.61	0.0138	0.735	0.763	18.45	0.0143	0.732
Wan-ACG+MCL	**0.765**	**18.73**	**0.0134**	**0.737**	**0.765**	**18.58**	**0.0138**	**0.739**
Wan-LACG	0.769	18.98	0.0127	0.744	0.782	19.31	0.0117	0.752
Wan-LACG+MCL	**0.772**	**18.99**	**0.0125**	**0.749**	**0.784**	**19.33**	**0.0115**	**0.758**

Protocol and methods. We evaluate on Robotwin [39] under the low-data target-embodiment setting (Table 2), reporting SSIM, PSNR, MSE, and last-frame SSIM-L on the target task (`place_a2b_left`) and the zero-shot transfer task (`place_a2b_right`). For each of the two regimes (Wan-ACG, Wan-LACG) we report the variant without MCL (only the host objective) and with MCL added; the structural claim is that MCL on the action conditioning improves both visual prediction quality and action-following over the matching no-MCL baseline in either regime.

6. Conclusion

Mechanism diversity provides a principled, parameter-free route to factored generative models: when conditioning sufficiently reshapes the generator, the latent space is provably factored. The framework strictly generalizes LRD, extends to self-supervised and multi-view regimes, and is operationalized by MCL as a model-agnostic objective for any conditional generator.

Limitations and future work. The framework is inspired by nonlinear ICA and targets identifiable, factored generative models. We see three open directions: extending mechanism diversity ***beyond factored generative models*** to broader conditional generators (large-scale text-to-image and video diffusion, visual world models); using ***MCL as a fine-tuning signal*** for large pretrained backbones, actively injecting mechanism shift on top of passive pretraining; and ***multi-objective, multi-auxiliary mechanism contrast***. Each direction is expanded in Appendix H.

References

[1] Anurag Ajay, Yilun Du, Abhi Gupta, Joshua Tenenbaum, Tommi Jaakkola, and Pulkit Agrawal. Is conditional generative modeling all you need for decision-making? In *International Conference on Learning Representations*, 2023. 24, 28

[2] Arthur Allshire, Roberto Martín-Martín, Charles Lin, Shawn Manuel, Silvio Savarese, and Animesh Garg. Laser: Learning a latent action space for efficient reinforcement learning. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6650–6656. IEEE, 2021. 1, 3

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. In *arXiv preprint arXiv:2311.15127*, 2023. 26

[4] Boyuan Chen, Tianyuan Zhang, Haoran Geng, Kiwhan Song, Caiyi Zhang, Peihao Li, William T Freeman, Jitendra Malik, Pieter Abbeel, Russ Tedrake, et al. Large video planner enables generalizable robot control. *arXiv preprint arXiv:2512.15840*, 2025. 3

[5] Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in neural information processing systems*, 2018. 2, 6, 20

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 6, 20

[7] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems*, 2023. 24, 28

[8] John B. Conway. *A Course in Functional Analysis*. Springer, 2 edition, 1994. 17

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2, 26, 27

[10] Nelson Dunford and Jacob T. Schwartz. *Linear Operators, Part III: Spectral Operators*. Wiley-Interscience, 1988. 17

[11] Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *International conference on learning representations*, 2018. 20

[12] Minghao Fu, Biwei Huang, Zijian Li, Yujia Zheng, Ignavier Ng, Guangyi Chen, Yingyao Hu, and Kun Zhang. Learning general causal structures with hidden dynamic process for climate analysis. 2025. 4

[13] Muhammad Waleed Gondal, Manuel Wuthrich, Djordje Miladinovic, Francesco Locatello, Martin Breidt, Valentin Volchkov, Joel Akpo, Olivier Bachem, Bernhard Schölkopf, and Stefan Bauer. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. In *Advances in Neural Information Processing Systems*, 2019. 6

[14] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010. 20

[15] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, 2018. 24

[16] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023. 3, 5, 24, 26

[17] Hermanni Hälvä, Sylvain Le Corff, Luc Lehéricy, Jonathan So, Yongjie Zhu, Elisabeth Gassiat, and Aapo Hyvarinen. Disentangling identifiable features from noisy data with structured nonlinear ica. *Advances in Neural Information Processing Systems*, 34:1624–1633, 2021. 2

[18] Nicklas Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. *arXiv preprint arXiv:2310.16828*, 2023. 5, 24

[19] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable GAN controls. In *Advances in neural information processing systems*, 2020. 6, 20

[20] William Harvey, Saeid Naderiparizi, and Frank Wood. Conditional image generation by conditioning variational autoencoders. *arXiv preprint arXiv:2102.12037*, 2021. 2

[21] Irina Higgins, Loïc 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 *ICLR*, 2017. 2, 6, 21, 27

[22] Kyle Hsu, Jubayer Ibn Hamid, Kaylee Burns, Chelsea Finn, and Jiajun Wu. Tripod: Three complementary inductive biases for disentangled representation learning. In *International Conference on Machine Learning*, 2024. 2, 5, 27

[23] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 28

[24] Yingyao Hu and Susanne M Schennach. Instrumental variable treatment of nonclassical measurement error models. *Econometrica*, 76(1):195–216, 2008. 4

[25] Aapo Hyvarinen and Hiroshi Morioka. Nonlinear ica of temporally dependent stationary sources. In *Artificial intelligence and statistics*, pages 460–469. PMLR, 2017. 2, 4, 20, 21, 27

[26] Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999. 2

[27] Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd international conference on artificial intelligence and statistics*, pages 859–868. PMLR, 2019. 1, 2, 3, 5, 20, 25

[28] Michael Janner, Yilun Du, Joshua Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *Proceedings of the 39th International Conference on Machine Learning*, pages 9902–9915. PMLR, 2022. 24, 28

[29] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International conference on*

CVPR #35

artificial intelligence and statistics, pages 2207–2217. PMLR, 2020. 2, 3, 6, 20, 21, 27

[30] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International conference on machine learning*, pages 2649–2658. PMLR, 2018. 2, 6, 20, 27

[31] David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. In *International Conference on Learning Representations*, 2021. 26, 27

[32] Lingjing Kong, Shaoan Xie, Weiran Yao, Yujia Zheng, Guangyi Chen, Petar Stojanov, Victor Akinwande, and Kun Zhang. Partial disentanglement for domain adaptation. In *International conference on machine learning*, pages 11455–11472. PMLR, 2022. 2

[33] Sébastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ica. In *Conference on Causal Learning and Reasoning*, pages 428–484. PMLR, 2022. 2, 5, 26, 27

[34] Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. *arXiv preprint arXiv:2503.00200*, 2025. 7

[35] Zinan Lin, Kiran Thekumparampil, Giulia Fanti, and Sewoong Oh. Infogan-cr and modelcentrality: Self-supervised model training and selection for disentangling gans. In *International conference on machine learning*, pages 6127–6139. PMLR, 2020. 6, 20

[36] Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Stratis Gavves. CITRIS: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning*, 2022. 26, 27

[37] 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 *Proceedings of the 36th International Conference on Machine Learning*, pages 4114–4124. PMLR, 2019. 1, 2

[38] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 1, 2

[39] Yao Mu, Tianxing Chen, Shijia Peng, Zanxin Chen, Zeyu Gao, Yude Zou, Lunkai Lin, Zhiqiang Xie, and Ping Luo. Robotwin: Dual-arm robot benchmark with generative digital twins (early version). In *European Conference on Computer Vision*, pages 264–273. Springer, 2024. 8

[40] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2, 5, 18, 19, 20, 27

[41] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 2, 24, 26

[42] William Peebles, John Peebles, Jun-Yan Zhu, Alexei Efros, and Antonio Torralba. The hessian penalty: A weak prior for unsupervised disentanglement. In *European Conference on Computer Vision*, 2020. 26, 27

[43] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. In *arXiv preprint arXiv:2204.06125*, 2022. 26, 27

[44] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. In *Advances in Neural Information Processing Systems*, 2019. 5, 26, 27

[45] Scott E Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. Deep visual analogy-making. In *Advances in neural information processing systems*, 2015. 6

[46] Patrik Reizinger, Alice Bizeul, Attila Juhos, Julia E Vogt, Randall Balestriero, Wieland Brendel, and David Klindt. Cross-entropy is all you need to invert the data generating process. *arXiv preprint arXiv:2410.21869*, 2024. 2, 5

[47] Patrik Reizinger, Balint Mucsanyi, Siyuan Guo, Benjamin Eysenbach, Bernhard Schölkopf, and Wieland Brendel. Skill learning via policy diversity yields identifiable representations for reinforcement learning. *arXiv preprint arXiv:2507.14748*, 2025. 2

[48] Xuanchi Ren, Tao Yang, Yuwang Wang, and Wenjun Zeng. Learning disentangled representation by exploiting pretrained generative models: A contrastive learning view. In *International Conference on Learning Representations*, 2021. 6, 20

[49] Geoffrey Roeder, Luke Metz, and Durk P Kingma. On linear identifiability of learned representations. In *Proceedings of the 38th International Conference on Machine Learning*, 2021. 1, 3

[50] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 4, 24, 26

[51] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 28

[52] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 24, 26

[53] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015. 1, 2, 26, 27

[54] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, 2017. 5, 26, 27

[55] 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. 1, 2

[56] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the GAN latent space. In *International conference on machine learning*, pages 9786–9796. PMLR, 2020. 6, 20

[57] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 7, 22

[58] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, 2020. 5, 15, 18, 19

[59] Yingheng Wang, Yair Schiff, Aaron Gokaslan, Weishen Pan, Fei Wang, Christopher De Sa, and Volodymyr Kuleshov. Infodiffusion: Representation learning using information maximizing diffusion models. *International conference on machine learning*, 2023. 26, 27

[60] Tao Yang, Yuwang Wang, Yan Lv, and Nanning Zheng. Disdiff: Unsupervised disentanglement of diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2023. 6, 20, 24, 26, 27

[61] Tao Yang, Cuiling Lan, Yan Lu, and Nanning Zheng. Diffusion model with cross attention as an inductive bias for disentanglement. *Advances in Neural Information Processing Systems*, 37:82465–82492, 2024. 6, 20, 21, 24, 26, 27

[62] Dingling Yao, Danru Xu, Sébastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. *arXiv preprint arXiv:2311.04056*, 2023. 1, 2

[63] Weiran Yao, Guangyi Chen, and Kun Zhang. Temporally disentangled representation learning. *Advances in Neural Information Processing Systems*, 35:26492–26503, 2022. 2, 26, 27

[64] Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026. 7

[65] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 833–842, 2021. 2

[66] Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: A general setting. *arXiv preprint arXiv:2402.05052*, 2024. 2

[67] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 1, 28

[68] Yujia Zheng, Ignavier Ng, and Kun Zhang. On the identifiability of nonlinear ica: Sparsity and beyond. *Advances in neural information processing systems*, 35:16411–16422, 2022. 2, 5, 25, 26, 27

[69] Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *Forty-second International Conference on Machine Learning*, 2025. 24, 26, 28

[70] 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*, pages 12979–12990. PMLR, 2021. 2, 5

Appendix

*Supplementary material for **Factored Generative Models through Mechanism Diversity***

Roadmap

A. Notation

Table 3. **Notation used throughout the paper.**

Symbol	Description
Generative model and data.	
$\mathbf{x}$	Observation (data sample).
$\hat{\mathbf{x}}$	Reconstructed/generated observation produced by the model.
$\mathbf{z}$	Ground-truth latent code.
$\hat{\mathbf{z}}$	Learned latent code; identifiability is up to permutation and componentwise invertible maps.
$\mathbf{u}$	Conditioning signal: text, class label, viewpoint, regime index, or action that reshapes the generator.
$\mathbf{o}_t$	World-model observation at time t.
$\mathbf{a}_t$	World-model action at time t, playing the role of $\mathbf{u}$ in the temporal GMD setting.
Generator and identifiability.	
$g(\mathbf{z}, \mathbf{u})$ or $g_{\mathbf{u}}(\mathbf{z})$	Conditional generator family indexed by $\mathbf{u}$; GMD studies how this family varies across conditioning values.
$\hat{g}(\hat{\mathbf{z}}, \mathbf{u})$	Observationally equivalent learned generator used in the identifiability argument.
$h_{\mathbf{u}} = \hat{g}_{\mathbf{u}}^{-1} \circ g_{\mathbf{u}}$	Relabeling map from true latents to learned latents at a fixed conditioning value.
J_g	Jacobian of the generator with respect to the latent variables.
J_h	Jacobian of the relabeling map with respect to the latent variables.
$m_i'(\mathbf{z}, \mathbf{u})$	First derivative $\partial_{z_i} \log \lvert J_g(\mathbf{z}, \mathbf{u})\rvert$.
$m_{ii}''(\mathbf{z}, \mathbf{u})$	Diagonal second derivative $\partial_{z_i}^2 \log \lvert J_g(\mathbf{z}, \mathbf{u})\rvert$.
$V(\mathbf{z}, \mathbf{u}^{(n)})$	GMD feature vector collecting the mechanism-derivative coordinates used in Assumption 2.
d_z	Latent dimension.
d_u	Conditioning dimension.
$\lvert\mathcal{U}\rvert$	Conditioning-set size; number of distinct conditioning mechanisms available for contrast.
MCL objective.	
c_ψ	Mechanism critic whose latent gradient is contrasted against the conditioning signal.
$\mathbf{g}(\mathbf{z}, \mathbf{u})$	Mechanism gradient $\nabla_{\mathbf{z}} c_\psi(G_\theta(\mathbf{z}, \mathbf{u}), \mathbf{z}, \mathbf{u})$.
Π_g	MCL projection head mapping the mechanism gradient into the embedding space.
Π_u	MCL projection head mapping the conditioning signal into the embedding space.
$\mathbf{q}(\mathbf{z}, \mathbf{u})$	L_2-normalized projected mechanism gradient (query).
$\mathbf{k}(\mathbf{u})$	L_2-normalized projected conditioning embedding (key).
τ	InfoNCE temperature in the MCL loss.
λ_{MCL}	MCL loss weight in the full training objective.

B. Formal Statements

The main text states the theory in compressed form to keep the paper readable. This appendix collects the fully quantified statements first; Appendix C then proves them in the same order.

B.1. Identifiability and Its Extensions

Proposition 1 (Full Statement: Capacity Separation of GMD and LRD). *Let $\Theta_{\text{LRD}} \subset \mathbb{R}^{d_L}$ and $\Theta_{\text{GMD}} \subset \mathbb{R}^{d_G}$ be open parameter domains. Let*

$$\Psi_{\text{LRD}} : \Theta_{\text{LRD}} \to \mathbb{R}^q, \qquad \Psi_{\text{GMD}} : \Theta_{\text{GMD}} \to \mathbb{R}^q$$

be finite separating coordinate charts for the induced conditional laws, obtained by evaluating a separating family of test functions at finitely many conditioning values. Assume both maps are real analytic, $\Psi_{\text{LRD}}(\Theta_{\text{LRD}}) \subseteq \Psi_{\text{GMD}}(\Theta_{\text{GMD}})$, and

there is a nonempty open set $\Omega \subset \Theta_{\mathrm{GMD}}$ *on which*

$$\operatorname{rank} D\Psi_{\mathrm{GMD}}(\theta_G) > \sup_{\theta_L \in \Theta_{\mathrm{LRD}}} \operatorname{rank} D\Psi_{\mathrm{LRD}}(\theta_L).$$

Then

$$\Theta_{\text{LRD-in-GMD}} := \{\theta_G \in \Theta_{\mathrm{GMD}} : \exists \theta_L \in \Theta_{\mathrm{LRD}} \text{ with } \Psi_{\mathrm{GMD}}(\theta_G) = \Psi_{\mathrm{LRD}}(\theta_L)\}$$

is contained in a proper analytic subset of Θ_{GMD} *and therefore has Lebesgue measure zero. Consequently, GMD strictly contains LRD at the level of conditional generative families.*

Theorem 1 (Full Statement: Identifiability under Sufficient GMD). *Let* $\mathcal{Z}, \widehat{\mathcal{Z}} \subset \mathbb{R}^{d_z}$ *be open connected sets, let* $p_{\mathbf{z}}$ *and* $p_{\hat{\mathbf{z}}}$ *be positive* C^2 *densities, and let* $\mathbf{z} \perp \mathbf{u}$*. For each* $\mathbf{u} \in \mathcal{U}$*, suppose*

$$g_{\mathbf{u}} : \mathcal{Z} \to \mathcal{X}, \qquad \hat{g}_{\mathbf{u}} : \widehat{\mathcal{Z}} \to \mathcal{X}$$

are C^3 *diffeomorphisms onto a common open image with nonzero Jacobian determinants. Assume observational equivalence,* $(g_{\mathbf{u}})_{\#}p_{\mathbf{z}} = (\hat{g}_{\mathbf{u}})_{\#}p_{\hat{\mathbf{z}}}$ *for every* $\mathbf{u}$*, and define*

$$h_{\mathbf{u}} := \hat{g}_{\mathbf{u}}^{-1} \circ g_{\mathbf{u}}, \qquad r_{\mathbf{u}} := h_{\mathbf{u}}^{-1}.$$

For $i = 1, \dots, d_z$*, define*

$$m_i'(\mathbf{z}, \mathbf{u}) = \partial_{z_i} \log |\det Dg_{\mathbf{u}}(\mathbf{z})|, \qquad m_{ii}''(\mathbf{z}, \mathbf{u}) = \partial_{z_i}^2 \log |\det Dg_{\mathbf{u}}(\mathbf{z})|,$$

and $V(\mathbf{z}, \mathbf{u}) = (m_1', \dots, m_{d_z}', m_{11}'', \dots, m_{d_z d_z}'') \in \mathbb{R}^{2d_z}$*. Assume:*
A1 ***(Sufficient GMD):*** *for every* $\mathbf{z} \in \mathcal{Z}$*, there exist* $\mathbf{u}^{(0)}, \dots, \mathbf{u}^{(2d_z)} \in \mathcal{U}$ *such that the* $2d_z$ *contrasts* $V(\mathbf{z}, \mathbf{u}^{(n)}) - V(\mathbf{z}, \mathbf{u}^{(0)})$*,* $n = 1, \dots, 2d_z$*, are linearly independent.*
A2 ***(Mixed-derivative separation):*** *for every distinct* k, l *and every* $n = 1, \dots, 2d_z$*,*

$$\left\langle V(\mathbf{z}, \mathbf{u}^{(n)}) - V(\mathbf{z}, \mathbf{u}^{(0)}),\ B_{kl}(\mathbf{z}, \mathbf{u}) \right\rangle = 0,$$

where, evaluated at $\hat{\mathbf{z}} = h_{\mathbf{u}}(\mathbf{z})$*,*

$$B_{kl}(\mathbf{z}, \mathbf{u}) = \left(\frac{\partial^2 r_{\mathbf{u},1}}{\partial \hat{z}_k \, \partial \hat{z}_l}, \dots, \frac{\partial^2 r_{\mathbf{u},d_z}}{\partial \hat{z}_k \, \partial \hat{z}_l}, \frac{\partial r_{\mathbf{u},1}}{\partial \hat{z}_k} \frac{\partial r_{\mathbf{u},1}}{\partial \hat{z}_l}, \dots, \frac{\partial r_{\mathbf{u},d_z}}{\partial \hat{z}_k} \frac{\partial r_{\mathbf{u},d_z}}{\partial \hat{z}_l} \right).$$

Then, for each $\mathbf{u}$*, there exist a permutation* $\pi_{\mathbf{u}}$ *and scalar* C^1 *diffeomorphisms* $\{h_{\mathbf{u},i}\}_{i=1}^{d_z}$ *such that*

$$\hat{z}_i = h_{\mathbf{u},i}(z_{\pi_{\mathbf{u}}(i)}), \qquad i = 1, \dots, d_z.$$

If $\mathbf{u} \mapsto h_{\mathbf{u}}$ *is continuous on a connected conditioning set, the permutation is independent of* $\mathbf{u}$*. Under the shared-latent-chart convention used in Definition 1, this is the factored form* $\hat{z}_i = h_i(z_{\pi(i)})$*.*

Proposition 2 (Full Statement: Self-Supervised GMD). *Let* $\mathcal{F}_t = \sigma(\mathbf{x}_{\le t})$ *and let* $\hat{\mathbf{z}}_t = e_\phi(\mathbf{x}_{\le t})$ *be an* $\mathcal{F}_t$*-measurable context representation. Suppose the next-step model has the form* $\mathbf{x}_{t+1} = g(\mathbf{z}_{t+1}, \hat{\mathbf{z}}_t)$*, with innovation exogeneity* $\mathbf{z}_{t+1} \perp \mathcal{F}_t$*, hence* $\mathbf{z}_{t+1} \perp \hat{\mathbf{z}}_t$*. If the support of* $\hat{\mathbf{z}}_t$ *contains the* $2d_z + 1$ *context values required by Theorem 1, and the conditional generator family satisfies the remaining regularity and mixed-derivative separation assumptions of that theorem with* $\mathbf{u} = \hat{\mathbf{z}}_t$*, then* $\mathbf{z}_{t+1}$ *is identifiable from* $\mathbf{x}_{t+1}$ *up to permutation and componentwise invertible transformations.*

Proposition 3 (Full Statement: Distribution-Level Invertibility from Three GMD Views). *Let* $\mathbf{x}^{(n)}$ *denote the observation generated under conditioning value* $\mathbf{u}^{(n)}$*,* $n = 0, 1, 2$*, with* $\mathbf{x}^{(n)} \perp \mathbf{x}^{(m)} \mid \mathbf{z}$ *for* $n \neq m$ *and* $\mathbf{z} \perp \mathbf{u}$*. Under the operator injectivity, bounded-density, non-redundancy, and smoothness assumptions stated in Assumptions 3–6, any observationally equivalent latent variable* $\hat{\mathbf{z}}$ *is related to* $\mathbf{z}$ *by an invertible differentiable relabeling map* h*, i.e.* $\hat{\mathbf{z}} = h(\mathbf{z})$*. Therefore, once the sufficient-GMD derivative condition of Theorem 1 holds after this distribution-level relabeling, the conclusion of Theorem 1 holds without assuming pointwise invertibility of* $g(\cdot, \mathbf{u})$*.*

B.2. MCL as a Mechanism-Diversity Objective

We use the notation of Appendix A and Section 4: the host generator G_θ, the mechanism critic c_ψ, the projection heads Π_g and Π_u into a common d-dimensional embedding space, the mechanism gradient $\mathbf{g}(\mathbf{z},\mathbf{u}) = \nabla_{\mathbf{z}} c_\psi(G_\theta(\mathbf{z},\mathbf{u}),\mathbf{z},\mathbf{u})$, and the L_2-normalized query and key $\mathbf{q}(\mathbf{z},\mathbf{u}) = \Pi_g(\mathbf{g}(\mathbf{z},\mathbf{u}))/\|\Pi_g(\mathbf{g}(\mathbf{z},\mathbf{u}))\|$ and $\mathbf{k}(\mathbf{u}) = \Pi_u(\mathbf{u})/\|\Pi_u(\mathbf{u})\|$. Replacing the in-batch negatives in the MCL loss by independent samples from $p(\mathbf{u})$ yields the population InfoNCE objective

$$\mathcal{L}^{\infty}_{\mathrm{MCL}}(\tau) = \mathbb{E}_{(\mathbf{z},\mathbf{u})\sim p(\mathbf{z})p(\mathbf{u})}\left[-\frac{\mathbf{q}(\mathbf{z},\mathbf{u})^\top \mathbf{k}(\mathbf{u})}{\tau} + \log \mathbb{E}_{\mathbf{u}'\sim p(\mathbf{u})}\exp\Big(\frac{\mathbf{q}(\mathbf{z},\mathbf{u})^\top \mathbf{k}(\mathbf{u}')}{\tau}\Big)\right]. \tag{2}$$

The next theorem shows that minimizers of $\mathcal{L}^{\infty}_{\mathrm{MCL}}$ realize the linear-independence content of Assumption 2.

A1 ***(Regularity):*** *G_θ and c_ψ are C^2 in $\mathbf{z}$; Π_g and Π_u are continuous; and $\mathbf{g}(\mathbf{z},\mathbf{u}) \neq 0$, $\Pi_g(\mathbf{g}(\mathbf{z},\mathbf{u})) \neq 0$, $\Pi_u(\mathbf{u}) \neq 0$ for $p(\mathbf{z})p(\mathbf{u})$-a.e. $(\mathbf{z},\mathbf{u})$.*

A2 ***(Capacity and spread):*** *$d \geq 2d_z + 1$, and the hypothesis class $\{(G_\theta, c_\psi, \Pi_g, \Pi_u)\}$ contains a configuration that simultaneously achieves alignment $\mathbf{q}(\mathbf{z},\mathbf{u}) = \mathbf{k}(\mathbf{u})$ a.e. and uniform spread $\mathbf{k}_{\#}\, p(\mathbf{u}) = \mathrm{Unif}(S^{d-1})$.*

A3 ***(Discriminative conditioning support):*** *there exist $\mathbf{u}^{(0)},\dots,\mathbf{u}^{(2d_z)} \in \operatorname{supp} p(\mathbf{u})$ such that $\Pi_u(\mathbf{u}^{(0)}),\dots,\Pi_u(\mathbf{u}^{(2d_z)})$ are linearly independent in $\mathbb{R}^d$.*

A2 is the standard alignment–uniformity capacity condition under which the population InfoNCE minimizer is unique [58]; without uniform spread, degenerate constant-$\mathbf{q}$ configurations can match the rescaled-loss infimum. A3 is the operational content of "enough distinguishable auxiliary inputs" and holds for any generic continuous Π_u whenever $\operatorname{supp} p(\mathbf{u})$ contains at least $2d_z + 1$ distinct points.

Theorem 2 (MCL minimizers satisfy mechanism-gradient diversity)**.** *Under* A1–A3*, for every fixed temperature $\tau > 0$ the population InfoNCE loss* (2)*, restricted to the hypothesis class of* A2*, attains its infimum at the alignment-and-spread configurations of* A2*; in particular every such minimizer satisfies $\mathbf{q}(\mathbf{z},\mathbf{u}) = \mathbf{k}(\mathbf{u})$ for $p(\mathbf{z})p(\mathbf{u})$-a.e. $(\mathbf{z},\mathbf{u})$. Consequently, for $p(\mathbf{z})$-a.e. $\mathbf{z}$ and the values $\{\mathbf{u}^{(n)}\}_{n=0}^{2d_z}$ supplied by* A3*, the projected mechanism-gradient contrasts*

$$\Delta_n(\mathbf{z}) \;:=\; \Pi_g(\mathbf{g}(\mathbf{z},\mathbf{u}^{(n)})) \;-\; \Pi_g(\mathbf{g}(\mathbf{z},\mathbf{u}^{(0)})), \qquad n = 1,\dots,2d_z, \tag{3}$$

are linearly independent in $\mathbb{R}^d$, realizing the linear-independence content of Assumption 2.

C. Proofs

C.1. Proof of Proposition 1

Proof. **(i) Inclusion.** Given any LRD model (f_L, g_L), construct a GMD model by setting $f_G(\boldsymbol{\epsilon}) := \boldsymbol{\epsilon}$ and $g_G(\mathbf{u},\mathbf{z}) := g_L(f_L(\mathbf{u},\mathbf{z}))$. Then $\mathbf{x} = g_G(\mathbf{u}, f_G(\boldsymbol{\epsilon})) = g_L(f_L(\mathbf{u},\boldsymbol{\epsilon}))$, identical to LRD. Hence $\mathcal{P}_{\mathrm{LRD}} \subseteq \mathcal{P}_{\mathrm{GMD}}$.

(ii) Zero-measure strictness. Let

$$\Phi_{\mathrm{LRD}} : \Theta_{\mathrm{LRD}} \to \mathcal{F}, \qquad \Phi_{\mathrm{GMD}} : \Theta_{\mathrm{GMD}} \to \mathcal{F}$$

denote the analytic maps from parameters to conditional data-distribution families, where $\mathcal{F}$ is any finite-dimensional coordinate chart obtained by evaluating a finite separating set of test functions and conditioning values. Because $P_{\boldsymbol{\mathcal{E}}}$ has a Lebesgue density, equality of the induced conditional laws implies equality of these separating coordinates. The LRD-realizable subset inside the GMD parameter space is therefore

$$\Theta_{\text{LRD-in-GMD}} = \Phi^{-1}_{\mathrm{GMD}}\big(\Phi_{\mathrm{LRD}}(\Theta_{\mathrm{LRD}})\big).$$

The inclusion construction above shows that this set is nonempty. Strictness means that the image $\Phi_{\mathrm{LRD}}(\Theta_{\mathrm{LRD}})$ is contained in a lower-dimensional analytic submanifold of $\Phi_{\mathrm{GMD}}(\Theta_{\mathrm{GMD}})$: LRD can only represent conditioning effects that factor through the latent distribution, whereas GMD parameters also include independent directions in which $\mathbf{u}$ changes the generator while $p(\mathbf{z})$ stays fixed. Under the nondegenerate analytic parameterization stated in the proposition, the analytic rank theorem then gives

$$\dim \Theta_{\text{LRD-in-GMD}} < \dim \Theta_{\mathrm{GMD}}.$$

Lower-dimensional analytic subsets of an open Euclidean parameter space have Lebesgue measure zero, proving the claim. □

C.2. Proof of Theorem 1

Proof. Denote $g_{\mathbf{u}}(\cdot) = g(\cdot, \mathbf{u})$, so $\mathbf{x} = g_{\mathbf{u}}(\mathbf{z})$. Observational equivalence defines

$$\hat{\mathbf{z}} = h_{\mathbf{u}}(\mathbf{z}) := \hat{g}_{\mathbf{u}}^{-1} \circ g_{\mathbf{u}}(\mathbf{z}),$$

with Jacobian $J_h(\mathbf{z}, \mathbf{u})$. By the change-of-variables formula,

$$\log p_{\hat{\mathbf{z}}}(\hat{\mathbf{z}}) = \log p_{\mathbf{z}}(\mathbf{z}) - \log |J_h(\mathbf{z}, \mathbf{u})|.$$

For distinct indices $k \neq l$, differentiate the identity twice with respect to $(\hat{z}_k, \hat{z}_l)$. The term $\log p_{\mathbf{z}}(\mathbf{z})$ is independent of $\mathbf{u}$ because $\mathbf{z} \perp \mathbf{u}$. After subtracting the equation at the reference value $\mathbf{u}^{(0)}$ from the equation at $\mathbf{u}^{(n)}$, the $\mathbf{u}$-independent density term cancels. The mixed-derivative separation condition in Theorem 1 separates true-generator and learned-generator derivative coordinates, leaving the homogeneous system

$$0 = \sum_{i=1}^{d_z} \Delta m''_{ii}(\mathbf{z}, n) \, \frac{\partial z_i}{\partial \hat{z}_l} \frac{\partial z_i}{\partial \hat{z}_k} + \sum_{i=1}^{d_z} \Delta m'_i(\mathbf{z}, n) \, \frac{\partial^2 z_i}{\partial \hat{z}_k \, \partial \hat{z}_l}, \tag{4}$$

where

$$\Delta m'_i(\mathbf{z}, n) = m'_i(\mathbf{z}, \mathbf{u}^{(n)}) - m'_i(\mathbf{z}, \mathbf{u}^{(0)}),$$
$$\Delta m''_{ii}(\mathbf{z}, n) = m''_{ii}(\mathbf{z}, \mathbf{u}^{(n)}) - m''_{ii}(\mathbf{z}, \mathbf{u}^{(0)}).$$

For fixed (k, l), define

$$B_{kl}(\mathbf{z}, \mathbf{u}) = \left(\frac{\partial^2 z_1}{\partial \hat{z}_k \, \partial \hat{z}_l}, \ldots, \frac{\partial^2 z_{d_z}}{\partial \hat{z}_k \, \partial \hat{z}_l}, \frac{\partial z_1}{\partial \hat{z}_l} \frac{\partial z_1}{\partial \hat{z}_k}, \ldots, \frac{\partial z_{d_z}}{\partial \hat{z}_l} \frac{\partial z_{d_z}}{\partial \hat{z}_k} \right).$$

Equation (4) is equivalently

$$\left\langle V(\mathbf{z}, \mathbf{u}^{(n)}) - V(\mathbf{z}, \mathbf{u}^{(0)}), B_{kl}(\mathbf{z}, \mathbf{u}) \right\rangle = 0, \qquad n = 1, \ldots, 2d_z.$$

The sufficient-GMD condition states that these $2d_z$ contrast vectors are linearly independent in $\mathbb{R}^{2d_z}$; hence they span $\mathbb{R}^{2d_z}$ and force $B_{kl}(\mathbf{z}, \mathbf{u}) = 0$ for every distinct pair $k \neq l$. In particular,

$$\frac{\partial z_i}{\partial \hat{z}_l} \frac{\partial z_i}{\partial \hat{z}_k} = 0 \quad \text{for all } i, \; k \neq l.$$

Thus each row of $J_{r_{\mathbf{u}}}(\hat{\mathbf{z}})$ has at most one nonzero entry. Since $r_{\mathbf{u}}$ is invertible, $J_{r_{\mathbf{u}}}(\hat{\mathbf{z}})$ is full rank, so each row and column has exactly one nonzero entry. Therefore there are a permutation $\pi_{\mathbf{u}}$ and scalar invertible functions $\rho_{\mathbf{u},i}$ such that

$$z_i = \rho_{\mathbf{u},i}(\hat{z}_{\pi_{\mathbf{u}}(i)}), \qquad i = 1, \ldots, d_z.$$

Inverting and reindexing gives $\hat{z}_i = h_{\mathbf{u},i}(z_{\pi_{\mathbf{u}}^{-1}(i)})$ for scalar invertible functions $h_{\mathbf{u},i}$. If $\mathbf{u} \mapsto h_{\mathbf{u}}$ is continuous in C^1 over a connected conditioning set, the discrete permutation $\pi_{\mathbf{u}}$ cannot change with $\mathbf{u}$ without making $J_{h_{\mathbf{u}}}$ discontinuous; hence $\pi_{\mathbf{u}}$ is constant across environments. □

C.3. Proof of Proposition 2

Proof. Let $\mathcal{F}_t = \sigma(\mathbf{x}_{\leq t})$ be the past-observation sigma-field and let $\hat{\mathbf{z}}_t = e_\phi(\mathbf{x}_{\leq t})$ be the context encoder output, so $\hat{\mathbf{z}}_t$ is $\mathcal{F}_t$-measurable. The predictive model has the form

$$\mathbf{x}_{t+1} = g(\mathbf{z}_{t+1}, \hat{\mathbf{z}}_t).$$

Assume the next-step innovation latent is conditionally exogenous:

$$\mathbf{z}_{t+1} \perp \mathcal{F}_t, \qquad \text{and therefore} \qquad \mathbf{z}_{t+1} \perp \hat{\mathbf{z}}_t.$$

Conditioning on the realized value of $\hat{\mathbf{z}}_t$ gives exactly the static GMD model of Theorem 1, with $\mathbf{u}$ replaced by $\hat{\mathbf{z}}_t$. If the support of $\hat{\mathbf{z}}_t$ contains the $2d_z + 1$ context values required by the sufficient-GMD condition, the theorem applies pointwise over those contexts and yields identifiability of $\mathbf{z}_{t+1}$ up to a permutation and componentwise invertible maps. When e_ϕ is learned jointly with g, the same conclusion holds at any population optimum whose encoder output satisfies the support and exogeneity conditions; when e_ϕ is fixed, these conditions are assumptions on the fixed representation. This proves the proposition. □

C.4. Proof of Proposition 3

The high-level argument and its multi-view interpretation are given in the main text after Proposition 3. Here we record the formal assumptions (Assumptions 3–6) and the proof.

We first introduce the operator machinery needed for the proof.

Definition 2 (Linear Operator)**.** *Consider two random variables a and b with support sets $\mathcal{A}$ and $\mathcal{B}$. The linear operator $L_{b|a}$ maps a density function p_a on $\mathcal{A}$ to a function on $\mathcal{B}$ via:*

$$[L_{b|a} \circ p_a](b) = \int_{\mathcal{A}} p(b \mid a)\, p_a(a)\, da. \tag{5}$$

Definition 3 (Diagonal Operator)**.** *Consider two random variables a and b with density functions p_a and p_b on support sets $\mathcal{A}$ and $\mathcal{B}$. The diagonal operator $D_{b|a}$ maps the density function p_a to another density function via pointwise multiplication of $p_{b|a}$ at a fixed point b:*

$$D_{b|a} \circ p_a = p_{b|a}(b \mid \cdot)\, p_a, \quad \textit{where } D_{b|a} = p_{b|a}(b \mid \cdot). \tag{6}$$

The proof requires the following additional assumptions.

Assumption 3 (Operator Injectivity)**.** *The linear operators $L_{\mathbf{x}^{(n)}|\mathbf{z}}$ defined by the conditional density $p(\mathbf{x} \mid \mathbf{z}, \mathbf{u}^{(n)})$ are injective for each conditioning value $\mathbf{u}^{(n)}$, $n = 0, 1, 2$.*

Assumption 4 (Bounded Density)**.** *The conditional densities $p(\mathbf{x} \mid \mathbf{z}, \mathbf{u}^{(n)})$ are uniformly bounded: $\sup_{\mathbf{z},\mathbf{x}} p(\mathbf{x} \mid \mathbf{z}, \mathbf{u}^{(n)}) < \infty$ for each n.*

Assumption 5 (Non-redundancy)**.** *For each conditioning value $\mathbf{u}^{(n)}$, distinct latent values produce distinct conditional densities: $\mathbf{z} \neq \mathbf{z}' \implies p(\mathbf{x} \mid \mathbf{z}, \mathbf{u}^{(n)}) \neq p(\mathbf{x} \mid \mathbf{z}', \mathbf{u}^{(n)})$ as functions of $\mathbf{x}$.*

Assumption 6 (Smoothness)**.** *The conditional density $p(\mathbf{x} \mid \mathbf{z}, \mathbf{u}^{(n)})$ is continuously differentiable in $\mathbf{z}$ for each n.*

Proof. Denote $\mathbf{x}^{(n)}$ as the observation generated under conditioning value $\mathbf{u}^{(n)}$, so $\mathbf{x}^{(n)} \sim p(\mathbf{x} \mid \mathbf{z}, \mathbf{u}^{(n)})$ for $n = 0, 1, 2$. Since $\mathbf{z} \perp \mathbf{u}$, the observations $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)}$ are conditionally independent given $\mathbf{z}$:

$$\begin{aligned} p(\mathbf{x}^{(0)} \mid \mathbf{x}^{(1)}, \mathbf{z}) &= p(\mathbf{x}^{(0)} \mid \mathbf{z}), \\ p(\mathbf{x}^{(2)} \mid \mathbf{x}^{(1)}, \mathbf{x}^{(0)}, \mathbf{z}) &= p(\mathbf{x}^{(2)} \mid \mathbf{z}). \end{aligned} \tag{7}$$

From $p(\mathbf{x}^{(2)}, \mathbf{x}^{(1)} \mid \mathbf{x}^{(0)})$, the conditional independence yields:

$$\begin{aligned} p(\mathbf{x}^{(2)}, \mathbf{x}^{(1)} \mid \mathbf{x}^{(0)}) &= \int_{\mathcal{Z}} p(\mathbf{x}^{(2)} \mid \mathbf{x}^{(1)}, \mathbf{z}, \mathbf{x}^{(0)})\, p(\mathbf{x}^{(1)}, \mathbf{z} \mid \mathbf{x}^{(0)})\, d\mathbf{z} \\ &= \int_{\mathcal{Z}} p(\mathbf{x}^{(2)} \mid \mathbf{z})\, p(\mathbf{x}^{(1)} \mid \mathbf{z})\, p(\mathbf{z} \mid \mathbf{x}^{(0)})\, d\mathbf{z}, \end{aligned} \tag{8}$$

where the second equality uses (7). Integrating both sides over $\mathbf{x}^{(0)}$ against $p(\mathbf{x}^{(0)})$ and expressing in operator form (Definitions 2 and 3):

$$L_{\mathbf{x}^{(1)};\mathbf{x}^{(2)}|\mathbf{x}^{(0)}} = L_{\mathbf{x}^{(2)}|\mathbf{z}}\, D_{\mathbf{x}^{(1)}|\mathbf{z}}\, L_{\mathbf{z}|\mathbf{x}^{(0)}}. \tag{9}$$

Integrating over $\mathbf{x}^{(1)}$ on both sides gives $L_{\mathbf{x}^{(2)}|\mathbf{x}^{(0)}} = L_{\mathbf{x}^{(2)}|\mathbf{z}}\, L_{\mathbf{z}|\mathbf{x}^{(0)}}$. By Assumption 3, $L_{\mathbf{x}^{(2)}|\mathbf{z}}$ is injective, so $L_{\mathbf{z}|\mathbf{x}^{(0)}} = L^{-1}_{\mathbf{x}^{(2)}|\mathbf{z}}\, L_{\mathbf{x}^{(2)}|\mathbf{x}^{(0)}}$. Substituting back into (9) and right-multiplying by $L^{-1}_{\mathbf{x}^{(2)}|\mathbf{x}^{(0)}}$, which exists and is densely defined by Assumption 3, gives

$$\begin{aligned} &L_{\mathbf{x}^{(1)};\mathbf{x}^{(2)}|\mathbf{x}^{(0)}} L^{-1}_{\mathbf{x}^{(2)}|\mathbf{x}^{(0)}} \\ &\qquad = L_{\mathbf{x}^{(2)}|\mathbf{z}}\, D_{\mathbf{x}^{(1)}|\mathbf{z}}\, L^{-1}_{\mathbf{x}^{(2)}|\mathbf{z}}. \end{aligned} \tag{10}$$

By Assumption 4, the left-hand side is a bounded operator. By the uniqueness of spectral decomposition [8, 10], the eigenvalues $D_{\mathbf{x}^{(1)}|\mathbf{z}}$ (the entries $\{p(\mathbf{x}^{(1)} \mid \mathbf{z})\}$) and the eigenfunctions in $L_{\mathbf{x}^{(2)}|\mathbf{z}}$ (the columns $\{p(\mathbf{x}^{(2)} \mid \mathbf{z})\}$) are unique up to standard indeterminacies:

$$\begin{aligned} L_{\mathbf{x}^{(2)}|\mathbf{z}} &= C\, L_{\mathbf{x}^{(2)}|\hat{\mathbf{z}}}\, P, \\ D_{\mathbf{x}^{(1)}|\mathbf{z}} &= P^{-1}\, D_{\mathbf{x}^{(1)}|\hat{\mathbf{z}}}\, P. \end{aligned} \tag{11}$$

where C is a nonzero scalar rescaling and P is a permutation operator. Since $\int p(\mathbf{x}^{(2)} \mid \mathbf{z})\, d\mathbf{x}^{(2)} = 1$ for every $\mathbf{z}$, the only solution to $\int C\, p(\mathbf{x}^{(2)} \mid \mathbf{z})\, d\mathbf{x}^{(2)} = 1$ is $C = 1$. From $D_{\mathbf{x}^{(1)}|\mathbf{z}} = P^{-1}\, D_{\mathbf{x}^{(1)}|\hat{\mathbf{z}}}\, P$, the sets of conditional densities must match: $\{p(\mathbf{x}^{(1)} \mid \mathbf{z})\}_{\mathbf{z}} = \{p(\mathbf{x}^{(1)} \mid \hat{\mathbf{z}})\}_{\hat{\mathbf{z}}}$. Since sets are unordered, a relabeling map h is needed to consistently match entries:

$$p(\mathbf{x}^{(1)} \mid h(\mathbf{z})) = p(\mathbf{x}^{(1)} \mid \hat{\mathbf{z}}), \quad \text{for all } \mathbf{z}, \hat{\mathbf{z}}. \tag{12}$$

By Assumption 5, distinct $\mathbf{z}$ values produce distinct conditional densities, so h is one-to-one (invertible). By Assumption 6, h is differentiable. Hence $\hat{\mathbf{z}} = h(\mathbf{z})$ with h invertible and differentiable.

Finally, if $d_{\hat{z}} > d_z$, then d_z latent components suffice to explain $\mathbf{x}$, and the remaining $d_{\hat{z}} - d_z$ components satisfy $p(\mathbf{x} \mid \mathbf{z}_{1:d_z}, \mathbf{z}'_{d_z+1:d_{\hat{z}}}) = p(\mathbf{x} \mid \mathbf{z}_{1:d_z}, \mathbf{z}''_{d_z+1:d_{\hat{z}}})$ for all $\mathbf{z}', \mathbf{z}''$, contradicting Assumption 5. If $d_{\hat{z}} < d_z$, then $d_z - d_{\hat{z}}$ latent variables are constant, contradicting that they are variables. Hence $d_{\hat{z}} = d_z$. □

C.5. Proof of Theorem 2

Proof. The argument has two threads: alignment of the projected mechanism gradient to the projected conditioning embedding, established by an exact decomposition of the population InfoNCE; and linear independence of the gradient contrasts at any fixed $\mathbf{z}$, which follows from the alignment-induced gauge equation together with A3.

The population InfoNCE (2) is the limit $\lim_{M\to\infty}[\mathcal{L}^M_{\mathrm{MCL}}(\tau) - \log M]$ of the standard M-negative InfoNCE estimator [40]. Substituting the unit-norm identities $\mathbf{q}^\top \mathbf{k}(\mathbf{u}) = 1 - \frac{1}{2}\|\mathbf{q} - \mathbf{k}(\mathbf{u})\|_2^2$ and $\mathbf{q}^\top \mathbf{k}(\mathbf{u}') = 1 - \frac{1}{2}\|\mathbf{q} - \mathbf{k}(\mathbf{u}')\|_2^2$ into (2) (the $\pm 1/\tau$ contributions cancel) yields the *exact* algebraic identity

$$\begin{aligned}\mathcal{L}^\infty_{\mathrm{MCL}}(\tau) = &\underbrace{\frac{1}{2\tau}\, \mathbb{E}_{(\mathbf{z},\mathbf{u})}\Big[\big\|\mathbf{q}(\mathbf{z},\mathbf{u}) - \mathbf{k}(\mathbf{u})\big\|_2^2\Big]}_{\mathcal{A}(\mathbf{q},\mathbf{k};\tau)\geq 0} \\ &+ \underbrace{\mathbb{E}_{(\mathbf{z},\mathbf{u})}\left[\log \mathbb{E}_{\mathbf{u}'\sim p(\mathbf{u})} \exp\left(\frac{-\|\mathbf{q}(\mathbf{z},\mathbf{u}) - \mathbf{k}(\mathbf{u}')\|_2^2}{2\tau}\right)\right]}_{\widetilde{\mathcal{U}}(\mathbf{q},\mathbf{k};\tau)}.\end{aligned} \tag{13}$$

Two observations make alignment fall out of (13) once the spread part of A2 holds. First, $\mathcal{A}(\mathbf{q},\mathbf{k};\tau) \geq 0$ with equality if and only if $\mathbf{q}(\mathbf{z},\mathbf{u}) = \mathbf{k}(\mathbf{u})$ for $p(\mathbf{z})p(\mathbf{u})$-a.e. $(\mathbf{z},\mathbf{u})$. Second, when $\mathbf{k}_{\#}\, p(\mathbf{u}) = \mathrm{Unif}(S^{d-1})$ the function $\mathbf{a} \mapsto \mathbb{E}_{\mathbf{u}'\sim p(\mathbf{u})} \exp(-\|\mathbf{a} - \mathbf{k}(\mathbf{u}')\|_2^2/(2\tau))$ is rotationally invariant on S^{d-1} (its value depends only on $\|\mathbf{a}\|_2$, by spherical symmetry of the uniform measure); for every unit vector $\mathbf{a}$ it equals the constant $I(\tau, d) := \int_{S^{d-1}} e^{-\|\mathbf{a}_0 - \mathbf{v}\|_2^2/(2\tau)}\, d\sigma(\mathbf{v})$, where $\mathbf{a}_0$ is any fixed unit vector and σ is the uniform measure on S^{d-1}.

Pick any $\mathbf{k}^\star$ realizing the spread part of A2; such a key map exists by A2. The rotational-invariance observation gives $\widetilde{\mathcal{U}}(\mathbf{q},\mathbf{k}^\star;\tau) = \log I(\tau, d)$ for every unit-norm $\mathbf{q}$, so substituting into (13) reduces the loss at $\mathbf{k}^\star$ to a constant plus the alignment term:

$$\mathcal{L}^\infty_{\mathrm{MCL}}(\mathbf{q},\mathbf{k}^\star;\tau) \;=\; \mathcal{A}(\mathbf{q},\mathbf{k}^\star;\tau) + \log I(\tau, d). \tag{14}$$

Minimizing over unit-norm $\mathbf{q}$ then selects the unique optimum $\mathbf{q}(\mathbf{z},\mathbf{u}) = \mathbf{k}^\star(\mathbf{u})$, achieving the value $L_{\mathrm{AU}}(\tau, d) := \log I(\tau, d)$. To rule out non-aligned minimizers elsewhere in the hypothesis class, observe that $L_{\mathrm{AU}}(\tau, d)$ is also the *global* infimum of $\mathcal{L}^\infty_{\mathrm{MCL}}$ over all unit-norm encoder pairs: by the joint-minimum analysis of population InfoNCE [58, Theorem 1 and Proposition 2] adapted to the two-encoder setup of (2), the global infimum is attained only on the gauge orbit $\{(\mathbf{q},\mathbf{k}) : \exists\, R \in O(d), \mathbf{q}(\mathbf{z},\mathbf{u}) = R\,\mathbf{k}(\mathbf{u})$ a.s. and $\mathbf{k}_{\#}\, p(\mathbf{u}) = \mathrm{Unif}(S^{d-1})\}$. The orthogonal R reflects the gauge invariance of the loss under simultaneous rotations $(\mathbf{q},\mathbf{k}) \mapsto (R\mathbf{q}, R\mathbf{k})$, and is fixed to the identity by the choice of architectures (Π_g, Π_u) once those are pinned down. Every minimizer inside the hypothesis class therefore satisfies $\mathbf{q}(\mathbf{z},\mathbf{u}) = \mathbf{k}(\mathbf{u})$ for $p(\mathbf{z})p(\mathbf{u})$-a.e. $(\mathbf{z},\mathbf{u})$.

With alignment in hand, the linear-independence claim is purely geometric. Fix $\mathbf{z}$ in the (full $p(\mathbf{z})$-measure) set on which $\mathbf{q}(\mathbf{z},\mathbf{u}) = \mathbf{k}(\mathbf{u})$ holds for $p(\mathbf{u})$-a.e. $\mathbf{u}$, and choose the $2d_z + 1$ values $\{\mathbf{u}^{(n)}\}_{n=0}^{2d_z}$ given by A3. Unrolling the definition of $\mathbf{q}, \mathbf{k}$,

$$\frac{\Pi_g(\mathbf{g}(\mathbf{z},\mathbf{u}^{(n)}))}{\|\Pi_g(\mathbf{g}(\mathbf{z},\mathbf{u}^{(n)}))\|} \;=\; \frac{\Pi_u(\mathbf{u}^{(n)})}{\|\Pi_u(\mathbf{u}^{(n)})\|},$$

so the two sides are positively collinear. By A1 the norms are strictly positive; setting

$$\alpha_n(\mathbf{z}) \;:=\; \frac{\|\Pi_g(\mathbf{g}(\mathbf{z},\mathbf{u}^{(n)}))\|}{\|\Pi_u(\mathbf{u}^{(n)})\|} \;>\; 0, \qquad n = 0, \ldots, 2d_z, \tag{15}$$

we obtain the gauge equation $\Pi_g(\mathbf{g}(\mathbf{z}, \mathbf{u}^{(n)})) = \alpha_n(\mathbf{z})\, \Pi_u(\mathbf{u}^{(n)})$. Substituting into (3),

$$\Delta_n(\mathbf{z}) \;=\; \alpha_n(\mathbf{z})\, \Pi_u(\mathbf{u}^{(n)}) \;-\; \alpha_0(\mathbf{z})\, \Pi_u(\mathbf{u}^{(0)}), \qquad n = 1, \dots, 2d_z.$$

Suppose $\sum_{n=1}^{2d_z} \beta_n\, \Delta_n(\mathbf{z}) = 0$ for scalars $\beta_n \in \mathbb{R}$. Expanding,

$$\sum_{n=1}^{2d_z} \beta_n\, \alpha_n(\mathbf{z})\, \Pi_u(\mathbf{u}^{(n)}) \;-\; \alpha_0(\mathbf{z}) \left(\sum_{n=1}^{2d_z} \beta_n \right) \Pi_u(\mathbf{u}^{(0)}) \;=\; 0,$$

which is a linear combination of the $2d_z + 1$ vectors $\{\Pi_u(\mathbf{u}^{(n)})\}_{n=0}^{2d_z}$. By A3 these vectors are linearly independent in $\mathbb{R}^d$; the embedding dimension $d \geq 2d_z + 1$ in A2 is what makes such a tuple feasible. Hence all coefficients vanish:

$$\beta_n\, \alpha_n(\mathbf{z}) \;=\; 0 \quad (n = 1, \dots, 2d_z), \qquad \alpha_0(\mathbf{z}) \left(\sum_{n=1}^{2d_z} \beta_n \right) = 0.$$

Since $\alpha_n(\mathbf{z}) > 0$ by (15), the first family forces $\beta_n = 0$ for $n = 1, \dots, 2d_z$. The second equation is then automatically satisfied. Therefore $\{\Delta_n(\mathbf{z})\}_{n=1}^{2d_z}$ is linearly independent in $\mathbb{R}^d$, completing the proof. □

C.6. From MCL to the GMD Condition

Theorem 2 establishes linear independence of the projected critic-gradient contrasts $\{\Delta_n(\mathbf{z})\}_{n=1}^{2d_z}$. Assumption 2 is stated in terms of $V(\mathbf{z}, \mathbf{u}^{(n)})$, the vector of first and second $\mathbf{z}$-derivatives of $\log |J_{G_\theta}(\mathbf{z}, \mathbf{u}^{(n)})|$. The two notions of mechanism diversity are not literally identical, but they coincide under a standard *score-matching* link between the critic and the generator. In each instantiation, MCL is paired with a host loss $\mathcal{L}_{\text{rec}}$ (variational ELBO for MD-VAE; conditional denoising score matching for MDDiff; latent-dynamics negative log-likelihood for ACG-WM) whose minimizer drives $c_\psi(\mathbf{x}, \mathbf{z}, \mathbf{u})$ to a $\mathbf{z}$-derivative-faithful proxy of the conditional log-density of the host model. For the diffusion case, the change-of-variables for the marginal $p_\theta(\mathbf{x} \mid \mathbf{u}) = p_{\mathbf{z}}(G_{\theta,\mathbf{u}}^{-1}(\mathbf{x})) \cdot |J_{G_{\theta,\mathbf{u}}}(G_{\theta,\mathbf{u}}^{-1}(\mathbf{x}))|^{-1}$ gives the score

$$\nabla_{\mathbf{x}} \log p_\theta(\mathbf{x} \mid \mathbf{u}) \;=\; J_{G_{\theta,\mathbf{u}}}^{-\top}(\mathbf{z}) \big[\nabla_{\mathbf{z}} \log p_{\mathbf{z}}(\mathbf{z}) - \nabla_{\mathbf{z}} \log |J_{G_{\theta,\mathbf{u}}}(\mathbf{z})| \big] \Big|_{\mathbf{z} = G_{\theta,\mathbf{u}}^{-1}(\mathbf{x})}, \tag{16}$$

so the conditioning-dependent piece of the critic gradient $\nabla_{\mathbf{z}} c_\psi$ contains exactly the first-order coordinates $m_i'(\mathbf{z}, \mathbf{u}) = \partial_{z_i} \log |J_{G_{\theta,\mathbf{u}}}(\mathbf{z})|$ of $V(\mathbf{z}, \mathbf{u})$, up to an additive $\mathbf{u}$-independent term that drops out of every contrast Δ_n. Linear independence of $\{\Delta_n(\mathbf{z})\}$ therefore transfers to linear independence of the first-order half of $V(\mathbf{z}, \mathbf{u}^{(n)}) - V(\mathbf{z}, \mathbf{u}^{(0)})$. The second-order coordinates m_{ii}'' are recovered by augmenting the projection head Π_g with the diagonal Hessian $\operatorname{diag}(\nabla_{\mathbf{z}}^2 c_\psi)$, after which the same gauge argument applies coordinate-wise. Thus, under the standard architectural pairing of c_ψ with the host generator's likelihood structure, Theorem 2 delivers Assumption 2 verbatim, and Theorem 1 then yields identifiability up to a permutation and componentwise transformations.

Scope. Theorem 2 concerns *population* minimizers; the standard caveats of finite-sample InfoNCE (capacity, optimization, batch-size dependence of the negative distribution) carry over from the broader contrastive-learning literature [40, 58]. The bridge in (16) is exact only when $G_{\theta,\mathbf{u}}$ is invertible; for non-invertible diffusion or world-model generators, the bridge is replaced by the score-matching identity at the noise level, and the linear-independence transfer holds whenever the score network is rich enough to recover the conditional score. Our experiments (Sections 5.2–5.3) verify that the population picture survives at realistic batch sizes.

C.7. Comparison with Contrastive Objectives

Contrastive losses share the InfoNCE template but differ in what they contrast and what guarantee they target. Table 4 compares MCL against four representative families. The structural distinction is that all prior families contrast *representations* (or representations of contexts/views) and target either a similarity geometry (SimCLR/MoCo), a mutual-information lower bound (CPC), or an LRD-style nonlinear-ICA identifiability (TCL, iVAE-contrastive). MCL is the only family that contrasts *mechanism gradients* $\nabla_{\mathbf{z}} c_\psi(G(\mathbf{z}, \mathbf{u}), \mathbf{z}, \mathbf{u})$ against the conditioning signal, with the explicit goal of certifying mechanism diversity (Assumption 2) and thus a factored generative model. Negatives are drawn by shuffling the conditioning index $\mathbf{u}$ within the batch (rather than perturbing $p(\mathbf{z})$), which is what couples the alignment objective to the GMD geometry rather than to a representational similarity.

Table 4. **Contrastive objective families.** Columns list the positive pair, the source of negatives, the object contrasted, and what the optimum targets, for five representative families.

Family	Positive pair	Negatives	Object contrasted	Optimum target
Augmentation [SimCLR/-MoCo; 6]	$(f(\mathbf{x}), f(\tilde{\mathbf{x}}))$, $\tilde{\mathbf{x}}$ an augmentation of $\mathbf{x}$	In-batch other images	Image representation $f(\mathbf{x})$	Augmentation-invariant similarity
Predictive [CPC; 40]	$(c_t, f(\mathbf{x}_{t+k}))$, context vs. future	In-batch other futures	Context vs. future representation	MI lower bound $I(c_t; \mathbf{x}_{t+k})$
Time-contrastive [TCL; 25]	$(\mathbf{x}$, segment label$)$	Other segment labels	Encoder of $\mathbf{x}$	Nonlinear ICA / LRD identifiability
Density-ratio / NCE [14]	Data $\mathbf{x}$ vs. noise $\tilde{\mathbf{x}}$	Noise samples	Log-density-ratio $\log p_{\text{data}}/p_{\text{noise}}$	Density / score estimation
Mechanistic MCL (ours)	$(\mathbf{g}(\mathbf{z}, \mathbf{u}),\ \mathbf{u})$	Shuffled $\mathbf{u}_j$ in batch	Mechanism gradient vs. conditioning	GMD identifiability, Theorem 2

D. Experimental Details

D.1. Disentanglement Metrics

FactorVAE score [30]. For each ground-truth factor k, sample observations varying only in k, encode them, and identify the latent dimension with minimal normalized variance: $d_k = \arg\min_j \sigma_j^2 / \sum_{j'} \sigma_{j'}^2$. The score is the accuracy of a majority-vote classifier predicting k from d_k. A score of 1.0 means perfect axis-alignment.

DCI disentanglement [11]. Train a predictor from $\mathbf{z}$ to each factor k and extract the feature importance matrix $R \in \mathbb{R}^{d \times K}$. The disentanglement score is

$$D = 1 - \frac{1}{d} \sum_{j=1}^{d} H\left(\frac{R_{j\cdot}}{\sum_k R_{jk}} \right) \Big/ \log K, \tag{17}$$

where $H(\cdot)$ is the entropy. $D = 1$ when each latent dimension encodes exactly one factor.

D.2. Baselines for Concept Disentanglement

We expand the baseline descriptions for Section 5.2 (Table 1).

VAE-based. **FactorVAE** [30] adds a total-correlation penalty on the aggregated posterior to encourage statistically independent latents. **β-TCVAE** [5] decomposes the KL term and upweights the total-correlation component, isolating the disentanglement-relevant pressure of β-VAE.

GAN-based. **InfoGAN-CR** [35] extends InfoGAN with a contrastive regulariser on the inferred factors. **GANSpace** [19] discovers disentangled directions *post-hoc* via PCA on early-layer GAN activations. **LatentDisco** [56] learns directions in a pretrained GAN's latent space using a self-supervised classifier. **DisCo** [48] is a contrastive method on a pretrained generator: it samples paired latent walks and trains an encoder to detect the active factor.

Diffusion-based. **DisDiff** [60] associates each latent with a per-factor sub-gradient field of the score network. **EncDiff** [61] routes each scalar latent through an independent MLP and injects the resulting tokens via cross-attention into the U-Net. **MDDiff (ours)** uses this tokenized diffusion scaffold, adds $\mathbf{u}$-conditioned decoding to form the MDDiff pretraining model, and then fine-tunes that checkpoint with MCL to enforce mechanism diversity through the cross-attention slots.

D.3. MD-VAE (Simulations)

Synthetic data-generating process. We follow the standard nonlinear ICA simulation protocol [27, 29] and instantiate it inside the GMD regime. For each run we draw a latent dimension $d_z \in \{5, 10, 20\}$ and a conditioning-set size $|\mathcal{U}| \in \{2, 4, 8, 16, 32\}$ (default $|\mathcal{U}| = 8$). Each simulated dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{u}_i)\}_{i=1}^N$ with $N = 50{,}000$ is generated as follows. The *conditioning signal* $\mathbf{u}_i$ is drawn uniformly from a fixed alphabet $\mathcal{U} = \{\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(|\mathcal{U}|)}\}$, where each $\mathbf{u}^{(n)} \in \mathbb{R}^{d_u}$ ($d_u = 16$) is sampled once at the start of the run from $\mathcal{N}(0, \mathbf{I}_{d_u})$. The *latent* $\mathbf{z}_i \sim \mathcal{N}(0, \mathbf{I}_{d_z})$ is sampled independently of $\mathbf{u}_i$, instantiating

the GMD assumption $\mathbf{z} \perp \mathbf{u}$; this is the key difference from the iVAE [29] simulation, in which $p(\mathbf{z} \mid \mathbf{u})$ shifts. The *mechanism-diverse generator* is $g(\mathbf{z}, \mathbf{u}) = h \circ \big(\mathbf{z} \odot \mathbf{m}(\mathbf{u}) + \mathbf{b}(\mathbf{u})\big)$, where $\mathbf{m}, \mathbf{b} : \mathbb{R}^{d_u} \to \mathbb{R}^{d_z}$ are two-layer MLPs (width 64, leaky-ReLU) initialized so that $\mathbf{m}(\mathbf{u}^{(n)})$ varies meaningfully across n, and h is a fixed invertible 3-layer leaky-ReLU MLP (width $\max(64, 4d_z)$); the componentwise product $\mathbf{z} \odot \mathbf{m}(\mathbf{u})$ implements per-coordinate mechanism reshaping, satisfying Assumption 2 for $|\mathcal{U}| \geq 2d_z + 1$. Finally, observations are $\mathbf{x}_i = g(\mathbf{z}_i, \mathbf{u}_i) + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, 0.01\,\mathbf{I})$. We use an $80\%/20\%$ train/test split, repeating each configuration over 30 random seeds.

Models. All three baselines share the same encoder–decoder backbone (3-layer leaky-ReLU MLPs, hidden width $\max(64, 4d_z)$, latent dim d_z) and differ only in how $\mathbf{u}$ is incorporated. **β-VAE** [21] is a standard VAE with KL weight $\beta = 4$ in which $\mathbf{u}$ is fed to neither encoder nor decoder. **iVAE** [29] uses a $\mathbf{u}$-conditional encoder $q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{u})$ and a $\mathbf{u}$-conditional Gaussian prior $p(\mathbf{z} \mid \mathbf{u}) = \mathcal{N}(\mu_\lambda(\mathbf{u}), \mathrm{diag}(\sigma^2_\lambda(\mathbf{u})))$ implemented by a 2-layer MLP, with a decoder that is unconditioned in $\mathbf{u}$ (this is exactly the LRD regime). **MD-VAE** (ours) keeps the prior $\mathbf{u}$-independent, $p(\mathbf{z}) = \mathcal{N}(0, \mathbf{I})$, and feeds $\mathbf{u}$ to both the encoder $q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{u})$ and the decoder $p_\theta(\mathbf{x} \mid \mathbf{z}, \mathbf{u})$, training with the conditional ELBO of Equation (1). We optimize with Adam (learning rate 10^{-3}, batch size 256) for 200 epochs.

Evaluation. After training we recover latents $\hat{\mathbf{z}}_i = \mathbb{E}_{q_\phi(\mathbf{z} \mid \mathbf{x}_i, \mathbf{u}_i)}[\mathbf{z}]$ on the test split and compute MCC by solving the linear assignment problem between $\hat{\mathbf{z}}$ and the true $\mathbf{z}$ in absolute Pearson correlation, following Hyvarinen and Morioka [25], Khemakhem et al. [29].

D.4. MDDiff (Image Disentanglement)

Architecture. MDDiff builds on EncDiff [61] with two structural additions that together make it a bona fide GMD model. The shared backbone is unchanged from EncDiff: an `Encoder4` CNN (hidden width 128, four stride-2 convolutional layers) maps a 64×64 RGB image to $K{=}20$ scalar concept factors; each scalar is expanded by an independent three-layer MLP ($1 \to 64 \to 128 \to 16$, ELU activations) into a 16-dim concept token, so each token corresponds to exactly one semantic factor. The denoising backbone is a U-Net with base channel width 64, channel multipliers $[1, 2, 4, 4]$, 2 residual blocks per resolution, and 8-head spatial cross-attention injecting the 20 concept tokens at all three downsampled resolutions.

On top of this backbone we make two additions required for the GMD interpretation. First, we condition the *VAE decoder* on the concept variable $\mathbf{u}$: the original EncDiff feeds the concept tokens only through the U-Net cross-attention, so the VAE decoder that maps the diffusion latent back to pixels is unconditional in $\mathbf{u}$ and the concept variable never enters the generator $g(\mathbf{z}, \mathbf{u})$ in the strict sense of Assumption 2. We therefore inject $\mathbf{u}$ as an auxiliary input to the VAE decoder so that it becomes $g(\mathbf{z}, \mathbf{u})$ with $\mathbf{z}$ the diffusion latent and $\mathbf{u}$ the concept token sequence; concretely, we concatenate the per-token features to the spatial feature map at each upsampling stage of the decoder, mirroring the cross-attention pattern in the U-Net. Second, we add a *mechanism critic* $c_\psi(\hat{\mathbf{x}}, \mathbf{z}, \mathbf{u})$ scoring reconstruction–latent–concept tuples, together with two MLP projection heads Π_g, Π_u ($\to$ 128-dim L_2-normalized embeddings) that map the critic gradient $\nabla_{\mathbf{z}} c_\psi$ and the concept variable into a shared embedding space, supporting the InfoNCE objective. The two additions together add roughly 265K parameters on top of the shared backbone.

MCL Objective. We compared five MCL variants: NCE, InfoNCE, Fisher score matching, denoising score matching, and Jacobian-based InfoNCE. InfoNCE consistently achieved the highest DCI disentanglement and FactorVAE scores across all three datasets. The other objectives either suffer from training instability (Fisher, Jacobian) or weaker gradient signal when the critic operates on high-dimensional image reconstructions (NCE, denoising SM). We therefore adopt InfoNCE as our final objective:

$$\mathcal{L}_{\mathrm{MCL}} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp\big(\Pi_g(\mathbf{g}_i)^\top \Pi_u(\mathbf{u}_i)/\tau\big)}{\sum_{j=1}^{B} \exp(\Pi_g(\mathbf{g}_i)^\top \Pi_u(\mathbf{u}_j)/\tau)}, \tag{18}$$

where $\mathbf{g}_i = \nabla_{\mathbf{z}_i} c_\psi(\hat{\mathbf{x}}_i, \mathbf{z}_i, \mathbf{u}_i)$ is the critic gradient with respect to the latent, $\tau{=}0.1$ is the temperature, and negatives are drawn from other samples in the same mini-batch via latent shuffling. The total training loss is

$$\mathcal{L} = \mathcal{L}_{\mathrm{diffusion}} + \lambda_{\mathrm{recon}}\, \mathcal{L}_{\mathrm{recon}} + \lambda_{\mathrm{MCL}}\, \mathcal{L}_{\mathrm{MCL}}, \tag{19}$$

where $\mathcal{L}_{\mathrm{recon}}$ is the pixel-level VQ-VAE reconstruction loss weighted by $\lambda_{\mathrm{recon}}{=}0.1$.

Training. Training proceeds in two stages on a single NVIDIA H100 GPU. *Stage 1* pretrains MDDiff as a conditional-generation model with the diffusion loss only for 10 epochs (AdamW, base learning rate 2×10^{-6}, 10,000-step linear warmup). *Stage 2* fine-tunes the pretrained MDDiff checkpoint with MCL, optimizing all components jointly under the full objective $\mathcal{L}$ for an additional 10 epochs (base learning rate 2×10^{-7}, 1,000-step warmup). Both stages use batch size 128 and T=1000 diffusion timesteps with a linear noise schedule ($\beta_{\min}$=0.0015, $\beta_{\max}$=0.0155). Stage 2 fine-tuning takes approximately 3 hours on Cars3D, 4 hours on Shapes3D, and 7 hours on MPI3D. We set λ_{MCL}=10^{-3} for Shapes3D and Cars3D, and λ_{MCL}=10^{-4} for MPI3D.

D.5. Action-Conditioned Generative World Model

Pipeline. Wan-ACG and Wan-LACG are built on a large pretrained video generative model [57], equipped with four components that together realize the forward action-conditioned generative model $\mathbf{x}_{t+1} = g_{\mathrm{fwd}}(\mathbf{z}_t, \mathbf{u}_t)$ on which MCL operates: (i) a frozen causal VAE encoder E_{vae} that maps a T-frame video window to a latent video sequence $v_{1:F} = E_{\mathrm{vae}}(\mathbf{x}_{1:T})$; (ii) an inverse dynamics model (IDM) I_ϕ that produces, for each frame transition, a diagonal-Gaussian posterior $q_\phi(z_t \mid \mathbf{x}_t, \mathbf{x}_{t+1}) = \mathcal{N}(\mu_t, \mathrm{diag}(\sigma_t^2))$ over a d_z-dimensional latent action; (iii) an action-interface module A_ψ that lifts the frame-rate latent action sequence $z_{1:T}$ to FDM-aligned conditioning signals $c_{1:F} = A_\psi(z_{1:T})$, injected into the FDM through AdaLN modulation; and (iv) a forward dynamics model (FDM) g_θ, initialized from the pretrained video-planner backbone, that performs flow matching $\hat{u}_\tau = g_\theta(\tilde{v}_\tau, \tau, z_{1:T})$ to predict future latent video dynamics conditioned on the latent action.

Mapping to the GMD setup. In our notation, the conditioning signal is the latent action, $\mathbf{u}_t := z_t$ (or, when an external command interface is required, a controller-predicted latent action $\hat{z}_t$), and the observation is the next latent video frame, $\mathbf{x}_{t+1} := v_{f+1}$. The action-conditioned FDM is the GMD generator $g(\mathbf{z}_t, \mathbf{u}_t)$: action-driven conditioning enters through AdaLN; the diffusion latent enters through $\tilde{v}_\tau$.

Our addition: MCL on the action conditioning. We attach the MCL objective from Section 4 to the action-conditioning channel of the FDM. A mechanism critic $c_\psi(\hat{\mathbf{x}}_{t+1}, z_t, \mathbf{u}_t)$ scores how strongly the realized forward mechanism depends on z_t, the mechanism gradient $\mathbf{g}_t = \nabla_{z_t} c_\psi(\hat{\mathbf{x}}_{t+1}, z_t, \mathbf{u}_t)$ is contrasted via InfoNCE against the projected latent-action embedding $\Pi_u(\mathbf{u}_t)$, and the resulting $\mathcal{L}_{\mathrm{MCL}}$ is added to the host objective with weight λ_{MCL}:

$$\mathcal{L}_{\mathrm{total}} \;=\; L_{\mathrm{rec}} + \beta L_{\mathrm{KL}} + \lambda_{\mathrm{adv}} L_{\mathrm{GRL}} + \lambda_{\mathrm{MCL}}\, \mathcal{L}_{\mathrm{MCL}},$$

where L_{rec} is the flow-matching reconstruction loss, L_{KL} is a KL regulariser on the IDM posterior, and L_{GRL} is a gradient-reversal embodiment-invariance term used to encourage the latent action to ignore embodiment-specific nuisance. The IDM, action interface, FDM, and MCL projection heads are trained jointly; the causal VAE remains frozen.

What MCL adds, and what we test. Without MCL, the action-conditioning channel is shaped only *passively* by the reconstruction loss; nothing forces distinct latent actions to induce distinct mechanisms in the FDM. MCL closes that gap: it makes mechanism diversity along $\mathbf{u}_t$ an *active* target, contrasting how the FDM's mechanism gradient changes when $\mathbf{u}_t$ varies across the in-batch action samples. The structural claim we will verify, against an MCL-free baseline and against a non-action-conditioned forward model, is that MCL on the action conditioning improves both visual prediction quality and action-following accuracy by enforcing sufficient GMD on the action variable.

E. PyTorch Implementation of MCL

For reproducibility we provide a self-contained PyTorch implementation of the Mechanistic Contrastive Learning loss. We first present the *primary* loss (Algorithm 1) that captures the core mechanism-gradient → projection → InfoNCE pipeline used throughout the paper, then give the *full* implementation (Algorithm 2) including the projection-head module `MLPProj` and the mechanism-critic module `MechanismCritic` that we use in the image and world-model experiments. The same loss applies to every host model in our experiments (MD-VAE, MDDiff, ACG-WM); only the host components change. The loss adds to the host's reconstruction objective with weight λ_{MCL} as $\mathcal{L} = \mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{MCL}}\, \mathcal{L}_{\mathrm{MCL}}$.

E.1. Primary MCL Loss

E.2. Full Code

The full implementation we use in practice adds the two-layer projection head `MLPProj`, the default mechanism critic `MechanismCritic`, and the InfoNCE wrapper used in our image and world-model experiments.

Algorithm 1. Primary PyTorch implementation of the MCL loss. The host generator `decoder_G` produces $\hat{\mathbf{x}}$ from a batch of latents and conditioning signals; the critic scores how strongly the realized mechanism depends on $\mathbf{z}$; the autograd of the critic w.r.t. $\mathbf{z}$ yields the per-sample mechanism gradient, which is contrasted via InfoNCE against the projected conditioning embedding.

```
def mcl_loss(decoder_G, z, u, *,
             critic, Pi_g, Pi_u, tau=0.1):
    """
    decoder_G : callable (z, u) -> x_hat
    z         : latent batch, shape (B, ...)
    u         : conditioning batch, shape (B, d_u)
    critic    : (x_hat, z, u) -> mechanism score per sample
    Pi_g, Pi_u: projection heads to shared embedding space
    """
    z = z.requires_grad_(True)
    x_hat = decoder_G(z, u)
    s = critic(x_hat, z, u)

    # mechanism gradient w.r.t. the latent
    g = torch.autograd.grad(s.sum(), z, create_graph=True)[0]

    # L2-normalized projections
    q = Pi_g(g.flatten(1));  q = q / q.norm(dim=1, keepdim=True)
    k = Pi_u(u);             k = k / k.norm(dim=1, keepdim=True)

    # B x B InfoNCE: positives on the diagonal
    logits = q @ k.t() / tau
    labels = torch.arange(q.size(0), device=q.device)
    return F.cross_entropy(logits, labels)
```

F. Additional Experimental Results

F.1. Effect of Mechanism Diversity

A central claim of Theorem 1 is that identifiability strengthens monotonically with the size of the conditioning alphabet $|\mathcal{U}|$: more distinct mechanisms produce more linearly-independent constraints on the generator Jacobian, eventually saturating the sufficient GMD condition (Assumption 2, which requires $|\mathcal{U}| \geq 2d_z + 1$). Table 5 reports MD-VAE's median MCC across $|\mathcal{U}| \in \{2, 4, 8, 16, 32\}$ at $d_z \in \{5, 10, 20\}$, fixing all other settings of Appendix D.3.

Table 5. **Effect of conditioning-set size $|\mathcal{U}|$ on MD-VAE's median MCC (↑, 30 seeds).**

	$\lvert\mathcal{U}\rvert = 2$	$\lvert\mathcal{U}\rvert = 4$	$\lvert\mathcal{U}\rvert = 8$	$\lvert\mathcal{U}\rvert = 16$	$\lvert\mathcal{U}\rvert = 32$
$d_z = 5$	0.51	0.79	0.96	0.97	0.97
$d_z = 10$	0.43	0.62	0.84	0.93	0.94
$d_z = 20$	0.39	0.51	0.71	0.86	0.92

The rows confirm two predictions of the theory: (i) for fixed d_z, MCC rises with $|\mathcal{U}|$, mirroring the increase in mechanism diversity; (ii) higher-dimensional latents require more conditioning values before performance saturates. The threshold $|\mathcal{U}| \geq 2d_z + 1$ in Assumption 2 is sufficient rather than necessary, so low-noise simulations can begin to saturate before the conservative bound is reached. The qualitative behaviour still matches the GMD identifiability principle: *contrasting more mechanisms produces more identifiable representations.*

F.2. Generation Quality of MDDiff

A representation is only useful if the generator that produces it remains a competent image model. Table 6 reports FID (sample quality) and LPIPS (reconstruction fidelity) for MDDiff and the strongest diffusion baseline EncDiff [61] on the three benchmarks of Section 5.2. MDDiff preserves generation quality while delivering the disentanglement gains reported in Table 1; on `Cars3D` it actually *improves* FID and LPIPS, and on `Shapes3D` the two methods are statistically tied. `MPI3D` shows a modest FID gap that we attribute to the smaller λ_{MCL} used for that dataset (Appendix D); LPIPS is unchanged.

Table 6. **Generation quality (FID ↓, LPIPS ↓).**

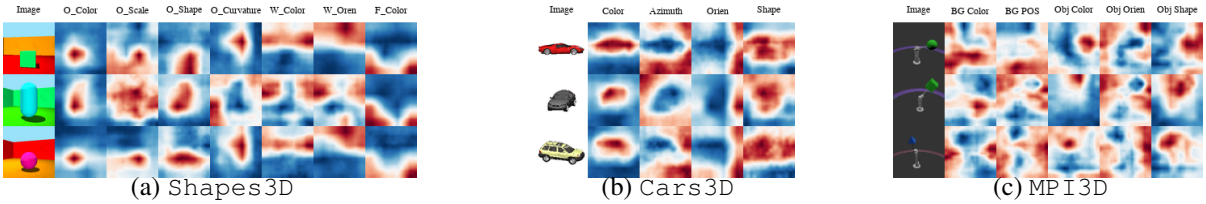
	Cars3D		Shapes3D		MPI3D	
Method	FID ↓	LPIPS ↓	FID ↓	LPIPS ↓	FID ↓	LPIPS ↓
EncDiff	62.31± 3.28	0.07± 0.03	5.33± 0.68	0.01± 0.01	4.06± 0.60	0.01± 0.01
MDDiff (Ours)	**39.32**± 6.67	**0.03**± 0.02	5.52± 0.78	0.01± 0.01	7.10± 2.12	0.01± 0.01

F.3. Latent Traversals

Latent traversals offer a complementary view of factoredness to the factor-swapping results in the main paper (Figure 6): holding all other dimensions fixed and sweeping one dimension across its range should produce smooth, single-factor change in the generated image. Figure 8 reports MDDiff traversals on all three benchmarks; each row corresponds to one latent dimension and isolates one semantic factor (wall colour, floor colour, object colour, shape, orientation, scale on `Shapes3D`; analogous factors on `Cars3D` and `MPI3D`).

F.4. Cross-attention Maps

Cross-attention maps complement latent traversals by exposing how mechanism diversity manifests *inside* the generator: each concept token (one per latent dimension) modulates the denoising U-Net through cross-attention, and each token's attention should localise to the spatial region of its factor. Figure 9 visualises the cross-attention weights between concept tokens (columns) and spatial locations across the three benchmarks; each token attends to the spatial region corresponding to its factor, evidence that MCL produces semantically meaningful, spatially-localised modulation in the generator.



(a) `Shapes3D` (b) `Cars3D` (c) `MPI3D`

Figure 9. **Cross-attention maps on all three benchmarks.** Columns show concept tokens; rows show spatial locations in the denoising network.

G. Extended Related Work

G.1. Why Existing Generative Models Can Be Factored

Theorem 1 explains why conditioning makes generative models learn factored representations, and when they fail. Two major families instantiate the GMD regime (Figure 3, bottom row). In *conditional image generation*, text-to-image models [41, 50, 52] and diffusion-based disentanglement methods [60, 61] let text or concept tokens reshape the denoising network at every layer; the latent code $\mathbf{z}$ encodes stochastic content while $\mathbf{u}$ controls semantics, creating mechanism diversity that factors the latent space. In *generative world models* [15, 16, 18, 69], actions $\mathbf{a}_t$ reshape the latent dynamics model $g(\mathbf{z}_t, \mathbf{a}_t)$; similarly, decision-as-generation methods [1, 7, 28] condition trajectory generation on returns or skills. Identifiability in world models is particularly important: a factored latent state means each dimension captures a physically meaningful quantity (object position, velocity, contact state), enabling compositional generalization to novel scenes, faithful physical reasoning, and interpretable planning. In both families, sufficient diversity in $\mathbf{u}$ provably factors the latent state by Theorem 1. However, none of these

models were designed with GMD in mind; whether Assumption 2 holds depends entirely on the data and architecture, with no explicit control. This motivates the active approach of Section 4.

G.2. Theoretical Foundations of LRD: Auxiliary-Variable Identifiability

The classical ***LRD*** identifiability result of Hyvarinen et al. [27] treats the case where an auxiliary variable $\mathbf{u}$ shifts the latent distribution while the generator g stays fixed. Translating their notation into ours ($\mathbf{z}$ for source, $\mathbf{u}$ for auxiliary, $\mathbf{x}$ for observation, g for mixing function), the setup is:

Setup. Observations are produced by an invertible mixing $\mathbf{x} = g(\mathbf{z})$, with $g \in \mathcal{C}^2$ and $\mathbf{z} \in \mathbb{R}^{d_z}$. The latent density is conditionally factorial given the auxiliary,

$$p(\mathbf{z} \mid \mathbf{u}) \;=\; \prod_{i=1}^{d_z} p_i(z_i \mid \mathbf{u}),$$

and each one-dimensional conditional belongs to an exponential family of order k:

$$\log p_i(z_i \mid \mathbf{u}) \;=\; \sum_{j=1}^{k} T_{i,j}(z_i)\, \lambda_{i,j}(\mathbf{u}) \;-\; \log Z_i(\mathbf{u}) \;+\; \log Q_i(z_i),$$

where $T_{i,j} : \mathbb{R} \to \mathbb{R}$ are the sufficient statistics, $\lambda_{i,j}(\mathbf{u}) \in \mathbb{R}$ the natural parameters, $Z_i(\mathbf{u})$ the normalizer, and Q_i a base measure. Stacking the natural parameters yields $\boldsymbol{\lambda}(\mathbf{u}) \in \mathbb{R}^{kd_z}$.

Sufficient diversity (Assumption of 27, Theorem 3). There exist $kd_z + 1$ values $\mathbf{u}^{(0)}, \mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(kd_z)}$ such that the matrix of contrasts

$$L \;=\; \big[\, \boldsymbol{\lambda}(\mathbf{u}^{(1)}) - \boldsymbol{\lambda}(\mathbf{u}^{(0)}),\; \ldots,\; \boldsymbol{\lambda}(\mathbf{u}^{(kd_z)}) - \boldsymbol{\lambda}(\mathbf{u}^{(0)}) \,\big] \;\in\; \mathbb{R}^{kd_z \times kd_z}$$

is invertible.

Result. Under the setup and the sufficient-diversity condition, the latent code $\mathbf{z}$ is identifiable from $\mathbf{x}$ up to a permutation and a componentwise invertible transformation [27, Theorem 3].

Why this is LRD. Identifiability hinges entirely on the auxiliary $\mathbf{u}$ ***shifting the latent distribution*** $p(\mathbf{z} \mid \mathbf{u})$ across enough values; the generator g is fixed and contributes no $\mathbf{u}$-dependent variability. This is exactly the LRD principle of Section 2. Our GMD condition (Theorem 1) replaces the diversity requirement on $p(\mathbf{z} \mid \mathbf{u})$ with an analogous condition on the generator Jacobian $J_g(\mathbf{z}, \mathbf{u})$, which is what makes our identifiability result strictly more expressive in the conditional generative regime (Proposition 1).

G.3. Theoretical Foundations of Regularization: Sparsity-Based Identifiability

The classical ***regularization-based*** identifiability result of Zheng et al. [68] obtains identifiability without any auxiliary variable, by imposing a structural sparsity constraint on the generator instead. We restate it in the notation of our paper.

Setup. Observations are produced by an invertible mixing $\mathbf{x} = g(\mathbf{z})$, $g \in \mathcal{C}^2$, with $\mathbf{z} \in \mathbb{R}^{d_z}$ drawn from a fixed prior $p(\mathbf{z})$ that does *not* depend on any auxiliary variable. Identifiability is recovered through structural assumptions on g and on $p(\mathbf{z})$.

Structural sparsity assumption (68, Definition 3). The Jacobian $J_g(\mathbf{z})$ of the mixing function has a structural sparsity pattern: there exists a binary support $M \in \{0,1\}^{d_x \times d_z}$ such that $\partial g_a / \partial z_i = 0$ whenever $M_{a,i} = 0$, for $p(\mathbf{z})$-a.e. $\mathbf{z}$. Each row of M is constrained to have at most a small number of non-zero entries, and the column supports are required to be *distinguishable* so that no two latent coordinates have the same set of dependent observation coordinates.

Non-degeneracy conditions (68, Theorem 3.5). The following two conditions ensure that latent coordinates are non-redundant: (i) the prior $p(\mathbf{z})$ is fully supported on $\mathbb{R}^{d_z}$ and admits a $\mathcal{C}^1$ density, and (ii) the second-order partial derivatives $\partial^2 g_a / \partial z_i \partial z_j$ satisfy a generic linear-independence condition across the support of $p(\mathbf{z})$, ruling out degenerate mixing functions.

Result. Under the setup, the structural sparsity assumption, and the non-degeneracy conditions, the latent code $\mathbf{z}$ is identifiable from $\mathbf{x}$ up to a permutation and a componentwise invertible transformation [68, Theorem 3.5].

Why this is regularization. Identifiability hinges on a structural ***constraint*** on the generator, in this case sparsity of J_g, rather than on auxiliary diversity. This is exactly the Regularization principle of Section 2, of which L_1/L_2 penalties on the generator Jacobian are the practical instantiation. Our GMD condition does not require sparsity of J_g; it instead asks that J_g vary sufficiently across $\mathbf{u}$, which is a complementary structural property and the basis of MCL.

G.4. Categorizing Methods by Principle

Table 7 categorizes representative methods according to which of the three principles for factored representations they employ (cf. Figure 2). A key distinction emerges: many conditional generative models *passively* operate in the GMD regime (the conditioning signal enters the generator architecturally), but none *actively* encourage sufficient mechanism diversity across conditioning signals. Our MCL is the first method that explicitly enforces the sufficient GMD condition (Assumption 2), turning passive architectural mechanism diversity into a controllable, provable guarantee.

LRD ∩ GMD. Methods in this intersection let $\mathbf{u}$ enter both the latent distribution and the generator. Conditional VAEs [53] have this structure by design: the prior $p(\mathbf{z} \mid \mathbf{u})$ and the decoder $p(\mathbf{x} \mid \mathbf{z}, \mathbf{u})$ both depend on $\mathbf{u}$. Classifier-guided diffusion [9] adds a classifier gradient to the sampling trajectory, shifting the effective latent distribution, on top of a class-conditional denoiser that reshapes the generator. DALL-E 2 [43] factorizes generation as $P(\mathbf{x} \mid \mathbf{y}) = P(\mathbf{x} \mid \mathbf{z}, \mathbf{y})\, P(\mathbf{z} \mid \mathbf{y})$: the text caption $\mathbf{y}$ enters both the diffusion prior over CLIP embeddings and the decoder. VQ-VAE-2 [44] combines all three principles: vector quantization, top-level codes conditioning the autoregressive prior over bottom-level codes, and top-level codes entering the feed-forward decoder.

Regularization ∩ GMD. Methods here apply explicit or implicit regularization to a conditional generator. EncDiff [61] conditions a latent diffusion model on concept tokens via cross-attention, with the diffusion information bottleneck serving as implicit regularization. DisDiff [60] decomposes the score into per-factor sub-gradient fields, a structural sparsity constraint, and conditions each field on a discovered factor. The Hessian Penalty [42] regularizes the off-diagonal Hessian of a GAN generator with respect to its latent input, encouraging axis-aligned disentanglement. VQ-VAE [54] combines vector quantization with a decoder conditioned on discrete codes. InfoDiffusion [59] adds a mutual-information maximization penalty to a latent-conditioned diffusion model to prevent posterior collapse.

LRD ∩ Regularization. Methods here combine auxiliary-variable-driven distributional shifts with structural constraints. Lachapelle et al. [33] use temporal or interventional auxiliary variables together with binary-mask sparsity on the causal graph; neither alone suffices for identifiability. SlowVAE [31] uses temporal context as the auxiliary variable and a Laplace sparse prior on temporal differences. CITRIS [36] observes intervention targets and constrains the transition model to respect a factored causal structure. TDRL [63] leverages distribution shifts across environments with a modular, sparse mechanism-shift assumption.

H. Recommendations for Future Work

This work introduces ***mechanism diversity*** as a sufficient condition for factored generative models, and operationalises it through Mechanistic Contrastive Learning (MCL). We see three open directions where the same principle has the potential to extend beyond the identifiability benchmarks studied here.

H.1. Beyond Factored Generative Models

Motivation. The notion of *factoredness* in Section 3.2, inherited from nonlinear ICA, is a stricter target than what most modern conditional generators are built or evaluated against. Large text-to-image diffusion [41, 50, 52], controllable video diffusion [3], visual world models [16, 69], and multi-modal generators routinely operate without ground-truth factors yet still benefit from compositional, controllable behaviour at the conditioning interface.

Table 7. **Categorizing methods by principle.** • = *active* (an explicit loss term encourages the principle); ○ = *passive* (satisfied by model architecture/design only, no dedicated loss).

Region	Method	LRD	Reg.	GMD	Mechanism
LRD only	iVAE [29]	○			Conditional prior $p(\mathbf{z}\|\mathbf{u})$ (architectural)
	TCL [25]	•			Contrastive classification loss over time segments
	CPC [40]	•			InfoNCE contrastive loss over temporal windows
Reg. only	β-VAE [21]		•		Upweighted KL penalty ($\beta > 1$)
	FactorVAE [30]		•		TC penalty via learned discriminator
	Sparse ICA [68]		•		L_1 sparsity on generator Jacobian
	Tripod [22]		•		Quantization + MI + Hessian penalty losses
Active GMD	MDDiff / MCL (Ours)			•	Contrastive loss on generator Jacobian gradients
	MD-VAE (Ours)			•	Conditional VAE with mechanism-diverse decoder
LRD ∩ GMD	CVAE [53]	○		○	Prior $p(\mathbf{z}\|\mathbf{u})$ + decoder $p(\mathbf{x}\|\mathbf{z},\mathbf{u})$ (both architectural)
	Classifier Guidance [9]	•		○	Classifier loss shifts sampling; denoiser cond. on y
	DALL-E 2 [43]	○		○	Diffusion prior $P(\mathbf{z}\|\mathbf{y})$ + decoder $P(\mathbf{x}\|\mathbf{z},\mathbf{y})$
Reg. ∩ GMD	EncDiff [61]		○	○	Architectural bottleneck + cross-attn conditioning
	DisDiff [60]		•	○	Disentangling loss + per-factor sub-gradient fields
	Hessian Penalty [42]		•	○	Hessian reg. loss + class-conditional GAN
	VQ-VAE [54]		•	○	VQ + commitment losses + code-conditioned decoder
	InfoDiffusion [59]		•	○	MI maximization loss + latent-conditioned diffusion
All three	VQ-VAE-2 [44]	○	•	○	VQ loss (active reg.) + cond. prior + cond. decoder
LRD ∩ Reg.	Mech. Sparsity [33]	○	•		Cond. prior (passive) + L_1 mask sparsity (active)
	SlowVAE [31]	○	•		Temporal prior (passive) + Laplace sparse penalty
	CITRIS [36]	○	•		Intervention prior (passive) + factored assignment loss
	TDRL [63]	○	•		Shift prior (passive) + modular mechanism loss

Conjecture. The mechanism-diversity principle should still apply in this regime: whenever $\mathbf{u}$ ***varies sufficiently*** across instances, the per-condition Jacobian family is structurally distinguishable, which is precisely what MCL contrasts. We expect an MCL-style objective added to a controllable generator to improve compositional generalization, prompt following, and conditioning-slot reusability, even without an identifiability proof, in the same way that mechanism contrast already improves

disentanglement scores at the benchmark scale (Table 1).

Open question. ***What should factoredness mean for a billion-parameter text-to-image model***, and which theoretical guarantees survive the absence of canonical ground-truth factors?

H.2. MCL as Fine-tuning of Large Pretrained Models

Motivation. Pretraining of large generative models is necessarily ***passive***: the conditioning signal is whatever caption, label, or action happens to accompany each datapoint at web scale, and no explicit objective enforces mechanism contrast. Fine-tuning is where domain-specific expert knowledge enters the model, through methods such as LoRA [23], ControlNet [67], and DreamBooth [51]. These methods add new conditioning channels but do not certify that distinct conditioning values produce distinct mechanisms.

Position. MCL is the natural ***active*** counterpart: a contrastive signal that explicitly enforces mechanism shift across conditioning values. This paper already demonstrates the recipe at the disentanglement scale: we first pretrain MDDiff as a $\mathbf{u}$-conditioned generator on disentangled concepts, and then apply MCL as a fine-tuning objective on top of that pretrained MDDiff checkpoint to drive both DCI and FactorVAE upward (Figure 5, Section 5.2). Scaling to frontier diffusion and world-model backbones should marry the ***data diversity*** captured by pretraining with the ***mechanism diversity*** needed for compositional control, yielding a sample-efficient route to factored controllable generators without retraining.

Open questions. Which layers of a U-Net or DiT should host the projection heads? How should MCL be balanced against the original denoising loss? Is parameter-efficient fine-tuning of the projection-head subspace alone sufficient?

H.3. Multi-objective, Multi-auxiliary Mechanism Contrast

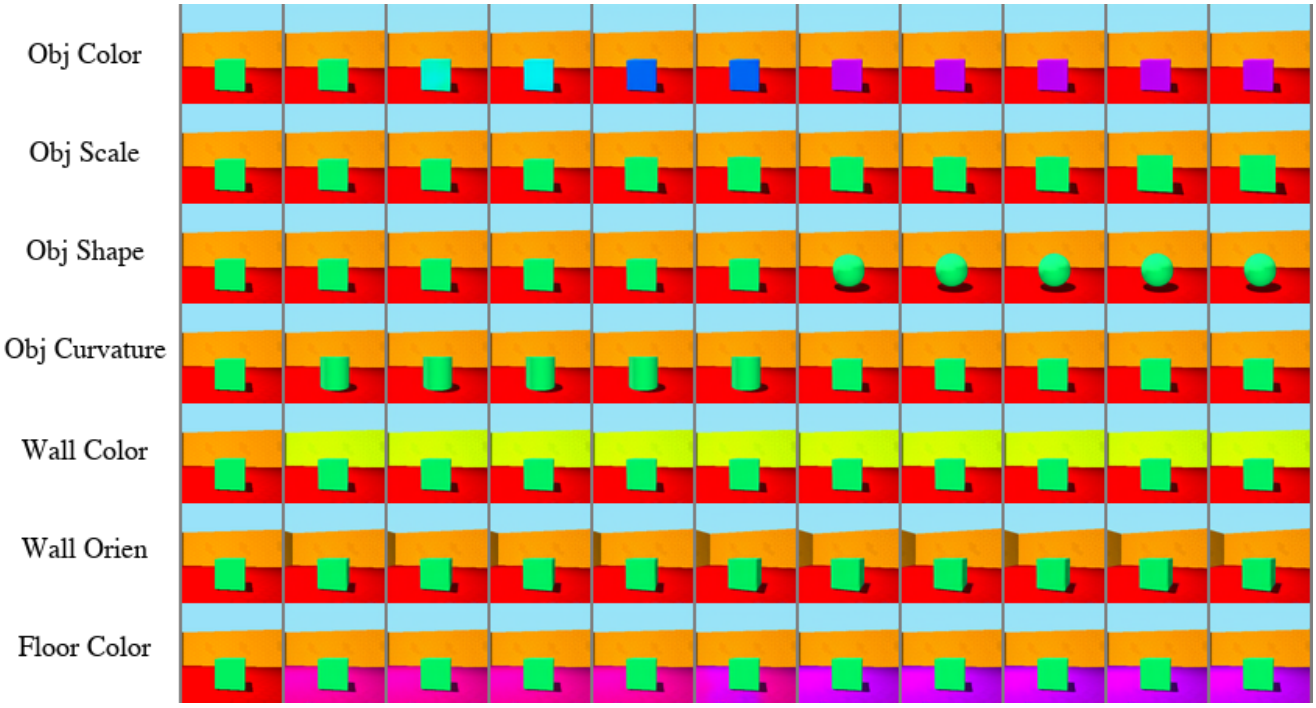
Motivation. Realistic generative pipelines pair ***several auxiliary variables*** (text caption, robot action, audio cue, regime index) with ***several generative objectives*** (image, video, control trajectory, language). Examples include video-action world models [69], manipulation diffusion policies [7], and decision-as-generation pipelines [1, 28]. MCL as defined in Section 4 contrasts one mechanism gradient against one conditioning embedding, which is well-matched to single-auxiliary, single-objective settings but leaves the multi-axis setting open.

Open questions. Should each auxiliary $\mathbf{u}^{(k)}$ carry its own projection head and InfoNCE term, or should the contrastive signal be ***jointly normalized*** across auxiliaries? Does the sufficient-GMD condition decompose objective-by-objective, or does identifiability require simultaneous diversity across all $\mathbf{u}^{(k)}$? How does the mechanism-shift principle behave when auxiliaries are ***causally entangled***, e.g. an action partially explains the next observation while a text caption only describes high-level semantics?

CVPR
#35

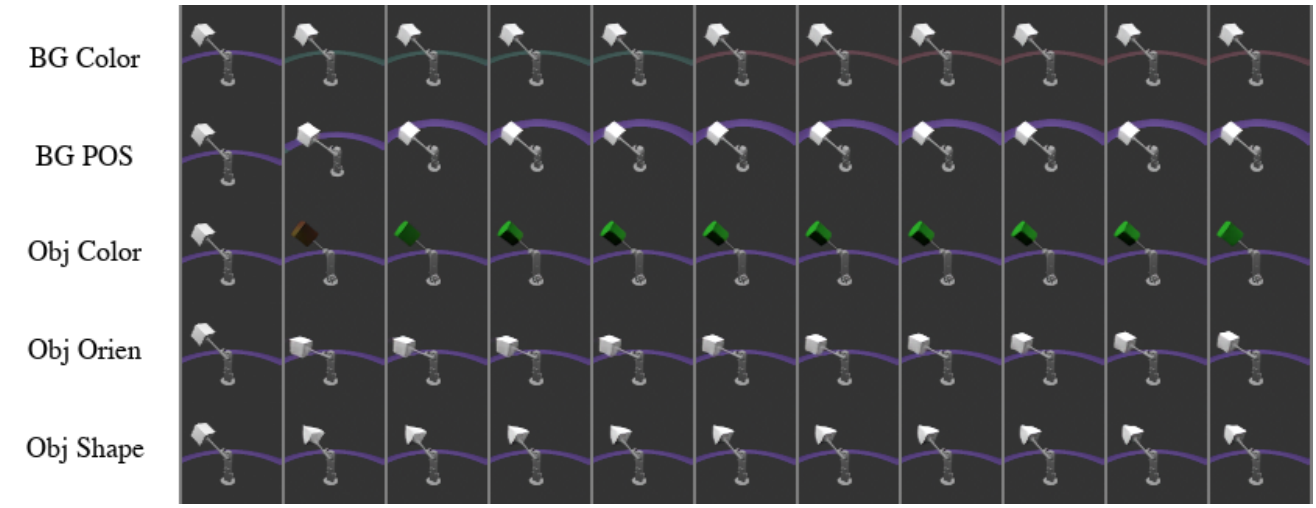
Algorithm 2. Full PyTorch implementation of MCL: the projection-head module, default mechanism-critic module, and the InfoNCE mechanism-gradient loss with optional separate keys.

```
import torch
import torch.nn as nn
import torch.nn.functional as F

def l2norm(x: torch.Tensor, eps: float = 1e-8) -> torch.Tensor:
    return x / (x.norm(dim=1, keepdim=True) + eps)

def info_nce_from_qk(q, k, tau: float = 0.1) -> torch.Tensor:
    logits = (q @ k.t()) / (tau + 1e-12)
    labels = torch.arange(q.size(0), device=q.device)
    return F.cross_entropy(logits, labels)

class MLPProj(nn.Module):
    def __init__(self, in_dim, out_dim=128, layernorm=False):
        super().__init__()
        layers = []
        if layernorm:
            layers.append(nn.LayerNorm(in_dim))
        layers += [nn.Linear(in_dim, out_dim),
                   nn.ReLU(inplace=True),
                   nn.Linear(out_dim, out_dim)]
        self.net = nn.Sequential(*layers)

    def forward(self, x):
        return self.net(x)

class MechanismCritic(nn.Module):
    def __init__(self, z_shape=(3, 16, 16), u_dim=20, hidden=256):
        super().__init__()
        zc, zh, zw = z_shape
        self.img = nn.Sequential(
            nn.Conv2d(3, 64, 4, 2, 1), nn.ReLU(True),
            nn.Conv2d(64, 128, 4, 2, 1), nn.ReLU(True),
            nn.AdaptiveAvgPool2d(1),
        )
        self.z_fc = nn.Linear(zc * zh * zw, hidden)
        self.u_fc = nn.Linear(u_dim, hidden)
        self.out  = nn.Sequential(nn.ReLU(True), nn.Linear(hidden, 1))

    def forward(self, x_hat, z, u):
        img_feat = self.img(x_hat).flatten(1)
        z_feat = self.z_fc(z.flatten(1))
        u_feat = self.u_fc(u)
        if img_feat.size(1) < z_feat.size(1):
            img_feat = F.pad(img_feat,
                             (0, z_feat.size(1) - img_feat.size(1)))
        else:
            img_feat = img_feat[:, :z_feat.size(1)]
        return self.out(z_feat + u_feat + img_feat).squeeze(1)

def mcl_infonce_mechgrad_loss(decoder_G, z, u_key, *,
                              u_for_G=None,
                              critic, Pi_g, Pi_u,
                              tau: float = 0.1,
                              create_graph: bool = False):
    """InfoNCE mechanism-gradient loss with optional separate keys.

    u_key   : auxiliary used as the contrastive key
    u_for_G : conditioning fed to decoder/critic (defaults to u_key)
    """
    u = u_key
    uG = u_for_G if u_for_G is not None else u
    z = z.requires_grad_(True)
    x_hat = decoder_G(z, uG)
    s = critic(x_hat, z, uG)
    g = torch.autograd.grad(s.sum(), z, create_graph=create_graph)[0]

    q = l2norm(Pi_g(g.flatten(1)))
    k = l2norm(Pi_u(u))
    return info_nce_from_qk(q, k, tau=tau)
```



(a) Shapes3D

(b) Cars3D



(c) MPI3D

Figure 8. **Latent traversals on all three benchmarks.** Each row varies a single MDDiff latent dimension while keeping the others fixed; columns sweep the dimension across its range.