

Causal AI for Transferable, Interpretable, and Controllable Machine Learning

Lingjing Kong

CMU-CS-26-127

August 2026

Computer Science Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

Thesis Committee:

Kun Zhang, Co-chair
Yuejie Chi, Co-chair (Yale University)
Eric Xing, Co-chair
Tom Mitchell
Kevin Murphy (Google DeepMind)

*Thesis submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science*

Copyright © 2026 Lingjing Kong

This research was supported in part by the National Science Foundation under grants CCF-1901199 and DMS-2134080, and by the Office of Naval Research under grant N00014-19-1-2404. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. government or any sponsoring organization.

Keywords: causal representation learning, causality, identifiability, compositional generalization, reasoning, trustworthy machine learning

Abstract

Traditional machine learning is largely organized around passive prediction: given observations from a training distribution, a model learns to predict labels, tokens, or future observations. Modern large models have changed the stakes of this paradigm. They now read, generate, reason, and assist decisions, yet they can remain brittle, opaque, difficult to control, and unreliable outside their training conditions. This thesis argues that trustworthy AI requires more than larger predictors. It requires representations that expose reusable mechanisms, preserve them across environments, and reveal when new evidence is needed.

The work is organized around four data regimes. In learning from multiple distributions, domain variation separates invariant and changing latent factors, enabling transfer without target-domain labels. In learning with unstructured distributions, masked prediction turns internal dependence into a signal for recovering hierarchical concepts. In generalization beyond training distributions, a novel prediction or generation is justified only when it can be decomposed into mechanisms already supported by training data. In learning through feedback and interaction, reward, multitask experience, and model verification identify reusable reasoning modules, compose dynamics across tasks, and direct new data collection toward weak parts of a world model.

Together, these regimes support the thesis that causal principles such as minimality, hierarchy, modularity, and structured feedback can turn representation learning from black-box fitting into mechanism discovery and support expansion. The dissertation develops this claim as a self-contained argument: each technical chapter states the available data signal, formalizes the latent structure, derives what can be recovered or generalized, and tests the resulting mechanism. The final chapter distinguishes the demonstrated results from the larger goal of trustworthy AI systems whose transfer, interpretation, control, and self-improvement can be diagnosed rather than merely hoped for.

Table of Contents

1	Introduction	1
1.1	From Passive Prediction to Mechanism-Aware AI	1
1.2	Large Models Raise the Stakes	2
1.3	Causal Representation Learning	3
1.4	Thesis Statement	4
1.5	Four Learning Regimes	5
1.6	Organization of the Thesis	7
2	Regime A: Learning from Multiple Distributions	9
2.1	Distributional Variation as Mechanism Signal	9
2.2	Partial Identifiability for Transfer	10
2.3	Causal Model and Minimal Change	10
2.4	Identifiability Assumptions	11
2.5	Identifiability Results	13
2.6	The iMSDA Method	14
2.7	Synthetic Experiments	15
2.8	Domain Adaptation Experiments	15
2.9	From Transfer to Mechanism Discovery	17
2.10	Limitations and Open Questions	19
3	Regime B: Learning with Unstructured Distributions	20
3.1	Learning Structure Without Explicit Labels	20
3.2	Masked Prediction in Hierarchical Latent Models	20
3.3	Hierarchical Latent Model for Images	21
3.4	Masks as Latent-Information Probes	22
3.5	Identifying Shared Latent Information	23
3.6	Masking Ratio and Patch Size	24
3.7	Empirical Evidence	24
3.8	Implications for Self-Supervised Pretraining	27
3.9	From Masked Prediction to Concept Learning	27
3.10	Limitations and Open Questions	28
4	Regime C: Generalization Beyond Training Distributions	30
4.1	Calibrated Generalization Outside Support	30
4.2	Compositional Generation	30
4.3	Compositional Generalization and Analogy	31
4.4	Hierarchical Concept Model	31

4.5	The Module-Support Condition	33
4.6	Identifying the Modules	33
4.7	Hierarchical Conditioning for Diffusion Models	34
4.8	Experimental Evidence	35
4.9	Extrapolation Beyond Source Support	37
4.10	From Composition to Interaction	37
4.11	Limitations and Open Questions	38
5	Regime D: Learning Through Feedback and Interaction	40
5.1	From Static Support to Interactive Discovery	40
5.2	Reward Feedback for Reusable Reasoning Modules	41
5.3	Empirical Evidence	45
5.4	Limitations of the Anchor Study	49
5.5	Supporting Extensions: Interactive World Models	50
5.6	From Reuse to Support Expansion	52
6	Self-Criticism, Reflection, and Future Directions	55
6.1	Freeing and Empowering Humans	55
6.2	Gaps Between Current Models and Human Empowerment	56
6.3	Assumptions and Diagnostics	58
6.4	Foundational Contributions	59
6.5	Causal Foundations and the Role of Academic Research	61
6.6	Diagnostics for Trustworthy Extension	65
6.7	Future Research: A Causal Loop for AI-Assisted Scientific Discovery	65
6.7.1	Research Thrust I: From Multimodal Data to Causal Scientific Insights	66
6.7.2	Research Thrust II: From Causal Insights to Informative Experiments	66
6.7.3	Research Thrust III: From Experiments to Targeted Scientific Data	67
6.8	Conclusion	68
	Bibliography	69
	Appendix A Formal Results and Proofs	78
	A.1 Regime A: Partial Identifiability for Transfer	78
	A.2 Regime B: Shared Information in Masked Prediction	81
	A.3 Regime C: Composition and Causal Module Identification	83
	Appendix B Additional Work Connected to the Four Regimes	86
	Appendix C Formal and Supplemental Results for Learning Through Feedback and Interaction	87
	C.1 Reasoning Module Identification and Local Witnesses	87
	C.2 Supplemental Evidence for Reward-Guided Reasoning	89
	C.3 Sparse Support in Compositional World Models	95
	C.4 Supplemental CWM Results	96
	C.5 Forward-Inverse Asymmetry and Verified Support Expansion	99
	C.6 Supplemental WAV Results	101

Chapter 1

Introduction

This chapter sets up the dissertation argument, moving from the limits of passive prediction to a four-regime causal representation learning program that identifies mechanisms, composes them, and expands their support through feedback.

1.1 From Passive Prediction to Mechanism-Aware AI

Traditional machine learning is largely organized around passive prediction. Given observations from a training distribution, the learner estimates a function that predicts labels, future observations, rewards, tokens, or other observable targets. This paradigm has been extraordinarily productive. It supports supervised classification, regression, ranking, sequence modeling, and many parts of modern representation learning. Yet the prediction view also deliberately brackets a harder question: what has the model learned about the mechanism that produced the data?

The distinction matters because a predictor can be accurate without being mechanism-aware. A classifier trained on photographs may exploit texture rather than object shape. A medical model may rely on hospital-specific measurement protocols rather than physiological state. A language model may output a convincing explanation without using that explanation to produce the answer. In each case, the model fits an observable relation, but the learned relation may not correspond to the latent causes, concepts, or modules that a human would expect to transfer. The result is a gap between performance on familiar samples and trust under new conditions.

This thesis takes that gap as its starting point. The goal is not to replace prediction. Prediction remains the observable evidence by which models are trained and evaluated. The goal is to place prediction inside a richer question: when can observed data reveal a reusable latent mechanism, and how should a learning algorithm exploit that mechanism? Causal representation learning offers a language for this question because it separates observations from the latent variables and structural relations that generate them. A causal representation

is valuable not merely because it compresses data, but because it identifies the pieces of the data-generating process that can be reused, inspected, and intervened on.

1.2 Large Models Raise the Stakes

The need for mechanism-aware learning has become more urgent with modern large models. Large language models and multimodal generative models no longer serve only as passive classifiers. They write, summarize, generate images, answer questions, call tools, reason through tasks, assist scientific workflows, and increasingly mediate decisions made by people. Their behavioral interface is active: a user can ask for a plan, an explanation, a diagnosis, a program, a composition of objects, or a chain of reasoning. These systems therefore create new demands on trust. We want them not only to produce plausible outputs, but to use the right latent structure when conditions change.

The community has documented several recurring failures. Large models can imitate falsehoods and hallucinate facts [1], produce toxic continuations from ordinary prompts [2], memorize and leak rare training examples [3], fail under transferable adversarial prompts [4], provide chain-of-thought explanations that are not faithful to the actual decision process [5], and become distracted by irrelevant context [6]. These problems are not identical, but they point to a shared limitation. Scaling improves the capacity to fit and generate, while leaving unresolved which mechanisms the model has internalized, which mechanisms are stable across contexts, and which interventions on the model will have predictable effects.

This motivates a sharper notion of trustworthy AI. A trustworthy model should transfer knowledge across domains without confusing nuisance variation for semantic content. It should expose representations that correspond to meaningful latent concepts rather than opaque statistical shortcuts. It should allow controlled interventions, such as changing sentiment while preserving topic, editing object attributes while preserving relations, or composing known skills in a new reasoning problem. It should also be calibrated about its own extrapolation: when a new request can be built from known modules, the model should exploit that structure; when the request requires a genuinely new mechanism, the model should not silently pretend that interpolation is enough.

Trustworthiness as Mechanism Recovery

Here, trustworthiness is organized around three requirements. Transferability asks whether a learned mechanism remains useful across distributions. Interpretability asks whether the learned variables or modules correspond to meaningful concepts. Controllability asks whether interventions on those variables or modules produce localized, predictable changes. Causal representation learning connects these requirements by treating observations as evidence about a latent data-generating process.

1.3 Causal Representation Learning

Causal representation learning studies when latent variables behind observed data can be recovered from samples. The central difficulty is non-identifiability. Without additional structure, many latent-variable models can explain the same observed distribution. A rotated, mixed, or otherwise transformed latent space may reconstruct the data just as well as the true latent variables. This is why the word “causal” must be used carefully. A causal representation claim should state the assumptions that break the ambiguity, the equivalence class that remains, and the task for which that equivalence class is sufficient.

Four terms retain the same meaning throughout the thesis. Training support is the set of observations, descriptors, or latent configurations that have positive probability in the training data. Module support is more local: every module state and parent context required by a proposed composition has appeared in some compatible training configuration. Identification means that any alternative model matching the observed distribution must recover the stated latent object up to the explicitly allowed relabeling or invertible transformation. Feedback is a signal observed after a prediction, trace, or action that can change which local events enter the learner’s data. These definitions distinguish supported recombination from unconstrained behavior outside all observed module contexts.

This thesis uses three structural principles: minimality, hierarchy, and modularity. Minimality says that a change in the data should be explained by the smallest latent mechanism consistent with the evidence. If domain shift changes only style, a good model should not let the entire representation move. Hierarchy says that high-dimensional observations are generated through multiple levels of abstraction. In images, pixels depend on parts, parts depend on objects, and objects depend on scene-level factors. In reasoning, tokens may be generated from local steps, local steps from skills, and skills from a high-level problem structure. Modularity says that mechanisms can be reused across contexts. A beak interacting with rice can be learned in one bird image and reused in another; a reasoning subroutine can be discovered in one problem and recombined in a new one. These principles are classical

in causal reasoning [7, 8], but the thesis develops them for high-dimensional representation learning.

The thesis also adopts a methodological standard. Each technical chapter begins by identifying a data signal. In some settings, the signal is explicit variation across domains. In others, it is the dependence revealed by masking. In still others, it is the absence of a requested composition from training support. In the final regime, feedback from rewards, task variation, or model verification creates new evidence about a module or failure. The chapter then states a structural principle that turns the signal into an identification, generalization, or support-expansion claim, translates that principle into an algorithmic design, and evaluates the design with experiments that probe the mechanism. This structure makes the dissertation self-contained. The papers provide provenance, but the chapters are written as parts of one thesis rather than as summaries of separate publications.

Each technical chapter develops one central study in full because that study carries the regime’s main problem, assumptions, method, measurements, evidence, and limitations. The central studies are partial domain adaptation in Regime A, masked autoencoding in Regime B, compositional image generation in Regime C, and reward-guided compositional reasoning in Regime D. Related publications appear as supporting extensions that test how far the same principle reaches; their specialized proofs and additional experiments are placed in the appendices when the chapter argument does not require them.

1.4 Thesis Statement

The thesis is: *causal principles can make high-dimensional representation learning more transferable, interpretable, and controllable by identifying reusable latent mechanisms from distributional variation and internal dependence, composing those mechanisms beyond observed support, and using feedback to verify and expand the mechanism library.*

The four regimes are not just four application areas. They are four ways in which data can provide information about latent mechanisms. Multiple distributions reveal what changes and what stays stable. Unstructured distributions reveal internal dependence when the learner creates missing information through a pretraining objective. Out-of-support requests reveal the limits of what can be justified by composition. Feedback and interaction reveal which modules are useful, which predictions are unreliable, and where new experience should be collected. Together, they form a sequence: learn mechanisms from external variation, learn mechanisms from internal structure, use learned mechanisms beyond observed support, and expand that support when static data is insufficient.

The dissertation contributes theory, algorithms, and empirical evidence in each regime. In Regime A, I study partially identifiable domain adaptation: domain variation separates invariant content from changing style, enabling transfer without target-domain labels [9]. In Regime B, I study masked autoencoders through hierarchical latent-variable models: masking selects the shared latent information required for reconstruction, explaining why particular masking regimes recover semantic representations [10]. In Regime C, I study compositional generalization: a novel composition is justified when it can be decomposed into latent modules whose parent supports have been observed in training [11]. In Regime D, I study how feedback turns static module support into an expanding mechanism library. The central study shows how reward-guided rollouts separate reusable reasoning skills and routing choices [12]. The supporting Compositional World Models (CWM) study uses sparse task routing and dynamics masks to compose world-model components across tasks [13]. The World Action Verifier (WAV) uses forward-inverse verification to direct interaction toward unreliable dynamics [14]. Other supporting work extends the four regimes to multi-domain generation [15], counterfactual text generation [16], biomedical multimodal analysis [17], modular vision-language alignment [18], discrete concept learning [19], concept-level controllable generation [20], white-box generative-model interpretation [21], and extrapolation from sparse latent shifts [22].

1.5 Four Learning Regimes

The four regimes below organize the body of the dissertation by the data signal available for mechanism recovery.

1.5.1 Regime A: Learning from Multiple Distributions

In the first regime, data arrive from several distributions: visual domains such as art, cartoons, photos, and sketches; hospital systems with distinct measurement protocols; or modalities such as image and text. The task is to use those distributional differences to learn a representation that transfers. The key danger is over-invariance. If all domain information is removed, the model may erase label-relevant changing factors. If all changing information is kept, the model may overfit to source-domain shortcuts. The causal principle is minimality: explain domain differences with a small changing block while preserving invariant content. This yields partial identifiability, with changing components recovered component-wise and invariant content recovered block-wise.

1.5.2 Regime B: Learning with Unstructured Distributions

In the second regime, the external domain signal is removed. Large-scale pretraining often begins with heterogeneous observations and no explicit concept labels. The task is to explain how useful structure can emerge anyway. Masked autoencoders provide a concrete setting. A model reconstructs masked patches from visible patches, but the learned representation supports semantic classification, detection, and segmentation. The causal principle is hierarchy: when local shortcuts are unavailable and enough global evidence remains, masked prediction forces the model to infer shared high-level latent variables. Masking ratio and patch size become knobs that choose which level of the hierarchy the objective makes recoverable.

1.5.3 Regime C: Generalization Beyond Training Distributions

In the third regime, the requested behavior lies outside the observed training support. A generative model may have seen peacocks and rice separately, but not a peacock eating rice. A predictor may see a target sample outside source support. A reasoning model may need to combine familiar skills in an unfamiliar order. The task is to determine when such behavior is justified. The causal principle is modularity. A new composition can be trusted when its required latent modules have been observed under some supported configuration, even if the high-level combination has not. The resulting goal is calibrated extrapolation, not unlimited extrapolation.

1.5.4 Regime D: Learning Through Feedback and Interaction

The fourth regime begins where the static support condition ends. If a needed module, interface, or action effect has not been constrained by data, a learner needs a way to acquire more informative evidence. The signal is feedback: task-level reward over generated reasoning traces, transition experience across many environments, or disagreement in a verification cycle. Feedback can expose which local reasoning choices are reusable, which entities and interactions govern a transition, and which action-conditioned predictions require new experience. The causal principle is interactive mechanism discovery: collect data that distinguishes a reusable mechanism or a model failure, then add that evidence to the supported library.

Regime Roadmap

Regime A uses explicit distributional variation as the signal. Minimal changes across domains identify changing factors and protect invariant content. The evidence comes from synthetic identifiability tests, PACS and Office-Home transfer, and ablations on loss terms and latent dimension.

Regime B uses missing-information prediction as the signal. Masking selects the latent variables shared between visible and hidden patches. The evidence comes from reconstruction metrics, attention analysis, linear probing, shape bias, robustness, and COCO transfer.

Regime C uses absent combinations as the signal. Hierarchical modules define which off-support requests are composable. The evidence comes from prompt-following benchmarks, ablations on time-dependent conditioning and sparse attention, qualitative analyses, and extrapolation theory.

Regime D uses feedback and interaction as the signal. Reward separates skills and routing choices inside reasoning traces, sparse routing composes world-model dynamics across tasks, and forward-inverse verification directs new interaction. The evidence comes from controlled reasoning tasks, more than 200 multitask control tasks, and nine world-model learning tasks.

1.6 Organization of the Thesis

Chapter 2 develops Regime A in detail. It begins with distributional variation as a signal for discovering mechanisms, then studies multi-source unsupervised domain adaptation. It introduces the latent content-style model, states the identifiability assumptions, presents the iMSDA algorithm, and interprets synthetic and real-world experiments as evidence for the minimal-change mechanism. The chapter then connects the same principle to multi-domain generation, counterfactual text generation, biomedical multimodal representation learning, and modular vision-language alignment.

Chapter 3 develops Regime B. It begins with the problem of recovering meaningful latent structure from unstructured observations, then studies masked image modeling. The chapter states the minimal shared latent set induced by a mask, proves why MAE can identify that shared information under invertibility and local invertibility assumptions, and connects the theory to the empirical behavior of masking ratio and patch size. It then situates the result within the broader concept-learning program, including discrete concepts, nonlinear hierarchical models, white-box generative models, SmartCLIP, and vision-language concepts.

Chapter 4 develops Regime C. It begins with the problem of calibrated behavior beyond training support, then studies compositional generation. It formalizes compositional generalization, states the module-support condition for a novel composition, proves component-wise identifiability of hierarchical modules under sufficient variation, and turns the theory into a diffusion-based text-to-image method. The chapter also connects the visual generation case to extrapolation, counterfactual control, compositional reasoning in diffusion language models, and the boundary between supported recombination and missing mechanisms.

Chapter 5 develops Regime D. It begins from the support boundary in Regime C, then presents reward-guided compositional reasoning as the central study. Language-model roll-outs separate atomic reasoning skills from information-routing choices, and controlled interventions test which supervised and rewarded trace supports make those modules reusable. CWM and WAV then serve as supporting extensions to environment interaction: sparse routing reuses world-model components across tasks, and forward-inverse verification directs new data collection. Detailed results for both extensions appear in Appendix C.

Chapter 6 is deliberately reflective. It frames the grand goal of AI as freeing and empowering humans, then asks why reliable embodied agents and causal scientific discovery remain beyond the reach of current models. It treats the papers in this thesis as a foundation rather than a completed solution, exposes the assumptions as practical contracts, and identifies future directions where interaction, communication, and causal structure can expand what models learn from scarce data.

Chapter 2

Regime A: Learning from Multiple Distributions

This chapter develops the first regime, where variation across observed distributions supplies the signal for separating invariant content from changing domain-specific factors.

2.1 Distributional Variation as Mechanism Signal

The first regime asks how multiple observed distributions can help recover latent mechanisms. A distribution shift is not only a nuisance. It is also an experiment performed by the world. When images move from photos to sketches, when clinical measurements move from one hospital to another, when a modality reveals only part of a biological system, or when a caption describes only part of an image, the observed change may expose which latent factors are stable and which factors are context-specific.

The standard instinct is to learn a domain-invariant representation. This is useful when all domain variation is nuisance variation, but it is not the right principle in general. Domain-changing factors may still be label-relevant. Removing all domain information can erase useful signal, while preserving all domain information can overfit to source-specific shortcuts. The thesis-level problem is therefore more precise: use distributional differences to separate what changes from what should transfer, and learn a prediction rule in the shared latent space where labels remain stable.

This regime treats multiple distributions as a weak form of intervention. Unlike randomized experiments, these distributions are observational and imperfectly controlled. Unlike a single static dataset, however, they provide contrasts. The research direction is to turn those contrasts into identification conditions, algorithms, and diagnostics: which part of the representation moves with the environment, which part remains usable across environments, and which changes can be used for control rather than merely removed.

2.2 Partial Identifiability for Transfer

The central technical setting is multi-source unsupervised domain adaptation. We observe labeled source domains $\{(\mathbf{x}_i, \mathbf{y}_i, \mathbf{u}_i) : \mathbf{u}_i \in \mathcal{U}_S\}$ and unlabeled target-domain examples $\{(\mathbf{x}_i, \mathbf{u}_i) : \mathbf{u}_i = \mathbf{u}^T\}$. The goal is to predict labels in the target domain without target labels. In visual benchmarks, the domains may be art, cartoon, photo, and sketch; in biomedical settings, they may be hospitals, assays, or measurement devices; in multimodal data, they may be complementary views of the same phenomenon.

The causal view treats domains as auxiliary variation. A domain index is not only a nuisance label to be hidden from a discriminator. It is also a source of statistical change that can reveal which latent variables move and which remain stable. This is the central idea in partial disentanglement for domain adaptation [9]. The chapter develops this idea as a self-contained mechanism: define a latent data-generating process, state when the process is identifiable, translate the assumptions into an algorithm, and test whether the learned representation supports target-domain transfer.

2.3 Causal Model and Minimal Change

Let $\mathbf{x} \in \mathcal{X}$ denote an observation, $\mathbf{y}$ a label, and $\mathbf{u}$ a domain index. The latent variable is partitioned as $\mathbf{z} = (\mathbf{z}_c, \mathbf{z}_s)$, where $\mathbf{z}_c$ is an invariant content block and $\mathbf{z}_s$ is a changing style block. The observation is generated by an invertible and smooth mixing function:

$$\mathbf{x} = g(\mathbf{z}_c, \mathbf{z}_s). \quad (2.1)$$

The domain affects only the changing block through an invertible domain-specific map:

$$\mathbf{z}_s = f_{\mathbf{u}}(\tilde{\mathbf{z}}_s), \quad (2.2)$$

where $\tilde{\mathbf{z}}_s$ is a high-level style variable shared across domains. The content distribution is invariant across domains, and the label depends on the high-level latent variables rather than directly on the raw domain:

$$\mathbf{y} \perp \mathbf{u} \mid (\mathbf{z}_c, \tilde{\mathbf{z}}_s). \quad (2.3)$$

Figure 2.1 shows the algorithmic version of this latent partition. The structural principle is minimal change. Domain variation should be explained by a low-dimensional changing block rather than by an unconstrained movement of the whole latent space. This principle is doing

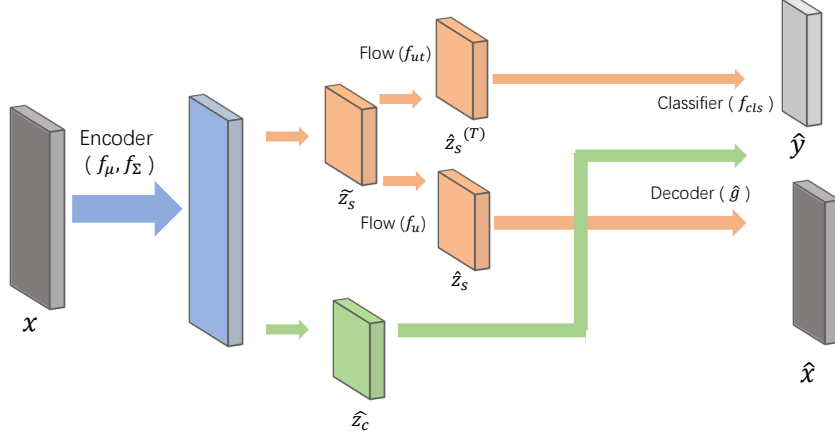


Figure 2.1. Regime A learns a partitioned latent representation. The changing component is mapped through a domain-specific invertible transformation into a shared high-level style variable, while invariant content is preserved for prediction.

real work. If the changing block is too large, it can absorb semantic content, making transfer unstable. If it is too small, the model cannot express the real shift. The theory below makes this tradeoff explicit: the number of domains needed for component-wise identifiability scales with the dimension of the changing block, not with the full latent dimension.

This model also clarifies the phrase partial disentanglement. The changing variables are identified component-wise, up to permutation and invertible scalar transformations. The invariant content variables are identified only as a block, up to an invertible transformation within the content subspace. This is not a weakness of the formulation. It matches the downstream task. For target-domain classification, we need the partition between stable content and aligned changing style to be correct; we do not need every content coordinate to have a unique semantic name.

2.4 Identifiability Assumptions

The theory compares the true model $(g, p_{\mathbf{z}_c}, p_{\mathbf{z}_s|\mathbf{u}})$ with an estimated model $(\hat{g}, p_{\hat{\mathbf{z}}_c}, p_{\hat{\mathbf{z}}_s|\mathbf{u}})$ that matches the observed marginals in every domain:

$$p_{\mathbf{x}|\mathbf{u}}(\mathbf{x}'|\mathbf{u}') = p_{\hat{\mathbf{x}}|\mathbf{u}}(\mathbf{x}'|\mathbf{u}'), \quad \forall \mathbf{x}' \in \mathcal{X}, \mathbf{u}' \in \mathcal{U}. \quad (2.4)$$

Matching marginals alone is not enough. The assumptions below are the contracts that make recovery possible.

2.4.1 Smooth Positive Densities and Conditional Independence

The latent density is smooth and positive on the latent domain. Conditional on $\mathbf{u}$, the latent coordinates factorize:

$$\log p_{\mathbf{z}|\mathbf{u}}(\mathbf{z}|\mathbf{u}) = \sum_{i=1}^n q_i(z_i, \mathbf{u}), \quad (2.5)$$

where q_i is the log density of coordinate z_i given the domain. This is the same type of regularity that appears throughout nonlinear independent component analysis: the domain must modulate the distribution of the latent variables in a way that can be measured through density changes.

2.4.2 Sufficient Variability of Changing Factors

For the changing block of dimension n_s , the theory requires that domain variation be rich enough to distinguish each changing coordinate. Define

$$\mathbf{w}(\mathbf{z}_s, \mathbf{u}) = \left(\frac{\partial q_{n_c+1}}{\partial z_{n_c+1}}, \dots, \frac{\partial q_n}{\partial z_n}, \frac{\partial^2 q_{n_c+1}}{\partial z_{n_c+1}^2}, \dots, \frac{\partial^2 q_n}{\partial z_n^2} \right), \quad (2.6)$$

where each derivative is evaluated at the corresponding $(z_i, \mathbf{u})$. For any $\mathbf{z}_s$, there must exist $2n_s + 1$ domain values such that the $2n_s$ difference vectors $\mathbf{w}(\mathbf{z}_s, \mathbf{u}_j) - \mathbf{w}(\mathbf{z}_s, \mathbf{u}_0)$ are linearly independent. Intuitively, the changing latent variables must change differently enough across domains. If two latent factors always shift in lockstep, no observational method can reliably separate them.

2.4.3 Domain Variability for the Invariant Block

The invariant block is protected by a second variability condition. Any nontrivial set in latent space that depends on the changing coordinates must receive different probability mass under at least two domains. This prevents a model from hiding changing information inside the estimated invariant block. If a set cannot be written as $B_{\mathbf{z}_c} \times \mathcal{Z}_s$, then some domains must disagree on its mass. The condition is abstract, but the intuition is simple: if a candidate invariant representation still contains changing information, the domain variation should expose it.

2.5 Identifiability Results

Key result: partial identifiability from multiple distributions

Assume that the data are generated according to Equations (2.1) and (2.2), and that the estimated model satisfies the marginal-matching condition in Equation (2.4). Further assume that the latent densities are smooth and positive, the latent coordinates are conditionally independent given the domain, the changing block satisfies the sufficient-variability condition, and the invariant block satisfies the domain-variability condition. Then:

- the changing variables $\mathbf{z}_s$ are identifiable up to permutation and component-wise invertible transformations;
- the invariant content variables $\mathbf{z}_c$ are identifiable block-wise, meaning $\hat{\mathbf{z}}_c = h_c(\mathbf{z}_c)$ for some invertible map h_c ;
- the target-domain joint distribution $p_{\mathbf{x}, \mathbf{y} | \mathbf{u}^T}$ is identifiable through the shared prediction rule $p_{\mathbf{y} | \mathbf{z}_c, \tilde{\mathbf{z}}_s}$.

The proof for the changing block follows the logic of nonlinear independent component analysis (nonlinear ICA). Because the estimated and true models match the observed marginals, the inverse maps from $\mathbf{x}$ to latent coordinates differ by an invertible transformation. Taking first and second derivatives of log densities across domains yields a linear system whose coefficients are the derivatives of this transformation. The sufficient variability condition makes the system full rank, forcing the transformation to be component-wise on the changing block. The invariant-block proof uses a contradiction argument. Suppose the estimated content block contains information from the true changing block. Then one can construct a latent set that is not expressible as a pure content set but that the estimated model treats as invariant. The domain-variability condition says such a set must have different probability mass in some domains, contradicting the matched marginal distributions. The result is block-wise recovery of content and separation from changing information.

Finally, identifiability of the target joint distribution follows from the label factorization. Once $\mathbf{z}_c$ and $\tilde{\mathbf{z}}_s$ are recovered, a classifier trained on source labels estimates $p_{\mathbf{y} | \mathbf{z}_c, \tilde{\mathbf{z}}_s}$. The target domain supplies unlabeled observations that identify $p_{\mathbf{x}, \mathbf{z}_c, \tilde{\mathbf{z}}_s | \mathbf{u}^T}$. Combining the two recovers $p_{\mathbf{x}, \mathbf{y} | \mathbf{u}^T}$. Formal statements and proofs are collected in Appendix A.1.

2.6 The iMSDA Method

The theory suggests an algorithm with three modules: an encoder-decoder that estimates the inverse of the mixing function, a domain-specific invertible flow that aligns the changing variables, and a classifier that operates in the shared high-level latent space. This is the identifiable Multi-Source Domain Adaptation method, abbreviated iMSDA.

The VAE encoder estimates a posterior over latent variables:

$$\hat{\mathbf{z}} \sim q_{f_\mu, f_\Sigma}(\hat{\mathbf{z}}|\mathbf{x}) := \mathcal{N}(f_\mu(\mathbf{x}), f_\Sigma(\mathbf{x})), \quad \hat{\mathbf{x}} = \hat{g}(\hat{\mathbf{z}}). \quad (2.7)$$

The latent vector is partitioned as $\hat{\mathbf{z}} = (\hat{\mathbf{z}}_c, \hat{\mathbf{z}}_s)$. The domain flow maps the changing part back to the shared high-level style variable:

$$\hat{\mathbf{z}}_s = \hat{f}_u^{-1}(\hat{\mathbf{z}}_s). \quad (2.8)$$

The classifier predicts from $(\hat{\mathbf{z}}_c, \hat{\mathbf{z}}_s)$:

$$\hat{\mathbf{y}} = f_{\text{cls}}(\hat{\mathbf{z}}_c, \hat{\mathbf{z}}_s). \quad (2.9)$$

Training combines reconstruction, latent alignment, source-label prediction, and target entropy minimization:

$$\mathcal{L}_{\text{iMSDA}} = \mathcal{L}_{\text{cls}} + \alpha_1 \mathcal{L}_{\text{ent}} + \alpha_2 \mathcal{L}_{\text{VAE}}. \quad (2.10)$$

The VAE term reconstructs observations and regularizes $(\hat{\mathbf{z}}_c, \hat{\mathbf{z}}_s)$ toward a simple shared prior. The classification term uses source labels to learn the stable prediction rule. The target entropy term encourages confident target predictions in the aligned latent space, serving as a practical constraint that the recovered space should support the same label mechanism across domains.

Algorithmic Interpretation

The flow in Equation (2.8) operationalizes domain alignment. The classifier in Equation (2.9) operationalizes the assumption that labels depend on $(\mathbf{z}_c, \tilde{\mathbf{z}}_s)$ rather than on the raw domain. The changing dimension n_s operationalizes minimality. Thus the algorithm is not a generic VAE plus a classifier; each piece corresponds to a structural assumption used in the identifiability claim.

Table 2.1. Synthetic identifiability evidence from Kong et al. [9]. More domains improve both component-wise recovery of changing variables and block-wise recovery of invariant content.

Metric	3 domains	5 domains	7 domains	9 domains
MCC for $\mathbf{z}_s$	0.67 ± 0.12	0.80 ± 0.15	0.90 ± 0.12	0.91 ± 0.09
R^2 for $\mathbf{z}_c$	0.59 ± 0.05	0.74 ± 0.09	0.84 ± 0.05	0.85 ± 0.06
Average	0.63 ± 0.08	0.77 ± 0.12	0.87 ± 0.09	0.88 ± 0.08

2.7 Synthetic Experiments

The synthetic experiments test identifiability directly. Latent variables have four dimensions, with $n_c = n_s = 2$. Content variables are sampled from a standard Gaussian. Changing variables are sampled from $\mathcal{N}(\mu_{\mathbf{u}}, \sigma_{\mathbf{u}}^2 I)$, where each domain has a randomly sampled mean and variance. A two-layer nonlinear MLP maps latent variables to observations. Because the true latents are known, component-wise recovery of the changing block is measured by mean correlation coefficient (MCC), and block-wise recovery of the content block is measured by R^2 .

Table 2.1 shows the theory’s signature pattern. As the number of domains increases, the sufficient-variability condition becomes more plausible, and both MCC and R^2 improve. The important point is not that nine domains is a universal requirement. The point is that the empirical curve matches the theorem: domain variation is the identifying signal, and additional distinct domains make the signal stronger.

2.8 Domain Adaptation Experiments

The real-world experiments evaluate multi-source adaptation on two image benchmarks that keep semantic classes constant while visual domains change. PACS contains 9,991 images of seven shared classes across Photo, Art Painting, Cartoon, and Sketch domains [23]. Office-Home contains approximately 15,500 images from 65 everyday office and household categories, such as mugs, keyboards, and chairs, across Art, Clipart, Product, and Real-World domains [24]. Its Product images are catalog-style objects with little or no background, while its Real-World images are ordinary camera photographs of objects in context. PACS emphasizes large style changes over a small class set, while Office-Home covers many more object categories and collection styles. Figure 2.2 shows these domain shifts. For each benchmark, each domain is held out in turn as the target, while the remaining domains are labeled sources. The method is compared against source-only training, adversarial align-



Figure 2.2. Representative images from the four visual domains in PACS [23] and Office-Home [24]. The top row shows PACS Art Painting, Cartoon, Photo, and Sketch examples; the bottom row shows Office-Home Art, Clipart, Product, and Real-World examples. Images are adapted from the original benchmark papers.

Table 2.2. PACS target-domain accuracy. iMSDA improves the average and is especially strong on the challenging sketch target.

Method	Art	Cartoon	Photo	Sketch	Avg.
Source only	74.9	72.1	94.5	64.7	76.6
M3SDA	89.3	89.9	97.3	76.7	88.3
T-SVDNet	90.43	90.61	98.50	85.49	91.25
iMSDA	93.75	92.46	98.48	89.22	93.48

ment, moment matching, and contemporary multi-source-domain adaptation methods. The thesis-relevant question is whether the same minimal-change mechanism improves target accuracy in natural visual domains.

Tables 2.2 and 2.3 show consistent average improvements. On PACS, the largest conceptual gain is Sketch, where the target domain differs strongly in style from photos, art, and cartoons. On Office-Home, iMSDA improves the average despite Clipart remaining difficult. This is a useful limitation: the method does not claim that every target can be solved by the same low-dimensional changing block. It claims that when the domain variation contains enough information to align a shared latent mechanism, the learned partition improves transfer.

Table 2.3. Office-Home target-domain accuracy. The method improves the average and gives large gains on Art, Product, and Real-World targets.

Method	Art	Clipart	Product	Real-World	Avg.
Source only	64.58	52.32	77.63	80.70	68.81
M3SDA	66.22	58.55	79.45	81.35	71.39
MIAN- γ	69.88	64.20	80.87	81.49	74.11
iMSDA	75.40	61.40	83.50	84.47	76.19

Table 2.4. Ablations from Kong et al. [9]. Both the target entropy term and the VAE alignment term matter, and the changing dimension has a non-monotone effect.

Setting	PACS average	Office-Home average
$\mathcal{L}_{\text{cls}}$ only	82.11	67.20
$+\alpha_1 \mathcal{L}_{\text{ent}}$	92.01	75.77
Full objective	93.48	76.19
PACS, $n_s = 8$	92.05	
PACS, $n_s = 4$	93.48	
PACS, $n_s = 2$	92.66	

The ablations in Table 2.4 are more important for the thesis than the headline averages. Adding target entropy provides a large gain, which supports the idea that the shared latent space must preserve the label mechanism. Adding the VAE alignment term gives a further improvement, which supports the generative alignment argument. The changing dimension has a sweet spot. This is exactly what minimal change predicts: too large a changing block violates the bias toward sparse domain change, while too small a block lacks the capacity to model the real shift.

Evidence: Regime A mechanism

The synthetic results test identifiability directly, the benchmark results test target transfer, and the ablations test the assumptions behind the method. The same pattern appears across them: distributional variation helps when it is channeled through an explicit changing block and aligned to a shared label-relevant space.

2.9 From Transfer to Mechanism Discovery

The domain-adaptation chapter is the first instance of a broader thesis direction: multiple distributions can reveal latent mechanisms when the difference between distributions is struc-

tured. The anchor paper uses this idea for target-domain classification, but the direction is larger. Across the supporting papers, Regime A becomes a language for controllable generation, counterfactual editing, multimodal scientific analysis, and modular vision-language alignment.

Multi-domain generation and translation. Multi-domain image generation and unpaired image translation pose a harder version of the same problem [15]. Instead of predicting labels in a target domain, the model must learn a joint distribution from multiple marginal image distributions. This is ill-posed because infinitely many joint distributions can share the same marginals. The supporting work uses the same minimal-change idea: domain information enters through restricted changing components, allowing the model to preserve content while translating style. Conceptually, this extends Regime A from transfer to controllable generation.

Counterfactual text generation. Counterfactual generation studies how to change style while preserving content in text [16]. It relaxes the overly clean assumption that content and style are independent. In language, content and style are often dependent, and the dependence itself varies across domains: positive restaurant reviews use different words from positive movie reviews. The contribution is to use relative influence sparsity to identify content and style even under domain-varying dependence. This strengthens the Regime A theme by showing that the changing and invariant factors need not be independent in a simplistic way for counterfactual control to be possible.

Biomedical multimodal representation learning. Biomedical datasets often contain multiple modalities, such as anthropometric measurements, sleep signals, genetic information, and clinical variables. The biomedical multimodal work treats these modalities as windows onto a shared physiological mechanism [17]. Its technical contribution is to move beyond coarse shared-subspace identification by using sparsity of cross-modal causal connections to identify individual latent components. This gives Regime A a scientific interpretation: distributional or modal differences are not merely obstacles to remove, but clues about latent physiology.

SmartCLIP and modular alignment. Vision-language contrastive learning faces information misalignment. A short caption may describe only one part of an image, while a long caption may entangle many concepts. SmartCLIP formulates alignment as a latent-variable identification problem and learns adaptive masks that align only the relevant visual and textual modules [18]. This is directly connected to Regime A because each caption creates a partial distributional view of the image. The method shows that the minimal-change idea can become a modular alignment rule: align the concepts that are shared by a pair, while preserving other concepts for other pairs.

Together these papers turn Regime A from a single adaptation problem into a general use of natural variation. Domain labels, modalities, captions, styles, and biomedical measurements all provide partial views of a hidden mechanism. The common technical move is to avoid both extremes: neither force everything to be invariant nor allow every factor to change freely. Instead, identify the small changing part, preserve the mechanism that should transfer, and validate the recovered factors through prediction, generation, intervention, or domain knowledge.

2.10 Limitations and Open Questions

Regime A depends on informative variation. If domains differ only weakly, the sufficient-variability conditions fail. If domain variation is dense and semantically entangled, a small changing block may be too restrictive. If the domain index itself is coarse, the model may merge multiple mechanisms into one estimated shift. These limitations are not incidental. They are the boundary of the identification claim.

The chapter also assumes that target-domain unlabeled samples are available. The method is therefore not a full theory of extrapolation. It adapts to a target distribution by using target observations to align the latent mechanism. Regime C will address the harder case in which the target request may be outside the training support and may not provide a full target distribution.

Finally, partial identifiability is not full human interpretability. Recovering a changing coordinate up to an invertible transformation does not automatically name that coordinate as pose, color, scanner effect, or sentiment. The supporting projects point toward the next layer of evidence: counterfactual edits, cross-modal grounding, biomedical validation, and modular alignment tests that ask whether the recovered factor behaves as a meaningful concept.

Transition: from external variation to internal structure

Regime A uses explicit differences across distributions as the causal signal. Regime B removes that signal. The next chapter asks whether a single unstructured distribution can still reveal latent mechanisms when the learner creates its own missing-information problem through masking.

Chapter 3

Regime B: Learning with Unstructured Distributions

This chapter develops the second regime, where the learner has no explicit domain labels and must recover structure from internal dependence induced by a self-supervised objective.

3.1 Learning Structure Without Explicit Labels

The second regime removes the explicit domain signal from Regime A. Large-scale pretraining often begins from unstructured observations: internet images, language corpora, videos, or multimodal pairs without clean labels for the underlying concepts. The learner may not know which examples belong to which domain, which latent factors should vary, or which concepts should be named. Yet modern self-supervised objectives often produce representations that support semantic downstream tasks.

The general problem is to explain when this success is more than accidental compression. A model trained on an unstructured distribution may learn texture statistics, co-occurrence patterns, or memorized templates. A trustworthy model needs something stronger: latent structure that can be inspected, grounded, and eventually controlled. In this regime, the learner cannot rely on domain labels or explicit interventions. The available signal is internal dependence within an observation or across paired modalities. The causal question is which objective turns that dependence into recoverable concepts.

3.2 Masked Prediction in Hierarchical Latent Models

Masked autoencoders (MAE) provide a focused technical setting [25]. MAE divides an image into patches, hides a random subset, encodes the visible patches, and reconstructs the hidden patches. The pretraining target is low-level pixel reconstruction, but the learned representation transfers to classification, detection, and segmentation. The empirical puzzle is why a pixel-level reconstruction task can learn high-level semantic structure.

This chapter gives a causal answer [10]. Masking turns an image into a missing-information problem. Depending on which patches are hidden and visible, reconstruction may require either local texture or global object-level information. Under a hierarchical latent data-generating process, each mask induces a minimal set of shared latent variables between the visible and hidden parts. MAE can identify that shared set under suitable conditions. The masking ratio and patch size therefore become concept-level knobs: they choose which dependencies are necessary for the reconstruction problem and, consequently, which latent variables the representation is pushed to recover.

3.3 Hierarchical Latent Model for Images

Images are high-dimensional, but their dependencies are structured. Neighboring pixels often share local texture and edge information. Distant regions can be dependent because they belong to the same object, part, scene, or global configuration. This suggests a hierarchy: high-level variables encode global semantics, intermediate variables encode parts and shapes, low-level variables encode local details, and pixels are generated at the leaves.

Formally, let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a directed acyclic graph with observable and latent variables

$$\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_m\}, \quad \mathcal{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_n\}. \quad (3.1)$$

Each variable may be multidimensional. There are no direct edges between two observables, and each variable is generated from its parents and independent exogenous noise:

$$\mathbf{z}_i = g_{\mathbf{z}_i}(\text{Pa}(\mathbf{z}_i), \varepsilon_i), \quad (3.2)$$

$$\mathbf{x}_j = g_{\mathbf{x}_j}(\text{Pa}(\mathbf{x}_j), \varepsilon_j). \quad (3.3)$$

The generating modules are assumed invertible in the relevant arguments. This does not mean that the actual graph is known to the learner. The graph is a theoretical representation of how observed dependence can be mediated by latent concepts.

Figure 3.1 gives the working intuition: masking can make high-level latent variables necessary for reconstruction when local shortcuts are removed. The no-observable-edge assumption is a useful idealization for images. Pixels do not directly cause one another in the physical scene; their statistical dependence is mediated by latent factors such as lighting, shape, material, object identity, and viewpoint. The hierarchy is deliberately general. It can represent latent variables at different abstraction levels and arbitrary directed relations among them, provided the graph remains acyclic.

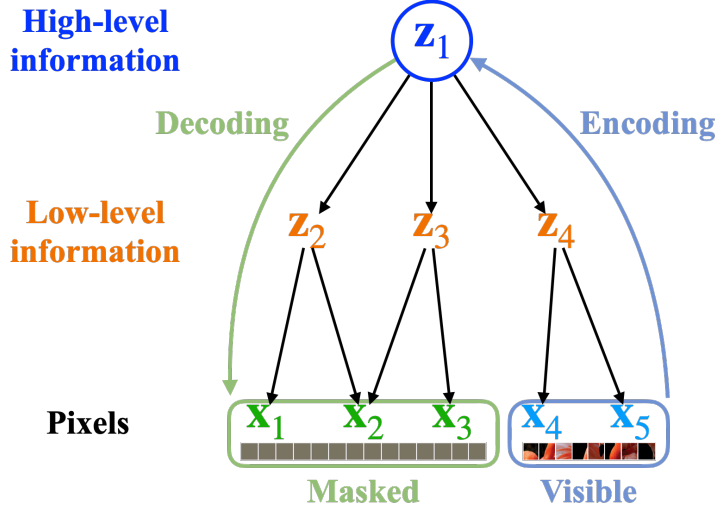


Figure 3.1. Masked prediction under a hierarchical generating process. A mask induces a shared latent set between visible and hidden patches. Intermediate masking can force the model to use object-level variables rather than local shortcuts.

3.4 Masks as Latent-Information Probes

For a mask $\mathbf{m}$, let $\mathbf{x}_{\mathbf{m}}$ denote the hidden patches and $\mathbf{x}_{\mathbf{m}^c}$ the visible patches. MAE samples masks from a distribution parameterized by masking ratio r and masking-unit size s . It trains an encoder $E_{\mathbf{m}^c}$ and decoder $D_{\mathbf{m}}$ with objective

$$\mathcal{L}(E, D) = \mathbb{E}_{\mathbf{m}, \mathbf{x}, \hat{\mathbf{s}}_{\mathbf{m}}} [\|D_{\mathbf{m}}(E_{\mathbf{m}^c}(\mathbf{x}_{\mathbf{m}^c}), \hat{\mathbf{s}}_{\mathbf{m}}) - \mathbf{x}_{\mathbf{m}}\|^2], \quad (3.4)$$

where $\hat{\mathbf{s}}_{\mathbf{m}}$ denotes mask tokens and positional information used by the decoder.

The causal interpretation is that each mask creates a two-view latent model. There is information specific to the masked part, information specific to the visible part, and a minimal shared set $\mathbf{c}(\mathbf{m})$ that makes reconstruction possible:

$$\mathbf{x}_{\mathbf{m}} = g_{\mathbf{x}_{\mathbf{m}}}(\mathbf{c}, \mathbf{s}_{\mathbf{m}}), \quad \mathbf{x}_{\mathbf{m}^c} = g_{\mathbf{x}_{\mathbf{m}^c}}(\mathbf{c}, \mathbf{s}_{\mathbf{m}^c}). \quad (3.5)$$

The shared set $\mathbf{c}$ is not fixed for all masks. A small local mask may share only low-level texture variables with nearby visible pixels. A mask that hides large object regions may force $\mathbf{c}$ to include object-level or scene-level variables. Thus masking is an information intervention: it changes which latent variables are necessary for prediction.

3.5 Identifying Shared Latent Information

Key result: mask-induced shared latent variables

For each specific mask $\mathbf{m}$ in a hierarchical latent graph, there exists a minimal latent set $\mathbf{c}(\mathbf{m})$ such that the masked and visible parts can be written as in Equation (3.5), the masked-specific information $\mathbf{s}_{\mathbf{m}}$ is independent of $(\mathbf{c}, \mathbf{s}_{\mathbf{m}^c})$, and no smaller latent set has these properties. This shared set is unique for the mask and can be located from the hierarchy by tracing ancestors of the masked and visible observables and pruning redundant ancestors.

The proof is graph-theoretic. Starting from the hidden observables, one walks upward through their latent parents. Any ancestor that also reaches the visible observables is a candidate shared variable. Ancestors that are already screened off by lower shared variables are pruned. The result is the smallest set that mediates the dependence between hidden and visible parts. The theorem explains why different masks select different abstraction levels. The next result connects this shared set to MAE training. The MAE model assumption has two parts. First, the encoder is embedded in an invertible map, so its output can be viewed as part of a lossless representation of the visible patches. Second, the decoder is locally invertible in the representation, so changes in the shared representation have recoverable effects on the masked output. Under these conditions, optimizing the reconstruction objective matches the joint distribution of visible and hidden parts in the shared-latent model.

Key result: MAE identification

For a fixed mask $\mathbf{m}$, given the dimensions of the shared and mask-specific variables, any MAE solution satisfying the invertibility and local invertibility conditions recovers all information in $\mathbf{c}(\mathbf{m})$ up to a one-to-one transformation. Equivalently, there exists an invertible map h such that $\hat{\mathbf{c}} = h(\mathbf{c})$, where $\hat{\mathbf{c}} = E_{\mathbf{m}^c}(\mathbf{x}_{\mathbf{m}^c})$ is the learned MAE representation.

The proof reduces MAE to an information-sharing latent model. Let $\mathbf{v}_1 = \mathbf{x}_{\mathbf{m}}$ and $\mathbf{v}_2 = \mathbf{x}_{\mathbf{m}^c}$. In the true model, $\mathbf{v}_1$ depends on $(\mathbf{c}, \mathbf{s}_1)$ and $\mathbf{v}_2$ depends on $(\mathbf{c}, \mathbf{s}_2)$. In the learned model, the decoder and the invertible completion of the encoder play the same roles. If the learned model matches the joint distribution, differentiating the equality of the two generative descriptions shows that the estimated shared representation cannot depend on $\mathbf{s}_1$ or $\mathbf{s}_2$. By symmetry, the true $\mathbf{c}$ cannot depend on the learned private variables either. Therefore the learned representation and the true shared set are in one-to-one correspondence. Formal statements and proofs are in Appendix A.2.

3.6 Masking Ratio and Patch Size

The theory predicts that masking aggressiveness should have a non-monotone effect on semantic representation quality. With a very low masking ratio, visible neighboring patches can often reconstruct the hidden patches through local interpolation. The shared set $\mathbf{c}(\mathbf{m})$ stays low-level. With an intermediate masking ratio, local shortcuts are broken while enough evidence remains to infer global object structure. The shared set moves upward in the hierarchy. With an extremely high masking ratio, too little evidence remains; reconstruction may again rely on low-level priors or incomplete signals rather than stable object-level information.

Patch size plays a related role. Small masking units allow local continuity to solve many reconstruction tasks. Larger units break local interpolation and encourage higher-level inference, but extremely large units can remove too much information or make the reconstruction target too coarse. The chapter’s key message is therefore not that more masking is always better. It is that masking should be strong enough to make high-level variables necessary and weak enough to keep them inferable.

3.7 Empirical Evidence

The experiments pretrain MAE models on ImageNet-1K with a ViT-Base backbone. One set fixes the masking unit size at 16 and varies the masking ratio over $\{0.10, 0.25, 0.50, 0.75, 0.90\}$. A second set fixes the masking ratio at 0.75 and varies the masking unit size over $\{8, 16, 32\}$ while decoupling the masking-unit size from the ViT patch size. The experiments then ask how masking aggressiveness changes the representation.

3.7.1 Reconstruction Metrics

The first evaluation compares high-level and low-level reconstruction metrics. Structural similarity (SSIM) and feature similarity (FSIM) measure whether the reconstruction preserves perceptual structure. Mean squared error and peak signal-to-noise ratio measure pixel-level fidelity. The empirical pattern matches the theory. Increasing masking from low to intermediate ratios improves high-level structural similarity, while low masking better preserves local pixel-level detail. Extremely high masking can hurt high-level structure and recover more low-level or background information. Larger masking units similarly push the task toward structure until the mask becomes too coarse.

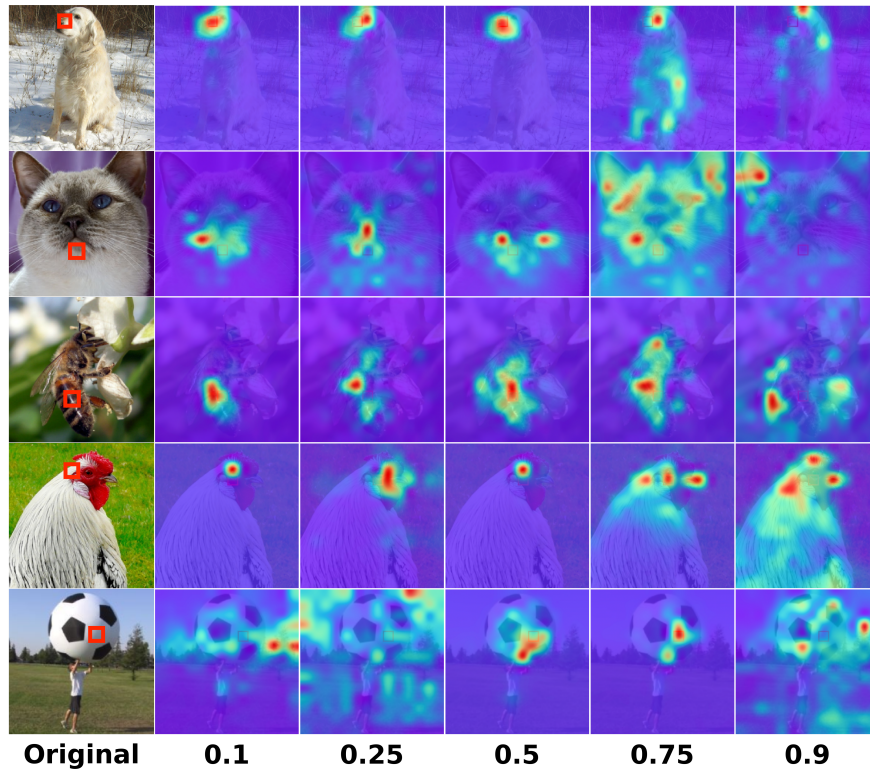


Figure 3.2. Attention visualizations from MAE models trained with different masking ratios. Intermediate masking encourages broader object-level attention, while extreme masking can return to local or incomplete behavior.

3.7.2 Attention and Object-Level Context

Attention visualizations test whether representations integrate information across object regions. The study examines both the [CLS] token and tokens associated with object parts. As masking increases from very low to intermediate values, attention becomes more object-centric and less restricted to local neighborhoods. At the extreme ratio of 0.90, attention can become less complete and more affected by background or missing evidence. Figure 3.2 illustrates this mechanism-level evidence.

3.7.3 Linear Probing and Robustness

Table 3.1 shows that linear probing and robustness improve as masking moves from 0.10 to 0.75, then decline at 0.90. The same pattern appears in t-SNE visualizations: intermediate masking produces more linearly separable class clusters. Robustness on ImageNet variants is especially important because it tests whether the recovered representation is semantic enough to transfer across shifts in rendition, sketch style, adversarially selected examples,

Table 3.1. Linear probing and robustness on ImageNet variants. IN1K is ImageNet-1K; IN-v2, IN-R, IN-A, and IN-S are ImageNet-v2, ImageNet-Rendition, ImageNet-Adversarial, and ImageNet-Sketch. Intermediate masking improves separability and robustness relative to extreme masking.

Mask ratio	IN1K	IN-v2	ObjectNet	IN-R	IN-A	IN-S
0.10	47.45	34.72	9.42	14.63	2.00	7.25
0.25	53.58	40.34	11.54	18.68	2.49	10.27
0.50	60.07	46.71	13.94	22.44	2.89	12.58
0.75	67.41	54.23	18.24	25.20	3.76	15.51
0.90	62.97	49.52	15.87	19.11	2.76	10.46

Table 3.2. Shape bias and Common Objects in Context (COCO) transfer results. A mask ratio near 0.75 favors high-level shape information and improves dense downstream tasks. AP denotes average precision.

Masking ratio	0.10	0.25	0.50	0.75	0.90
Shape bias	0.135	0.255	0.246	0.256	0.201
COCO box AP	30.47	32.38	34.87	39.72	37.17
COCO mask AP	28.24	29.95	32.11	36.35	34.35

and object context.

3.7.4 Shape Bias and Dense Transfer

Shape bias measures whether a model relies more on object shape than texture. Common Objects in Context (COCO) detection and segmentation use box and mask average precision (AP) to test whether the representation supports dense semantic downstream tasks. Table 3.2 again shows a non-extreme optimum. The agreement among linear probing, robustness, attention, shape bias, and COCO transfer is the chapter’s main empirical evidence. A single accuracy number could have many explanations. The collection of probes points back to the same hierarchical mechanism.

Evidence: Regime B mechanism

Masking creates a missing-information problem. Low masking leaves local shortcuts; extremely high masking removes too much evidence; intermediate masking makes object-level latent variables useful and recoverable. The evidence appears across reconstruction metrics, attention, separability, robustness, shape bias, and detection or segmentation transfer.

3.8 Implications for Self-Supervised Pretraining

The theory also suggests ways to improve masking-based pretraining. First, boundary pixels between visible and masked regions often encourage local interpolation. If the downstream task values high-level semantics, one can reduce the contribution of near-boundary pixels in the reconstruction loss. Second, random masking is only one way to choose $\mathbf{c}(\mathbf{m})$. A masking scheme informed by latent structure could deliberately target object parts or regions whose reconstruction requires higher-level variables. Third, different downstream tasks may prefer different abstraction levels. Pixel-level restoration may benefit from lower-level shared information, while classification and detection benefit from higher-level shared information. This turns pretraining from a fixed recipe into a mechanism-design problem.

3.9 From Masked Prediction to Concept Learning

Regime B is the bridge from domain-based identification to foundation-model interpretability. The MAE analysis shows that a self-supervised objective can select useful latent information, but the broader direction asks for more: concept graphs, grounding, and interventions inside high-dimensional models. Several supporting projects deepen this bridge.

Discrete concepts in latent hierarchical models. The discrete-concept work formalizes concepts as discrete latent causal variables organized in a hierarchy [19]. It establishes conditions under which discrete concepts and their relations can be identified from high-dimensional continuous observations. This strengthens the MAE chapter by moving from shared latent information to named abstraction levels: object classes, shapes, breeds, parts, and other discrete concepts can be treated as latent variables whose recoverability is a formal question.

Nonlinear hierarchical identification. The nonlinear hierarchy project [26] addresses a more structural version of the same problem: can latent variables and their causal graph be identified when observed variables are generated by nonlinear, causally related latent variables? Its contribution is an identification criterion and estimation procedure for nonlinear latent hierarchical models with more general graph structures than latent trees. This work

supplies the graph-learning counterpart to the MAE analysis. MAE explains why a pretraining objective can recover a mask-induced shared set; nonlinear hierarchy identification asks how the broader latent hierarchy itself may be recovered.

White-box generative models. The white-box generative-model work asks how internal concepts in diffusion and language models can be made interpretable and controllable [21]. It uses causal minimality to connect sparsity or compression constraints to component-wise identifiable concept variables. This extends Regime B from explaining a pretraining objective to extracting concept graphs inside already trained generative models. It also sharpens the control goal: a recovered concept should function as a lever for model steering.

Vision-language concepts. The vision-language concept work studies controllable image generation by identifying atomic textual and visual concepts and sparse connections between them [20]. This is the multimodal counterpart of the MAE story. Instead of masks selecting shared information between visible and hidden patches, text-image pairs create shared and modality-specific concept variables. The technical role in the thesis is to connect unsupervised visual concept recovery to controllable generation through language grounding.

SmartCLIP and partial cross-modal observations. SmartCLIP contributes a complementary route to the same direction [18]. Contrastive vision-language learning usually treats an image and caption as a fully aligned pair, but real captions are partial. They mention some objects, attributes, or relations and ignore others. SmartCLIP uses adaptive masking to identify the modules actually shared by the two modalities. This turns Regime B from single-modality masked prediction into cross-modal concept selection: the model must decide which latent concepts are visible to language and which remain private to the image. Together, these papers make Regime B a research program on concept formation. MAE explains how an objective can pull out shared information from a single image distribution. Discrete and nonlinear hierarchy work asks how the concepts and their relations can be formally identified. White-box and vision-language work ask how those concepts become editable levers inside generative models. SmartCLIP shows that partial multimodal supervision can ground only the relevant modules. The common theme is that interpretability is not a post-hoc label on a black-box representation; it is a structural property of how the representation is learned.

3.10 Limitations and Open Questions

The MAE theory covers a simplified objective and relies on invertibility or local invertibility assumptions. Real pretrained models use finite capacity, optimization heuristics, augmentations, regularization, and data distributions that only approximately match the theory. The

claim is therefore not that the theorem fully explains every feature in every masked model. The claim is that it identifies a concrete mechanism that generates testable predictions about masking ratio, patch size, abstraction level, and downstream semantics.

Another limitation is naming. The theorem identifies shared latent information up to an invertible transformation. Human interpretability requires more: grounding to language, interventions, behavioral tests, and stability across datasets. This is why the supporting work on discrete concepts, white-box generative models, and vision-language alignment is not peripheral. It supplies the missing path from recovered shared variables to usable concepts. Finally, Regime B learns from observed distributions. It can explain how concepts emerge inside the support of the data, but it does not by itself say when those concepts can be recombined outside that support. This is the role of Regime C.

Transition: from recovered concepts to new combinations

Regime B shows how an objective can recover latent concepts inside an unstructured distribution. Regime C asks when those concepts or modules can be recombined for a request that never appeared in training.

Chapter 4

Regime C: Generalization Beyond Training Distributions

This chapter develops the third regime, where trustworthy behavior requires knowing when an off-support prediction, generation, or reasoning step is supported by reusable latent modules.

4.1 Calibrated Generalization Outside Support

The third regime asks for prediction or generation beyond the observed training support. This is the most ambitious form of trustworthiness in the thesis. In Regime A, the learner sees unlabeled target examples and uses them to align a latent mechanism. In Regime B, the learner stays inside an unstructured distribution and uses masking to recover latent concepts. Regime C asks what can be done when the requested behavior itself may be absent from training.

The general danger is unsupported confidence. Large models are often asked to answer a novel question, generate a novel scene, solve a new reasoning problem, or act in a new environment. When the request is outside the support of the training data, a high-quality surface form can hide a missing mechanism. A trustworthy model must therefore distinguish two cases: the request is a new combination of known modules, or the request requires a mechanism that has not been learned. Regime C studies the first case and marks the boundary with the second.

4.2 Compositional Generation

The motivating example is compositional generation. A model may have seen peacocks, chickens, rice, and chickens eating rice, but never a peacock eating rice. The requested text prompt is off support at the level of high-level concept combinations. Without structural assumptions, the problem is unidentifiable. Two conditional generative models can agree

perfectly on every observed combination and disagree arbitrarily on the missing one. Thus the question is not whether a larger neural network can interpolate better. The question is what latent structure makes the off-support conditional distribution determined by mechanisms already observed on support.

This chapter develops the answer from hierarchical concept models [11]. A novel composition is justified when it can be decomposed into reusable low-level modules, and every required module has already appeared under some supported training configuration. This is the chapter’s central discipline: generalization beyond support should be calibrated by module support. The model can recombine known mechanisms, but it should not pretend to know a mechanism that the training data never constrained.

4.3 Compositional Generalization and Analogy

Let $\mathbf{d}$ denote discrete high-level concepts, such as objects, attributes, and relations specified by text. Let $\mathbf{x}$ denote the observation, such as an image. Training identifies $p(\mathbf{x}|\mathbf{d})$ only for concept vectors in the source support $\mathcal{D}_S$. A vector $\mathbf{d} \notin \mathcal{D}_S$ is composable if the conditional distribution $p(\mathbf{x}|\mathbf{d})$ is nevertheless determined by mechanisms learned from $\mathcal{D}_S$.

The human analogy is decomposition. We can imagine a peacock eating rice because “peacock” and “chicken” share low-level bird modules such as beak, wings, and feet, while “chicken eating rice” teaches a reusable interaction between beak and rice. The target composition is novel at the high level, but its low-level modules are familiar. Causal modularity says these modules are autonomous enough to be reused across contexts. Minimal change says related high-level concepts share many low-level modules and differ only in a small number of distinguishing features.

Figure 4.1 illustrates this point. The new prompt is not solved by a single global interpolation between images. It is solved by identifying which modules are needed, locating where those modules have appeared in training, and recombining them with sparse interactions. This is why the chapter begins with a latent hierarchy rather than with a diffusion architecture. The architecture is only justified once the module-support condition is stated.

4.4 Hierarchical Concept Model

The model contains discrete high-level variables $\mathbf{d}$, continuous latent concepts across levels $\mathbf{z}_1, \dots, \mathbf{z}_L$, and an observation $\mathbf{x}$. For exposition, $\mathbf{z}_0$ denotes the discrete concept vector and $\mathbf{z}_{L+1}$ denotes the observation. A top-level concept variable is generated from its corresponding discrete specification, and every lower variable is generated from its parents and exogenous noise:

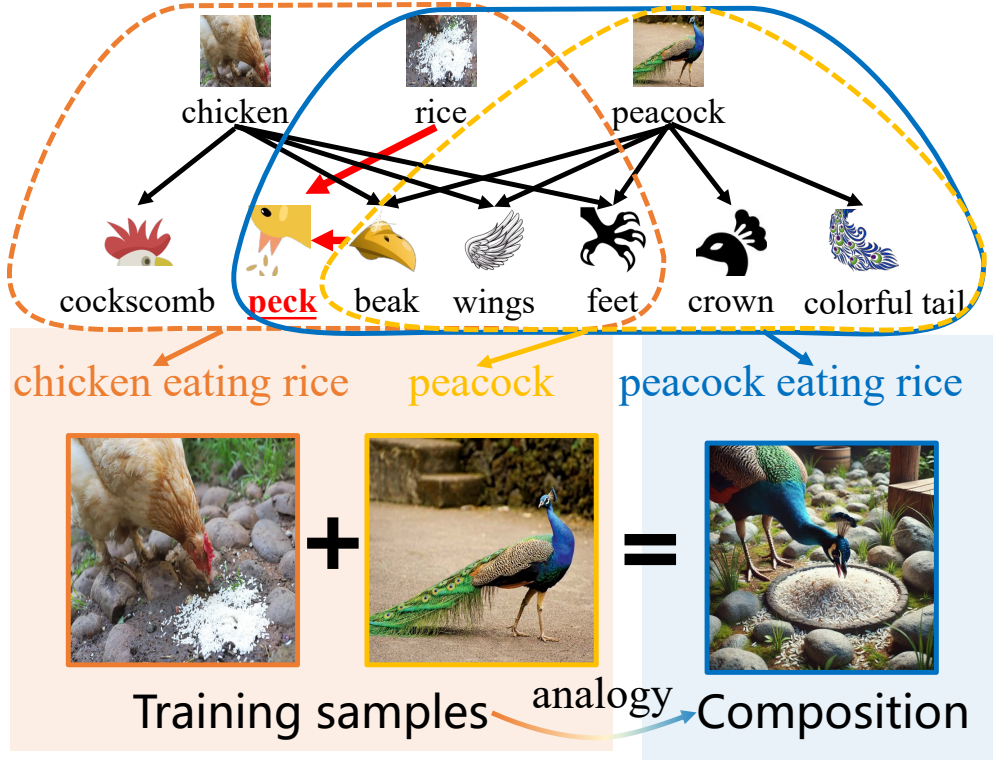


Figure 4.1. Composition by analogy. A novel high-level composition can be generated by decomposing source concepts into reusable low-level modules and recombining the modules whose supports have been observed.

$$z_{1,i} \sim p(z_{1,i}|d_i), \quad (4.1)$$

$$v = g_v(\text{Pa}(v), \varepsilon_v), \quad v \in \{\mathbf{z}_2, \dots, \mathbf{z}_L, \mathbf{x}\}. \quad (4.2)$$

The graph is hierarchical, but not necessarily a tree. A low-level module can be shared by several high-level concepts, and a downstream module can encode an interaction among multiple parents. This is essential for visual relations. An “eating” relation is not the additive sum of a bird pixel pattern and a rice pixel pattern; it is an interaction module whose inputs must be supported.

The model also encodes absence. If a discrete concept $d_i = 0$ indicates that an object or attribute is absent, then the corresponding top-level latent variable is degenerate. Nonzero values of d_i specify varieties of the same concept while sharing a support for the associated latent variable. This lets the theory distinguish a concept being absent from a concept appearing in a new variety.

4.5 The Module-Support Condition

Key result: module-support condition for composition

Under the hierarchical process in Equation (4.2), a discrete concept vector $\mathbf{d}$ is composable if, for every continuous latent module z required to generate $\mathbf{x}|\mathbf{d}$, the support of its parents under $\mathbf{d}$ is contained in the parent support observed under some training concept vector $\tilde{\mathbf{d}} \in \mathcal{D}_S$:

$$\text{supp}(\text{Pa}(z)|\mathbf{d}) \subseteq \text{supp}(\text{Pa}(z)|\tilde{\mathbf{d}}). \quad (4.3)$$

The supporting $\tilde{\mathbf{d}}$ may differ across modules. This is the formal version of analogy: one module can be borrowed from one training example and another module from a different training example.

This condition is less restrictive than requiring the full high-level combination to appear in training. It is also more expressive than additive or polynomial interaction assumptions. Additive models make composition possible by restricting the form of interaction. The hierarchical module view makes composition possible by requiring the relevant interactions to be supported somewhere in the training data. Sparse graphs help because each module has fewer parents, making Equation (4.3) easier to satisfy.

The proof is based on modular replacement. If every module required by $\mathbf{d}$ receives parent inputs that were observed under some supported configuration, then two models that agree on the source support must agree on the output of that module under the target configuration. Applying this argument through the hierarchy shows agreement on $p(\mathbf{x}|\mathbf{d})$. If a required module receives an unsupported parent configuration, the target conditional is not determined by the source support.

4.6 Identifying the Modules

The composition condition assumes access to the latent hierarchy. The second theoretical question is whether such a hierarchy can be learned from observed image-text data. The thesis uses component-wise identifiability: a learned latent variable $\hat{z}_i$ identifies a true latent variable $z_{\pi(i)}$ up to an invertible scalar transformation.

The identification assumptions are the familiar contracts from causal representation learning, adapted to a hierarchy. First, the observation must preserve the information in each latent level through an invertible map. Second, conditional densities between adjacent levels must be smooth. Third, components at a lower level are conditionally independent given the higher level, so dependence among siblings is explained by their parents. Fourth, the parent

variables must create sufficiently diverse changes in the distributions of their children. In derivative form, for each level there must be enough parent values to make the first- and second-derivative vectors of $\log p(\mathbf{z}_{l+1}|\mathbf{z}_l)$ linearly independent.

Key result: hierarchical module identifiability

Under invertibility, smooth density, conditional independence across variables at the same level, and sufficient variability from parent concepts, the latent variables $\mathbf{z}_l$ at every level are component-wise identifiable up to permutation within the level. Once the variables are identified, standard causal discovery reasoning can recover the graph structure among them up to the corresponding graphical equivalence.

The proof is recursive. Textual variables provide the first source of variation, which identifies the adjacent high-level latent concepts. Once those concepts are identified, they serve as auxiliary variables for identifying the next level. The argument repeats down the hierarchy until the low-level modules are identified. This matters for composition because subspace-level recovery is not enough. If “beak”, “tail”, and “rice interaction” are mixed in one block, the model cannot reliably recombine them as separate modules.

4.7 Hierarchical Conditioning for Diffusion Models

The theory motivates a text-to-image method with two design principles. First, conditioning should be hierarchical across diffusion steps. High-noise steps represent high-level layout and relation information; low-noise steps represent local object and attribute detail. Second, local concepts should interact sparsely, because unnecessary overlap between local modules undermines compositionality.

The method begins by decomposing a global prompt $\mathbf{y}$ into M local descriptions $\{\mathbf{y}_0^{(m)}\}_{m=1}^M$. In the experiments, a language model produces these descriptions at test time, while training uses a dataset with global and local descriptions. The language-model decomposer is not the main source of the gain: a simple heuristic decomposer obtains nearly the same score in the appendix analysis. Its role is to instantiate the decomposition step that the theory requires. At diffusion step t , the model interpolates between cross-attention from the global prompt and cross-attention from local prompts:

$$\mathbf{A}_t = (1 - s(t))\mathbf{A}_{T-1} + \frac{s(t)}{M} \sum_{m=1}^M \mathbf{A}_0^{(m)}, \quad (4.4)$$

Table 4.1. DPG-Bench prompt-following results. The method improves global, entity, attribute, relation, and other prompt-understanding metrics.

Method	Avg.	Global	Entity	Attribute	Relation	Other
SD v2	68.09	77.67	78.13	74.91	80.72	80.66
PixArt- α	71.11	74.97	79.32	78.60	82.57	76.96
Playground v2	74.54	83.61	79.91	82.67	80.62	81.22
ELLA	74.91	84.03	84.61	83.48	84.03	80.79
Ours	79.28	85.77	85.15	86.98	86.82	87.77

where $s(0) = 1$ and $s(T - 1) = 0$. Thus high-noise steps are guided by global conditioning, while low-noise steps inject local descriptions. The specific implementation uses a cosine schedule for $s(t)$.

To encourage sparse interactions, the method penalizes overlap among local attention maps:

$$\mathcal{L}_n = \sum_{m \neq n} D(\mathbf{H}_0^{(m)}, \mathbf{H}_0^{(n)}), \quad (4.5)$$

where D is a Dice-style overlap loss. The full objective is

$$\mathcal{L} = \mathcal{L}_d + \lambda \mathcal{L}_n, \quad (4.6)$$

with $\mathcal{L}_d$ the diffusion loss. The method is therefore a direct translation of the theory: decompose the prompt into modules, inject the right granularity at the right step, and discourage unnecessary module overlap.

4.8 Experimental Evidence

The main implementation fine-tunes from Stable Diffusion v1.5 and uses a frozen strong text encoder for text understanding. Training uses a layout-oriented dataset with global prompts and local descriptions. Evaluation focuses on complex prompt following with multiple objects, attributes, and relations. DPG-Bench measures this prompt following through global, entity, attribute, relation, and other-category scores; higher values indicate closer agreement between the requested and generated content.

Table 4.1 shows the main result. The method improves the average DPG-Bench score from ELLA’s 74.91 to 79.28 and improves every metric. The gains are strongest on attributes, relations, and the “Other” category, which contain many failures of binding and long-prompt

Table 4.2. Ablations on DPG-Bench. TD denotes time-dependent conditioning and SR denotes sparsity regularization.

Metric	w/o TD	w/o SR	Ours
Global	85.15 ± 2.26	83.09 ± 2.49	85.77 ± 1.21
Entity	84.95 ± 0.52	86.44 ± 0.58	85.15 ± 0.66
Attribute	86.31 ± 0.24	86.56 ± 0.82	86.98 ± 0.09
Relation	86.29 ± 1.03	87.14 ± 0.59	86.82 ± 0.86
Other	85.64 ± 0.90	85.56 ± 0.60	87.77 ± 1.45

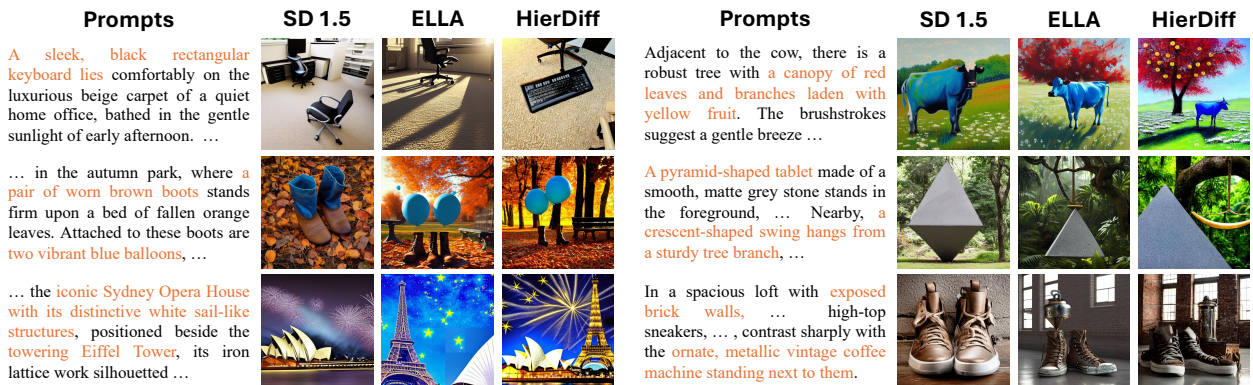


Figure 4.2. Representative text-to-image results. Step-wise conditioning and sparse local attention improve long, compositional prompt following.

following. The method also improves a second compositional generation benchmark, reaching 0.53 overall versus 0.45 for ELLA and 0.43 for Stable Diffusion 1.5.

Table 4.2 shows that both time-dependent conditioning and sparsity regularization contribute, though not uniformly on every metric. Removing time dependence weakens the hierarchical injection principle: the model uses a less structured conditioning signal across diffusion steps. Removing sparsity regularization weakens local control: attention maps can cover multiple objects or blend modules that should be separate. The qualitative results show the failure modes more clearly than the aggregate numbers.

Figure 4.2 illustrates the errors targeted by the method: missing objects, incorrect attribute binding, and collapsed relations. These are not generic image-quality failures. They are failures to assign the right module to the right concept and relation. Additional attention analyses show that local attention maps shift from dispersed global regions to more focused local regions over diffusion steps, which is consistent with the timestep-hierarchy interpretation. Sparse-autoencoder checks in the supporting analysis further support the claim that different diffusion steps correspond to different abstraction levels.

Evidence: Regime C mechanism

The mechanism is supported by three types of evidence. Benchmark gains show better compositional prompt following. Ablations show that hierarchical conditioning and sparse local interaction both matter. Qualitative and attention analyses show the specific failure modes addressed by the theory: object omission, attribute misbinding, relation collapse, and excessive overlap among local concept maps.

4.9 Extrapolation Beyond Source Support

The same regime includes prediction beyond support. In the extrapolation work, the target may be a single out-of-support sample rather than a full target distribution [22]. The latent model again separates an invariant variable c from a changing variable s :

$$\mathbf{x} = g(c, s), \quad \mathbf{y} = g_y(c). \quad (4.7)$$

The target sample is off support because s moves outside the source support, while c remains within the source content support. Extrapolation becomes the problem of identifying c from $\mathbf{x}$ even when s is out of support.

The theory distinguishes dense and sparse influence. If s can influence all observed dimensions, extrapolation requires class manifolds to be separable and the target shift to remain close enough to the source support. If s influences only a sparse subset of dimensions, then c can remain identifiable regardless of the distance of s from the source support. This result complements the composition theory. Both say that off-support behavior is possible only when the latent mechanism has structure. Dense unsupported changes are dangerous; sparse or modular changes can be handled.

4.10 From Composition to Interaction

Regime C connects the trustworthiness goals back to causal representation learning. Regime A identifies stable mechanisms across observed distributions. Regime B recovers latent concepts from unstructured data. Regime C asks when those mechanisms and concepts can be used beyond the observed support.

Extrapolation as sparse latent shift. The extrapolation work provides the classification-side theory for Regime C [22]. It asks whether invariant content can remain identifiable when changing factors move outside source support. The key lesson is that not all off-support shifts are equal. Dense shifts can destroy identifiability unless the target remains close enough to source support, while sparse shifts can leave invariant content recoverable even far away

from the training distribution. This gives the regime a diagnostic principle: before trusting extrapolation, identify which latent factors changed and whether their influence is sparse or dense.

Counterfactual text generation as local off-support control. The text-generation and counterfactual work from Regime A also belongs here from a control perspective [16]. A counterfactual text edit is a small off-support move in representation space: change tense, sentiment, or style while preserving content. The method’s relative-influence sparsity assumption is a language analogue of module support. A counterfactual is reliable only when the edited factor can be changed without dragging all other factors with it.

Hierarchical reasoning as supported refinement. Step-aware policy optimization extends the same question from visual modules to reasoning structure in diffusion language models [27]. Iterative denoising should follow incremental logical progress rather than refine all tokens without structure. This remains a Regime C question because the method organizes how supported reasoning steps are composed. The next regime asks the different question of how feedback can reveal reusable reasoning modules and expand the interfaces that training supports.

Together these results make Regime C a boundary condition for extrapolation. The point is not to celebrate arbitrary novelty. It is to make extrapolation conditional, modular, and diagnosable. A model should recombine when support exists. When support is missing, structure alone cannot supply the absent mechanism; the learner needs feedback or interaction that provides new evidence.

4.11 Limitations and Open Questions

The strongest limitation is support. The method can compose only what the latent module supports. If a prompt requires genuinely new physics, a new visual relation, or an interaction never constrained by data, the theory does not justify confident generation. This is a feature, not a bug. It marks the boundary between recombination and unsupported invention.

The second limitation is decomposition quality. The current method relies on a prompt decomposer and attention-based regularization. These are practical surrogates for latent modules, not direct proof that the full causal hierarchy has been recovered in a large diffusion model. Future work should build diagnostics that estimate whether a requested module is supported and whether the generated attention or representation actually localizes that module.

The third limitation is interaction density. Natural scenes and reasoning tasks can contain dense interactions. Sparse modules make composition easier, but not all mechanisms are

sparse. Sparsity therefore functions as a contract: when the interaction graph is sparse enough, composition is recoverable; when it is dense, the model needs more data, stronger supervision, or a warning that extrapolation is not justified.

Transition: from supported composition to interactive learning

Regime C establishes when a model may generalize beyond its training distribution by recombining mechanisms that the data already support. The same result also exposes a boundary: when a required module or interface has not been observed, compositional structure alone cannot justify extrapolation. Regime D asks how a learner can cross that boundary. Feedback and interaction can expose reusable modules, test whether their predictions are reliable, and collect new evidence when the current mechanism library is incomplete.

Chapter 5

Regime D: Learning Through Feedback and Interaction

The first three regimes ask what a learner can recover from a fixed collection of observations. Regime A uses changes across distributions, Regime B uses dependence within an observation, and Regime C uses modular support to decide which new combinations are justified. The support condition in Regime C also marks a limit. If the data never constrain a required module or interface, a model cannot justify the missing mechanism by recombination alone. This chapter studies how feedback can expand that boundary. Its central study uses reward-guided compositional reasoning because that setting makes the support-expansion mechanism unusually explicit: the learner generates traces, receives an outcome reward, and can be tested on known skills joined through new local interfaces. Two supporting studies then show how the same principle extends to environment interaction. A compositional world model (CWM) shares sparse dynamics across control tasks, while World Action Verifier (WAV) selects new interactions by checking plausible futures against action-conditioned predictions. Together, the studies turn a fixed collection of latent mechanisms into a library that can be identified, reused, checked, and expanded.

5.1 From Static Support to Interactive Discovery

Interaction in this chapter has a broad but concrete meaning. In language-model reasoning, the learner generates alternative traces and receives a task-level reward. In massively multitask control, the learner observes transitions across environments that reuse some objects, actions, and interaction rules. In world-model improvement, the learner compares a proposed future with an action-conditioned rollout and then executes one selected action in the environment. In all three cases, the learner does more than fit a fixed joint distribution. It changes which local events appear in its data.

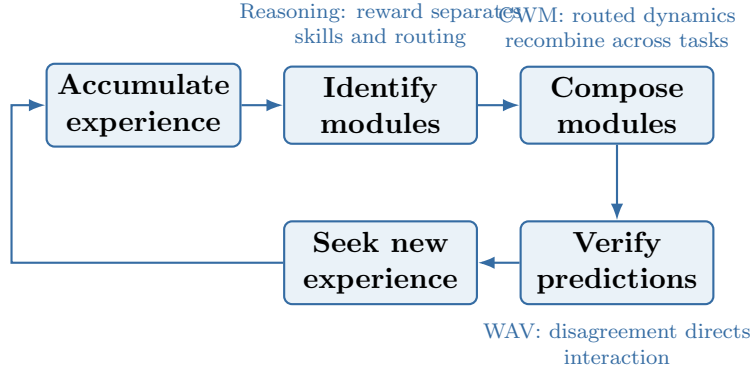


Figure 5.1. The feedback cycle that organizes Regime D. The reasoning work uses reward to identify modules inside generated traces, CWM composes routed and masked dynamics across tasks, and WAV uses verification error to select new interaction. Feedback is useful when it reveals a reusable choice or an unsupported prediction.

The common cycle is shown in Figure 5.1. Experience first exposes candidate modules. Structural assumptions separate the modules from their usual contexts. The learner composes the recovered pieces under a new task, checks whether the resulting prediction is credible, and seeks new experience when verification reveals a gap. The cycle does not guarantee that every mechanism can be discovered. It states what useful feedback must do: make a reusable module or a model failure distinguishable in the data.

Feedback as Support Expansion

Static data determine which modules and local interfaces are supported. Feedback can expand this set only when the learner can reach an informative event and the observed signal distinguishes that event from alternatives. Reward, task variation, and forward-inverse disagreement instantiate this principle in the three studies below.

5.2 Reward Feedback for Reusable Reasoning Modules

Post-training commonly combines supervised fine-tuning (SFT) with reinforcement learning (RL). SFT supplies correct reasoning traces, while RL samples model-generated traces and rewards successful answers. The combined recipe often transfers better to unfamiliar reasoning problems than supervised fine-tuning alone, but an accuracy gain does not explain what the two stages learn. The reasoning study asks a mechanism-level question: can feedback separate reusable local operations and information-routing choices that remain entangled inside canonical supervised traces [12]?

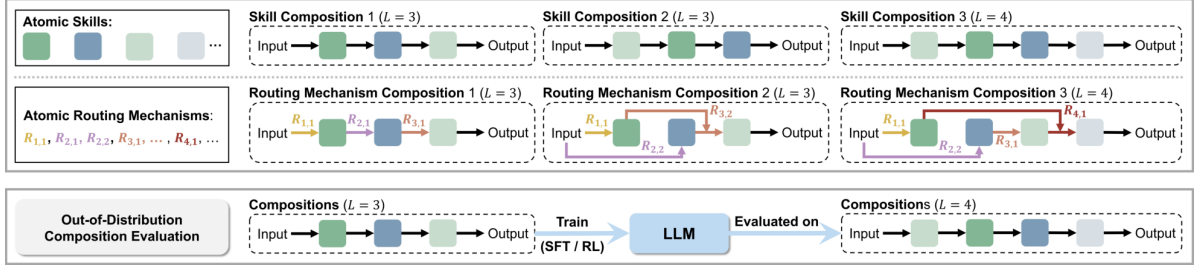


Figure 5.2. Atomic skills and routing mechanisms form two compositional axes. Evaluation changes the global skill arrangement while retaining locally familiar operations and interfaces. Adapted from Kong et al. [12].

5.2.1 Reasoning as Hierarchical Module Selection

Let P denote a problem descriptor and $D = (D_1, \dots, D_T)$ an observed reasoning trace. The descriptor summarizes the constraints in the problem statement. Between P and the trace lies a hierarchy of discrete latent selection variables

$$\mathbf{S} = [\mathbf{S}^{(1)}, \dots, \mathbf{S}^{(L)}]. \quad (5.1)$$

A realization of $\mathbf{S}$ specifies which reusable modules are selected and how they are connected. The observable conditional distribution is therefore generated through

$$P \longrightarrow \mathbf{S}^{(1)} \longrightarrow \dots \longrightarrow \mathbf{S}^{(L)} \longrightarrow D. \quad (5.2)$$

The hierarchy contains two kinds of modules. *Atomic skills* perform local operations, such as applying a string rewrite, substituting an expression, or comparing two quantities. *Routing mechanisms* determine which intermediate information is used by the next skill. A routing choice may forward the most recent result, retrieve an earlier expression, skip a branch, or combine two intermediate values. The distinction matters because two traces can use the same skills but connect them differently, and a model that memorizes a complete trace template has not necessarily recovered either type of reusable module. Figure 5.2 shows the two axes and the held-out composition test.

Let Ω_{supp} be the support of training descriptors. A larger set Ω_{comp} contains descriptors that form valid new combinations of attributes observed during training. Compositional generalization asks whether matching the true trace distribution on Ω_{supp} also determines the trace distribution on Ω_{comp} . The answer depends on whether the latent modules and the local interfaces needed by a new descriptor have been identified.

5.2.2 Identifying the Latent Selection Hierarchy

The identification result relies on local differences that connect a latent choice to an observable consequence. Three conditions give the main intuition. First, different module states need different observable signatures. Second, the valid choices in a local context must occupy a restricted subset of all combinations. Third, the data must show each module in enough distinct neighborhoods to separate it from its usual companions. A sorting operation, for example, is hard to distinguish from a fused “sort then select the median” template if the operation never appears before any other continuation.

Key result: identification of reasoning modules

Assume that distinct module states have distinct observable trace signatures, each local context permits a restricted set of module states, and the observed descriptors vary each module across enough neighboring contexts to distinguish it from fused alternatives. Under these conditions, the discrete latent selection variables and their adjacency structure are identifiable from $p(D | P)$, up to a separate relabeling of the states of each latent variable. Identification therefore recovers both the selected atomic modules and the local structure that composes them, within the stated relabeling ambiguity.

The result is conditional rather than automatic. If two local choices always produce the same observable continuation, their states cannot be separated. If two skills always occur together, a model can represent them as one fused template. The theorem turns trace diversity into an identification requirement: the data must contain local contrasts that isolate a module from its usual context. Appendix C.1 gives the formal statement and proof structure.

5.2.3 Why Supervised Traces Leave Hidden Support

Correct supervised traces supply the raw material from which modules could be learned. They show successful skills and routing choices in composition. A curated dataset often provides only one or a few canonical traces for each problem, however, so useful alternatives remain outside the observed trace support.

For a descriptor p , let $\mathcal{A}_{\text{SFT}}(p)$ be the subset of valid traces revealed by supervised fine-tuning, and let

$$s(p) = \sum_d p^*(d | p) \mathbf{1}\{d \in \mathcal{A}_{\text{SFT}}(p)\} \quad (5.3)$$

be the true probability mass covered by that subset. Renormalizing the true trace distribution inside $\mathcal{A}_{\text{SFT}}(p)$ produces a second model that agrees with every supervised observation

but differs on the hidden trace mass. For a prompt marginal $\mu^\star$ and the corresponding joint laws $Q^\star$ and $\tilde{Q}$, the difference is

$$\left\|Q^\star - \tilde{Q}\right\|_1 = 2 \mathbb{E}_{p \sim \mu^\star}[1 - s(p)]. \quad (5.4)$$

An identification-critical event that appears only outside the supervised set cannot be certified from supervised observations alone. This is a support limit, not a claim that supervised fine-tuning never learns useful modules.

5.2.4 How Reward Enriches Useful Local Events

Reinforcement learning can expose a hidden event when two conditions hold. Let

$$\mathcal{E} = \{U = u, \text{Ch}(U) = \mathbf{v}_{\text{Ch}(U)}\} \quad (5.5)$$

denote a latent module choice together with the child context that makes it meaningful. The current policy must reach $\mathcal{E}$ with probability strictly between zero and one. The task reward must also prefer traces containing the event:

$$\Delta_{\mathcal{E}}(p) = \Pr(R = 1 \mid \mathcal{E}, P = p) - \Pr(R = 1 \mid \mathcal{E}^c, P = p) > 0. \quad (5.6)$$

Under these conditions, successful rollouts contain the event more often than the pre-update policy:

$$\Pr(\mathcal{E} \mid R = 1, P = p) - \hat{p}_0(\mathcal{E} \mid P = p) > 0. \quad (5.7)$$

The enrichment magnitude is proportional to $\hat{p}_0(\mathcal{E} \mid p)(1 - \hat{p}_0(\mathcal{E} \mid p))\Delta_{\mathcal{E}}(p)$. Reward therefore separates a useful local choice only when the model sometimes generates it and the verifier can distinguish its value. An unreachable skill or an uninformative reward produces no such signal.

5.2.5 Local Witnesses for New Compositions

Identification alone does not guarantee generalization. A new problem can make familiar modules meet at an interface never observed during training. For a latent node U and child context $\mathbf{v}_{\text{Ch}(U)}$, write $p^\star(u, \mathbf{v}_{\text{Ch}(U)} \mid p)$ for the probability that $U = u$ and $\text{Ch}(U) = \mathbf{v}_{\text{Ch}(U)}$ under prompt p . The training-witnessed compatibility set is

$$\mathcal{W}_U(\mathbf{v}_{\text{Ch}(U)}) = \{u : \exists p \in \Omega_{\text{supp}}, p^*(u, \mathbf{v}_{\text{Ch}(U)} \mid p) > 0\}. \quad (5.8)$$

Key result: compositional generalization through local witnesses

Consider a new descriptor $p_{\text{new}} \in \Omega_{\text{comp}}$. If every latent family induced by the descriptor selects a parent value $u \in \mathcal{W}_U(\mathbf{v}_{\text{Ch}(U)})$, then the locally learned constraints can be composed consistently under the new descriptor. The condition requires training evidence for every local interface used by the new trace; it does not require the full high-level trace to have appeared before.

This result connects reward enrichment to generalization. Supervised traces expose a set of successful compositions, but they can omit local alternatives. When reward-guided rollouts reach and favor a hidden event, they enlarge the local witness sets. The learner can then reuse those local interfaces inside a new global trace. If two modules have never met in a compatible local context, the witness condition fails even when each module is individually familiar.

5.3 Empirical Evidence

The evidence proceeds from the common benchmark and measurements to four interventions that isolate module recovery, interface exposure, routing, and the relationship between supervised and reward-guided data.

5.3.1 Controlled Benchmark, Training Stages, and Measurements

The experiments use deterministic string transformations because every local operation, routing choice, and composition depth can be controlled directly. The benchmark contains 24 atomic skills and 10 routing mechanisms. A skill changes a string, while a router decides which earlier outputs become the next skill’s input. Nesting them produces a trace of depth L . This construction makes it possible to remove a known component, restore it through different forms of data, and test whether the component becomes reusable.

Training has two stages over the same problem family. During supervised fine-tuning (SFT), ten reasoning-inclusive responses are generated for each training problem, and only responses with the correct final answer are retained. All controlled experiments fine-tune Llama-3.1-8B-Instruct: SFT runs for two epochs, and RL uses on-policy rollout batches of 16 with a binary reward based only on whether the final transformed string is correct. Appendix Table C.1 gives the complete shared setup. Both stages can contain isolated modules or compound

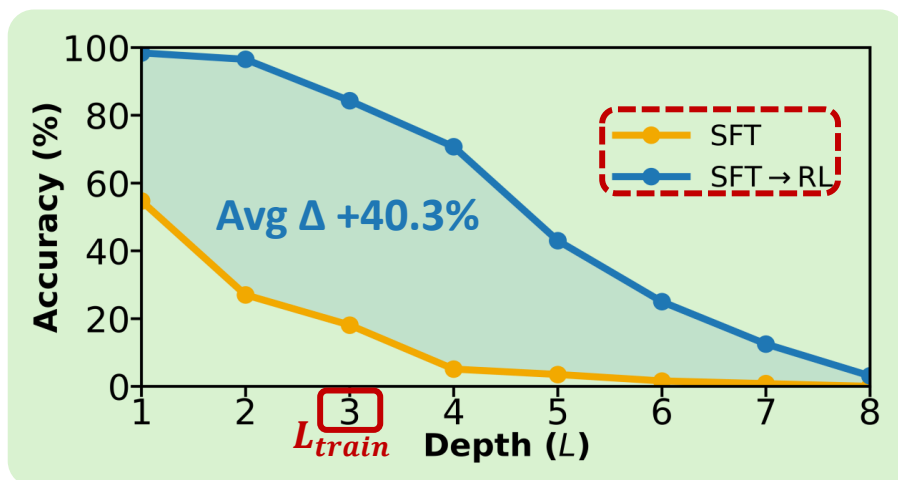


Figure 5.3. Final-string accuracy after depth-3 training. RL improves atomic skills and held-out compositions through $L = 8$, averaging a 40.3-point gain on the latter.

traces. Unless stated otherwise, the model trains on depth-3 compositions and the unseen split contains skill-and-router combinations absent from both SFT and RL. Evaluation reports exact final-string accuracy at depths $L = 1$ through $L = 8$. The reintroduction and SFT–RL relationship experiments report means over two independent random seeds, with shaded regions showing their empirical variation. Comparisons without bands are treated as controlled mechanism tests rather than precise estimates of population-level uncertainty.

5.3.2 Reward Recovers Modules and Extends Composition Depth

Figure 5.3 tests the module-decomposition claim. Both models see compound depth-3 traces, yet the SFT-plus-RL model is also better on the isolated skills at $L = 1$. RL adds 43.7 percentage points at that depth, which shows that outcome feedback can strengthen operations that were present only as parts of longer traces. Across unseen depths $L = 2$ through $L = 8$, the average gain is 40.3 percentage points. Accuracy declines as traces grow longer, so the result supports reuse of recovered modules without implying unlimited depth extrapolation. The seen-versus-unseen comparison in Figure 5.4 separates better fitting from support expansion. SFT reaches 78.2% on seen compositions and 18.1% on unseen ones. Adding RL raises these values to 99.1% and 84.3%, respectively. The unseen gain of 66.2 points is more than three times the 20.9-point seen gain. This contrast is evidence that generated, rewarded traces expose useful arrangements beyond the canonical supervised traces.

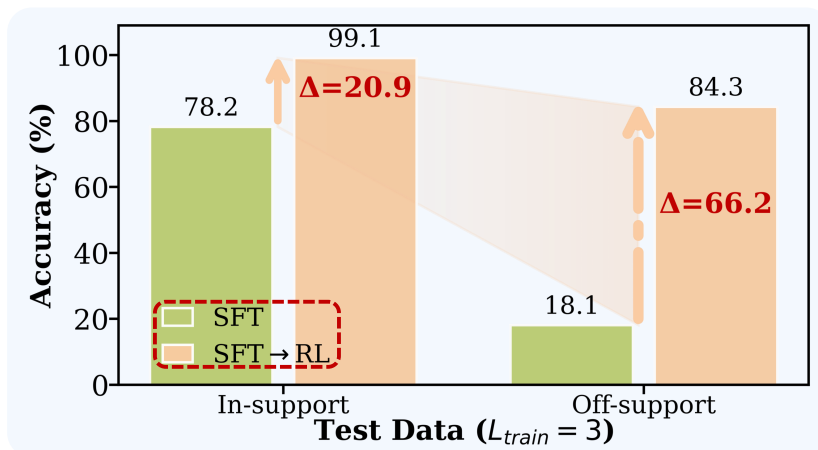


Figure 5.4. Seen and held-out composition accuracy. RL improves both groups, but the larger held-out gain indicates expansion beyond supervised trace support.

Table 5.1. Effect of reinforcement-learning data structure on unseen depth-4 compositions. Compound traces expose interfaces among modules, so they improve accuracy more than isolated atomic tasks under the same supervised initialization and update budget.

Training setting	Accuracy (%)	Gain <i>vs.</i> SFT
Supervised fine-tuning baseline	4.8	–
Reinforcement learning on atomic modules	14.8	10.0
Reinforcement learning on compound traces	42.6	37.8

5.3.3 Compound Rollouts Supply the Missing Interfaces

Table 5.1 holds the supervised initialization, the 300-step RL budget, and the depth-4 unseen evaluation fixed. RL on isolated atomic modules improves accuracy from 4.8% to 14.8%, but RL on depth-3 compound traces raises it to 42.6%. The 27.8-point advantage of compound over atomic RL supports the local-witness account: the learner needs examples of modules connected through interfaces, not only additional practice on each module separately.

5.3.4 Skills and Routing Require Compositional Exposure

The removal experiments make the interface requirement more direct. Eight of the 24 skills are excluded from the base RL compositions. The experiment then adds either isolated examples of a held-out skill or compound traces that use that skill. Figure 5.5 shows that isolated examples can recover the skill itself but do not restore deeper traces. Compound reintroduction restores substantially more accuracy because it shows how the held-out skill

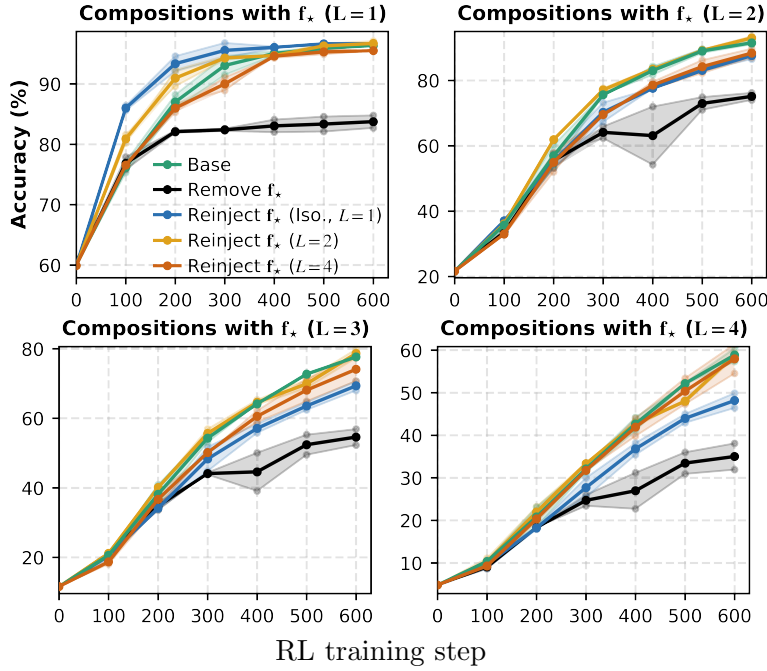


Figure 5.5. Held-out-skill accuracy. Compound-trace reintroduction restores deeper compositions more effectively than isolated depth-1 examples.

connects to neighboring modules. This is the empirical counterpart of enlarging the local witness set in Equation 5.8.

Routing choices show the same behavior. The study removes one of the ten routing mechanisms from RL compositions, restores it through lower-depth compositions, and evaluates held-out depth-3 and depth-4 traces. Figure 5.6 shows that removing the router lowers accuracy and that compositions containing the router recover performance. Lower-depth restoration is especially effective because it isolates the routing choice from unnecessary trace complexity. The parallel skill and routing results support one shared learning account: both local operations and information-flow choices become reusable through compositional exposure.

5.3.5 Reinforcement Learning Is Most Useful Outside Supervised Support

The last controlled intervention changes the relationship between the SFT and RL composition sets while retaining depth-3 training. The four settings make the RL set a subset, a superset, a partial overlap, or a disjoint set relative to the supervised compositions. Figure 5.7 reports seen accuracy, unseen accuracy, and depth extrapolation. The choice has

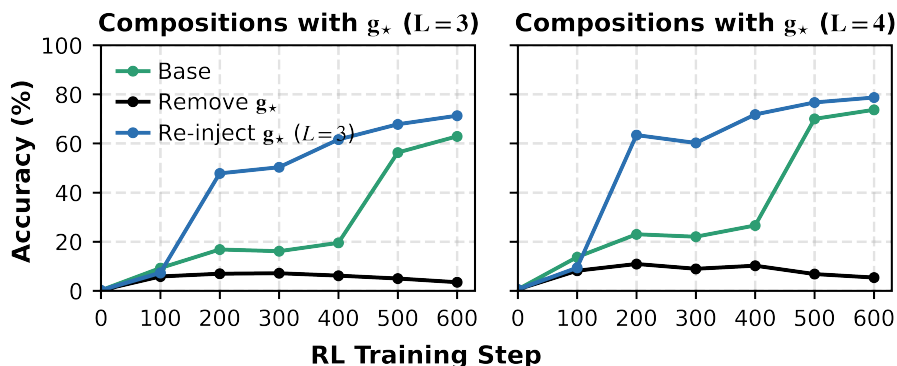


Figure 5.6. Accuracy after a routing mechanism is retained, removed, or reintroduced through lower-depth compositions. Reintroduction recovers unseen depth-3 and depth-4 performance, showing that routing and skill modules both require compositional exposure.

little effect on seen compositions, but it changes unseen performance substantially. RL restricted to compositions already covered by SFT performs worst on unseen combinations, while disjoint SFT and RL composition sets perform best. Gray, purple, cyan, and magenta show RL contained in SFT, SFT contained in RL, partial overlap, and disjoint sets, respectively. The result supports a concrete data-design rule: SFT should establish the atomic inventory through successful compound traces, while RL should explore compositions outside that supervised trace support.

Evidence: Reward separates modules when it changes local support

Supervised traces supply successful compound behavior. Reward-guided rollouts add a distinct signal only when they reach alternative local events and success depends on those events. Compound rollout data then improves both atomic recovery and unseen composition because it reveals reusable module interfaces, not merely isolated skills.

5.4 Limitations of the Anchor Study

The cleanest evidence comes from synthetic string transformations. These tasks make skills, routing, and composition depth observable, but natural reasoning contains ambiguous steps, multiple valid abstractions, and noisy evaluation. An open-source model analysis in Appendix C.2 shows compatible local recombination behavior rather than direct identification of the latent hierarchy. Longer training runs and sensitivity checks in the same appendix test whether the controlled findings persist beyond the main training budget.

The theory also places two strict requirements on reinforcement learning. The current policy must already reach an informative event, and reward must distinguish traces containing it.

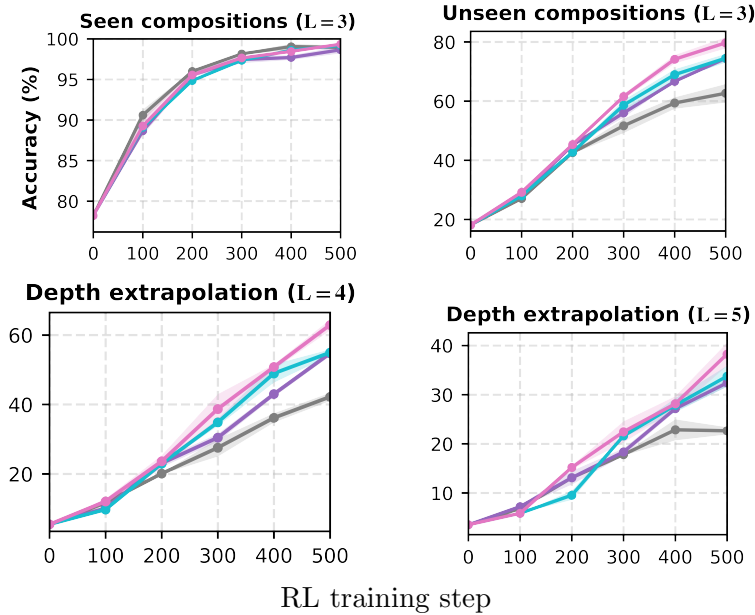


Figure 5.7. Accuracy under four SFT–RL composition-set relationships. Disjoint sets give the strongest unseen and depth generalization. Lines are two-seed means; shading shows variation.

Reinforcement learning cannot recover a skill that the model never generates, and a sparse or misspecified reward can enrich the wrong event. Finally, local witnesses guarantee only compositions whose interfaces have evidence in training. They do not justify a genuinely new operation or an interface absent from every rollout.

Transition: from reasoning traces to environment interaction

The anchor study shows how generated and rewarded traces enlarge the set of supported module interfaces. CWM and WAV extend this principle to sequential environments: one reuses sparse dynamics across tasks, while the other chooses new actions that expose a world-model error.

5.5 Supporting Extensions: Interactive World Models

5.5.0.1 Compositional world models.

A world model predicts how an environment changes after an action. The compositional world model (CWM) study asks what should be shared when one model learns more than 200 control tasks [13]. A dense task-conditioned transition can entangle unrelated objects, embodiments, and action effects. CWM instead divides the latent state into entities, routes each task to a small set of dynamics experts, and masks pairwise interactions. A target entity

therefore depends only on the task, entities, and interaction links needed for its next-state prediction.

This design extends the anchor study’s module-support principle. Reasoning skills and routers are reusable only after their local interfaces have been observed; CWM likewise reuses a dynamics link only after task variation has made that link visible. Its formal result separates support recovery on training tasks from reuse on a held-out combination: active links must vary enough to affect the target, inactive links must remain irrelevant, and the held-out task may combine only links and source values supported in training. The architecture and formal statement appear in Appendix C.3.

Under the same 100-million-step interaction budget as its NewT backbone, CWM raises the aggregate normalized score from 0.38 ± 0.03 to 0.51 ± 0.04 across ten domains and more than 200 tasks. It also adapts faster on held-out tasks that recombine familiar embodiments, objects, goals, and interactions. Appendix C.4 gives the per-domain learning curves, held-out adaptation curves, component-removal study, and per-domain numerical comparison. These results show wider reach for sparse module reuse, but they do not establish that every learned soft mask is a causal support.

5.5.0.2 World Action Verifier.

A learned forward model can be confidently wrong in regions with little data. WAV replaces direct self-evaluation with two smaller checks: whether a proposed future state is plausible, and whether an inverse model can recover an action that reaches it [14]. For current state s_t , WAV proposes plausible futures, infers an action for each one, predicts each action’s outcome with the forward model, and selects the candidate with the largest disagreement. The executed transition is then added to the training set. A learned feature mask limits the inverse model to the state coordinates most relevant to action.

This cycle relates to reward-based identification because both methods require a reachable event and a signal that distinguishes it. In the reasoning study, successful traces enrich a useful local choice. In WAV, disagreement between a plausible future and an action-conditioned prediction identifies a transition whose outcome can separate competing world-model hypotheses. The finite-sample analysis predicts the largest benefit when the action-relevant state is small, uncontrolled visual variation is large, and labeled interactions are scarce. Appendix C.5 gives the method diagram and formal result.

Across RoboMimic and ManiSkill, WAV reaches comparable prediction quality with approximately half as many labeled interactions as the strongest comparison methods. Separate MiniGrid tests show that the inverse verifier remains more accurate as labeled data shrink, scenes contain more objects, or observation noise increases. Under object and interaction

shifts, WAV’s selected data also improve world-model prediction and the reward of policies trained with imagined rollouts. Appendix C.6 reports the complete 200-sample table with standard errors, sample-efficiency curves, controlled MiniGrid tests, shifted-setting results, and policy results. The extension depends on the candidate generator and inverse verifier remaining trustworthy; action ambiguity, a shifted action-relevant feature set, or unsafe exploration can invalidate the acquisition signal.

5.6 From Reuse to Support Expansion

The three studies answer different parts of one question: how can a learner move beyond the support supplied by a fixed dataset without treating every new case as unconstrained extrapolation? Table 5.2 compares their answers.

Table 5.2. Three forms of feedback in Regime D.

Study	Reusable unit	Signal that expands support	Boundary or failure
Reasoning modules	Atomic skills and local routing interfaces	Reward enriches successful generated traces that contain a useful local event.	The policy must reach the event, reward must distinguish it, and every new interface needs a local witness.
Compositional world models	Task experts, entity dynamics, and pairwise interactions	Variation across many tasks exposes sparse source-target links that a shared interaction can recombine.	New links, source values outside training support, or a regime-specific extension remain unjustified.
World Action Verifier	Plausible subgoals and action-relevant transition features	Forward-inverse disagreement selects an action whose observed result adds informative transition data.	Action aliasing, verifier shift, candidate misspecification, and unsafe exploration can invalidate the signal.

The progression is deliberate. Reward changes the frequency of latent choices inside generated reasoning traces. Multitask experience reveals which dynamics components can be reused across environments. Verification chooses a new physical interaction when reuse is no longer enough. In each case, feedback is valuable because it creates a local contrast.

It separates a successful module from alternatives, an active dynamics link from irrelevant sources, or an unreliable prediction from a plausible and reachable future.

This view also clarifies what the chapter does not claim. Feedback does not remove support requirements. It changes them. A learner can expand support only through events it can reach, signals it can interpret, and structural assumptions that remain valid after expansion. Reward hacking, dense entangled dynamics, and a shifted verifier are three versions of the same failure: the observed signal no longer isolates the mechanism that the learner intends to add.

The combined framework also suggests four practical checks. First, measure *reachability*: can the current generator or policy produce the event whose effect must be learned? Second, measure *discrimination*: does reward, task variation, or verifier disagreement change when that event occurs? Third, measure *local coverage*: have all interfaces required by the proposed composition appeared under a compatible context? Fourth, measure *stability after update*: does adding the new evidence improve the target mechanism without erasing supported behavior elsewhere?

These checks make the static and interactive regimes continuous. Regime C tests whether a requested composition lies inside the envelope of supported modules. Regime D supplies methods for enlarging that envelope and for testing whether the enlargement was informative. The remaining challenge is to make the cycle safe, interpretable, and useful to people in open-ended physical and scientific settings.

Transition: from interactive learning to reflection

The four regimes now form one sequence. External variation reveals changing mechanisms, internal prediction reveals shared latent structure, modular support licenses new combinations, and feedback expands support when a required module or interface is missing. The final chapter evaluates what this sequence contributes, where its assumptions remain fragile, and how it can support reliable embodied agents and scientific discovery.

Chapter 6

Self-Criticism, Reflection, and Future Directions

This chapter steps back from the technical development and asks what the thesis contributes to the larger goal of trustworthy, human-empowering artificial intelligence.

6.1 Freeing and Empowering Humans

The grand goal of AI is to free and empower humans. This framing has deep roots in the history of computing. Licklider imagined computers as partners in formulative thinking rather than only machines for executing predetermined procedures [28]. Engelbart framed computing as a way to augment human intellect, giving people better tools for understanding complex situations and solving difficult problems [29]. Modern AI makes this vision newly concrete. It can automate routine work, compress knowledge, generate hypotheses, and coordinate perception, language, and action. Yet the same power raises the standard for reliability.

There are two especially important forms of human empowerment. The first is freedom from unwanted physical labor. Reliable embodied agents could help in manufacturing, agriculture, elder care, household work, hazardous cleanup, disaster response, and scientific laboratories. The second is expansion of the human knowledge horizon. AI systems could help scientists discover mechanisms in biology, medicine, materials, climate, and human behavior, not only by fitting predictive functions but by revealing interpretable principles that humans can understand, test, and build upon.

Both pursuits require trust. An embodied agent acts in the physical world, so a mistake can damage objects, injure people, or create unsafe cascades. A scientific discovery system acts in the epistemic world, so a mistake can produce false hypotheses, misleading explanations, or brittle correlations that look like mechanisms. The thesis argues that causal representation learning is a natural foundation for both pursuits because it studies the hidden mechanisms

behind observations. It aims to recover variables, modules, and dependencies that support transfer, interpretation, control, and calibrated generalization.

6.2 Gaps Between Current Models and Human Empowerment

The distance between this grand goal and current machine learning is still large. The papers in this thesis address important pieces of this distance, but they do not yet deliver fully reliable embodied agents or fully autonomous scientific discovery systems. The gaps below define the next scientific frontier.

6.2.1 Reliability in the Physical World

Current foundation models are powerful pattern learners, but embodied systems need more than pattern completion. They must perceive a changing world, infer physical affordances, choose actions, respect safety constraints, and recover from unexpected outcomes. The safety literature already identifies practical failure modes such as side effects, reward hacking, scalable supervision, unsafe exploration, and distribution shift [30]. Safe reinforcement learning studies related issues when policies must perform well while respecting constraints during learning and deployment [31].

Embodied AI makes these problems more urgent. A language model can hallucinate a procedure; a robot can execute it. A vision model can misread an object; a manipulator can move toward it with force. A planner can choose a shortcut; a mobile robot can enter a physically dangerous state. Projects such as SayCan, RT-1, RT-2, and Open X-Embodiment show the promise of combining language, vision, and action for real-world control [32–35]. They also expose the core tension: web-scale language and vision knowledge is abundant, but reliable action knowledge is sparse, embodiment-specific, and expensive to collect.

6.2.2 The Scarcity of Embodied Data

The success of language and vision models rests partly on the scale of internet data. Embodied data does not scale in the same way. Robot demonstrations require hardware, environments, safety checks, time, and often human supervision. Physical interaction cannot be copied as cheaply as text. Even large robot datasets cover only a tiny fraction of possible objects, tasks, morphologies, contacts, and failure modes.

This scarcity changes the learning problem. If reliable embodied agents must learn only by scaling static datasets, progress will be slow and brittle. The alternative is to learn beyond existing data through interaction, communication, and compositional reuse. A robot should

recognize when a task is a recombination of known mechanisms, when a human instruction supplies a missing constraint, and when exploration is needed to identify a new mechanism. This is exactly where the thesis principles point forward. Regime A uses environmental variation to separate stable and changing factors. Regime B uses objectives such as masking to recover latent structure without explicit labels. Regime C states when recombination beyond support is justified by reusable modules and when additional data is needed. Regime D then supplies three concrete ways to act on that boundary: reward can reveal useful reasoning interfaces, multitask experience can expose reusable dynamics, and forward-inverse verification can choose an interaction that resolves a world model failure.

6.2.3 Causal Interpretability for Scientific Discovery

The second grand direction is scientific discovery. AI has already changed the scientific landscape, with AlphaFold as a striking example of how learned models can transform a difficult scientific prediction problem [36]. Broader reviews describe AI as increasingly central to hypothesis generation, experiment design, simulation, and interpretation of scientific data [37]. Earlier work on robot scientists already showed that automated systems can generate and test hypotheses in functional genomics [38].

Yet prediction is not the same as discovery. A scientific discovery is valuable because it reveals a principle that humans can understand, test, communicate, and use in new settings. In many domains, the desired output is not merely a label, image, or ranking. It is an interpretable mechanism: which biological pathway regulates a phenotype, which latent physiological factor links modalities, which material property causes a behavior, or which intervention would change an outcome. Causality is the natural language for this goal. It connects variables to mechanisms, mechanisms to interventions, and interventions to counterfactual reasoning [7, 8, 39].

6.2.4 The Scarcity of Scientific Data

Scientific domains also face data scarcity. Biomedical measurements may be expensive, private, noisy, or collected under hidden selection bias. Materials experiments may be slow and costly. Climate and environmental systems cannot be randomized at will. Laboratory data often come from specialized instruments and protocols rather than broad internet-scale corpora. In such settings, simply scaling a black-box model may be impossible or scientifically unsatisfying.

The response is to understand the structure of the data more deeply. Which variables are shared across modalities? Which shifts are caused by instruments or environments? Which latent factors are sparse, modular, or hierarchical? Which interventions would distinguish competing mechanisms? Causal representation learning offers a framework for asking these questions in the language of identifiable variables, structured variation, conditional inde-

pendence, and support. This is why the biomedical, multimodal, and white-box generative-model papers are not side projects. They point toward AI systems that help humans discover interpretable world knowledge rather than only fit high-dimensional correlations.

6.2.5 Foundations and Remaining Gaps

The papers mostly study controlled regimes: multiple distributions with structured changes, masked objectives under hierarchical assumptions, compositional generation with explicit support conditions, generated reasoning traces with verifiable outcomes, and simulated control environments with known task boundaries. They do not yet solve safe embodied exploration, full scientific experiment design, or open-ended discovery of new causal mechanisms. They also rely on assumptions such as invertibility, sufficient variability, sparsity, modular support, informative reward, and a verifier that remains reliable under shift. All are approximations in real systems.

Nevertheless, the papers provide a strong foundation. They replace vague desires for interpretability, controllability, and generalization with mathematical contracts. They show how different data signals can identify different parts of a latent mechanism. They connect theory to algorithms and experiments across domain adaptation, masked representation learning, controllable generation, biomedical analysis, language editing, reasoning, and world-model learning. Regime D further shows how a learner can collect evidence that expands a supported mechanism library. The remaining task is not to abandon these contracts, but to extend them into richer settings where agents interact, communicate, and discover safely.

6.3 Assumptions and Diagnostics

Table 6.1 connects the assumptions used in the four technical chapters to practical checks. For each regime, it states the central contract and the evidence that should be examined before treating the learned representation as causally meaningful.

Table 6.1. Assumptions as contracts across the four regimes.

Regime	Contract	What to Diagnose
Learning from multiple distributions	Domain variation is informative, low-dimensional enough to satisfy minimal change, and sufficiently diverse to identify changing factors.	Vary the changing dimension, test domain alignment without label collapse, measure whether inferred changing factors track domain-specific changes rather than semantic content.

Table 6.1. Continued.

Regime	Contract	What to Diagnose
Learning with unstructured distributions	The observation is generated by a latent hierarchy, and masking makes useful shared latent variables necessary for reconstruction.	Compare low, intermediate, and extreme masks; examine reconstruction structure, attention span, shape bias, linear separability, and dense downstream transfer.
Generalization beyond training distributions	The target request can be decomposed into modules whose parent supports have appeared in training.	Check whether each requested module is supported, whether local attention or representations remain disentangled, and whether failure modes correspond to missing modules or dense unsupported interactions.
Learning through feedback and interaction	The learner can reach an informative event, the feedback distinguishes its effect, and the reusable module or verifier remains stable after the update.	Measure event reachability, reward or verification separation, local interface coverage, sparse dynamics support, and whether new data improve the target failure without erasing supported behavior.

Table 6.1 reframes the theory as practical diagnostic language. The table is intentionally modest. It does not claim that every real system satisfies the assumptions. It says what must be inspected before a causal interpretation becomes credible.

6.4 Foundational Contributions

Table 6.2 explains how the supporting papers extend the four main technical chapters. For each paper, it identifies its role in the thesis and the technical contribution that connects it to the broader argument.

Table 6.2. Supporting work by thesis role.

Work	Thesis Role	Technical Contribution
Multi-domain image generation [15]	Regime A, controllable generation	Uses minimal domain change to learn a joint distribution from marginal multi-domain image distributions, extending transfer-oriented partial identifiability to generation and translation.
Counterfactual text generation [16]	Regime A and Regime C, language control	Identifies content and style under domain-varying dependence using relative influence sparsity, enabling label-free interventions in text.
Biomedical multi-modal causal representation learning (CRL) [17]	Regime A, scientific mechanisms	Uses cross-modal causal sparsity to identify individual latent biomedical components rather than only shared subspaces, connecting CRL to interpretable scientific analysis.
SmartCLIP [18]	Regime A and Regime B, modular alignment	Treats image-caption mismatch as a latent identification problem and aligns only caption-relevant visual and textual modules through adaptive masking.
Discrete concept learning [19]	Regime B, concept hierarchy	Formalizes discrete concepts as latent hierarchical variables and gives identification conditions for concept relations from high-dimensional observations.
Nonlinear hierarchy identification [26]	Regime B, graph recovery	Provides identification criteria and an estimation procedure for nonlinear latent hierarchical causal models, supplying the graph-learning counterpart to objective-level analyses such as MAE.
White-box generative models [21]	Regime B, interpretation and control	Uses causal minimality to connect sparsity or compression constraints to component-wise identifiable concepts inside diffusion and language models.
Vision-language concepts [20]	Regime B and Regime C, grounded concepts	Identifies atomic textual and visual concepts and sparse cross-modal connections for controllable text-to-image generation and editing.

Table 6.2. Continued.

Work	Thesis Role	Technical Contribution
Extrapolation [22]	Regime C, prediction outside support	Characterizes when invariant content remains identifiable from one or a few off-support target samples, distinguishing dense from sparse latent shifts.
Diffusion reasoning [27]	Regime C, generative reasoning	Uses step-aware denoising to align diffusion language-model training with the latent structure of multistep reasoning.
Reasoning traces to reusable modules [12]	Regime D, reward feedback	Identifies atomic skills and local routing choices inside a hierarchical trace model, then shows how successful generated traces can enlarge the local support needed for unseen compositions.
Compositional world models [13]	Regime D, reusable dynamics	Combines sparse task routing with entity and interaction masks, improving one shared world model across more than 200 tasks and on held-out task recombinations.
World Action Verifier [14]	Regime D, targeted interaction	Uses plausible subgoals and a sparse inverse model to verify forward predictions, then collects interactions that improve world-model and policy performance under limited data.

The four main chapters carry the spine of the thesis, while the supporting papers widen the arc. Together they show the same causal principles moving across tasks: from adaptation to generation, from masked prediction to concept graphs, from compositional image generation to reasoning, and from reusable dynamics to targeted interaction.

6.5 Causal Foundations and the Role of Academic Research

Commercial laboratories now lead resource-intensive development of general-purpose AI models. Industry produced the majority of notable frontier models in 2025, while most AI models designed specifically for science came from academic and government institutions

[40]. I see this contrast as a reason to make foundational questions more central, not less. Larger empirical systems create powerful predictors and generators, but their scale does not by itself explain which learned features correspond to stable parts of the world or how those parts would respond to change. The research line behind this thesis studies that gap. Its organizing question is how high-dimensional observations can support the recovery of causal variables and mechanisms, rather than prediction alone.

Classical causal inference begins with variables whose scientific meaning is already specified. A causal model connects the distribution we observe to questions about interventions, which ask what happens when a variable is set, and counterfactuals, which compare what happened with what would have happened under another action [7, 8]. Causal discovery then uses observable constraints, such as conditional independence, to narrow the possible graph or its equivalence class [41]. The equivalence class records an important limit: observational evidence may determine some connections while leaving other directions unresolved. These traditions establish the language for mechanisms and the assumptions needed to reason about them. At the same time, they often start after variables such as a treatment, gene, or phenotype have already been represented at a scientifically meaningful level. Images, language, and multimodal measurements move that representation step inside the causal problem.

With high-dimensional observations, the learner must determine which hidden features should count as variables before it can discover how those variables are related. This problem is ambiguous without additional information. Many different latent coordinate systems can generate the same observations, and unsupervised objectives alone do not select the factors a scientist intended. The impossibility result in Locatello et al. [42], which was established earlier in the nonlinear ICA literature [43], makes this point precise for disentangled representations: recovery requires assumptions about the data or model. Causal representation learning takes the resulting challenge as a research program. It seeks high-level variables whose changes and dependencies can support reasoning about mechanisms, transfer, and interventions, while stating what information makes those variables recoverable [39]. The key shift is therefore from asking only whether a representation predicts well to asking what fixes its meaning and which remaining transformations the data cannot distinguish.

Several identification results map different ways to break this ambiguity. Nonlinear independent component analysis uses observed context, time, domain, labels, or other auxiliary information to create changes in hidden-factor distributions. Under explicit independence and variability conditions, these changes can separate factors that a single observational distribution would leave mixed [9, 44–46]. Auxiliary information is not the only route. Restrictions on the hidden distribution and on how the decoder maps hidden variables to

observations can also yield levels of recovery from observations alone [47, 48]. This does not contradict the earlier impossibility result. The results concern different model classes and impose different assumptions. Together they show that an identification claim must name both what breaks the ambiguity and what equivalence remains, whether that is a permutation, a component-wise change of coordinates, or a broader transformation. Theory is valuable here because it maps exactly which information or structural restriction supports each level of interpretation.

Identifying hidden variables is only one part of scientific mechanism discovery. A second task is to identify which hidden variables cause which others. These goals are distinct: a method may recover useful latent factors without determining their causal graph. Mixture structure can provide enough information to recover both latent representations and their causal relations under stated conditions [49]. Rank constraints offer a different signal for identifying linear latent hierarchies beyond simple latent trees [50], and related rank-based work allows hidden variables to occupy more general positions and to be causally related to one another [51]. My work extends this progression to nonlinear latent mechanisms and more general hierarchical graphs [26], then treats discrete concepts and their relations as identifiable variables in a hierarchy [19]. Each step enlarges the structures that can be recovered, but each also changes the assumptions and the ambiguity left by the data. Scientific use requires both parts: variables that correspond to meaningful factors and relations that express a testable mechanism among them.

Against this background, the thesis is organized by the signals that data make available rather than by a single model family. Distributional variation separates stable and changing mechanisms for transfer and control [9, 15]. Masking and internal dependence expose shared variables, while hierarchical structure makes their relations recoverable [10, 26]. Cross-modal sparsity can isolate physiological factors rather than only a shared subspace [17], and causal minimality links sparse components to interpretation and control in current generative models [21]. The later regimes use local support and feedback to mark when known mechanisms can be recombined and when new evidence is needed. Across these settings, the same discipline recurs: identify the data signal, state the recoverable object and remaining ambiguity, translate the condition into an algorithm, and diagnose when it fails. This recurring paradigm makes recovered mechanisms useful to existing AI systems in three concrete ways. It supports transferability by separating stable knowledge from environment-specific variation, so a model can carry useful structure across data domains or modalities. It supports interpretability by tying internal variables to meaningful factors and relations, helping users detect shortcuts and judge whether an explanation reflects the model’s learned structure. It supports controllability by isolating components whose modification should produce a local-

ized effect, enabling targeted text or image edits and more predictable reuse of learned skills. For scientific discovery, however, mechanisms are fundamental. Prediction can forecast an outcome or rank candidate molecules, but discovery requires an account of which variables and relations produce the outcome, how an intervention would change it, and which experiment could distinguish that account from competing explanations. This necessity became concrete for me while working with biomedical collaborators on multimodal causal representation learning [17]. Their central need was not another predictive score. They wanted identifiable physiological factors that connected measurements across modalities, so they could reason about a shared mechanism rather than only predict an outcome from each data source. That distinction changed how I understood the value of the theory. A model may predict accurately while using correlations that do not reveal which underlying variable is shared, which measurement reflects it, or what would happen if part of the system changed. For scientific work, those questions determine whether a learned representation can support an explanation and help shape the next experiment. My research on identifiability asks what information is sufficient to separate such factors and which ambiguities the data leave unresolved. In this sense, causality offers one foundation for moving from prediction toward mechanism: it provides language for variables, dependencies, and interventions, while the identification results state when those objects can be recovered from complex observations. It is not a complete solution. Domain knowledge is needed to decide which variables are scientifically meaningful; measurement design is needed to create informative variation; and experiments are needed to test whether a proposed mechanism survives intervention. I see the theory as useful because it makes these requirements visible and connects them to concrete learning problems, not because it removes the need for scientific judgment. This is the deeper reason I pursue causal hidden-variable research: it can help turn patterns in complex data into explanations that scientists can challenge, refine, and test.

My view is that academic research has a responsibility to protect fundamental, long-horizon inquiry whose value may not yet be commercial. Commercial large models can provide infrastructure for representing, generating, and simulating scientific data. Academia can build on that infrastructure while pursuing open theory about what can be recovered, developing independent evaluation of whether claimed mechanisms hold, and making assumptions and limits explicit. They can sustain public-interest and domain-specific questions that require close work with scientists but may not fit a near-term product. These roles are complementary: commercial systems can expand the scale and quality of empirical tools, while university researchers can study the causal layer that identifies variables, tests mechanisms, and guides experiments with domain scientists. Openness is important because assumptions, failure conditions, and competing explanations must be available for criticism and revision.

This position does not require universities to stand apart from frontier systems. It asks them to engage with those systems as research objects and collaborators while preserving questions whose answers may take years to mature. I place this thesis in that shared ecosystem, contributing conditional foundations for interpreting powerful models and using them in scientific discovery. Its role is to connect the empirical reach of modern systems with open, testable accounts of what their hidden variables and mechanisms mean.

6.6 Diagnostics for Trustworthy Extension

For embodied agents, the key diagnostic is whether a candidate action is supported by learned mechanisms and physical affordances. A model should know which object, relation, and skill modules are being reused, which assumptions about the scene are uncertain, and which missing observations require human communication or active exploration. CWM supplies a test for whether the needed dynamics links are supported, while WAV supplies a way to select new data when the forward prediction conflicts with a plausible and reachable future.

For scientific discovery, the key diagnostic is whether a learned representation corresponds to a testable mechanism. A useful system should separate predictive correlations from causal variables, connect latent factors to measurements or interventions, and propose experiments that can distinguish competing explanations.

For foundation models, the key diagnostic is support. A system should estimate whether a requested prediction, generation, or reasoning step lies inside a compositional envelope supported by data. When support is missing, the trustworthy behavior is not confident generation. It is abstention, request for clarification, targeted data collection, or safe exploration. The reasoning study adds a further check: generated traces help only when reward reliably enriches the local skills or routing interfaces needed by the new problem.

6.7 Future Research: A Causal Loop for AI-Assisted Scientific Discovery

My future research will develop AI systems that participate in scientific discovery by organizing and refining causal explanations, rather than only predicting from accumulated observations. Scientific data, scientific insight, and experiments form a coupled loop: measurements suggest possible mechanisms, mechanisms guide the next interventions, and experimental outcomes reshape what should be measured and believed. A causal model is the shared representation at the center of this loop. It can connect heterogeneous observations to scientific variables, record uncertainty and competing explanations, and make explicit which

interventions would distinguish them.

The four regimes in this thesis provide complementary principles for building such a system. Variation across environments helps separate stable mechanisms from changing context. Structured prediction within high-dimensional data reveals reusable variables. Compositional support determines when familiar pieces can justify a novel hypothesis. Feedback and interaction supply new evidence when passive data leave a mechanism unresolved. The following three research thrusts develop these principles in parallel, with the causal model carrying information among them.

6.7.1 Research Thrust I: From Multimodal Data to Causal Scientific Insights

Modern science increasingly produces images, molecular sequences, time series, text, and instrument records at a scale that no single investigator can inspect directly. Large pre-trained models offer a powerful interface to these data: they can turn raw measurements into semantic features that already capture objects, structures, and relations. Yet semantic usefulness alone does not establish a scientific variable or a causal mechanism. I plan to place causal representation learning above these large-model features so that the resulting variables are organized by how they change across datasets, sites, instruments, and experimental conditions.

This direction extends Regime A’s use of distributional variation and my work on multimodal biomedical causal representation learning [17]. The larger goal is to transform diverse scientific records into causal maps that scientists can inspect, challenge, and refine. A practical step toward that goal will be an open-source platform for causal scientific-data analysis. Rather than presenting one learned graph as a final answer, the platform will expose candidate variables, relationships, assumptions, and uncertainty, allowing feedback from practitioners to become part of the learning process.

6.7.2 Research Thrust II: From Causal Insights to Informative Experiments

Once scientific variables and relationships have been organized, a causal model should do more than summarize existing findings. It should help imagine which hypotheses, designs, and interventions are worth pursuing next. Many scientific search spaces are combinatorial, and their most valuable candidates may join familiar parts in arrangements that have never been observed. My work on compositional generation and reward-guided reasoning [11, 12] suggests a way to reduce this complexity: recover reusable components, represent

the conditions under which they can be combined, and extrapolate only where those local relationships are supported.

Protein and drug design provide a compelling example, although the principle is broader. Generative models could simulate large numbers of new proteins, molecules, materials, or interventions, while a causal model organizes those possibilities by feasibility, predicted effect, and the scientific uncertainty their outcomes could resolve. The long-term ambition is a causal generative system that does not merely produce plausible candidates. It distinguishes novel combinations that reuse established mechanisms from those whose value depends on an unresolved interaction, then directs scarce experimental effort toward the most consequential unknowns.

6.7.3 Research Thrust III: From Experiments to Targeted Scientific Data

Scientific explanations ultimately become useful by surviving interventions. The third thrust therefore closes the loop: use the current causal model to identify experiments whose outcomes would separate competing explanations, collect the corresponding data, and revise both the variables and their causal relationships. This is a scientific form of the support-expansion principle in Regime D. Reward-guided reasoning shows that feedback is informative when it distinguishes a reachable local event; CWM and WAV extend the same idea to reusable dynamics and targeted interaction [13, 14].

The ambition is to carry this principle into real discovery settings, where experiments are costly and scientific judgment cannot be reduced to an acquisition score. Collaborations with experimental scientists will contribute domain knowledge, measurement practice, feasibility constraints, and the interpretation of unexpected results. Experimental outcomes will then return to the causal model, reshape the representations built in the first thrust, and focus the generative search in the second. In this way, data collection is not the final stage of a pipeline but the mechanism by which the entire system continues to learn.

The three thrusts will proceed in parallel through collaborations within and across departments. The open-source platform will provide common infrastructure, while real scientific partnerships will reveal which causal assumptions are useful, which abstractions remain understandable, and which experiments can create genuine impact. The goal is not an autonomous scientist that replaces human judgment. It is a collaborative discovery system that makes causal reasoning computationally scalable, empirically revisable, and easier to connect across data, models, and experiments.

Transition: from reflection to scientific discovery

The thesis begins with passive prediction and ends with a coupled discovery loop. Multimodal data inform causal models, causal models organize hypotheses and experiments, and experimental evidence refines both the model and the next collection of data. Open tools and scientific collaboration make this loop a shared research program rather than a closed automated pipeline.

6.8 Conclusion

The thesis advances one claim across four data regimes: trustworthy learning depends on recovering mechanisms at the level justified by the available signal. Distribution changes identify what varies and what remains invariant. Masked prediction identifies shared latent structure within an observation. Modular support determines which unseen combinations are licensed. Feedback and interaction enlarge that support by revealing useful reasoning choices, reusable dynamics, or missing action effects.

This claim is deliberately conditional. Every identification or generalization result comes with a contract about variation, hierarchy, sparsity, support, reward, or verification. The experiments show that these contracts can guide effective algorithms, but they do not turn assumptions into facts. A reliable system must test the contract it relies on, expose when the test fails, and seek new evidence only when doing so is safe and informative.

The resulting research direction joins causal representation learning with interactive learning. Its goal is not simply a larger predictive model. It is a learner that knows which mechanisms it has recovered, which compositions are supported, which failures require more evidence, and how a new observation changes that boundary. Such a learner would be a stronger foundation for both embodied agents that act responsibly and scientific systems that help people discover understandable principles.

Bibliography

- [1] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2022. URL <https://arxiv.org/abs/2109.07958>.
- [2] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.findings-emnlp.301. URL <https://aclanthology.org/2020.findings-emnlp.301/>.
- [3] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. *arXiv preprint arXiv:2012.07805*, 2021. URL <https://arxiv.org/abs/2012.07805>.
- [4] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023. URL <https://arxiv.org/abs/2307.15043>.
- [5] Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don’t always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract-Conference.html.
- [6] Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR, 2023. URL <https://proceedings.mlr.press/v202/shi23a.html>.
- [7] Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009.
- [8] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.

- [9] Lingjing Kong, Shaoan Xie, Weiran Yao, Yujia Zheng, Guangyi Chen, Petar Stojanov, Victor Akinwande, and Kun Zhang. Partial disentanglement for domain adaptation. In *Proceedings of the 39th International Conference on Machine Learning*, pages 11455–11472, 2022.
- [10] Lingjing Kong, Martin Q. Ma, Guangyi Chen, Eric P. Xing, Yuejie Chi, Louis-Philippe Morency, and Kun Zhang. Understanding masked autoencoders via hierarchical latent variable models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
- [11] Lingjing Kong, Shaoan Xie, Yang Jiao, Yetian Chen, Yanhui Guo, Simone Shao, Yan Gao, Yi Sun, Eric P. Xing, Guangyi Chen, and Kun Zhang. Understanding compositional generalization via hierarchical concept models, 2026. Preprint.
- [12] Lingjing Kong, Xin Liu, Guangyi Chen, Martin Q. Ma, Xiangchen Song, Yuekai Sun, Mikhail Yurochkin, Taylor W. Killian, Ruslan Salakhutdinov, Kun Zhang, Eric P. Xing, and Zhengzhong Liu. From reasoning traces to reusable modules: Understanding compositional generalization in language model reasoning. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL <https://arxiv.org/abs/2606.18089>.
- [13] Fan Feng, Lingjing Kong, Kevin Murphy, and Kun Zhang. Learning modular world models by composing reusable dynamics. Under review. Fan Feng and Lingjing Kong co-led the project and contributed equally, 2026.
- [14] Yuejiang Liu, Fan Feng, Lingjing Kong, Weifeng Lu, Jinzhou Tang, Kun Zhang, Kevin Murphy, Chelsea Finn, and Yilun Du. World action verifier: Self-improving world models via forward-inverse asymmetry, 2026. URL <https://arxiv.org/abs/2604.01985>.
- [15] Shaoan Xie, Lingjing Kong, Mingming Gong, and Kun Zhang. Multi-domain image generation and translation with identifiability guarantees. In *International Conference on Learning Representations*, 2023.
- [16] Hanqi Yan, Lingjing Kong, Lin Gui, Yuejie Chi, Eric Xing, Yulan He, and Kun Zhang. Counterfactual generation with identifiability guarantees. In *Advances in Neural Information Processing Systems*, 2023.
- [17] Yuewen Sun, Lingjing Kong, Guangyi Chen, Loka Li, Gongxu Luo, Zijian Li, Yixuan Zhang, Yujia Zheng, Mengyue Yang, Petar Stojanov, Eran Segal, Eric P. Xing, and Kun Zhang. Causal representation learning from multimodal biomedical observations. In *International Conference on Learning Representations*, 2025.

- [18] Shaoan Xie, Lingjing Kong, Yujia Zheng, Yu Yao, Zeyu Tang, Eric P. Xing, Guangyi Chen, and Kun Zhang. Smartclip: Modular vision-language alignment with identification guarantees. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.
- [19] Lingjing Kong, Guangyi Chen, Biwei Huang, Eric Xing, Yuejie Chi, and Kun Zhang. Learning discrete concepts in latent hierarchical models. *Advances in Neural Information Processing Systems*, 37:36938–36975, 2024.
- [20] Shaoan Xie, Lingjing Kong, Yujia Zheng, Zeyu Tang, Eric P. Xing, Guangyi Chen, and Kun Zhang. Learning vision and language concepts for controllable image generation. In *International Conference on Machine Learning*, 2025.
- [21] Lingjing Kong, Shaoan Xie, Guangyi Chen, Yuewen Sun, Xiangchen Song, Eric P. Xing, and Kun Zhang. Beyond the black box: Identifiable interpretation and controllability in generative models via causal minimality, 2025. Preprint.
- [22] Lingjing Kong, Guangyi Chen, Petar Stojanov, Haoxuan Li, Eric P. Xing, and Kun Zhang. Towards understanding extrapolation: A causal lens. In *Advances in Neural Information Processing Systems*, 2024.
- [23] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5542–5550, 2017.
- [24] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027, 2017.
- [25] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021.
- [26] Lingjing Kong, Biwei Huang, Feng Xie, Eric P. Xing, Yuejie Chi, and Kun Zhang. Identification of nonlinear latent hierarchical models. In *Advances in Neural Information Processing Systems*, 2023.
- [27] Shaoan Xie, Lingjing Kong, Xiangchen Song, Xinshuai Dong, Guangyi Chen, Eric P. Xing, and Kun Zhang. Step-aware policy optimization for reasoning in diffusion large language models, 2026. Preprint.
- [28] J. C. R. Licklider. Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics*, HFE-1(1):4–11, 1960.

- [29] Douglas C. Engelbart. Augmenting human intellect: A conceptual framework. Technical Report AFOSR-3223, Stanford Research Institute, 1962.
- [30] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety, 2016.
- [31] Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(42):1437–1480, 2015.
- [32] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jau-regui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jodilyn Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as i can, not as i say: Grounding language in robotic affordances, 2022.
- [33] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Keerthana Gopalakrishnan, Suyog Han, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, et al. Rt-1: Robotics transformer for real-world control at scale, 2022.
- [34] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Keerthana Gopalakrishnan, Suyog Han, Karol Hausman, Alex Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Eric Jang, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023.
- [35] Open X-Embodiment Collaboration et al. Open x-embodiment: Robotic learning datasets and rt-x models, 2023.
- [36] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583–589, 2021. doi: 10.1038/s41586-021-03819-2.

- [37] Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, Anima Anandkumar, Karianne Bergen, Carla P. Gomes, Shirley Ho, Pushmeet Kohli, Joan Lasenby, Jure Leskovec, Tie-Yan Liu, Arjun Manrai, Debora Marks, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620:47–60, 2023. doi: 10.1038/s41586-023-06221-2.
- [38] Ross D. King, Kenneth E. Whelan, Ffion M. Jones, Philip G. K. Reiser, Christopher H. Bryant, Stephen H. Muggleton, Douglas B. Kell, and Stephen G. Oliver. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427: 247–252, 2004. doi: 10.1038/nature02236.
- [39] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954.
- [40] Stanford Institute for Human-Centered Artificial Intelligence. The 2026 AI index report. Technical report, Stanford University, 2026. URL <https://hai.stanford.edu/ai-index/2026-ai-index-report>.
- [41] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.
- [42] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019.
- [43] Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999. ISSN 0893-6080. doi: 10.1016/S0893-6080(98)00140-3. URL <https://www.sciencedirect.com/science/article/pii/S0893608098001403>.
- [44] 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.
- [45] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International conference on artificial intelligence and statistics*, pages 2207–2217. PMLR, 2020.
- [46] Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: A general setting. *arXiv preprint arXiv:2402.05052*, 2024.

- [47] Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Identifiability of deep generative models without auxiliary information. *Advances in Neural Information Processing Systems*, 35:15687–15701, 2022.
- [48] Yujia Zheng and Kun Zhang. Generalizing nonlinear ica beyond structural sparsity. *Advances in Neural Information Processing Systems*, 36:13326–13355, 2023.
- [49] Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Learning latent causal graphs via mixture oracles. *Advances in Neural Information Processing Systems*, 34:18087–18101, 2021.
- [50] Biwei Huang, Charles Low, Feng Xie, Clark Glymour, and Kun Zhang. Latent hierarchical causal structure discovery with rank constraints. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL <https://openreview.net/forum?id=IeuKiTZsLY>.
- [51] Xinshuai Dong, Biwei Huang, Ignavier Ng, Xiangchen Song, Yujia Zheng, Songyao Jin, Roberto Legaspi, Peter Spirtes, and Kun Zhang. A versatile causal discovery framework to allow causally-related hidden variables. *arXiv preprint arXiv:2312.11001*, 2023.

- [52] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aur'elien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Cantón Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab A. AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriele Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guanglong Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Ju-Qing Jia, Kalyan Vasuden Alwala, K. Upasani, Kate Plawiak, Keqian Li, Ken-591 neth Heafield, Kevin R. Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuen-ley Chiu, Kunal Bhalla, Lauren Rantala-Yearly, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melissa Hall Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri S. Chatterji, Olivier Duchenne, Onur cCelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasić, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ron nie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sa hana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Chandra Raparthi, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuwei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yiqian Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zhengxu Yan, Zhengxing Chen, Zoe Papakipos, Aaditya K. Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adi Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein,

Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Po-Yao (Bernie) Huang, Beth Loyd, Beto de Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Shangwen Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkan Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzm'an, Frank J. Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory G. Sizov, Guangyi Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Han Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kaixing(Kai) Wu, U KamHou, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, A Lavender, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthias Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Sathnam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollár, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Kumar Gupta, Sung-Bae Cho, Sunny Virk, Suraj Subrama-

nian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Andrei Poenaru, Vlad T. Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xia Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models. *ArXiv*, abs/2407.21783, 2024.

- [53] Lifan Yuan, Weize Chen, Yuchen Zhang, Ganqu Cui, Hanbin Wang, Ziming You, Ning Ding, Zhiyuan Liu, Maosong Sun, and Hao Peng. From $f(x)$ and $g(x)$ to $f(g(x))$: Llms learn new skills in rl by composing old ones. *arXiv preprint arXiv:2509.25123*, 2025.

Appendix A

Formal Results and Proofs

This appendix collects the formal statements and proofs that support the theoretical results used in the main text. The material is adapted from the source papers for the first three regimes, with notation harmonized to the thesis. Regime D results are collected separately in Appendix C.

A.1 Regime A: Partial Identifiability for Transfer

Regime A follows the latent data-generating process used by Kong et al. [9]. Observations are generated as

$$\mathbf{z}_c \sim p_{\mathbf{z}_c}, \quad \mathbf{z}_s \sim p_{\mathbf{z}_s|\mathbf{u}}, \quad \mathbf{x} = g(\mathbf{z}_c, \mathbf{z}_s), \quad (\text{A.1})$$

where g is smooth and injective, $\mathbf{z}_c \in \mathcal{Z}_c$ is invariant across domains, and $\mathbf{z}_s \in \mathcal{Z}_s$ changes with the domain index $\mathbf{u}$. A learned model $(\hat{g}, p_{\hat{\mathbf{z}}_c}, p_{\hat{\mathbf{z}}_s|\mathbf{u}})$ is assumed to match the conditional observation distribution in every domain:

$$p_{\mathbf{x}|\mathbf{u}}(\mathbf{x}'|\mathbf{u}') = p_{\hat{\mathbf{x}}|\mathbf{u}}(\mathbf{x}'|\mathbf{u}'), \quad \forall \mathbf{x}' \in \mathcal{X}, \mathbf{u}' \in \mathcal{U}. \quad (\text{A.2})$$

Theorem A.1 (Changing-part identifiability). *Assume the process in Equation A.1 and the matching condition in Equation A.2. Suppose:*

- the latent density $p_{\mathbf{z}|\mathbf{u}}$ is smooth and strictly positive on $\mathcal{Z} \times \mathcal{U}$;
- conditioned on $\mathbf{u}$, the latent components are mutually independent, so $\log p_{\mathbf{z}|\mathbf{u}}(\mathbf{z}|\mathbf{u}) = \sum_i q_i(z_i, \mathbf{u})$;
- for every $\mathbf{z}_s \in \mathcal{Z}_s$, there exist $2n_s + 1$ domain values $\mathbf{u}_0, \dots, \mathbf{u}_{2n_s}$ such that the $2n_s$ vectors $\mathbf{w}(\mathbf{z}_s, \mathbf{u}_j) - \mathbf{w}(\mathbf{z}_s, \mathbf{u}_0)$, $j = 1, \dots, 2n_s$, are linearly independent, where

$$\mathbf{w}(\mathbf{z}_s, \mathbf{u}) = \left(\frac{\partial q_{n_c+1}(z_{n_c+1}, \mathbf{u})}{\partial z_{n_c+1}}, \dots, \frac{\partial q_n(z_n, \mathbf{u})}{\partial z_n}, \frac{\partial^2 q_{n_c+1}(z_{n_c+1}, \mathbf{u})}{\partial z_{n_c+1}^2}, \dots, \frac{\partial^2 q_n(z_n, \mathbf{u})}{\partial z_n^2} \right). \quad (\text{A.3})$$

Then the changing latent variables $\mathbf{z}_s$ are component-wise identifiable: there is a permutation π and invertible scalar functions h_i such that $\hat{z}_{s,i} = h_i(z_{s,\pi(i)})$.

Proof. The proof follows the source proof in Kong et al. [9]. Let $h = g^{-1} \circ \hat{g}$ be the smooth invertible map from estimated latents to true latents. Matching the observation distribution gives, for every $\mathbf{u}$,

$$p_{\mathbf{z}|\mathbf{u}}(\mathbf{z}|\mathbf{u}) = p_{\hat{\mathbf{z}}|\mathbf{u}}(h^{-1}(\mathbf{z})|\mathbf{u}) |\det \mathbf{J}_{h^{-1}}(\mathbf{z})|. \quad (\text{A.4})$$

Taking logarithms and writing $q_i = \log p_{z_i|\mathbf{u}}$ and $\hat{q}_i = \log p_{\hat{z}_i|\mathbf{u}}$ yields

$$\sum_{i=1}^n q_i(z_i, \mathbf{u}) + \log |\det \mathbf{J}_h| = \sum_{i=1}^n \hat{q}_i(\hat{z}_i, \mathbf{u}). \quad (\text{A.5})$$

Differentiate Equation A.5 twice with respect to $\hat{z}_k$ and $\hat{z}_q$ for $k \neq q$. With $h'_{i,(k)} = \partial z_i / \partial \hat{z}_k$ and $h''_{i,(k,q)} = \partial^2 z_i / (\partial \hat{z}_k \partial \hat{z}_q)$, the mixed derivative equation is

$$\sum_{i=1}^n \left[\frac{\partial^2 q_i}{\partial z_i^2} h'_{i,(k)} h'_{i,(q)} + \frac{\partial q_i}{\partial z_i} h''_{i,(k,q)} \right] + \frac{\partial^2 \log |\det \mathbf{J}_h|}{\partial \hat{z}_k \partial \hat{z}_q} = 0. \quad (\text{A.6})$$

Subtracting the equation for $\mathbf{u}_0$ from the equations for $\mathbf{u}_1, \dots, \mathbf{u}_{2n_s}$ eliminates the invariant content terms, because $p_{\mathbf{z}_c}$ does not vary across domains. The remaining $2n_s$ equations involve only the changing coordinates $i = n_c + 1, \dots, n$. By the linear independence assumption, the only solution is

$$h'_{i,(k)} h'_{i,(q)} = 0, \quad h''_{i,(k,q)} = 0, \quad i = n_c + 1, \dots, n, \quad k \neq q. \quad (\text{A.7})$$

Thus each true changing coordinate can depend on at most one estimated coordinate. Since h is invertible, the Jacobian block mapping estimated changing variables to true changing variables must have exactly one nonzero entry in each changing row. Invariant estimated variables have domain-invariant marginals, whereas the true changing variables vary with $\mathbf{u}$, so a true changing coordinate cannot depend on an estimated invariant coordinate. Hence each true changing variable corresponds to one estimated changing variable up to permuta-

tion and an invertible scalar transformation. $\square$

Theorem A.2 (Invariant-block identifiability). *Assume the process in Equation A.1, the matching condition in Equation A.2, and the first two assumptions of Theorem A.1. Suppose additionally that for any measurable set $A_{\mathbf{z}} \subseteq \mathcal{Z}$ with nonzero probability that cannot be written as $B_{\mathbf{z}_c} \times \mathcal{Z}_s$, there exist two domain values $\mathbf{u}_1, \mathbf{u}_2 \in \mathcal{U}$ such that*

$$\int_{\mathbf{z} \in A_{\mathbf{z}}} p_{\mathbf{z}|\mathbf{u}}(\mathbf{z}|\mathbf{u}_1) d\mathbf{z} \neq \int_{\mathbf{z} \in A_{\mathbf{z}}} p_{\mathbf{z}|\mathbf{u}}(\mathbf{z}|\mathbf{u}_2) d\mathbf{z}. \quad (\text{A.8})$$

Then $\mathbf{z}_c$ is block-wise identifiable: there exists an invertible map $h_c : \mathcal{Z}_c \rightarrow \mathcal{Z}_c$ such that $\hat{\mathbf{z}}_c = h_c(\mathbf{z}_c)$.

Proof. The proof follows the four-step source proof in Kong et al. [9]. Define $\bar{h} = \hat{g}^{-1} \circ g$, so $\hat{\mathbf{z}} = \bar{h}(\mathbf{z})$, and write $\bar{h}_c$ for its content block. Because the learned content distribution is invariant to $\mathbf{u}$, for every measurable $A_{\mathbf{z}_c} \subseteq \mathcal{Z}_c$ and every $\mathbf{u}_1, \mathbf{u}_2$,

$$\mathbb{P}(\{\bar{h}_c(\mathbf{z}) \in A_{\mathbf{z}_c}\} \mid \mathbf{u} = \mathbf{u}_1) = \mathbb{P}(\{\bar{h}_c(\mathbf{z}) \in A_{\mathbf{z}_c}\} \mid \mathbf{u} = \mathbf{u}_2). \quad (\text{A.9})$$

Equivalently,

$$\int_{\mathbf{z} \in \bar{h}_c^{-1}(A_{\mathbf{z}_c})} p_{\mathbf{z}|\mathbf{u}}(\mathbf{z}|\mathbf{u}_1) d\mathbf{z} = \int_{\mathbf{z} \in \bar{h}_c^{-1}(A_{\mathbf{z}_c})} p_{\mathbf{z}|\mathbf{u}}(\mathbf{z}|\mathbf{u}_2) d\mathbf{z}. \quad (\text{A.10})$$

It remains to show that $\bar{h}_c$ cannot depend on $\mathbf{z}_s$. Suppose it did. Then for some open ball $A_{\mathbf{z}_c}$, the preimage $\bar{h}_c^{-1}(A_{\mathbf{z}_c})$ would not have the product form $B_{\mathbf{z}_c} \times \mathcal{Z}_s$. This preimage has nonzero probability by positivity and continuity. The domain-variability assumption therefore gives two domains $\mathbf{u}_1, \mathbf{u}_2$ for which the two integrals in Equation A.10 differ, contradicting the invariance implied by the estimated content block. Hence $\bar{h}_c$ depends only on $\mathbf{z}_c$.

The final step uses invertibility. The Jacobian of $\bar{h}$ can be written in block form as

$$\mathbf{J}_{\bar{h}} = \begin{bmatrix} \partial \hat{\mathbf{z}}_c / \partial \mathbf{z}_c & \partial \hat{\mathbf{z}}_c / \partial \mathbf{z}_s \\ \partial \hat{\mathbf{z}}_s / \partial \mathbf{z}_c & \partial \hat{\mathbf{z}}_s / \partial \mathbf{z}_s \end{bmatrix}. \quad (\text{A.11})$$

The previous paragraph shows the upper-right block is zero. Since $\bar{h}$ is invertible, the upper-left block must be nonsingular, and the inverse map cannot mix $\hat{\mathbf{z}}_s$ into $\mathbf{z}_c$. Thus $\hat{\mathbf{z}}_c = \bar{h}_c(\mathbf{z}_c)$ through an invertible block map. $\square$

Proposition A.3 (Target joint distribution in the shared space). *If Theorems A.1 and A.2 hold, and the label depends on the latent state through $(\mathbf{z}_c, \tilde{\mathbf{z}}_s)$ rather than directly through the*

domain index, then the target-domain joint distribution $p_{\mathbf{x}, \mathbf{y} | \mathbf{u}^\tau}$ is identifiable from unlabeled target-domain observations and labeled source-domain observations.

Proof. Together, Theorems A.1 and A.2 recover $\mathbf{z}_c$ and $\mathbf{z}_s$ from $\mathbf{x}$ up to the indeterminacies stated above. Since the domain-specific transformation $f_{\mathbf{u}}$ is invertible component-wise, the shared style variable $\tilde{\mathbf{z}}_s = f_{\mathbf{u}}^{-1}(\mathbf{z}_s)$ is also identifiable up to the same benign indeterminacies. Source labels identify $p_{\mathbf{y} | \mathbf{z}_c, \tilde{\mathbf{z}}_s}$, and this conditional does not change across domains by assumption. For the target domain,

$$p_{\mathbf{x}, \mathbf{y} | \mathbf{u}^\tau} = \int p_{\mathbf{x}, \mathbf{z}_c, \tilde{\mathbf{z}}_s | \mathbf{u}^\tau} p_{\mathbf{y} | \mathbf{z}_c, \tilde{\mathbf{z}}_s} d\mathbf{z}_c d\tilde{\mathbf{z}}_s. \quad (\text{A.12})$$

Both factors on the right-hand side are identifiable in the shared latent space, which gives the target joint distribution. $\square$

A.2 Regime B: Shared Information in Masked Prediction

Regime B follows the hierarchical latent-variable model of Kong et al. [10]. For a mask $\mathbf{m}$, write $\mathbf{x}_{\mathbf{m}}$ for the masked part and $\mathbf{x}_{\mathbf{m}^c}$ for the visible part. The key latent object is the minimal shared information $\mathbf{c}$ between these two parts.

Theorem A.4 (Locating shared information). *For each mask $\mathbf{m}$ in a hierarchical latent-variable graph $\mathbf{G}$, there exists a minimal set of latent variables $\mathbf{c}$ such that the generating process can be written as*

$$\mathbf{x}_{\mathbf{m}} = g_{\mathbf{m}}(\mathbf{c}, \mathbf{s}_{\mathbf{m}}), \quad \mathbf{x}_{\mathbf{m}^c} = g_{\mathbf{m}^c}(\mathbf{c}, \mathbf{s}_{\mathbf{m}^c}), \quad (\text{A.13})$$

where both maps are invertible, $\mathbf{s}_{\mathbf{m}} \perp (\mathbf{c}, \mathbf{s}_{\mathbf{m}^c})$, and no lower-dimensional latent set can satisfy these properties. For a fixed mask, the minimal $\mathbf{c}$ and the corresponding $\mathbf{s}_{\mathbf{m}}$ are unique.

Proof. The source proof constructs $\mathbf{c}$ directly from the hierarchy. Starting from the masked observables, recursively trace each variable to its parents. Whenever a parent is also an ancestor of a visible observable, place it in the candidate set for $\mathbf{c}$ and stop tracing through that branch. All remaining exogenous variables needed to reconstruct $\mathbf{x}_{\mathbf{m}}$ are placed in $\mathbf{s}_{\mathbf{m}}$. The candidate set is then pruned: if a candidate variable lies above another candidate on every directed path to the visible part, the higher variable is removed because the lower one already blocks the shared path.

Invertibility of the original hierarchical generating process implies that the backtracking procedure is lossless. Before pruning, the variables collected in $(\mathbf{c}, \mathbf{s}_{\mathbf{m}})$ contain exactly the information needed to reconstruct $\mathbf{x}_{\mathbf{m}}$. Pruning does not lose information, because each removed variable can be recovered by backtracking from a retained descendant on the same directed path. The same argument applied to the visible side produces some $\mathbf{s}_{\mathbf{m}^c}$ such that $(\mathbf{c}, \mathbf{s}_{\mathbf{m}^c})$ reconstructs $\mathbf{x}_{\mathbf{m}^c}$.

The independence $\mathbf{s}_{\mathbf{m}} \perp\!\!\!\perp (\mathbf{c}, \mathbf{s}_{\mathbf{m}^c})$ follows from the graphical construction. If an element of $\mathbf{s}_{\mathbf{m}}$ had a directed path into either $\mathbf{c}$ or $\mathbf{s}_{\mathbf{m}^c}$, then the first child on that path would also be an ancestor of both masked and visible observables, contradicting the rule that such variables are selected into $\mathbf{c}$. Minimality follows because each retained element of $\mathbf{c}$ lies on a path whose dependence cannot be blocked by any lower-dimensional choice. Uniqueness follows from the same pathwise argument: for a fixed mask, each required shared path has a unique lowest blocking variable under the hierarchical ordering. $\square$

Assumption A.5 (MAE model). For each mask $\mathbf{m}$, the MAE decoder $D_{\mathbf{m}}(\hat{\mathbf{c}}, \hat{\mathbf{s}}_{\mathbf{m}})$ has a non-singular Jacobian matrix almost everywhere. There exists an invertible map $\tilde{g}_{\mathbf{m}^c}$ such that $E_{\mathbf{m}^c}(\cdot) = [\tilde{g}_{\mathbf{m}^c}^{-1}(\cdot)]_{1:d_{\mathbf{c}}}$, where the first $d_{\mathbf{c}}$ dimensions correspond to $\mathbf{c}$. Moreover, $(D_{\mathbf{m}}, \tilde{g}_{\mathbf{m}^c})$ forms an invertible map between $(\hat{\mathbf{c}}, \hat{\mathbf{s}}_{\mathbf{m}}, \hat{\mathbf{s}}_{\mathbf{m}^c})$ and $(\mathbf{x}_{\mathbf{m}}, \mathbf{x}_{\mathbf{m}^c})$.

Theorem A.6 (Shared-variable identifiability). *Consider a generating process*

$$\mathbf{v}_1 = g_1(\mathbf{c}, \mathbf{s}_1), \quad \mathbf{v}_2 = g_2(\mathbf{c}, \mathbf{s}_2), \quad [\mathbf{v}_1, \mathbf{v}_2] = g(\mathbf{c}, \mathbf{s}_1, \mathbf{s}_2), \quad (\text{A.14})$$

where g is invertible and g_1, g_2 are smooth with nonsingular Jacobian matrices almost everywhere. If an estimated model with the same structural form matches the joint distribution of $(\mathbf{v}_1, \mathbf{v}_2)$, then the shared variable $\mathbf{c}$ is identifiable up to a one-to-one transformation.

Proof. Matched joint distributions imply that, for samples from the true process,

$$\mathbf{v}_1 = g_1(\mathbf{c}, \mathbf{s}_1) = \hat{g}_1(\hat{\mathbf{c}}, \hat{\mathbf{s}}_1), \quad (\text{A.15})$$

$$\mathbf{v}_2 = g_2(\mathbf{c}, \mathbf{s}_2) = \hat{g}_2(\hat{\mathbf{c}}, \hat{\mathbf{s}}_2), \quad (\text{A.16})$$

$$(\hat{\mathbf{c}}, \hat{\mathbf{s}}_1, \hat{\mathbf{s}}_2) = \hat{g}^{-1}(g(\mathbf{c}, \mathbf{s}_1, \mathbf{s}_2)) := h(\mathbf{c}, \mathbf{s}_1, \mathbf{s}_2). \quad (\text{A.17})$$

Substituting the last relation into the first equation gives $g_1(\mathbf{c}, \mathbf{s}_1) = \hat{g}_1(h_{\mathbf{c}, \mathbf{s}_1}(\mathbf{c}, \mathbf{s}_1, \mathbf{s}_2))$. Differentiating with respect to a coordinate of $\mathbf{s}_2$ gives a homogeneous linear system whose coefficient matrix is the Jacobian of $\hat{g}_1$ with respect to $(\hat{\mathbf{c}}, \hat{\mathbf{s}}_1)$. This matrix is nonsingular almost everywhere, so $(\hat{\mathbf{c}}, \hat{\mathbf{s}}_1)$ cannot depend on $\mathbf{s}_2$. Repeating the symmetric argument with $\mathbf{v}_2$ shows that $(\hat{\mathbf{c}}, \hat{\mathbf{s}}_2)$ cannot depend on $\mathbf{s}_1$. Therefore $\hat{\mathbf{c}}$ depends only on $\mathbf{c}$.

The same argument can be applied after swapping the true and estimated models, because the assumptions are symmetric under matched distributions. Thus $\mathbf{c}$ depends only on $\hat{\mathbf{c}}$. The two one-way dependences imply a one-to-one mapping between $\mathbf{c}$ and $\hat{\mathbf{c}}$. $\square$

Theorem A.7 (MAE identifiability). *Under Theorem A.4 and Assumption A.5, the MAE encoder $E_{\mathbf{m}^c}(\mathbf{x}_{\mathbf{m}^c})$ recovers the shared latent information $\mathbf{c}$ induced by the mask $\mathbf{m}$. Equivalently, there exists a one-to-one map h such that $h(\mathbf{c}) = \hat{\mathbf{c}}$, where $\hat{\mathbf{c}} = E_{\mathbf{m}^c}(\mathbf{x}_{\mathbf{m}^c})$.*

Proof. Apply Theorem A.6 with $\mathbf{v}_1 = \mathbf{x}_{\mathbf{m}}$ and $\mathbf{v}_2 = \mathbf{x}_{\mathbf{m}^c}$. The estimated visible-side map is $\tilde{g}_{\mathbf{m}^c}$, whose inverse contains the MAE encoder in its first d_c coordinates, and the estimated masked-side map is the decoder $D_{\mathbf{m}}(\cdot, \hat{\mathbf{s}}_{\mathbf{m}})$. Assumption A.5 ensures that these maps satisfy the invertibility and local nonsingularity conditions required by Theorem A.6. Hence the MAE estimate $\hat{\mathbf{c}} = E_{\mathbf{m}^c}(\mathbf{x}_{\mathbf{m}^c})$ is in one-to-one correspondence with the true shared information $\mathbf{c}$. $\square$

A.3 Regime C: Composition and Causal Module Identification

Regime C follows the hierarchical concept model in Kong et al. [11]. Let $\mathbf{d}$ denote a discrete high-level concept combination, $\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(L)}$ the continuous latent levels, and $\mathbf{x}$ the observation. Each continuous latent variable z is generated by a module $g_z(\text{pa}(z), \varepsilon_z)$ from its parents and an exogenous variable.

Theorem A.8 (Composition generalization). *Assume the hierarchical data-generating process above. A concept combination $\mathbf{d}$ is composable, meaning $\mathbf{d} \in \mathcal{D}_{\text{comp}}$, if for every continuous latent variable z needed to generate $\mathbf{x}|\mathbf{d}$, there exists a supported training combination $\tilde{\mathbf{d}} \in \mathcal{D}_s$ such that*

$$\text{supp}(\text{pa}(z) \mid \mathbf{d}) \subseteq \text{supp}(\text{pa}(z) \mid \tilde{\mathbf{d}}). \quad (\text{A.18})$$

The supporting $\tilde{\mathbf{d}}$ may differ across latent variables.

Proof. By definition, $\mathbf{d}$ is composable when any two model specifications that match on the training support $\mathcal{D}_s$ also agree on the modules needed to generate $\mathbf{x}|\mathbf{d}$. For a latent variable z , the relevant input support under $\mathbf{d}$ is $\text{supp}(\text{pa}(z) \mid \mathbf{d}) \times \text{supp}(\varepsilon_z)$. The exogenous support is invariant to $\mathbf{d}$. If Equation A.18 holds, then every parent configuration needed by module g_z under $\mathbf{d}$ has already appeared under some supported $\tilde{\mathbf{d}}$. Since the two model specifications agree on $\mathcal{D}_s$, they agree on g_z over that supported input set, and therefore over the input

set required by $\mathbf{d}$. This argument applies to every module in the hierarchy. The equality of all required modules implies equality of the full conditional generation process for $\mathbf{x}|\mathbf{d}$, so $\mathbf{d}$ is composable. $\square$

Definition A.9 (Componentwise identifiability). Two latent vectors $\mathbf{z}$ and $\hat{\mathbf{z}}$ are componentwise identifiable if there exists a permutation π and invertible scalar maps h_i such that $\hat{z}_i = h_i(z_{\pi(i)})$ for every coordinate i .

Condition A.10 (Identification conditions). For each adjacent level $\mathbf{z}^{(\ell)} \rightarrow \mathbf{z}^{(\ell+1)}$ in the hierarchy:

- the observation preserves the information in each latent level through a smooth invertible map from $(\mathbf{z}^{(\ell)}, \boldsymbol{\varepsilon}_\ell)$ to $\mathbf{x}$;
- the conditional density $p(\mathbf{z}^{(\ell+1)}|\mathbf{z}^{(\ell)})$ is smooth;
- components of $\mathbf{z}^{(\ell+1)}$ are conditionally independent given $\mathbf{z}^{(\ell)}$;
- for each value of $\mathbf{z}^{(\ell+1)}$, there exist $2n(\mathbf{z}^{(\ell+1)}) + 1$ values of $\mathbf{z}^{(\ell)}$ such that the corresponding derivative-variation vectors

$$\mathbf{w}(\mathbf{z}^{(\ell+1)}, \mathbf{z}^{(\ell)}) = \left(\partial_{z_1} \log p(\mathbf{z}^{(\ell+1)}|\mathbf{z}^{(\ell)}), \dots, \partial_{z_n} \log p(\mathbf{z}^{(\ell+1)}|\mathbf{z}^{(\ell)}), \right. \\ \left. \partial_{z_1}^2 \log p(\mathbf{z}^{(\ell+1)}|\mathbf{z}^{(\ell)}), \dots, \partial_{z_n}^2 \log p(\mathbf{z}^{(\ell+1)}|\mathbf{z}^{(\ell)}) \right) \quad (\text{A.19})$$

have $2n(\mathbf{z}^{(\ell+1)})$ linearly independent differences relative to a reference value.

Lemma A.11 (Single-level identification). *Consider a one-level process*

$$\mathbf{z} \sim p(\mathbf{z}|\mathbf{u}), \quad \boldsymbol{\varepsilon} \sim p(\boldsymbol{\varepsilon}), \quad \mathbf{x} = g(\mathbf{z}, \boldsymbol{\varepsilon}), \quad (\text{A.20})$$

where $\boldsymbol{\varepsilon}$ is independent of $\mathbf{z}$. Under the one-level instance of Condition A.10, $\mathbf{z}$ is componentwise identifiable.

Proof. This is the single-level nonlinear-ICA argument used by Kong et al. [9]. The proof repeats the mixed-derivative calculation in Theorem A.1: matched observation distributions imply an invertible indeterminacy map between true and estimated latents; conditional independence makes the log density additive across components; and sufficient variability in the auxiliary variable $\mathbf{u}$ forces the mixed derivatives of the indeterminacy map to vanish. Each estimated coordinate therefore depends on only one true coordinate, up to permutation and invertible scalar transformations. $\square$

Theorem A.12 (Causal module identification). *Under Condition A.10, the latent variables at each level of the hierarchy are component-wise identifiable, and the graph structure is identifiable up to permutation of variables within each level.*

Proof. Treat the observed discrete concept vector $\mathbf{d}$ as the auxiliary variable for the top latent level $\mathbf{z}^{(1)}$. Lemma A.11 identifies $\mathbf{z}^{(1)}$ component-wise. Once $\mathbf{z}^{(1)}$ is identified, it can be used as the auxiliary variable for the next level, so the same single-level argument identifies $\mathbf{z}^{(2)}$. Repeating this procedure from top to bottom identifies every latent level component-wise, with only permutation indeterminacy within each level.

After the variables are identified, their conditional independences can be read from the recovered latent distribution. Standard causal discovery over the identified latent variables then recovers the graphical structure up to the same within-level permutations. This proves that the modular hierarchy required by Theorem A.8 is recoverable under the stated conditions. $\square$

Appendix B

Additional Work Connected to the Four Regimes

Related projects extend the four-regime thesis spine in several directions.

Extension Map

Regime A extensions include multi-domain image generation and translation, counterfactual text generation, biomedical multimodal causal representation learning, and modular vision-language alignment. Their role is to show that multiple distributions can reveal latent mechanisms beyond classification, including generation, multimodal scientific data, and modular alignment.

Regime B extensions include discrete concept learning in hierarchical models, concept-graph extraction from foundation models, and atomic vision-language concept alignment for controllable generation. Their role is to extend the MAE story from one pretraining objective to broader concept discovery, naming, alignment, and control.

Regime C extensions include causal extrapolation, counterfactual control, and compositional reasoning in diffusion language models. Their role is to extend compositional generation toward calibrated behavior beyond the training distribution.

Regime D extensions include reward-guided discovery of reusable reasoning modules, compositional world models, and forward-inverse verification. Their role is to show how generated traces, multitask transitions, and selected interaction can expand the support available to a learner.

Appendix C

Formal and Supplemental Results for Learning Through Feedback and Interaction

C.1 Reasoning Module Identification and Local Witnesses

This section records the formal structure behind the reasoning results in Chapter 5. Let P denote the observed problem descriptor, D the observed reasoning trace, and $\mathbf{S}$ the hierarchy of finite discrete selection variables,

$$\mathbf{S} = (\mathbf{S}^{(1)}, \dots, \mathbf{S}^{(L)}). \quad (\text{C.1})$$

The directed graph is acyclic and the joint law is Markov to the hierarchy

$$p(D, \mathbf{S} \mid P) = p(D \mid \mathbf{S}^{(L)}) \prod_{\ell=1}^L p(\mathbf{S}^{(\ell)} \mid \mathbf{S}^{(\ell-1)}), \quad \mathbf{S}^{(0)} := P. \quad (\text{C.2})$$

The following conditions state every requirement used by the identification claim rather than leaving them implicit in Kong et al. [12].

C.1.1 Identification Conditions for the Selection Hierarchy

The identification theorem uses the following eight conditions.

1. **Faithfulness.** Every conditional independence in the observed distribution follows from separation in the latent directed graph; no extra independence appears through a special parameter cancellation.
2. **Rank faithfulness.** Every nonnegative-rank constraint in a conditional probability table of observed variables follows from the graph and the stated finite latent cardinalities; no accidental rank drop imitates a smaller latent bottleneck.

3. **Restricted choice sets.** Each selection variable realizes a strict subset of the Cartesian product obtained by freely combining the states of its immediate constituents. Thus, valid local compositions do not fill every logically possible configuration.
4. **Deterministic coarsest selection.** Each latent selection variable is the unique coarsest deterministic statistic of its immediate parents that preserves its local predictive role. Splitting a state adds no predictive information, while merging two states removes some.
5. **Neighborhood coverage.** Every latent selection S has an observed diagnostic parent D^* that is a parent of no other latent selection and has more observable states than S . After conditioning on the other observed parents of S , the observed variables outside this local neighborhood retain at least $|\text{supp}(S)|$ supported configurations. This makes the rank drop created by S visible rather than a consequence of an undersized comparison table.
6. **Observable module signatures.** For any two supported configurations of the observed parents of S , the configurations produce the same value of S if and only if they induce the same conditional law for the observed trace variables outside S 's local neighborhood. This excludes both merging distinct module states and assigning different labels to the same state in different contexts.
7. **No twins.** No two distinct latent variables have exactly the same observed and latent neighbors. Otherwise, the observed law could not determine which copy generated a local effect.
8. **Maximality.** No latent variable can be split into several variables while preserving both the Markov property and the no-twins condition. This excludes artificial refinements of the hierarchy.

Theorem C.13 (Identification of the reasoning hierarchy). *Consider the class of hierarchical latent selection models that satisfy the eight conditions above. If both the data-generating model and an alternative model in this class induce the same conditional law $p(D \mid P)$, then their latent selection variables and adjacency structures agree up to component-wise relabeling of discrete states.*

Proof structure. The proof proceeds from the trace-adjacent latent layer toward the descriptor. Observable signatures distinguish the states of each bottom latent variable. Restricted local choice and neighborhood coverage then separate a state from the states with which it

commonly co-occurs. After identifying one layer, its identified variables become the observed children for the next layer. Repeating the same local argument identifies the hierarchy and its adjacency, with a separate label permutation remaining at each latent variable. $\square$

For a latent node U and child configuration $\mathbf{v}_{\text{Ch}(U)}$, define the witnessed compatibility set

$$\mathcal{W}_U(\mathbf{v}_{\text{Ch}(U)}) = \{u : \exists p \in \Omega_{\text{supp}} \text{ with } p^*(U = u, \text{Ch}(U) = \mathbf{v}_{\text{Ch}(U)} \mid P = p) > 0\}. \quad (\text{C.3})$$

Theorem C.14 (Composition through local witnesses). *Assume the identified latent selection model above. If every local family induced by a new descriptor p_{new} selects a value $u \in \mathcal{W}_U(\mathbf{v}_{\text{Ch}(U)})$, then the learned local constraints compose consistently under p_{new} .*

Proof structure. Each witnessed value satisfies the child-imposed constraints for its local family because that family has positive probability under a compatible training descriptor. Beginning at the trace-adjacent layer, select a witnessed value for every local family. The selections agree on shared variables by the identified adjacency and local compatibility conditions. Induction over the hierarchy therefore constructs a globally consistent latent assignment for the new descriptor. If any local witness set is empty, this construction stops at the unsupported interface, which is why identification alone is insufficient. $\square$

C.2 Supplemental Evidence for Reward-Guided Reasoning

The following analyses test whether the Chapter 5 findings persist in a released model pair, over longer RL runs, and across the reported training choices.

C.2.1 Shared Controlled-Experiment Setup

Table C.1 records the settings shared by the controlled experiments. The interventions change which skills, routers, or compositions appear in SFT and RL while holding this implementation fixed.

C.2.2 Open-Source Model Analysis

The controlled benchmark identifies skills and routers by construction. A complementary analysis asks whether a related local-recombination signature is visible in existing language models. The analysis collects 643 problems from MATH-500, AIME 2024 and 2025, AMC, and GSM8K. Qwen3-4B and its Qwen3-4B-Thinking-2507 counterpart generate one reasoning trace per problem. A language model first builds an inventory of mathematical skills

Table C.1. Shared setup for the controlled reasoning experiments. Correct final strings determine both supervised filtering and the outcome-only RL reward; no step-level reasoning labels are used during RL.

Item	Setting
Base model	Llama-3.1-8B-Instruct [52]
Task family	Synthetic string transformations with 24 atomic skills from Yuan et al. [53] and 10 routing mechanisms
SFT data	Ten reasoning-inclusive responses sampled per problem; responses with an incorrect final answer are discarded
SFT training	2 epochs, learning rate 2×10^{-5} , batch size 128
RL reward	Final-string exact match; no step-level labels
RL rollouts	On-policy responses, with successful and unsuccessful samples retained for reward-based updates
Default RL settings	Learning rate 1×10^{-6} , maximum length 8192 tokens, temperature 1.0, rollout batch size 16
Filtering	Prompts whose rollouts are all correct or all incorrect are omitted; KL and entropy coefficients are both zero
Evaluation	Exact final-string accuracy on seen compositions, unseen compositions, and greater composition depths

from the combined traces and then maps each trace to a sequence of skill identifiers. The comparison counts distinct contiguous skill sequences, called n-grams.

Figure C.1 shows that the reasoning-tuned model produces more distinct two-skill and three-skill sequences, while the difference narrows for longer sequences. Short sequences measure whether a local operation appears beside different neighbors, which is the observable pattern expected when module interfaces become more diverse. The analysis does not establish a causal effect of RL because the two released models may differ in other post-training details, and the skill tokenizer is itself model-assisted.

C.2.3 Longer Runs and Training Sensitivity

The main skill-reintroduction and SFT–RL support experiments use the training budgets described in Chapter 5. To check whether their ordering is a short-run effect, the study continues both experiments to 1000 RL steps. Figure C.2 retains the advantage of reintroducing held-out skills in compound traces. Figure C.3 retains the advantage of using RL compositions outside the supervised composition set. These plots support persistence over the tested horizon, not convergence to a unique asymptotic solution.

Figure C.4 varies reward shaping, Kullback–Leibler regularization, entropy regularization, rollouts per prompt, and sampling temperature. The curves average over the tested values

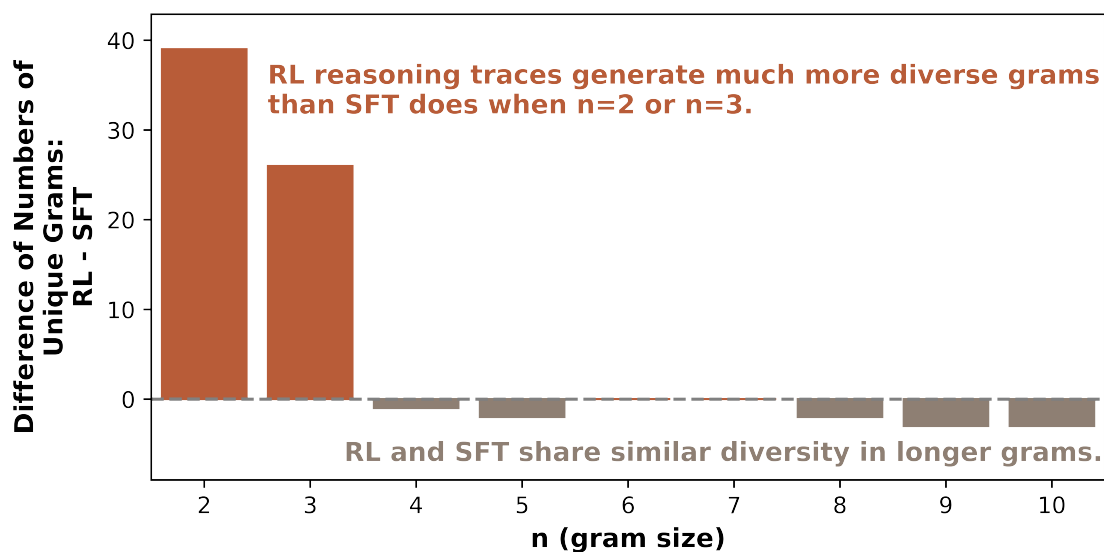


Figure C.1. Difference in the number of distinct contiguous skill sequences produced by Qwen3-4B-Thinking-2507 and Qwen3-4B on 643 mathematics problems. Positive values mean that the reasoning-tuned model produces more distinct sequences. The largest differences occur for two-skill and three-skill sequences, which is compatible with greater local recombination but is not direct proof of latent-module identification.

within each group, and the bands show variation across those settings. The qualitative learning trend remains visible across these interventions, although the figure is a sensitivity analysis rather than an exhaustive search over RL algorithms.

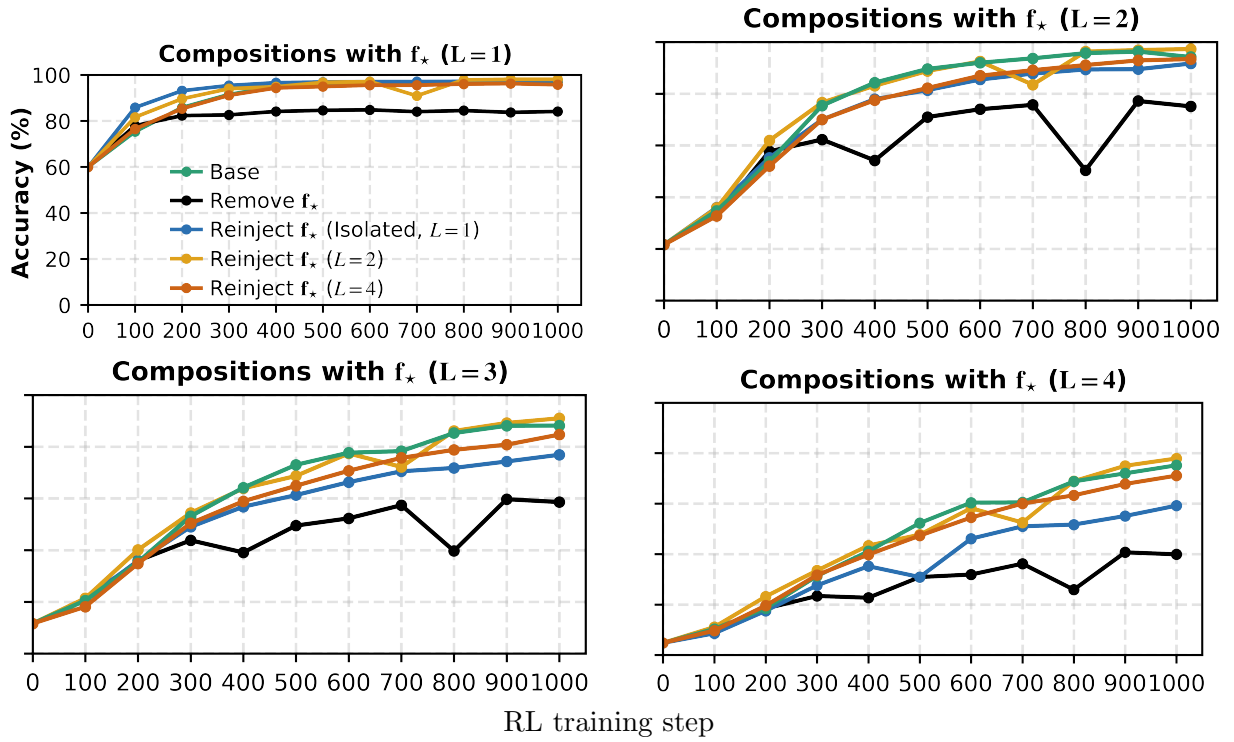


Figure C.2. Held-out-skill reintroduction extended to 1000 RL steps. The vertical axis reports accuracy on traces containing removed skills. The advantage of restoring skills through compound traces persists after the shorter main-study budget.

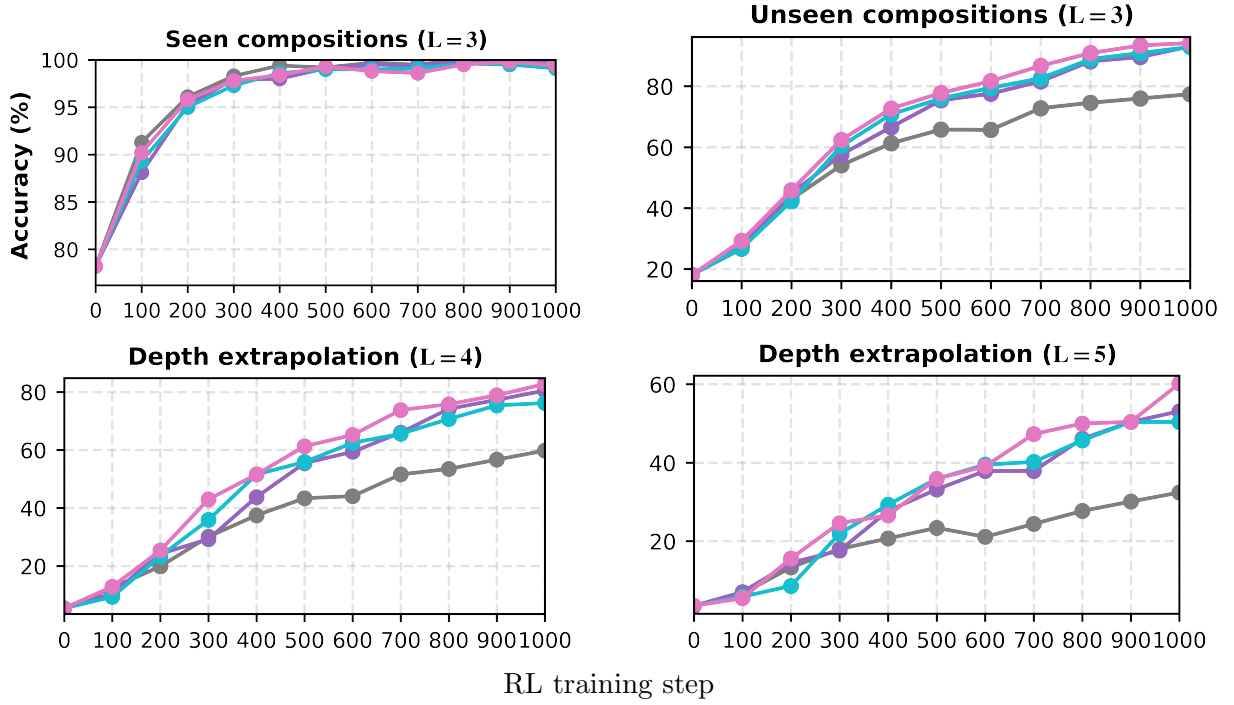


Figure C.3. SFT–RL composition-set relationships extended to 1000 RL steps. Accuracy is shown for seen and unseen compositions. The relative ordering persists: RL compositions outside supervised support yield stronger unseen performance than RL compositions contained within supervised support. Gray, purple, cyan, and magenta denote, respectively, RL compositions contained in SFT, SFT compositions contained in RL, partial overlap, and disjoint composition sets.

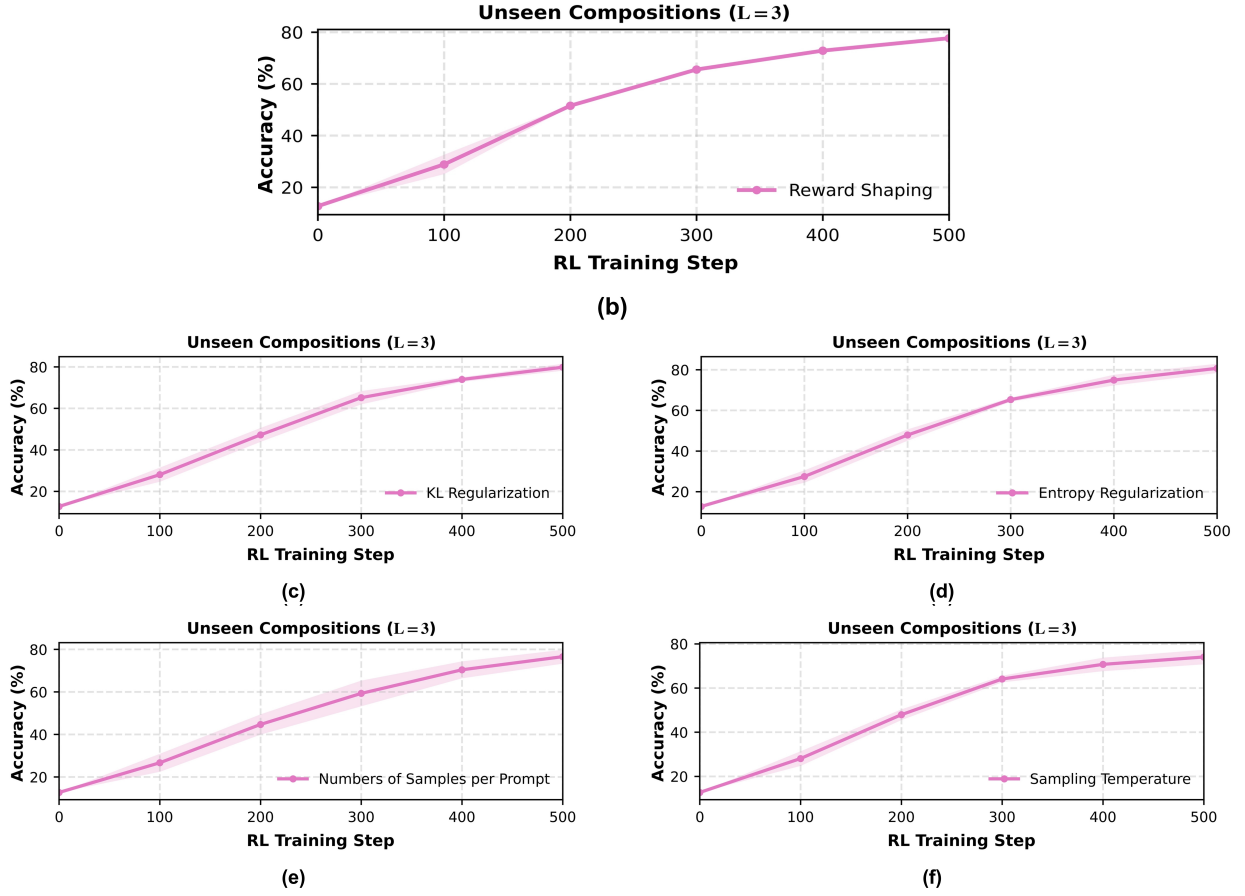


Figure C.4. Sensitivity of unseen-composition accuracy to five training choices: reward shaping, Kullback–Leibler and entropy penalties, rollouts per prompt, and sampling temperature. Solid curves average the tested values and shaded regions show their variation. The five displayed panels are labeled (b)–(f) because panel (a), whose setting was specified only through a cross-reference, is omitted.

Assumption C.15 (Local excitation and invariance). For every training regime c , target partition k , and required source $i \in I_c^{(k)}$, the support contains two transitions with the same $(z_{-i,t}, a_t, g_c)$ but different $z_{i,t}$ that produce different $z_{k,t+1}$. For every inactive source $j \notin I_c^{(k)}$, the target $z_{k,t+1}$ is invariant to $z_{j,t}$ when $(z_{-j,t}, a_t, g_c)$ is fixed on the support of regime c .

Let $\mathcal{S}_i^{\text{tr}}$ be the union of training supports for source partition i in regimes where it is active. The recombination domain $\mathcal{X}_{\text{rec}}$ contains candidate transitions whose active links have all appeared in training and whose selected source values belong to the corresponding $\mathcal{S}_i^{\text{tr}}$.

Assumption C.16 (Reusable-extension selection). Among zero-error predictors with minimum total support cardinality, the complexity functional Ω selects a shared interaction that agrees with the oracle on $\mathcal{X}_{\text{rec}}$. Equivalently, the selected predictor extends the observed dynamics through one reusable source-target rule rather than through a separate lookup rule for each regime.

Theorem C.17 (Support recovery with reusable-extension selection). *Suppose the oracle sparse dynamics are realizable. Under Assumption C.15, every minimum-support zero-error solution recovers $M_c = M_c^*$ for each training regime. If the selected solution also satisfies Assumption C.16, then its composed predictor is consistent with the oracle on $\mathcal{X}_{\text{rec}}$. Thus, for any $(a_t, z_t, M_\star^*) \in \mathcal{X}_{\text{rec}}$, the learned transition satisfies $\hat{z}_{k,t+1} = z_{k,t+1}$ for every $k \in [K]$.*

Proof structure. First, local excitation implies that omitting an oracle-active link makes zero-error prediction impossible because two otherwise matched transitions have different targets. Every zero-error support must therefore include all oracle-active links. Second, invariance implies that adding an oracle-inactive link cannot reduce prediction error, so minimum support excludes all such links. These two steps recover every training support matrix. Finally, the reusable-extension assumption selects, among the recovered-support solutions, the shared interaction that agrees with the oracle on valid recombinations. The restriction to $\mathcal{X}_{\text{rec}}$ is essential because the argument does not identify unseen links or source values outside their training supports. $\square$

C.4 Supplemental CWM Results

MM-Bench contains more than 200 tasks across ten control domains. Every multitask method trains one shared agent for 100 million environment steps. Scores are normalized within each domain, averaged over that domain’s tasks, and then averaged across domains. Table C.2 reports means and standard deviations across three seeds for CWM and the NewT backbone whose transition model it changes.

Table C.2. Final normalized scores after 100 million environment steps. Values are domain means with standard deviations across three seeds. CWM and NewT use the same training budget and planning pipeline; CWM changes task routing and the entity and interaction structure of the transition.

Domain	NewT	CWM	Domain	NewT	CWM
DMControl	.50 $\pm$.04	.57$\pm$.03	Box2D	.33 $\pm$.03	.41$\pm$.03
DMControl-Ext.	.37 $\pm$.05	.47$\pm$.04	RoboDesk	.38 $\pm$.04	.70$\pm$.06
Meta-World	.50 $\pm$.03	.71$\pm$.05	OGBench	.30 $\pm$.03	.44$\pm$.03
ManiSkill	.69 $\pm$.04	.77$\pm$.03	MiniArcade	.44 $\pm$.03	.48$\pm$.05
MuJoCo	.20 $\pm$.03	.25$\pm$.05	Atari	.13 $\pm$.04	.28$\pm$.05
Aggregate average				.38 $\pm$.03	.51$\pm$.04

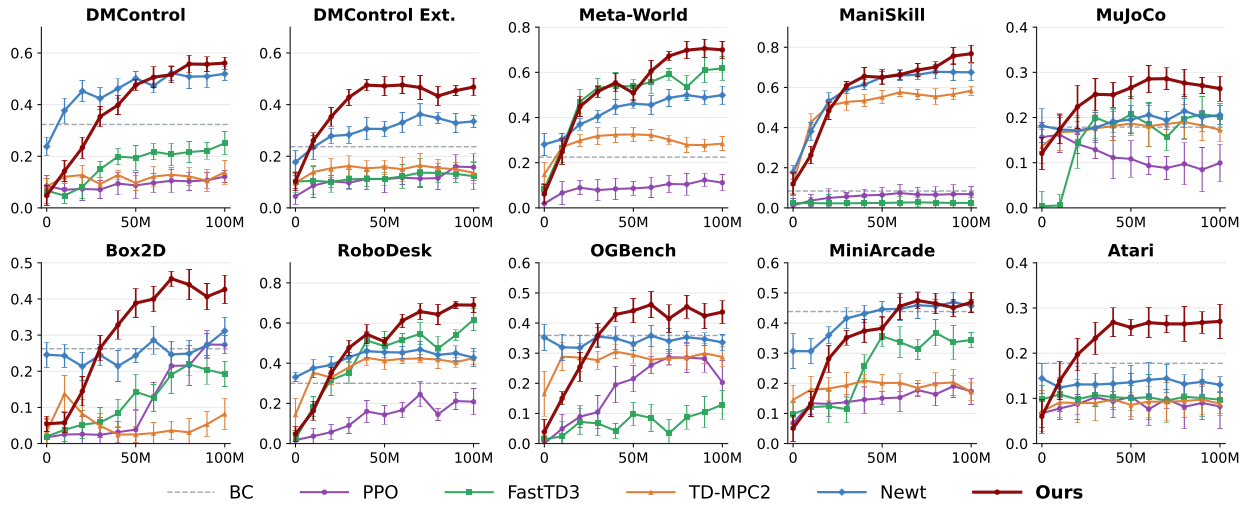


Figure C.6. Normalized multitask learning curves across the ten MM-Bench domains. Curves are means over three seeds and bands are standard errors. All methods use the same 100-million-step budget. CWM improves on NewT in every domain at the final evaluation.

Figure C.6 gives the full learning curves. CWM improves the final score in every reported domain. These are multitask results, so they show that sparse routing and masking strengthen the shared model before any held-out adaptation.

Held-out tasks recombine embodiments, objects, goals, initial states, and interaction patterns that appeared during the 200-task training stage but not in the same task configuration. Figure C.7 shows fine-tuning from the shared checkpoint. CWM provides faster adaptation and a higher final score than NewT and the component-removal variants on the reported tasks. This evaluation is inside the recombination domain described by Theorem C.17; it does not test unseen physical laws.

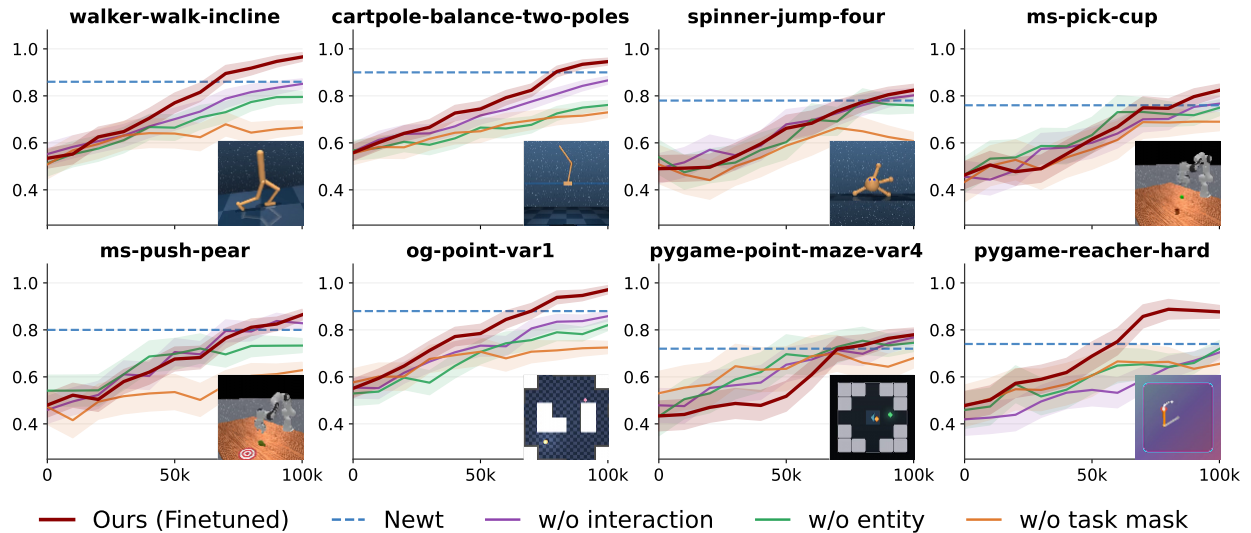


Figure C.7. Fine-tuning on held-out compositional tasks after shared 200-task training. Each task recombines factors seen during training but not previously observed together. Curves are three-seed means and bands are standard errors. CWM adapts faster and reaches the highest final score on the displayed tasks.

Figure C.8 removes the interaction mask, entity mask, or task router while keeping the remaining system fixed. Every removal lowers the average score in all ten domains. Removing task routing causes the largest general decline; entity masking matters especially in object-centered domains, and interaction masking gives a smaller but consistent gain. The result supports complementary roles for the three components rather than attributing the aggregate gain only to a larger transition network.

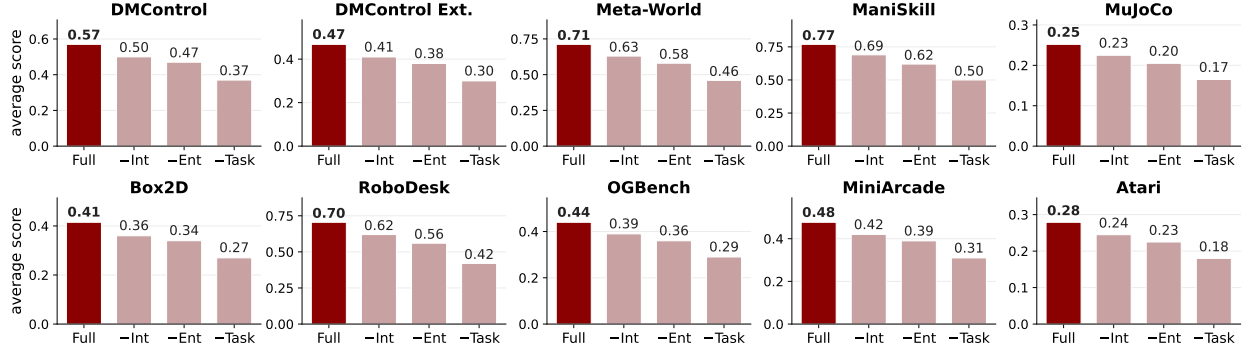


Figure C.8. Component-removal results by MM-Bench domain. The variants remove the interaction mask, entity mask, or task router from CWM. Lower scores after every removal show that all three components contribute, with task routing producing the largest overall effect.

C.5 Forward-Inverse Asymmetry and Verified Support Expansion

WAV begins from the factorization

$$p(s_{t+1} \mid s_t, a_t) \propto p(s_{t+1} \mid s_t)p(a_t \mid s_t, s_{t+1}). \quad (\text{C.7})$$

An action-free transition prior estimates whether a proposed future is plausible. A sparse inverse model estimates whether an action can reach it. For K candidate futures, the acquisition cycle is

$$s_t \xrightarrow{g_\phi} \tilde{s}_{t+1}^{1:K} \xrightarrow{h_\psi} \hat{a}_t^{1:K} \xrightarrow{f_\theta} \hat{s}_{t+1}^{1:K} \xrightarrow{\ell} \hat{e}_{1:K} \xrightarrow{\max} a_t^*. \quad (\text{C.8})$$

Here g_ϕ proposes plausible futures, h_ψ infers actions, f_θ predicts each action’s outcome, and ℓ measures the conflict between the proposed and predicted future. The highest-conflict action is executed and its observed transition is added to the labeled set. Figure C.9 displays this ordering and the data returned to the world model.

The first formal result quantifies the finite-sample advantage of a sparse inverse model. Consider the linear-Gaussian dynamics

$$s' = As + Ba + \xi, \quad \xi \sim \mathcal{N}(0, \sigma_s^2 I_{d_s}), \quad (\text{C.9})$$

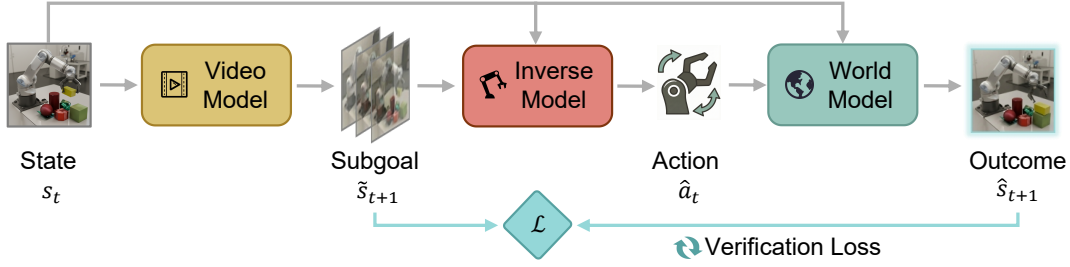


Figure C.9. WAV acquisition cycle. An action-free prior proposes plausible futures, a masked inverse model infers a reaching action, and the forward world model predicts that action’s outcome. The proposed-versus-predicted discrepancy selects the next environment interaction.

and an inverse model that recovers a from $[z; z'] \in \mathbb{R}^{2d_z}$ with $z = Ms$ and action noise variance σ_a^2 . Define $\lambda = \|B\|_{\text{op}}$.

Proposition C.18 (Forward-inverse error comparison). *Fit the dense forward and sparse inverse regressions by ordinary least squares on n independent transitions with isotropic Gaussian covariates. If $n > d_s + d_a + 1$ and $n > 2d_z + 1$, then*

$$\mathbb{E}[E_F] = \sigma_s^2 \frac{d_s + d_a}{n - (d_s + d_a) - 1}, \quad (\text{C.10})$$

$$\mathbb{E}[E_I] \leq \lambda^2 \frac{d_a}{d_s} \sigma_a^2 \frac{2d_z}{n - 2d_z - 1}. \quad (\text{C.11})$$

Consequently, the ratio obeys

$$\frac{\mathbb{E}[E_F]}{\mathbb{E}[E_I]} \geq \left(\frac{d_s + d_a}{2d_z} \frac{d_s}{d_a} \right) \left(\frac{\sigma_s}{\lambda \sigma_a} \right)^2 \left(\frac{n - 2d_z - 1}{n - (d_s + d_a) - 1} \right). \quad (\text{C.12})$$

Proof. For each coordinate of s' , ordinary least squares estimates a scalar regression with $d_s + d_a$ inputs and noise variance σ_s^2 . The exact isotropic-design excess risk is $\sigma_s^2(d_s + d_a)/(n - (d_s + d_a) - 1)$. Each action coordinate in the inverse model uses $2d_z$ inputs and has excess risk $\sigma_a^2(2d_z)/(n - 2d_z - 1)$. Summing over the d_a action coordinates and mapping action error through B contributes at most $\lambda^2 d_a/d_s$ in normalized state-space error. Dividing the two bounds gives the stated ratio. $\square$

The second result explains how inverse verification can label transitions outside the full seed support. Let the latent state contain blocks $z = (z_1, \dots, z_k)$. A set S_{src} is insulated when the next-state blocks it contains have no parents in the complementary current-state blocks. Let S_{act} contain blocks affected by action and define the verification subset $S = S_{\text{src}} \cap S_{\text{act}}$.

Theorem C.19 (Identifiability of verified self-improvement). *Assume that the latent blocks are identified up to a fixed permutation and element-wise invertible transformations. Assume also that the restricted pair (z_S, a) remains on the seed support for every transition to be labeled, even when the full pair (z, a) is outside that support. Finally, assume that action is identifiable from $(z_{S,t}, z_{S,t+1})$ on the restricted seed support. Then an inverse model trained on labeled seed transitions uniquely recovers the missing action for each such unlabeled transition, up to a fixed label permutation resolved by the labeled data. The newly labeled full transition can therefore expand the action-labeled support.*

Proof structure. The identified representation aligns the learned and true latent blocks while preserving the verification subset. The support-gap assumption keeps the restricted transition on support, so the seed data determine the inverse map on the input used for verification. Action identifiability then recovers the unique action label from the restricted current and next states. Attaching this label to the full observation transition adds a correct labeled example even when the full state-action pair was not represented in the seed data. $\square$

C.6 Supplemental WAV Results

The following results separate data efficiency, controlled verifier robustness, adaptation under two shifts, and usefulness for downstream policy refinement.

C.6.1 Sample Efficiency and the 200-Sample Comparison

The robotic evaluation covers RoboMimic Can, Lift, and Square and ManiSkill PokeCube, PullCube, and LiftPeg. Every method starts from the same warm-start budget, selects additional interactions, and trains the same world-model architecture. Prediction quality is the mean squared error over 32 predicted frames after two acquisition rounds. Figure C.10 reports three-seed means over warm-start budgets from 200 to 1000 samples. Across the six tasks, WAV reaches the error attained by the vanilla acquisition baseline with 1000 samples using fewer than 500. This is the basis for the approximately twofold sample-efficiency summary in Chapter 5.

Table C.3 gives the complete 200-sample comparison, including the standard error of every entry. WAV has the lowest mean error among the non-oracle methods on all six tasks. The masked inverse verifier also improves over the unmasked inverse-dynamics baseline on every task, which isolates the value of restricting verification to action-relevant features.

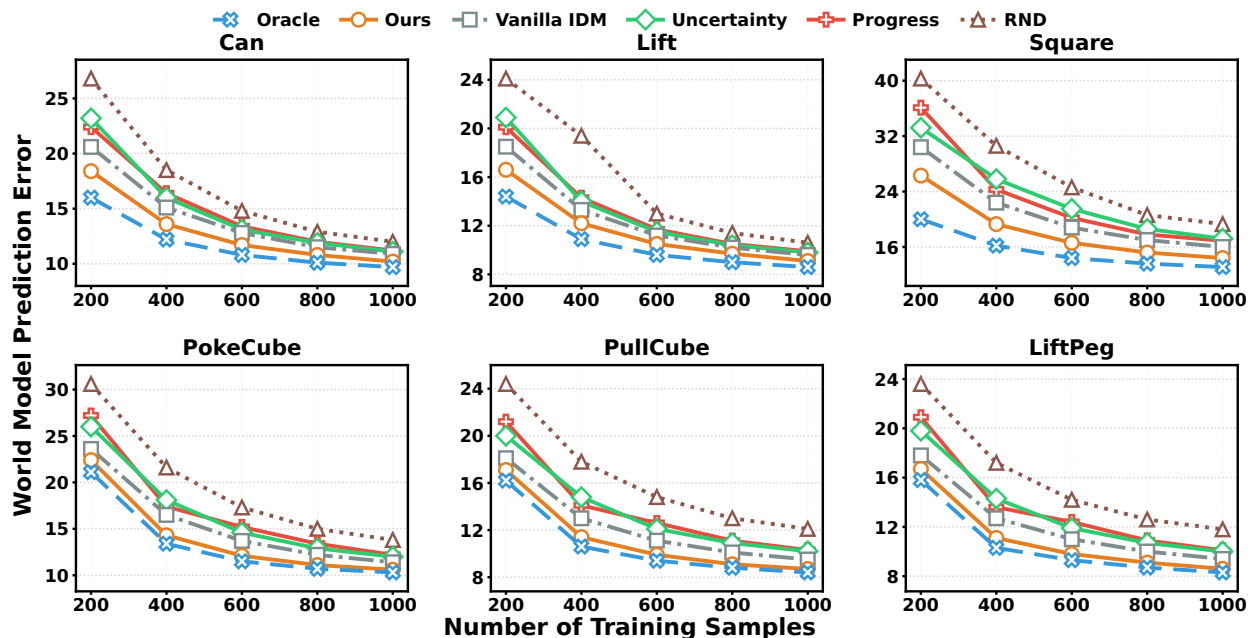


Figure C.10. World-model prediction error across six manipulation tasks and five warm-start budgets. Lower is better. Each point is the mean over three seeds after two acquisition rounds. WAV reaches the vanilla baseline’s 1000-sample error with fewer than 500 samples on the reported tasks.

C.6.2 Controlled Robustness of the Inverse Verifier

MiniGrid permits separate tests of the three factors in Equation C.12. Models train on environments with six objects. The first panel of Figure C.11 varies labeled transitions from 400 to 2000. The second increases test-time objects from 6 to 14, which increases state complexity while preserving the action set. The third increases the number of floor tiles whose colors change after every action, which adds observation noise without changing the action imprint. Forward and inverse models are compared in the same next-state space by executing each inferred action in the simulator. The inverse models retain higher dynamics accuracy under smaller data, greater scene complexity, and more observation noise; the sparse inverse model is the strongest verifier.

C.6.3 Shifted Settings and Policy Usefulness

The shifted-setting evaluation begins from one world model pretrained on the original RoboMimic Can task. A visual variant changes background and embodiment color while preserving task dynamics. An object and interaction variant adds objects and mixes demonstrations of different quality. Before acquisition, each target domain supplies action-free observations

Table C.3. World-model prediction error with 200 training samples. Values are means with standard errors; lower is better. All methods use the same downstream world-model architecture and differ in how they select interaction data.

Method	Can	Lift	Square	PokeCube	PullCube	LiftPeg
Random	26.5 $\pm$ 0.7	23.3 $\pm$ 0.8	39.7 $\pm$ 0.8	30.1 $\pm$ 0.9	24.6 $\pm$ 0.8	23.4 $\pm$ 0.7
Uncertainty	22.9 $\pm$ 0.6	20.7 $\pm$ 0.6	33.0 $\pm$ 0.7	25.2 $\pm$ 0.7	19.9 $\pm$ 0.5	19.4 $\pm$ 0.6
Progress	22.0 $\pm$ 0.6	19.6 $\pm$ 0.6	35.1 $\pm$ 0.7	27.3 $\pm$ 0.7	21.1 $\pm$ 0.6	20.5 $\pm$ 0.8
Vanilla inverse model	20.2 $\pm$ 0.2	18.3 $\pm$ 0.5	29.8 $\pm$ 0.7	23.4 $\pm$ 0.6	18.2 $\pm$ 0.5	17.5 $\pm$ 0.5
WAV	18.2$\pm$0.5	16.4$\pm$0.4	26.0$\pm$0.3	22.4$\pm$0.5	17.1$\pm$0.4	16.8$\pm$0.5

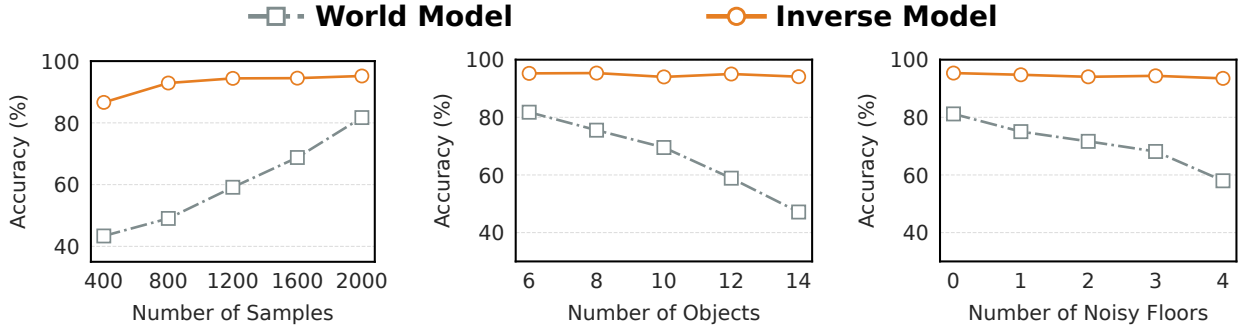


Figure C.11. MiniGrid robustness tests for forward and inverse prediction. From left to right, the panels vary labeled sample count, number of objects, and stochastic floor tiles. Dynamics accuracy measures only state elements that change over time. The sparse inverse model is more stable than the action-conditioned forward model in all three controlled shifts.

but no action labels. Every strategy receives the same budget of 200 labeled target-domain trajectories and the same fine-tuning schedule.

Figure C.12 reports normalized prediction error and downstream reward after adaptation. WAV has the lowest error and highest reward among the displayed non-oracle methods in both variants, with the larger gain under object and interaction shift. The figure does not include the unadapted source model, so it supports a comparison among equal-budget acquisition methods but does not by itself measure absolute improvement over no adaptation. Finally, Figure C.13 tests whether prediction gains remain useful for control. For each learned world model, the same base diffusion policy is refined through imagined rollouts under a fixed 1000-trajectory budget. Policies refined with WAV world models earn higher reward than those using the displayed non-oracle comparison models on all six tasks. The figure reports per-task results and does not provide an uncertainty-bearing aggregate, so the thesis uses it as supporting evidence rather than as a precise average effect estimate.

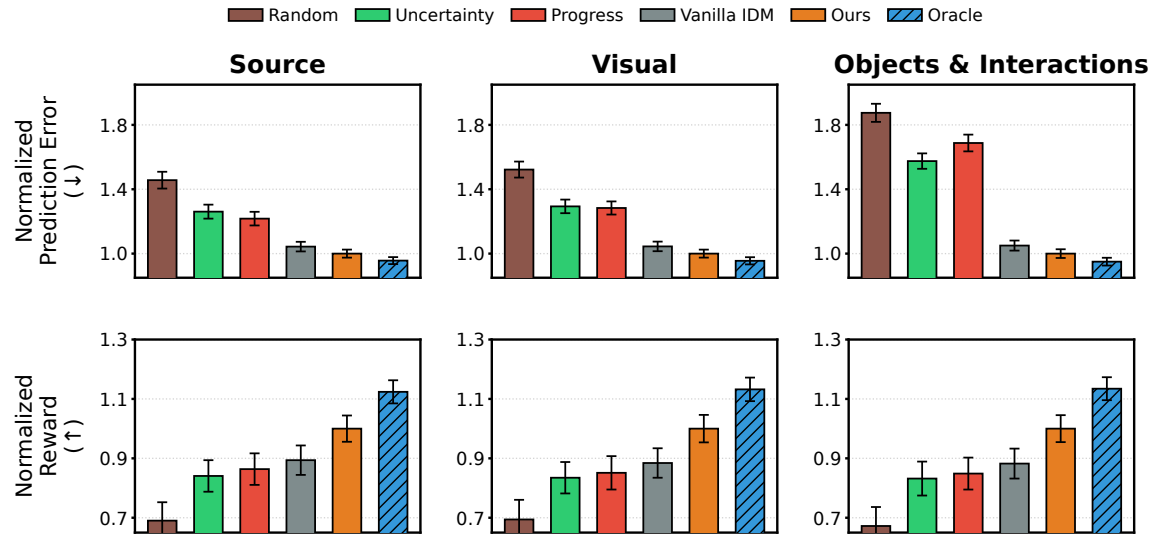


Figure C.12. Equal-budget adaptation under visual shift and under object and interaction shift. Each method receives 200 labeled target-domain trajectories. The top row reports normalized prediction error, where lower is better; the bottom row reports downstream policy reward, where higher is better. Error bars are standard errors over three seeds.

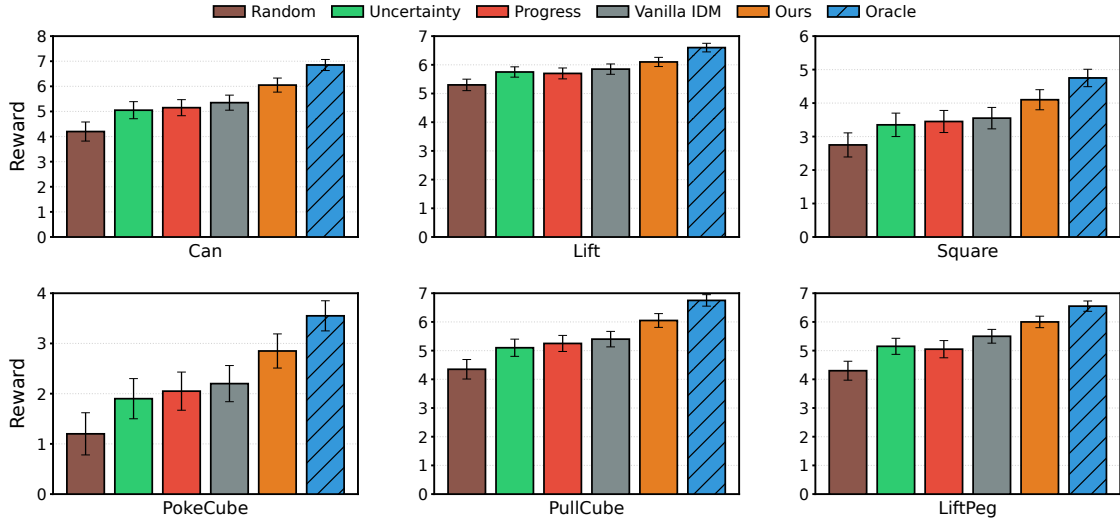


Figure C.13. Task reward after imagination-based refinement of the same base diffusion policy with world models learned by different acquisition methods. Every method uses a fixed 1000-trajectory budget. Higher reward for WAV on each displayed task shows that its prediction gains transfer to the policy-learning stage.