From Comparison to Composition: Towards Understanding Machine Cognition of Unseen Categories

Anonymous CVPR submission

Paper ID 22

Abstract

*Humans acquire visual concepts through a natural compare-then-compose process, enabling effortless generalization to novel categories. Whether machines can achieve similar genuine semantic generalization—recognizing unseen categories without any training exposure—remains a fundamental open question. We formalize the cognitive compare-then-compose mechanism into **C**omparison–**C**omposition **C**ognition ($\boldsymbol{C}^3$), an identifiability framework grounded in two complementary conditions:* comparison *requires sufficient cross-category contrast for latent concept identification;* composition *requires disjoint concept supports for reliable unseen category recognition. Under mild, nonparametric assumptions, we prove these conditions yield both necessary and sufficient guarantees. On eight fine-grained benchmarks under a genuine generalization protocol, our instantiation achieves $+3.8\%$ average accuracy over state-of-the-art, with ablations confirming each component's contribution. C^3 provides the first principled characterization of when and why learned representations generalize to unseen categories.*

1. Introduction

Modern recognition systems achieve impressive closed-world performance [42] yet systematically fail on categories absent from training [64, 88]. Even billion-parameter foundation models do not resolve this [9, 70, 85]. The core challenge is *genuine semantic generalization*: recognizing novel categories without any exposure during training.

Humans routinely recognize unfamiliar objects by transferring visual concepts from known ones [2, 46, 59]. Cognitive science characterizes this as a *compare-then-compose* process: structure-mapping extracts discriminative features via cross-instance comparison [17, 53], while recognition-by-components emphasizes composition of reusable primitives [2, 45]. However, deep networks learn entangled features plagued by spurious correlations [26]: as shown in Fig. 1, correlated attributes like *white head* and *black wings* become coupled, preventing transfer to unseen combinations.

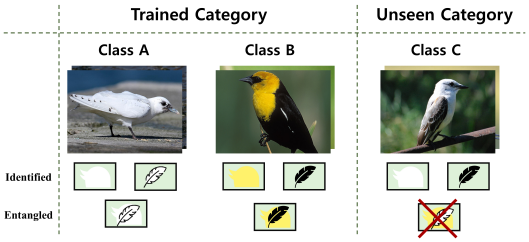

Figure 1. Semantic cognition with identified concepts vs. entangled features. Disentangled concepts enable compositional transfer to unseen categories.

Despite progress on category discovery [72, 76], genuine semantic generalization without novel-category access remains theoretically underexplored [11, 86]. Prior methods either assume exposure to target semantics [5, 72] or lack guarantees [48, 70]. This highlights: *under what conditions can learned representations provably generalize to unseen categories?*

We answer this through **C**omparison–**C**omposition **C**ognition ($\mathbf{C}^3$), formalizing the human compare-then-compose mechanism into an identifiability framework. Our analysis: (1) under sufficient cross-category variation, contextual and semantic factors separate (Thm. 2.1); (2) with additional contrast and sparse structure, individual semantic concepts are recovered (Thm. 2.2); (3) when concept supports are disjoint, unseen categories are reliably recognized as novel compositions (Thm. 2.3). These conditions are nonparametric and naturally satisfied in classification settings. Our contributions:

- A theoretical framework ($\mathbf{C}^3$) providing identifiability guarantees for unseen category cognition under assumption-light conditions.
- A theory-to-method mapping validated on synthetic data.

- $+3.8\%$ average accuracy on eight fine-grained benchmarks under On-the-fly Category Discovery [11].

2. The C^3 Framework

2.1. Problem Setup

Our framework grounds on two cognitive principles: **(1)** *Dissociable Representation*: the human visual system maintains separable processing for identity semantics and scene context [14, 30]; **(2)** *Sparse Diagnostic Features*: recognition relies on a small subset of diagnostic visual features [15, 19]. We formalize these principles through a latent-concept data-generating process (Fig. 2):

$$\mathbf{x} = g(\mathbf{z}, \mathbf{c}), \quad \mathbf{z} = f(\mathbf{z}, y, \boldsymbol{\epsilon}), \quad \mathbf{c} \sim p_{\mathbf{c}}, \quad \boldsymbol{\epsilon} \sim p_{\boldsymbol{\epsilon}}, \quad (1)$$

where observations $\mathbf{x} \in \mathbb{R}^{d_x}$ are generated from category-invariant *contextual concepts* $\mathbf{c} \in \mathbb{R}^{d_c}$ (*e.g.*, background, surroundings) and category-dependent *semantic concepts* $\mathbf{z} \in \mathbb{R}^{d_z}$ (*e.g.*, textures, shapes), modulated by label y and low-level patterns $\boldsymbol{\epsilon}$ through mechanism f [54]. Notably, $\mathbf{z}$ inside $f(\cdot)$ captures interactions among semantic concepts, modeled as a latent Markov network $\mathcal{M}$. This connects the cognitive decomposition of objects into components [2] with identifiability theory [16, 60].

2.2. Identifiability Theory

We establish three identifiability results for unseen category generalization. Full proofs and formal definitions are provided in the supplementary.

Theorem 2.1 (**Contextual Recovery**). *Under the DGP in Eq. 1, assume* ***(A1)*** *the joint density $p_{\mathbf{z},\mathbf{c}|y}$ is smooth and positive, and* ***(A2)*** *there exist $d_z + 1$ label values with linearly independent semantic score vectors $\boldsymbol{v}(\mathbf{z}, y_j) - \boldsymbol{v}(\mathbf{z}, y_0)$, where $\boldsymbol{v}(\mathbf{z}, y) = \left(\frac{\partial \log p(z_1|y)}{\partial z_1}, \ldots, \frac{\partial \log p(z_{d_z}|y)}{\partial z_{d_z}}\right)$. Then contextual concepts are recovered: $\hat{\mathbf{c}} = h(\mathbf{c})$ for some invertible h.*

Remark. A1 is a standard regularity condition. **A2** requires semantic concepts to vary sufficiently across categories—a natural property in classification tasks where categories exhibit distinct visual characteristics. Crucially, our conditions do not require $\mathbf{z}$ to be conditionally independent given y [38] nor parametric distributional assumptions [77].

Theorem 2.2 (**Semantic Recovery**). *Under the conditions of Thm. 2.1, additionally assume* ***(A3)*** *there exist $2d_z + |\mathcal{M}| + 1$ labels with linearly independent first- and second-order semantic score vectors, and* ***(A4)*** *the estimated Markov network is at least as sparse as the true one: $|\hat{\mathcal{M}}| \leq |\mathcal{M}|$. Then:* ***(i)*** *isolated concepts ($|\Psi_{\mathcal{M}}(z_i)|=0$, where $\Psi_{\mathcal{M}}$ denotes the maximum clique) are recovered component-wise: $\hat{z}_i = h(z_{\pi(i)})$;* ***(ii)*** *coupled concepts are recovered modularly: $\hat{z}_i = h(z_{\pi(i)} \cup \Psi_{\mathcal{M}}(z_{\pi(i)}))$.*

Remark. A3 extends **A2** to second-order statistics. **A4** encodes the standard prior that most semantic concepts are conditionally independent [56]. Sparsity serves a dual purpose: enabling component-wise identifiability and reducing effective dimensionality, making **A3** achievable with realistic numbers of training categories.

Theorem 2.3 (**Compositional Generalization**). *Additionally assume* ***(A5)*** *different concept values have disjoint supports: $z_i^{(k)} \neq z_i^{(l)} \Rightarrow \mathcal{S}_i(z_i^{(k)}) \cap \mathcal{S}_i(z_i^{(l)}) = \varnothing$, and* ***(A6)*** *all concept values appear with positive probability during training. Then for any unseen tuple $\mathbf{z}^q$, the product region $\mathcal{R}(\mathbf{z}^q) = \bigtimes_{i=1}^{d_z} \mathcal{S}_i(z_i^q)$ is disjoint from all seen categories.*

Remark. A5 specifies that distinct concept values (*e.g.*, "red" vs. "yellow") occupy separate support regions; without this, semantic ambiguity arises. **A6** requires concept primitives to appear during training, though in different combinations—a natural compositionality premise. Together, Thm. 2.1–2.3 guarantee that comparison identifies faithful concepts and composition enables recognition of novel categories as combinations of identified primitives.

3. Methodology

We instantiate C^3 for On-the-fly Category Discovery (OCD) [11], demanding genuine generalization without novel-category access. The overall architecture is shown in Fig. 3. Each component operationalizes theoretical conditions from Sec. 2.

Problem Setting. Given N training samples $\mathcal{D}_S = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ from seen categories $\mathcal{Y}_S$, we categorize query samples from $\mathcal{Y}_Q \supseteq \mathcal{Y}_S$. Only $\mathcal{D}_S$ is used for training; queries arrive individually at test time [11, 86].

Contextual-Semantic Separation (A1–A2). Features from a frozen DINO encoder [6] are transformed into intermediate ß $= \hat{g}^{-1}(\mathbf{x})$, then partitioned via learnable Bernoulli mask with Gumbel-Softmax [27]: $[\hat{\mathbf{z}}, \hat{\mathbf{c}}] = M(\text{ß})$. Contextual $\hat{\mathbf{c}}$ aligns with a standard Gaussian prior; semantics $\hat{\mathbf{z}}$ follows a label-conditioned prior via $\mathcal{L}_{\text{ELBO},c}$.

Semantic Concept Identification (A3–A4). Sparsity is enforced through ℓ_1 regularization $\mathcal{L}_s = \|\hat{\mathbf{z}}\|_1$ and a sparsely-gated MoE [62] that learns adjacency $M_{\mathbf{z}}$ via top-k masking. A normalizing flow $f_{\text{flow}}(\cdot, y)$ [58] models label-conditioned density on masked representations:

$$\mathcal{L}_{\text{flow}} = -\log p_{\epsilon}\big(f_{\text{flow}}(M_{\mathbf{z}} \odot \hat{\mathbf{z}}, y)\big) - \log\left|\det \tfrac{\partial f_{\text{flow}}}{\partial \hat{\mathbf{z}}}\right|. \quad (2)$$

Prototypical learning [7, 80] provides cross-category contrast for **A3**.

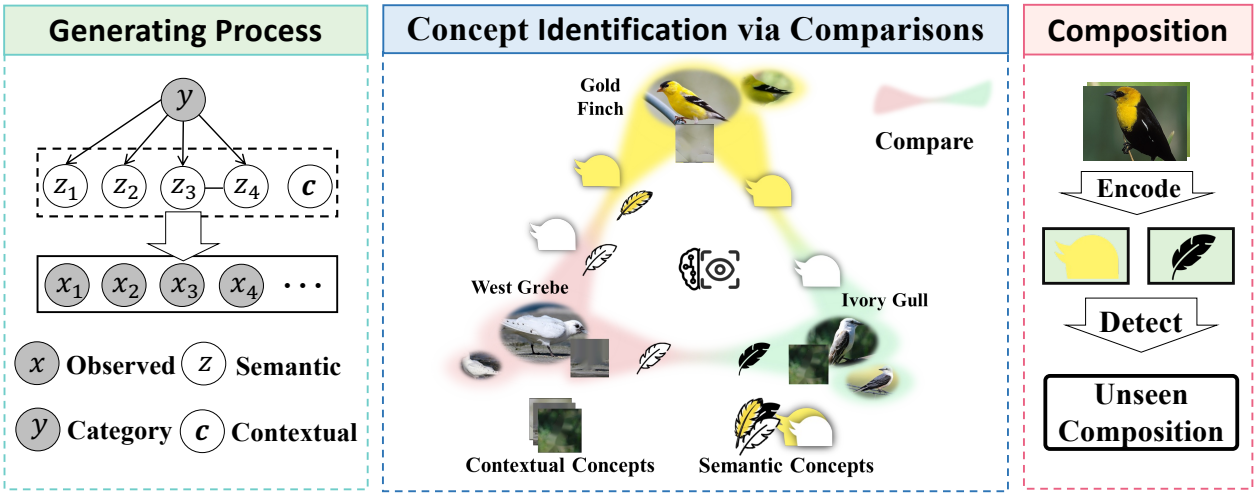

Figure 2. **The C^3 Framework.** *Left:* Data-generating process from labels y through semantic concepts $\mathbf{z}$ and contextual factors $\mathbf{c}$ to observations $\mathbf{x}$. *Middle:* Concept identification via comparison—cross-category variations separate semantic concepts (category-dependent) from contextual factors (category-invariant). *Right:* Compositional inference encodes queries into identified concepts and detects whether the resulting combination corresponds to a seen or unseen category.

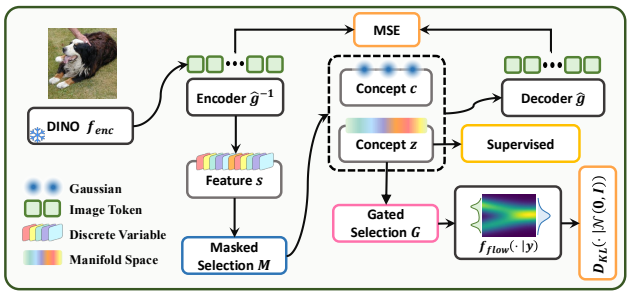

Figure 3. **Architecture** of C^3. Features from a frozen DINO encoder are partitioned into contextual $\hat{\mathbf{c}}$ and semantic $\hat{\mathbf{z}}$ via masked selection. Sparse gating learns Markov structure; normalizing flow models label-conditioned density.

Compositional Generalization (A5–A6). Discrete hash coding enforces support separation. Residual integration $\hat{\mathbf{a}} = \hat{\mathbf{z}} + f_{\text{residual}}(M_{\mathbf{z}} \odot \hat{\mathbf{z}})$ handles modular recovery per Thm. 2.2.

Learning Objective.

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{sup}}}_{\substack{\text{comparison}\\(\text{A2,A3,A5})}} + \underbrace{\mathcal{L}_s + \mathcal{L}_{\text{flow}}}_{\substack{\text{structure}\\(\text{A3,A4})}} + \underbrace{\mathcal{L}_{\text{ELBO},c}}_{\substack{\text{generative}\\(\text{A1,A2})}}, \tag{3}$$

where $\mathcal{L}_{\text{sup}}$ combines prototype and hash losses, $\mathcal{L}_s$ promotes sparsity, and $\mathcal{L}_{\text{ELBO},c}$ ensures reconstruction fidelity.

4. Experimental Results

4.1. Setup

We evaluate on eight fine-grained benchmarks: CUB-200 [74], Stanford Cars [41], Oxford-IIIT Pet [55], Food-101 [3], and four iNaturalist [71] subsets (Fungi, Arachnida, Animalia, Mollusca). Following Du et al. [11], we split each dataset into seen/unseen categories with 50% of seen-category samples for training, and adopt clustering accuracy with Strict-Hungarian matching. We compare against SLC [22], RankStat [21], WTA [28], SMILE [11], and PHE [83]. All methods use DINO ViT-B/16 [6].

4.2. Main Results

Tab. 1 presents results on four standard benchmarks. C^3 consistently outperforms all baselines, achieving $+3.8\%$ average accuracy over PHE across CUB, Stanford Cars, Oxford Pets, and Food101. Gains are observed on both seen (`Old`) and unseen (`New`) categories, indicating that identifiability-guided concept learning improves representation quality without sacrificing known-category performance. The $+2.1\%$ average improvement on `New` categories confirms genuine semantic transfer to unseen compositions. On four additional iNaturalist benchmarks (supplementary Tab. S1), C^3 achieves a further $+1.5\%$ average gain, validating scalability.

4.3. Ablation Study

We systematically ablate each component on CUB and Stanford Cars (Tab. 2). Removing $\mathcal{L}_{\text{proto}}$ (**A2**, **A3**) and $\mathcal{L}_{\text{ELBO},c}$ (**A1**, **A2**) causes the largest drops (-4.0% and

Table 1. **Main results** on four standard fine-grained benchmarks. Best / second-best: **bold** / underline. Results on four additional iNaturalist subsets are provided in the supplementary (Tab. S1).

Method	CUB (%)			Stanford Cars (%)			Oxford Pets (%)			Food101 (%)			Average (%)		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
SLC [22]	31.3	48.5	22.7	24.0	45.8	13.6	35.5	41.3	33.1	20.9	48.6	6.8	27.9	46.1	19.1
RankStat [21]	27.6	46.2	18.3	18.6	36.9	9.7	33.2	42.3	28.4	22.3	50.7	7.8	25.4	44.0	16.1
WTA [28]	26.5	20.0	38.8	10.6	24.4	13.6	35.2	46.3	29.3	18.2	40.5	6.1	25.0	42.7	15.8
SMILE [11]	32.2	50.8	22.9	26.1	46.6	16.2	41.2	42.1	40.7	24.0	54.6	8.4	30.9	48.6	22.1
PHE [83]	36.4	55.8	27.0	31.3	61.9	16.8	48.3	53.8	45.4	29.1	64.7	11.1	36.3	59.1	25.1
C^3 (Ours)	**40.1**	**62.1**	**29.5**	**34.1**	**69.0**	**17.8**	**54.7**	**63.9**	**49.6**	**31.5**	**68.3**	**11.9**	**40.1**	**65.8**	**27.2**

Setup	Assump.	CUB-200 (%)			SCars (%)		
		All	Old	New	All	Old	New
Full	–	**40.1**	**62.1**	**29.5**	**34.1**	**69.0**	**17.8**
$-\mathcal{L}_{\text{flow}}$	A3,A4	$38.5^{-1.6}$	$61.3^{-0.8}$	$27.0^{-2.5}$	$30.2^{-3.9}$	$59.6^{-9.4}$	$16.0^{-1.8}$
$-\mathcal{L}_s$	A4	$39.3^{-0.8}$	$62.4^{+0.3}$	$27.7^{-1.8}$	$31.1^{-3.0}$	$59.0^{-10.0}$	$17.6^{-0.2}$
$-\mathcal{L}_{\text{ELBO},c}$	A1,A2	$36.9^{-3.2}$	$59.8^{-2.3}$	$24.5^{-5.0}$	$27.6^{-6.5}$	$54.2^{-14.8}$	$15.5^{-2.3}$
$-\mathcal{L}_{\text{proto}}$	A2,A3	$36.1^{-4.0}$	$60.7^{-1.4}$	$24.3^{-5.2}$	$28.3^{-5.8}$	$55.8^{-13.2}$	$15.2^{-2.6}$
$-f_{\text{res}}$	Thm 2.3	$39.5^{-0.6}$	$61.8^{-0.3}$	$28.5^{-1.0}$	$33.2^{-0.9}$	$67.8^{-1.2}$	$17.1^{-0.7}$

Table 2. **Ablation study.** Removing any module degrades performance. Comparison components (**A2**, **A3**) show larger effects; composition components provide complementary gains.

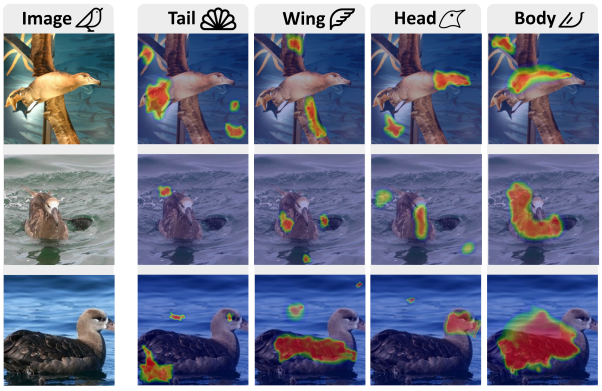

Figure 4. **Concept visualization** on CUB-200. Each column corresponds to a specific interpretable concept activated via prototypical attention, highlighting sparse, semantically meaningful regions.

-3.2% on CUB All), confirming that cross-category contrast and well-posed latent structure are the most critical conditions for concept identification. The flow module $\mathcal{L}_{\text{flow}}$ (**A3**, **A4**) shows moderate impact (-1.6%), while sparsity $\mathcal{L}_s$ (**A4**) and context separation $\mathcal{L}_{\text{ctx}}$ (**A2**) contribute smaller but consistent gains. Overall, components addressing comparison (**A2**, **A3**) exhibit larger effects than those for composition (**A4**), yet both are essential—confirming the complementary nature of the two cognitive mechanisms.

4.4. Concept Visualization

To provide evidence that learned representations correspond to semantically meaningful concepts, Fig. 4 shows probed concept activations on CUB-200. Different latent dimensions activate interpretable regions (*e.g.*, head, wings, tail, body), with activations appearing visually sparse and consistent across instances—supporting the sparse structure assumption (**A4**). We further validate via concept-guided proxy classification among visually confusing categories: distinguishing *Black-footed* from *Sooty Albatross*, discriminative regions (head: 95.0%, body: 92.3%) substantially outperform non-discriminative ones (tail: 64.1%), confirming that the learned concepts encode semantically meaningful distinctions. Synthetic experiments further validate our theory: identifiability metrics reach MCC > 0.9 and $R^2 > 95\%$ when the category count satisfies $2d_z + |\mathcal{M}| + 1 \leq n_s$ (supplementary Fig. S1).

5. Conclusion

We introduced C^3, a theoretical framework formalizing how machines generalize to unseen categories by identifying transferable concepts through cross-category comparison and recognizing novel classes as compositional recombinations. Our identifiability guarantees—sufficient contrast for concept separation, sparse structure for component recovery, and disjoint supports for compositional discrimination—are assumption-light and naturally satisfied. Experimental validation on eight benchmarks demonstrates both theoretical soundness and practical effectiveness ($+3.8\%$ average accuracy).

Limitations and Future Work. Our framework assumes concept primitives in unseen categories appear during training (**A6**); relaxing this via generative extrapolation is a promising direction. Integrating C^3 with end-to-end trainable or foundation-scale architectures and extending to multi-modal concept discovery are important avenues.

References

[1] Michel Besserve, Rémy Sun, Dominik Janzing, and Bernhard Schölkopf. A theory of independent mechanisms for extrapolation in generative models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6741–6749, 2021. 11

[2] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2): 115, 1987. 1, 2

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pages 446–461. Springer, 2014. 3, 19

[4] Jack Brady, Julius von Kügelgen, Sébastien Lachapelle, Simon Buchholz, Thomas Kipf, and Wieland Brendel. Interaction asymmetry: A general principle for learning composable abstractions. *arXiv preprint arXiv:2411.07784*, 2024. 11

[5] Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning, 2022. 1, 10

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2, 3

[7] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019. 2, 17

[8] Sua Choi, Dahyun Kang, and Minsu Cho. Contrastive mean-shift learning for generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23094–23104, 2024. 10

[9] Alessandro Conti, Massimiliano Mancini, Enrico Fini, Yiming Wang, Paolo Rota, and Elisa Ricci. On large multimodal models as open-world image classifiers. *arXiv preprint arXiv:2503.21851*, 2025. 1

[10] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 20

[11] Ruoyi Du, Dongliang Chang, Kongming Liang, Timothy Hospedales, Yi-Zhe Song, and Zhanyu Ma. On-the-fly category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11691–11700, 2023. 1, 2, 3, 4, 9, 10, 19, 20

[12] Yilun Du, Shuang Li, and Igor Mordatch. Compositional visual generation with energy based models. *Advances in Neural Information Processing Systems*, 33:6637–6647, 2020. 11

[13] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. *Advances in neural information processing systems*, 32, 2019. 20

[14] Russell Epstein and Nancy Kanwisher. A cortical representation of the local visual environment. *Nature*, 392(6676): 598–601, 1998. 2

[15] Jerry A Fodor and Zenon W Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988. 2

[16] Minghao Fu, Biwei Huang, Zijian Li, Yujia Zheng, Ignavier Ng, Guangyi Chen, Yingyao Hu, and Kun Zhang. Learning general causal structures with hidden dynamic process for climate analysis. *arXiv preprint arXiv:2501.12500*, 2025. 2, 11

[17] Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170, 1983. 1

[18] Yunchao Gong, Svetlana Lazebnik, Albert Gordo, and Florent Perronnin. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(12):2916–2929, 2013. 18

[19] Frédéric Gosselin and Philippe G Schyns. Bubbles: a technique to reveal the use of information in recognition tasks. *Vision research*, 41(17):2261–2271, 2001. 2

[20] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8401–8409, 2019. 10

[21] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Autonovel: Automatically discovering and learning novel visual categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6767–6781, 2021. 3, 4, 9, 20

[22] John A. Hartigan. *Clustering Algorithms*. John Wiley & Sons, Inc., New York, 1975. 3, 4, 9, 20

[23] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017. 10

[24] Biwei Huang, Charles Jia Han Low, Feng Xie, Clark Glymour, and Kun Zhang. Latent hierarchical causal structure discovery with rank constraints. *Advances in neural information processing systems*, 35:5549–5561, 2022. 11

[25] Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999. 11

[26] Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew G Wilson. On feature learning in the presence of spurious correlations. *Advances in Neural Information Processing Systems*, 35:38516–38532, 2022. 1

[27] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 2

[28] Xuhui Jia, Kai Han, Yukun Zhu, and Bradley Green. Joint representation learning and novel category discovery on single-and multi-modal data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 610–619, 2021. 3, 4, 9, 20

[29] Yangqing Jia, Joshua T Abbott, Joseph L Austerweil, Tom Griffiths, and Trevor Darrell. Visual concept learning: Combining machine vision and bayesian generalization on concept hierarchies. *Advances in Neural Information Processing Systems*, 26, 2013. 11

[30] Nancy Kanwisher, Josh McDermott, and Marvin M Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, 17(11):4302–4311, 1997. 2

[31] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 11

[32] 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. 10

[33] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 18

[34] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International conference on machine learning*, pages 2649–2658. PMLR, 2018. 10

[35] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 11

[36] David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. *arXiv preprint arXiv:2007.10930*, 2020. 10

[37] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020. 11

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

[39] Lingjing Kong, Biwei Huang, Feng Xie, Eric Xing, Yuejie Chi, and Kun Zhang. Identification of nonlinear latent hierarchical models. *Advances in Neural Information Processing Systems*, 36:2010–2032, 2023. 11

[40] Lingjing Kong, Guangyi Chen, Biwei Huang, Eric P Xing, Yuejie Chi, and Kun Zhang. Learning discrete concepts in latent hierarchical models. *arXiv preprint arXiv:2406.00519*, 2024. 11

[41] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 3, 19

[42] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 1

[43] Sébastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ica. In *Conference on Causal Learning and Reasoning*, pages 428–484. PMLR, 2022. 11

[44] Sébastien Lachapelle, Divyat Mahajan, Ioannis Mitliagkas, and Simon Lacoste-Julien. Additive decoders for latent variables identification and cartesian-product extrapolation. *Advances in Neural Information Processing Systems*, 36: 25112–25150, 2023. 11

[45] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015. 1

[46] Loka Li, Wong Yu Kang, Minghao Fu, Guangyi Chen, Zhenhao Chen, Gongxu Luo, Yuewen Sun, Salman Khan, Peter Spirtes, and Kun Zhang. Personax: Multimodal datasets with llm-inferred behavior traits. *arXiv preprint arXiv:2509.11362*, 2025. 1

[47] Wenbin Li, Zhichen Fan, Jing Huo, and Yang Gao. Modeling inter-class and intra-class constraints in novel class discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3449–3458, 2023. 10

[48] Ziyun Li, Jona Otholt, Ben Dai, Christoph Meinel, Haojin Yang, et al. A closer look at novel class discovery from the labeled set. *arXiv preprint arXiv:2209.09120*, 2022. 1, 11

[49] Zijian Li, Minghao Fu, Junxian Huang, Yifan Shen, Ruichu Cai, Yuewen Sun, Guangyi Chen, and Kun Zhang. Towards identifiability of hierarchical temporal causal representation learning. *arXiv preprint arXiv:2510.18310*, 2025. 11

[50] Juan Lin. Factorizing multivariate function classes. *Advances in neural information processing systems*, 10, 1997. 13

[51] 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. 11

[52] Andrei Margeloiu, Matthew Ashman, Umang Bhatt, Yanzhi Chen, Mateja Jamnik, and Adrian Weller. Do concept bottleneck models learn as intended? *arXiv preprint arXiv:2105.04289*, 2021. 11

[53] Arthur B Markman and Dedre Gentner. Thinking. *Annual Review of Psychology*, 51(1):223–247, 2000. 1

[54] David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010. 2

[55] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 3, 19

[56] William Peebles, John Peebles, Jun-Yan Zhu, Alexei Efros, and Antonio Torralba. The hessian penalty: A weak prior for unsupervised disentanglement. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 581–597. Springer, 2020. 2

[57] Sarah Rastegar, Hazel Doughty, and Cees Snoek. Learn to categorize or categorize to learn? self-coding for generalized category discovery. *Advances in Neural Information Processing Systems*, 36, 2024. 10

[58] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015. 2

[59] Eleanor H Rosch. Natural categories. *Cognitive psychology*, 4(3):328–350, 1973. 1

[60] 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. 2

[61] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 18

[62] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017. 2

[63] Sungbin Shin, Yohan Jo, Sungsoo Ahn, and Namhoon Lee. A closer look at the intervention procedure of concept bottleneck models. In *International Conference on Machine Learning*, pages 31504–31520. PMLR, 2023. 11

[64] Lei Shu, Hu Xu, and Bing Liu. Unseen class discovery in open-world classification. *arXiv preprint arXiv:1801.05609*, 2018. 1

[65] Alexander J Smola, A Gretton, and K Borgwardt. Maximum mean discrepancy. In *13th international conference, ICONIP*, pages 3–6, 2006. 11

[66] Yiyou Sun, Zhenmei Shi, Yingyu Liang, and Yixuan Li. When and how does known class help discover unknown ones? provable understanding through spectral analysis. *arXiv preprint arXiv:2308.05017*, 2023. 10, 11

[67] Yiyou Sun, Zhenmei Shi, and Yixuan Li. A graph-theoretic framework for understanding open-world semi-supervised learning. *Advances in Neural Information Processing Systems*, 36, 2024. 11

[68] Luyao Tang, Kunze Huang, Chaoqi Chen, Yuxuan Yuan, Chenxin Li, Xiaotong Tu, Xinghao Ding, and Yue Huang. Dissecting generalized category discovery: Multiplex consensus under self-deconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 297–307, 2025. 11

[69] Luyao Tang, Yuxuan Yuan, Chaoqi Chen, Zeyu Zhang, Yue Huang, and Kun Zhang. Ocrt: Boosting foundation models in the open world with object-concept-relation triad. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25422–25433, 2025. 11

[70] Vishaal Udandarao, Ameya Prabhu, Adhiraj Ghosh, Yash Sharma, Philip Torr, Adel Bibi, Samuel Albanie, and Matthias Bethge. No" zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1

[71] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 3, 19

[72] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7492–7501, 2022. 1, 10

[73] Sagar Vaze, Andrea Vedaldi, and Andrew Zisserman. No representation rules them all in category discovery. *Advances in Neural Information Processing Systems*, 36, 2024. 11

[74] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical report, Computation & Neural Systems Technical Report, 2011. 3, 19

[75] Yair Weiss, Antonio Torralba, and Rob Fergus. Spectral hashing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2009. 18

[76] Xin Wen, Bingchen Zhao, and Xiaojuan Qi. Parametric classification for generalized category discovery: A baseline study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16590–16600, 2023. 1, 10

[77] Thaddäus Wiedemer, Jack Brady, Alexander Panfilov, Attila Juhos, Matthias Bethge, and Wieland Brendel. Provable compositional generalization for object-centric learning. *arXiv preprint arXiv:2310.05327*, 2023. 2

[78] Jiajun Wu, Joshua B Tenenbaum, and Pushmeet Kohli. Neural scene de-rendering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 699–707, 2017. 11

[79] Shengzhou Xiong, Yihua Tan, and Guoyou Wang. Explore visual concept formation for image classification. In *Proceedings of the 38th International Conference on Machine Learning*, pages 11470–11479. PMLR, 2021. 11

[80] Mengqi Xue, Qihan Huang, Haofei Zhang, Lechao Cheng, Jie Song, Minghui Wu, and Mingli Song. Protopformer: Concentrating on prototypical parts in vision transformers for interpretable image recognition. *arXiv preprint arXiv:2208.10431*, 2022. 2, 17

[81] Weiran Yao, Guangyi Chen, and Kun Zhang. Temporally disentangled representation learning. *Advances in Neural Information Processing Systems*, 35:26492–26503, 2022. 11, 14

[82] Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. *arXiv preprint arXiv:2205.15480*, 2022. 11

[83] Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: A general setting. *arXiv preprint arXiv:2402.05052*, 2024. 3, 4, 9, 11, 13

[84] Sheng Zhang, Salman Khan, Zhiqiang Shen, Muzammal Naseer, Guangyi Chen, and Fahad Shahbaz Khan. Promptcal: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3479–3488, 2023. 10

[85] Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification? *arXiv preprint arXiv:2405.18415*, 2024. 1

[86] Haiyang Zheng, Nan Pu, Wenjing Li, Nicu Sebe, and Zhun Zhong. Prototypical hash encoding for on-the-fly fine-grained category discovery. *arXiv preprint arXiv:2410.19213*, 2024. 1, 2, 10, 17, 18, 19, 20

[87] Yujia Zheng, Shaoan Xie, and Kun Zhang. Nonparametric identification of latent concepts. In *Forty-second International Conference on Machine Learning*. 11

[88] Fei Zhu, Shijie Ma, Zhen Cheng, Xu-Yao Zhang, Zhaoxiang Zhang, and Cheng-Lin Liu. Open-world machine learning: A review and new outlooks, 2024. 1

From Comparison to Composition: Towards Understanding Machine Cognition of Unseen Categories

Supplementary Material

Contents: **A.** Extended Theory | **B.** Additional Results | **C.** Related Work | **D.** Proofs | **E.** Experimental Details

A. Extended Theory

A.1. Formal Identification Criteria

Let $\mathbf{X} = \{\mathbf{x}_i\}_{\mathcal{X}}$ be observations from the true model $(f, g, p(\epsilon))$ in Eq. 1. A learned model $(\hat{f}, \hat{g}, \hat{p}(\boldsymbol{\epsilon}))$ is *observationally equivalent* if $p_{\hat{f},\hat{g},\hat{p}(\boldsymbol{\epsilon})}(\hat{\mathbf{x}}) = p_{f,g,p(\epsilon)}(\mathbf{x})$. We seek conditions ensuring recovery up to a bijective transformation h:

i. *Contextual Subspace:* $\hat{\mathbf{c}} = h(\mathbf{c})$—estimated context contains no semantic information.
ii. *Semantic Component:* isolated concepts satisfy $\hat{z}_i = h_i(z_{\pi(i)})$; coupled concepts satisfy $\hat{z}_i = h(z_{\pi(k)}, z_{\pi(l)})$.
iii. *Unseen Category:* for query $y^q \in Y_Q$, the concept tuple $\phi(y^q) \notin \Phi_S$ iff $\mathbf{z}^q \in \mathcal{R}(\mathbf{z}^q)$ and $\mathcal{R}(\mathbf{z}^q) \cap \mathcal{R}(\mathbf{z}^s) = \varnothing$ for all seen $\mathbf{z}^s \in \Phi_S$.

A.2. Proof Sketches

Theorem 2.1 (Contextual Recovery). The key observation is that contextual factors $\mathbf{c}$, by definition, have category-invariant distributions, while semantic factors $\mathbf{z}$ vary with label y. Under observational equivalence, any learned model must capture this invariance-variance structure, which forces separation of $\mathbf{c}$ and $\mathbf{z}$. The full derivation proceeds via log-density matching and first-order differentiation under **A2**'s linear independence condition.

Theorem 2.2 (Semantic Recovery). Building on subspace separation, we differentiate the log-density matching condition under **A3**. This yields Jacobian constraints: $h'_{i,l}h'_{i,k} = 0$, $h'_{j,l}h'_{i,k} = 0$, and $h''_{i,kl} = 0$, implying each true concept z_i depends on at most one estimated coordinate $\hat{z}_k$. The sparsity constraint **A4** further prevents adjacent Markov nodes from mapping to the same coordinate, enabling component-wise or modular recovery depending on clique structure.

Theorem 2.3 (Compositional Generalization). We leverage factorization: conditioned on y, the latent density decomposes as $p(\mathbf{z}|y) = \prod_{i=1}^{d_z} p_i(z_i|y, \Psi_{\mathcal{M}}(z_i))$. Combined with disjoint supports (**A5**), this ensures the joint concept space is a Cartesian product of coordinate-wise supports. Coverage (**A6**) guarantees each region is learnable. Thus, any unseen tuple $\mathbf{z}^q$ lies in a product region provably disjoint from all seen categories.

B. Additional Results

B.1. Results on iNaturalist

Table S1. **Results on four iNaturalist subsets.** Best / second-best: **bold** / underline.

Method	Fungi			Arachnida			Animalia			Mollusca			Average		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
SLC [22]	27.7	60.0	13.4	25.4	44.6	11.4	32.4	**61.9**	19.3	31.1	59.8	15.0	29.2	56.6	14.8
RankStat [21]	23.8	50.5	12.0	26.6	51.0	10.0	31.4	54.9	21.6	29.3	55.2	15.5	27.8	52.9	14.8
WTA [28]	27.5	65.6	12.0	28.1	55.5	10.9	33.4	59.8	22.4	30.3	55.4	17.0	29.8	59.1	15.6
SMILE [11]	29.3	64.6	13.6	29.9	57.9	12.2	35.9	49.4	30.3	33.3	44.5	27.2	32.1	54.1	20.8
PHE [83]	31.4	67.9	15.2	37.0	**75.7**	12.6	40.3	55.7	31.8	39.9	65.0	26.5	37.2	66.1	21.5
C^3 (Ours)	**32.9**	**69.8**	**16.2**	**38.1**	72.0	**13.1**	**42.0**	60.1	**32.2**	**41.9**	**70.2**	**27.3**	**38.7**	**68.1**	**22.2**

B.2. Synthetic Validation

To verify our theoretical results under controlled conditions, we generate data following Eq. 1 and compare against disentangled representation learning baselines: i-VAE [32], FactorVAE [34], SlowVAE [36], and β-VAE [23]. As shown in Fig. S1, C^3 consistently outperforms all baselines. Both MCC and R^2 improve monotonically with training categories, reaching high identifiability (MCC > 0.9, $R^2 > 95\%$) when $2d_z + |\mathcal{M}| + 1 \leq n_s$, validating the theoretical predictions.

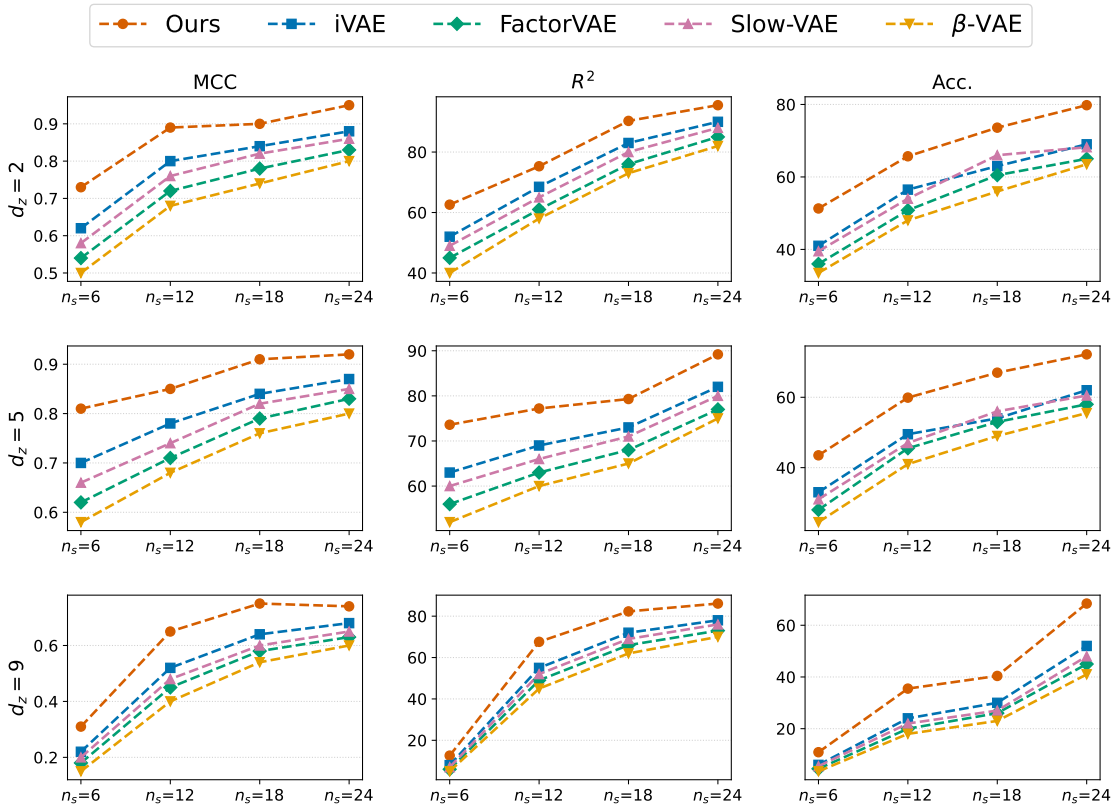

Figure S1. **Synthetic data comparisons.** C^3 achieves high identifiability (MCC > 0.9, $R^2 > 95\%$) when the number of training categories exceeds the theoretical threshold.

C. Related Work

C.1. On-the-fly Category Discovery

Task Evolution. *Novel Category Discovery* (NCD), originally formulated in [20], aims to cluster unlabeled data from novel categories by leveraging knowledge from labeled known categories. *Generalized Category Discovery* (GCD) [5, 72] extends NCD by assuming unlabeled data may contain both known and novel categories. Despite progress in NCD/GCD [8, 20, 47, 57, 66, 76, 84], these settings assume access to unlabeled data from target categories during training, which is often violated in practice.

On-the-fly Category Discovery (OCD). OCD [11] addresses this limitation through two key modifications: `(i)` exclusion of unlabeled novel-category data during training, and `(ii)` streaming inference, where query instances arrive individually and require instant feedback. This setting captures the *genuine* challenge of semantic generalization: learning representations from known categories that transfer to unseen ones without any exposure. Prior OCD methods, including SMILE [11] using hash-based prototypes, and PHE [86] enhancing discrete code discriminability, demonstrate empirical progress but lack theoretical grounding for *why* and *when* such transfer succeeds. Our work bridges this gap by formalizing the human *compare-then-compose* mechanism into assumption-light identifiability conditions, offering both principled guarantees and actionable guidelines for when and why such transfer succeeds.

C.2. Theoretical Perspectives on Semantic Transfer

Existing Theoretical Work. Several studies have explored theoretical aspects of category discovery. Li et al. [48] examines transfer from known to novel categories using maximum mean discrepancy metrics [65]. Sun et al. [66] derives generalization bounds based on semi-supervised contrastive representations, while Sun et al. [67] analyzes spectral graph-based contrastive learning for GCD. Complementing these theoretical results, Vaze et al. [73] empirically shows that physically-grounded concepts (color, shape, material) enhance transferability.

Limitations and Our Contribution. These results share a critical limitation: they rely on unlabeled novel-category data during training, and thus are inapplicable to real-world OCD, where such data is unavailable. Our framework addresses this by establishing identifiability conditions under which semantic concepts can be recovered from seen-category comparisons alone and composed to recognize unseen ones. This provides the *first* theoretical foundation for genuine semantic generalization without novel-category exposure.

C.3. Latent Variable Identification

Core Challenge. Latent variable identification (LVI) aims to recover underlying factors from observations when direct measurement is unavailable [16, 81]. A fundamental challenge is that latent variables remain unidentifiable in the non-parametric setting, *i.e.*, under arbitrary nonlinear mixing [25], even when independence is assumed [51]. This motivates our search for additional structure that enables recovery in our problem setting.

Identification via Auxiliary Information and Sparsity. Recent advances establish identifiability through auxiliary information [83] or structural constraints. Lachapelle et al. [43] shows that sparsity mechanisms enable concept-level interpretation, while hierarchical approaches [24, 39, 40, 49] uncover multi-level latent structures through rank constraints or distributional assumptions. However, these methods typically rely on linear structure, interventional data, or parametric mixing functions, assumptions that rarely hold in visual recognition tasks.

Connection to Our Framework. Our identifiability results depart from this line by leveraging *label-induced distributional changes*: we require only that semantic concepts vary sufficiently across categories, without parametric constraints on the mixing function. With a flexible sparse Markov assumption, this yields component-wise recovery guarantees applicable to the nonlinear entanglement prevalent in visual domains.

C.4. Concept Learning and Compositional Generalization

Concept-Based Representations. Visual concept learning spans retrieval [35], captioning [31], scene understanding [78], and explainable classification [29, 79]. Concept Bottleneck Models (CBMs) [37, 63, 82] use human-annotated concepts for interpretable intermediate bottleneck representations before classifiers. However, existing CBMs face three critical limitations that hinder their applicability and scalability: `(i)` reliance on expensive expert annotations or textual supervision on known categories, `(ii)` inability to capture patterns beyond human intuition [52], and `(iii)` lack of adaptability to novel categories without test-time concept inspection or annotation. Recent work [68, 69] decomposes representations into primitive components but assumes novel-category access during training or lacks faithfulness guarantees.

Compositional Generalization. Prior work approaches composition from empirical, theoretical, and structural perspectives. Du et al. [12] uses energy-based models for Cartesian-product extrapolation but assumes extremely low latent dimensionality, *i.e.*, single-factor variation. Besserve et al. [1] formalizes out-of-distribution samples as decoder-layer transformations. Lachapelle et al. [44] achieves identification through additive decoders, while Brady et al. [4] formalizes composition via interaction asymmetry. Zheng et al. [87] identifies factors through sparse connectivity diversity.

Our Distinction. Our framework establishes a unified foundation for identifiability and compositionality under assumption-light conditions. In contrast to CBMs, our concepts emerge from data without manual annotation yet satisfy faithfulness recovery guarantees. In contrast to prior work on compositional generalization, we impose neither parametric nor factor dimensionality constraints, with only natural semantic diversity and disjoint concept supports assumptions. Together, these conditions establish reliable mechanisms of unseen category generalization, bridging the gap between identifiability theory and open-world scenarios.

D. Proofs

To better understand our proof, we first present some useful definitions regarding the graphical model.

D.1. Markov Network

A Markov network (or Markov random field) is a graphical model that represents the joint distribution of a set of random variables using an undirected graph.

Definition D.1 (Markov Network)**.** Markov network is an undirected graph $G = (V, E)$ with a set of random variables $X_{v\in V}$, where any two non-adjacent variables are conditionally independent given all other variables. That is,

$$X_a \perp X_b | X_{V\setminus\{a,b\}}, \quad \forall (a,b) \notin E. \tag{S1}$$

Markov Networks and Directed Acyclic Graphs (DAGs) are both graphical models employed to represent joint distributions and to illustrate conditional independence properties.

D.2. Isomorphism of Markov networks

Definition D.2 (Isomorphism of Markov networks)**.** We let the $V(\cdot)$ be the vertical set of any graphs, an isomorphism of Markov networks M and $\hat{M}$ is a bijection between the vertex sets of M and $\hat{M}$

$$f : V(M) \to V(\hat{M})$$

such that any two vertices u and v of M are adjacent in G if and only if $f(u)$ and $f(v)$ are adjacent in $\hat{M}$.

D.3. Proof of Theorem 2.1

Proof. We begin with the matched marginal distribution $p_{\mathbf{x}|y}$ to bridge the relation between $\mathbf{z}$ and $\hat{\mathbf{z}}$. For brevity, we use y to represent the labels of known categories, $y^s \in \mathcal{Y}_s$. Suppose that $\hat{g} : \mathcal{Z} \times \mathcal{C} \to \mathcal{X}$ is an invertible estimated generating function, we have Eq. S2.

$$\forall y \in \mathcal{Y}_s, \quad p_{\hat{\mathbf{x}}|y} = p_{\mathbf{x}|y} \Longleftrightarrow p_{\hat{g}(\hat{\mathbf{z}},\hat{\mathbf{c}})|y} = p_{g(\mathbf{z},\mathbf{c})|y}. \tag{S2}$$

Sequentially, by using the change of variables formula, we can further obtain Eq. S3

$$p_{\hat{g}(\hat{\mathbf{z}},\hat{\mathbf{c}}|y)} = p_{g(\mathbf{z},\mathbf{c}|y)} \Longleftrightarrow p_{g^{-1}\circ g(\hat{\mathbf{z}},\hat{\mathbf{c}})|y}|\boldsymbol{J}_{g^{-1}}| = p_{\mathbf{z},\mathbf{c}|y}|\boldsymbol{J}_{g^{-1}}| \Longleftrightarrow p_{h(\hat{\mathbf{z}},\hat{\mathbf{c}})|y} = p_{\mathbf{z},\mathbf{c}|y}, \tag{S3}$$

where $h := g^{-1} \circ g$ is the transformation between the ground-true and the estimated latent variables, respectively. $\boldsymbol{J}_{g^{-1}}$ denotes the absolute value of Jacobian matrix determinant of g^{-1}. Since we assume that g and $\hat{g}$ are invertible, $|\boldsymbol{J}_{g^{-1}}| \neq 0$ and h is also invertible.

According to A2 (conditional independent assumption), we can have Eq. S4.

$$p_{\mathbf{z}|y}(\mathbf{z}|y) = p_{\mathbf{c}|y}(\mathbf{c}|y) \cdot p_{\mathbf{z}|y}(\mathbf{z}|y); \quad p_{\hat{\mathbf{z}}|y}(\hat{\mathbf{z}}|y) = p_{\hat{\mathbf{z}}|y}(\hat{\mathbf{z}}|y) \cdot p_{\hat{\mathbf{c}}|y}(\hat{\mathbf{c}}|y). \tag{S4}$$

For convenience, we take the logarithm on both sides of Eq. S4 and further let $q_s = \log p_{\mathbf{z}|y}(\mathbf{z}|y), q_c = \log p_{\mathbf{c}|y}(\mathbf{c}|y), p_s = \log p_{\hat{\mathbf{c}}|y}(\hat{\mathbf{c}}|y), p_c = \log p_{\hat{\mathbf{c}}|y}(\hat{\mathbf{c}}|y)$. Hence we have:

$$\log p_{\mathbf{z}|y}(\mathbf{z}|y) = q_s + q_c; \quad \log p_{\hat{\mathbf{z}}|y}(\hat{\mathbf{z}}|y) = p_s + p_c. \tag{S5}$$

By combining Eq. S5 and Eq. S3, we have:

$$p_{\mathbf{z}|y} = p_{h(\hat{\mathbf{z}}|y)} \Longleftrightarrow p_{\hat{\mathbf{z}}|y} = p_{\mathbf{z}|y}|\boldsymbol{J}_{h^{-1}}| \Longleftrightarrow q_s + q_c + \log|\boldsymbol{J}_{h^{-1}}| = p_s + p_c, \tag{S6}$$

where $\boldsymbol{J}_{h^{-1}}$ are the Jacobian matrix of h^{-1}.

Sequentially, we take the first-order derivative with $\hat{z}_j$ on Eq. (S6), where $\hat{z}_j$ is from $\mathbf{c}$, and have

$$\sum_{z_i\in\mathbf{z}} \frac{\partial q_s}{\partial z_i} \cdot \frac{\partial z_i}{\partial \hat{z}_j} + \sum_{z_i\in\mathbf{c}} \frac{\partial q_c}{\partial z_i} \cdot \frac{\partial z_i}{\partial \hat{z}_j} + \frac{\partial \log|\boldsymbol{J}_{h^{-1}}|}{\partial \hat{z}_j} = \frac{\partial p_s}{\partial \hat{z}_j} + \frac{\partial p_c}{\partial \hat{z}_j}. \tag{S7}$$

Suppose $y = y_0, y_1, \cdots, y_{n_z}$, we subtract the Eq. S7 corresponding to y_k with that corresponds to y_0, and we have:

$$\begin{aligned} \sum_{z_i\in\mathbf{z}} \left(\frac{\partial q_s(y_k)}{\partial z_i} - \frac{\partial q_s(y_0)}{\partial z_i}\right) \cdot \frac{\partial z_i}{\partial \hat{z}_j} + \sum_{z_i\in\mathbf{c}} \left(\frac{\partial q_c(y_k)}{\partial z_i} - \frac{\partial q_c(y_0)}{\partial z_i}\right) \cdot \frac{\partial z_i}{\partial \hat{z}_j} \\ = \frac{\partial \hat{q}_s(y_k)}{\partial \hat{z}_j} - \frac{\partial \hat{q}_s(y_0)}{\partial \hat{z}_j} + \frac{\partial \hat{q}_c(y_k)}{\partial \hat{z}_j} - \frac{\partial \hat{q}_c(y_0)}{\partial \hat{z}_j}. \end{aligned} \tag{S8}$$

Since the distribution of estimated $\hat{z}_j$ does not change across different categories, $\frac{\partial \hat{q}_s(y_k)}{\partial \hat{z}_j} - \frac{\partial \hat{q}_s(y_0)}{\partial \hat{z}_j} = 0$. Since $\frac{\partial q_s(y_k)}{\partial z_i}$ does not change across different categories, $\frac{\partial q_c(y_k)}{\partial z_i} = \frac{\partial q_c(y_0)}{\partial z_i}$, $\frac{\partial q_s(y_k)}{\partial z_i} = \frac{\partial q_s(y_0)}{\partial z_i}$ for $z_i \in \mathcal{Z}_s$. So we have

$$\sum_{i \in \mathbf{c}} \left(\frac{\partial q_s(y_k)}{\partial z_i} - \frac{\partial q_s(y_0)}{\partial z_i} \right) \cdot \frac{\partial z_i}{\partial \hat{z}_j} = 0. \tag{S9}$$

Based on the linear independence assumption (A3), the linear system is a $n_z \times n_z$ full-rank system. Therefore, the only solution is $\frac{\partial z_i}{\partial \hat{z}_j} = 0$.

Since $h(\cdot)$ is smooth over $\mathcal{Z}$, its Jacobian can be formalized as follows

$$\boldsymbol{J}_h = \left[\begin{array}{c|c} \mathbf{A} := \frac{\partial \mathbf{z}}{\partial \hat{\mathbf{z}}} & \mathbf{B} := \frac{\partial \mathbf{z}}{\partial \hat{\mathbf{c}}} \\ \hline \mathbf{C} := \frac{\partial \mathbf{c}}{\partial \hat{\mathbf{z}}} & \mathbf{D} := \frac{\partial \mathbf{c}}{\partial \hat{\mathbf{c}}}. \end{array} \right] \tag{S10}$$

Note that $\frac{\partial z_i}{\partial \hat{z}_j} = 0$ for $z_i \in \mathcal{Z}$ and $z_j \in \mathcal{Z}$ means that $\mathbf{B} = 0$. Since $h(\cdot)$ is invertible, $\boldsymbol{J}_h$ is a full-rank matrix. Therefore, for each $\mathbf{z}$, there exists a h_i such that $\mathbf{z} = h_i(\hat{\mathbf{z}})$. □

D.4. Proof of Theorem 2.2

We begin by presenting a useful lemma from [83], which connects group-wise transformations to component-wise transformations in a Markov network. This lemma is instrumental for the subsequent proof, in particular, it enables us to first recover the latent variables within groups of adjacent nodes in the Markov network.

Lemma D.3 (Identifiability of Hidden Causal Variables)**.** *If z_i is a function of at most one of $\hat{z}_k$ and $\hat{z}_l$, and given that z_i and z_j are adjacent in Markov network $\mathcal{M}_{\mathbf{z}}$, at most one of them is a function of $\hat{z}_k$ or $\hat{z}_l$. Then, there exists a permutation π of the estimated hidden variables, denoted as $\hat{z}_\pi$, such that each $\hat{z}_{\pi(i)}$ is a function of (a subset of) the variables in $\{\mathbf{z}_i\} \cup \Psi_{\mathbf{z}_i}$.*

Proof. Step 0 (Setup and change of variables). By Theorem 2.1, there exists an invertible, dimension-preserving h such that

$$h(\hat{\mathbf{z}}) = \mathbf{z} \implies p_{h(\hat{\mathbf{z}})} = p_{\mathbf{z}}.$$

Let J_h be the Jacobian of h and $J_{h^{-1}}$ that of h^{-1}. By the change-of-variables formula,

$$p(\hat{\mathbf{z}} \mid \hat{y}^s) \left| \det J_{h^{-1}}(\hat{\mathbf{z}}) \right| = p(\mathbf{z} \mid y) \quad \Longrightarrow \quad \log p(\hat{\mathbf{z}} \mid \hat{y}^s) = \log p(\mathbf{z} \mid y) + \log \left| \det J_h(\mathbf{z}) \right|. \tag{S11}$$

Suppose $\hat{z}_k \perp\!\!\!\perp \hat{z}_l \mid \hat{z}_{[n] \setminus \{k,l\}}$ (i.e., k, l are non-adjacent in the Markov network over $\hat{\mathbf{z}}$). Then for each $\hat{y}^s$, by [50],

$$\frac{\partial^2}{\partial \hat{z}_k \, \partial \hat{z}_l} \log p(\hat{\mathbf{z}} \mid \hat{y}^s) = 0. \tag{S12}$$

Differentiate Eq. S11 w.r.t. $\hat{z}_k$:

$$\frac{\partial}{\partial \hat{z}_k} \log p(\hat{\mathbf{z}} \mid \hat{y}^s) = \sum_{i=1}^{n} \frac{\partial \log p(\mathbf{z} \mid y)}{\partial z_i} \frac{\partial z_i}{\partial \hat{z}_k} + \frac{\partial}{\partial \hat{z}_k} \log \left| \det J_h(\mathbf{z}) \right|.$$

Introduce the shorthand

$$\eta(y) := \log p(\mathbf{z} \mid y), \quad \eta_i'(y) := \frac{\partial \log p(\mathbf{z} \mid y)}{\partial z_i}, \quad \eta_{ij}''(y) := \frac{\partial^2 \log p(\mathbf{z} \mid y)}{\partial z_i \partial z_j}, \quad h_{i,l}' := \frac{\partial z_i}{\partial \hat{z}_l}, \quad h_{i,kl}'' := \frac{\partial^2 z_i}{\partial \hat{z}_k \partial \hat{z}_l}.$$

Differentiating again w.r.t. $\hat{z}_l$ and using Eq. S12 yields

$$\begin{aligned} 0 &= \sum_{j=1}^{n} \sum_{i=1}^{n} \eta_{ij}''(y) \, h_{j,l}' \, h_{i,k}' + \sum_{i=1}^{n} \eta_i'(y) \, h_{i,kl}'' + \frac{\partial^2}{\partial \hat{z}_k \partial \hat{z}_l} \log \left| \det J_h(\mathbf{z}) \right| \\ &= \sum_{i=1}^{n} \eta_{ii}''(y) \, h_{i,l}' \, h_{i,k}' + \sum_{j=1}^{n} \sum_{i : \{z_j, z_i\} \in \mathcal{E}(\mathcal{M}_{\mathbf{z}})} \eta_{ij}''(y) \, h_{j,l}' \, h_{i,k}' + \sum_{i=1}^{n} \eta_i'(y) \, h_{i,kl}'' + \frac{\partial^2}{\partial \hat{z}_k \partial \hat{z}_l} \log \left| \det J_h(\mathbf{z}) \right|. \end{aligned} \tag{S13}$$

Here $\mathcal{E}(\mathcal{M}_{\mathbf{z}})$ denotes the edges of the Markov network over $\mathbf{z}$.

By Assumption **A3**, pick $2d_z + |\mathcal{M}_{\mathbf{z}}| + 1$ values $y^{(u)}$, $u = 0, \ldots, 2d_z + |\mathcal{M}_{\mathbf{z}}|$, so that Eq. S13 holds. Subtract the $u = 0$ instance from each $u \geq 1$, the Jacobian term cancels, yielding constraints below.

From the linear independence condition (Assumption **A3**), we deduce that for any edge $\{i, j\} \in \mathcal{E}(\mathcal{M}_{\mathbf{z}})$,

$$h'_{i,k}h'_{i,l} = 0, \qquad h'_{i,k}h'_{j,l} + h'_{j,k}h'_{i,l} = 0, \qquad h''_{i,kl} = 0.$$

The constraints imply that each z_i can depend on at most one of $\hat{z}_k, \hat{z}_l$. By contradiction: if $h'_{i,k}h'_{j,l} \neq 0$, then $h'_{i,l} = 0$ by the first constraint, which makes the second constraint force $h'_{i,k}h'_{j,l} = 0$, contradiction. Thus at most one of an adjacent pair (z_i, z_j) depends on a given recovered coordinate. Hence, isolated z_i yield component-wise identifiability ($\hat{z}_{\pi(i)} = h_i(z_i)$), while nodes in a clique can only be recovered modularly. Sparsity ensures most concepts are identifiable as individual components.

By Lemma D.3, there exists a permutation π such that each $\hat{z}_{\pi(i)}$ is a function of $\{z_i\} \cup \Psi_{z_i}$, where Ψ_{z_i} are the neighbors of z_i in $\mathcal{M}_{\mathbf{z}}$. In sparse cases this reduces to invertible component-wise identifiability.

□

Illustrative Examples of Assumptions This assumption characterizes the discriminative component of the model. The condition about the linear independence implies that there exists a unique characteristic of the concept that cannot be linearly represented by other variables. To clarify this assumption, we provide two examples [81] to demonstrate scenarios where the assumption holds and where it does not. Let $\eta_k = \frac{\partial q_k(d_k, y)}{\partial d_k}$.

Example 1: Violation of the Assumption (Additive Gaussian Noise) Consider a case where the assumption is violated due to the presence of additive Gaussian noise. Let y denote the label, and let $d_k = q_k(y) + \epsilon_k$, where $\epsilon_k \sim N(0, 1)$. In this scenario, we have: $\eta_k = -\log\sqrt{2\pi} - \frac{(d_k - q_k(y))^2}{2}$, and $\frac{\partial^2 \log P(d_k|y)}{\partial^2 d_k} = 0$. This result violates the assumption because the second derivative of the log-likelihood with respect to d_k is zero, indicating a lack of discriminative power in the latent variables.

Example 2: Validation of the Assumption (Generalized Normal Distribution) Conversely, consider a case where the assumption holds. Let ϵ_k follow a zero-mean generalized normal distribution: $P(\epsilon_k) \propto e^{-\lambda|\epsilon_k|^\beta}$, where $\lambda > 0$, $\beta > 2$, and $\beta \neq 3$. Let $d_k = q_k(y) + \epsilon_k$, where q is a linear function. If, for each d_k, there exists at least one l such that $c_{kl} = \frac{\partial d_k}{\partial y_l} \neq 0$, the assumption must hold.

In this case, we derive the following:

$$\frac{\partial^3 \eta_k}{\partial^2 d_k \partial y_l} = -\lambda \operatorname{sgn}(\epsilon_k)\, \beta(\beta - 1)(\beta - 2)|\epsilon_k|^{\beta-3} c_{kl},$$

and

$$\frac{\partial^2 \eta_k}{\partial d_k \partial y_l} = -\lambda\beta(\beta - 1)|\epsilon_k|^{\beta-2} c_{kl}.$$

Here, $|\epsilon_k|^{\beta-2}$ and $|\epsilon_k|^{\beta-3}$ are linearly independent because their ratio, $|\epsilon_k|$, is not constant. Furthermore, the functions $|\epsilon_{lt}|^{\beta-2}$ and $|\epsilon_{lt}|^{\beta-3}$, for $l = 1, 2, \ldots, n$, are $2n$ linearly independent functions due to their distinct arguments.

Suppose there exist coefficients α_{l1} and α_{l2} for $l = 1, 2, \ldots, n$ such that the weighted sum with respect to $\mathbf{w}_{l,t}$ is zero:

$$\alpha_{k1}c_{kl}|\epsilon_k|^{\beta-2} + \alpha_{k2}c_{kl}|\epsilon_k|^{\beta-3} + \sum_{l \neq k} \left(\alpha_{l1}c_{ll}|\epsilon_{lt}|^{\beta-2} + \alpha_{l2}c_{ll}|\epsilon_{lt}|^{\beta-3}\right) = 0.$$

Since $|\epsilon_k|^{\beta-2}$ and $|\epsilon_k|^{\beta-3}$ are linearly independent and $c_{kl} \neq 0$, the only way for the above Eq. to hold is if $\alpha_{k1} = \alpha_{k2} = 0$ for all k. This implies that α_{l1} and α_{l2} must be zero for all $l = 1, 2, \ldots, n$. Consequently, the set $\{\mathbf{w}_{lt}\}$ is linearly independent, confirming that the assumption holds in this case.

D.5. Proof of Theorem 2.3

Proof. By Theorem 2.1, there exist a permutation π of $\{1, \ldots, d_z\}$ and invertible, dimension-preserving maps $h_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_i}$ such that

$$\hat{z}_{\pi(i)} = h_i(z_i), \qquad i = 1, \ldots, d_z. \tag{S14}$$

For a fixed coordinate i and a value z_i, let $\mathcal{S}_i(z_i) \subset \mathbb{R}^{d_i}$ denote the support set from Assumption **A5**. Define the pushed-forward supports in the recovered coordinates by

$$\widehat{\mathcal{S}}_{\pi(i)}(z_i) \;:=\; h_i\big(\mathcal{S}_i(z_i)\big) \;\subset\; \mathbb{R}^{d_i}. \tag{S15}$$

For a full concept tuple $z = (z_1, \ldots, z_{d_z})$, define the product (rectangle) regions

$$\mathcal{R}(z) \;:=\; \prod_{i=1}^{d_z} \mathcal{S}_i(z_i), \qquad \widehat{\mathcal{R}}(z) \;:=\; \prod_{i=1}^{d_z} \widehat{\mathcal{S}}_{\pi(i)}(z_i). \tag{S16}$$

By Assumption **A5**, for any two distinct values $u \neq u'$ of the i-th concept, $\mathcal{S}_i(u) \cap \mathcal{S}_i(u') = \emptyset$. Since h_i in Eq. S14 is bijective, it preserves set disjointness:

$$\widehat{\mathcal{S}}_{\pi(i)}(u) \cap \widehat{\mathcal{S}}_{\pi(i)}(u') = h_i(\mathcal{S}_i(u)) \cap h_i(\mathcal{S}_i(u')) = h_i\big(\mathcal{S}_i(u) \cap \mathcal{S}_i(u')\big) = \emptyset.$$

Hence, for each coordinate i, the family $\{\widehat{\mathcal{S}}_{\pi(i)}(u)\}_u$ is pairwise disjoint.

For each i, define a decoder $\psi_{\pi(i)} : \mathbb{R}^{d_i} \to$(value set of z_i) by membership in the disjoint sets:

$$\psi_{\pi(i)}(\hat{z}_{\pi(i)}) \;=\; u \quad\Longleftrightarrow\quad \hat{z}_{\pi(i)} \in \widehat{\mathcal{S}}_{\pi(i)}(u). \tag{S17}$$

The right-hand side determines u uniquely by Step 1, so $\psi_{\pi(i)}$ is well-defined (ties can only occur on set boundaries, which have probability zero under standard absolute continuity assumptions). Moreover, Eq. S14 and the definition in Eq. S15 give, for any realization from the true model,

$$\hat{z}_{\pi(i)} \;=\; h_i(z_i) \in h_i\big(\mathcal{S}_i(z_i)\big) \;=\; \widehat{\mathcal{S}}_{\pi(i)}(z_i),$$

and therefore $\psi_{\pi(i)}(\hat{z}_{\pi(i)}) = z_i$ almost surely.

Define $\psi : \mathbb{R}^{d_z} \to$(value set of $\mathbf{z}$) by

$$\psi(\hat{\mathbf{z}}) \;:=\; \big(\psi_{\pi(1)}(\hat{z}_{\pi(1)}), \ldots, \psi_{\pi(d_z)}(\hat{z}_{\pi(d_z)})\big).$$

Applying Step 2 coordinate-wise gives $\psi(\hat{\mathbf{z}}) = z$ almost surely for any sample generated by the true model and mapped by Eq. S14. Thus the tuple $\mathbf{z}$ is a function of $\hat{\mathbf{z}}$ via support membership.

Let $z \neq z'$ be two tuples. Then there exists some index i such that $z_i \neq z'_i$. By Assumption **A5**, $\mathcal{S}_i(z_i) \cap \mathcal{S}_i(z'_i) = \emptyset$. Consequently,

$$\mathcal{R}(z) \cap \mathcal{R}(z') \;=\; \Big(\prod_{j \neq i} \mathcal{S}_j(z_j) \cap \mathcal{S}_j(z'_j)\Big) \;\times\; \big(\mathcal{S}_i(z_i) \cap \mathcal{S}_i(z'_i)\big) \;=\; \emptyset,$$

where $\mathcal{R}$ is defined in Eq. S16. Hence the rectangles $\{\mathcal{R}(z)\}$ are pairwise disjoint. The same argument on the pushed-forward sets shows $\{\widehat{\mathcal{R}}(z)\}$ are pairwise disjoint.

Fix an unseen tuple $z^q = (z_1^q, \ldots, z_{d_z}^q)$ and suppose a test latent $z^\sharp$ lies in the rectangle $\mathcal{R}(z^q)$, i.e., $z_i^\sharp \in \mathcal{S}_i(z_i^q)$ for all i. Then $\hat{z}^\sharp_{\pi(i)} = h_i(z_i^\sharp) \in h_i(\mathcal{S}_i(z_i^q)) = \widehat{\mathcal{S}}_{\pi(i)}(z_i^q)$, so by Eq. S17 we obtain $\psi_{\pi(i)}(\hat{z}^\sharp_{\pi(i)}) = z_i^q$ for each i, and therefore $\psi(\hat{\mathbf{z}}^\sharp) = z^q$. By Step 4, $\widehat{\mathcal{R}}(z^q)$ is disjoint from $\widehat{\mathcal{R}}(z')$ for any $z' \neq z^q$, which implies that the decoding to z^q is unique on $\widehat{\mathcal{R}}(z^q)$.

Assumption **A6** states that for the learned model on seen categories, each coordinate value that may occur at test time is realized by at least one training label through the learned generator $\hat{f}$. Operationally, this ensures that every transformed support $\widehat{\mathcal{S}}_{\pi(i)}(\cdot)$ appearing in Step 2 is estimable from training data in the recovered space, so that the membership-based decoders $\{\psi_{\pi(i)}\}$ can be implemented (*e.g.*, by empirical support estimation or consistent plug-in rules). This bridges the population-level identifiability shown in Steps 1–5 with a practical decoding rule learned on seen categories.

□

Proposition D.4 (Prototype learning $\Rightarrow$ A3 with a Spectral Bound)**.** *Assume Eq. 1, the conditions of Theorem 2.1, and observational equivalence. Let the prototype layer g_p contain m learnable prototype vectors*

$$\mathbf{P} = \{_j\}_{j=1}^{m}, \qquad _j \in \mathbb{R}^{d_z},$$

and define the prototype similarity as

$$s_j(\mathbf{z}) = \log \frac{\|\mathbf{z} - _j\|_2^2 + 1}{\|\mathbf{z} - _j\|_2^2 + \epsilon}, \qquad 0 < \epsilon < 1.$$

Let the class potentials be linear in similarities:

$$\phi_y(\mathbf{z}) = b_y + \sum_{j=1}^{m} U_{jy}\, s_j(\mathbf{z}), \qquad p_\theta(y \mid \mathbf{z}) = \frac{e^{\phi_y(\mathbf{z})}}{\sum_{y'} e^{\phi_{y'}(\mathbf{z})}}.$$

*Suppose the supervised prototype loss L_{proto} converges so that $p_\theta(y \mid \mathbf{z}) = p(y \mid \mathbf{z})$ (Fisher consistency). If (1) $m \geq 2d_z + |\mathcal{M}|$ and the prototypes $\{_j\}$ are in general position, and (2) there exist $K := 2d_z + |\mathcal{M}|$ labels whose coefficient differences $\beta_y := U_{\cdot y} - U_{\cdot 0} \in \mathbb{R}^m$ are linearly independent, then for any $\mathbf{z}$, Assumption **A3** (Semantic Comparison) holds. Moreover, letting*

$$W(\mathbf{z}) := \big[\, \boldsymbol{w}(\mathbf{z}, y_1) - \boldsymbol{w}(\mathbf{z}, 0) \;\cdots\; \boldsymbol{w}(\mathbf{z}, y_K) - \boldsymbol{w}(\mathbf{z}, 0) \,\big],$$

where

$$\boldsymbol{w}(\mathbf{z}, y) = \Big(\tfrac{\partial \log p(\mathbf{z}|y)}{\partial z_1}, \ldots, \tfrac{\partial \log p(\mathbf{z}|y)}{\partial z_{d_z}}, \tfrac{\partial^2 \log p(\mathbf{z}|y)}{\partial z_1^2}, \ldots, \tfrac{\partial^2 \log p(\mathbf{z}|y)}{\partial z_{d_z}^2} \Big) \oplus \Big(\tfrac{\partial^2 \log p(\mathbf{z}|y)}{\partial z_i \partial z_j} \Big)_{(i,j)\in\mathcal{M}},$$

we have

$$\sigma_{\min}\big(W(\mathbf{z})\big) \;\geq\; \sigma_{\min}\big(G(\mathbf{z})\big)\, \sigma_{\min}(B) \;>\; 0,$$

where $G(\mathbf{z}) \in \mathbb{R}^{(2d_z+|\mathcal{M}|)\times m}$ and $B := [\, \beta_{y_1} \;\cdots\; \beta_{y_K} \,] \in \mathbb{R}^{m\times K}$.

Proof. By Bayes' rule, $\log p(\mathbf{z} \mid y) = \log p(y \mid \mathbf{z}) + \log p(\mathbf{z}) - \log p(y)$. Differentiating and subtracting the base class $y = 0$ removes the y-independent term:

$$\nabla\big(\log p(\mathbf{z} \mid y) - \log p(\mathbf{z} \mid 0) \big) = \nabla\big(\log p(y \mid \mathbf{z}) - \log p(0 \mid \mathbf{z}) \big).$$

Since $\log p(y \mid \mathbf{z}) = \phi_y(\mathbf{z}) - \log \sum_{y'} e^{\phi_{y'}(\mathbf{z})}$, the partition term cancels in differences, so

$$\nabla\big(\log p(\mathbf{z} \mid y) - \log p(\mathbf{z} \mid 0) \big) = \sum_{j=1}^{m} \beta_{y,j}\, \nabla s_j(\mathbf{z}), \tag{S18}$$

$$\nabla^2\big(\log p(\mathbf{z} \mid y) - \log p(\mathbf{z} \mid 0) \big) = \sum_{j=1}^{m} \beta_{y,j}\, \nabla^2 s_j(\mathbf{z}). \tag{S19}$$

with $\beta_y = U_{\cdot y} - U_{\cdot 0}$. For $r_j^2 = \|\mathbf{z} - _j\|_2^2$,

$$\nabla s_j(\mathbf{z}) = 2s_j'(r_j^2)\,(\mathbf{z} - _j), \qquad \nabla^2 s_j(\mathbf{z}) = 2s_j'(r_j^2) I + 4s_j''(r_j^2)\,(\mathbf{z} - _j)(\mathbf{z} - _j)^\top,$$

and $s_j'(t) = (\epsilon - 1)/((t+1)(t+\epsilon)) \neq 0$. Stack the entries required by A3 in

$$g_j(\mathbf{z}) := \Big(\nabla s_j(\mathbf{z}),\; \mathrm{diag}\big(\nabla^2 s_j(\mathbf{z})\big),\; (\nabla^2 s_j(\mathbf{z}))_{(i,k)\in\mathcal{M}} \Big)^\top \in \mathbb{R}^{2d_z+|\mathcal{M}|},$$

and define $G(\mathbf{z}) := [\, g_1(\mathbf{z}) \;\cdots\; g_m(\mathbf{z}) \,]$. Then for each y,

$$\boldsymbol{w}(\mathbf{z}, y) - \boldsymbol{w}(\mathbf{z}, 0) = G(\mathbf{z})\, \beta_y. \tag{S20}$$

Stacking Eq. S20 across the K labels yields $W(\mathbf{z}) = G(\mathbf{z})B$.

Because $\{_j\}$ are in general position and $m \geq 2d_z + |\mathcal{M}|$, some $(2d_z + |\mathcal{M}|) \times (2d_z + |\mathcal{M}|)$ minor of $G(\mathbf{z})$ is nonzero (its analytic expression is not identically zero), hence $\operatorname{rank}[G(\mathbf{z})] = 2d_z + |\mathcal{M}|$. The discriminative prototype training induces class-wise exclusivity so B has full column rank K. Therefore $\operatorname{rank}[W(\mathbf{z})] = 2d_z + |\mathcal{M}|$, implying the vectors $\{\boldsymbol{w}(\mathbf{z}, y_i) - \boldsymbol{w}(\mathbf{z}, 0)\}_{i=1}^{K}$ are linearly independent and **A3** holds. Finally,

$$\sigma_{\min}\big(W(\mathbf{z})\big) \;\geq\; \sigma_{\min}\big(G(\mathbf{z})\big)\,\sigma_{\min}(B) \;>\; 0$$

by sub-multiplicativity of singular values, meaning that we can leverage prototype learning to sufficiently learn the distinctive concepts. □

Remarks. $\sigma_{\min}(G(\mathbf{z}))$ captures geometric diversity of prototypes, while $\sigma_{\min}(B)$ reflects class diversity induced by the prototype loss. Their product lower-bounds the degree of semantic comparison, linking prototype learning convergence to **A3**.

E. Methodology Details

The identifiability theory in Section 2.2 establishes the conditions guarantees compositional generalization to unseen categorie. This section details how each condition is operationalized in our architectural design. Our design choices are consequences of satisfying specific identifiability requirements, following the theoretical hierarchy from subspace separation (Thm. 2.1) to component recovery (Thm. 2.2) and compositional generalization (Thm. 2.3).

E.1. Encourage Discriminative Subspace

Theorem 2.2 requires that the local geometry of $\log p(\hat{\mathbf{z}}|y)$, characterized by gradients and Hessians, varies in a linearly independent directions across category labels (Assumption A3). A standard softmax classifier optimizes a single global projection of $\hat{\mathbf{z}}$, offering no mechanism for label-wise variation in local curvature. Thus, high accuracy may coexist with geometrically degenerate representations, violating A3.

To induce position-dependent curvature tied to category identity, we exploit a prototype layer g_p that transforms $\hat{\mathbf{z}}$ into a similarity score vector $\mathbf{s} \in \mathbb{R}^m$ via m learnable anchor points $\{\mathbf{p}_1, \ldots, \mathbf{p}_m\}$. Unlike standard prototype networks [7] designed for metric-based classification, our formulation specifically targets the geometric diversity condition in A3. Drawing on the similarity computation from ProtoPFormer [80] and its OCD adaptation [86], we define:

$$s_{i\to j} = g_{p_j}(\hat{\mathbf{z}}) = \log\left(\frac{\|\hat{\mathbf{z}} - \mathbf{p}_j\|_2^2 + 1}{\|\hat{\mathbf{z}} - \mathbf{p}_j\|_2^2 + \epsilon}\right), \tag{S21}$$

where $\hat{\mathbf{z}}$ represents the semantic subspace associated with sample i, and ϵ ensures numerical stability. This logarithmic formulation ensures that gradients and Hessians with respect to $\hat{\mathbf{z}}$ depend explicitly on prototype locations, creating position-dependent curvature. When prototypes are category-assigned and trained via $\mathcal{L}_{\text{proto}}$, different categories necessarily induce different local geometries. The following proposition formalizes this guarantee:

Proposition E.1 (Prototype Learning Ensures Semantic Comparison)**.** *Under Eq. 1 and the conditions of Thm. 2.1, consider a prototype layer g_p with m learnable prototypes $\mathbf{P} = \{_j\}_{j=1}^{m}$ and similarity $s_j(\mathbf{z}) = \log \frac{\|\mathbf{z}-_j\|_2^2+1}{\|\mathbf{z}-_j\|_2^2+\epsilon}$ for $0 < \epsilon < 1$. If* (i) *$m \geq 2d_z + |\mathcal{M}|$ with $\{_j\}$ in general position, and* (ii) *there exist $K := 2d_z + |\mathcal{M}|$ labels whose coefficient differences $\beta_y := U_{\cdot y} - U_{\cdot 0}$ are linearly independent, then after convergence of $\mathcal{L}_{proto}$, Assumption **A3** holds with spectral bound $\sigma_{\min}(W(\mathbf{z})) \geq \sigma_{\min}(G(\mathbf{z}))\,\sigma_{\min}(B) > 0$, where $W(\mathbf{z}) = [\,\boldsymbol{w}(\mathbf{z}, y_i) - \boldsymbol{w}(\mathbf{z}, 0)\,]_{i=1}^{K}$, $G(\mathbf{z}) \in \mathbb{R}^{(2d_z+|\mathcal{M}|)\times m}$, and $B = [\,\beta_{y_1} \;\cdots\; \beta_{y_K}\,]$.*

We use k prototypes per category, yielding $m = k \cdot |\mathcal{Y}|$ total prototypes sufficient to satisfy condition (i). Multiple prototypes per category capture intra-class variation while the supervised loss ensures inter-class separation. Additionally, the prototype mechanism separates label-discriminative information into $\hat{\mathbf{z}}$ while leaving category-agnostic variation in $\hat{\mathbf{c}}$, thereby supporting context invariance (A2).

E.2. Encourage Discriminative Component

Theorem 2.3 establishes that compositional generalization requires different values of the same semantic concept to occupy disjoint support regions (Assumption A5): $k \neq l \Rightarrow \mathcal{S}_i(z_i^{(k)}) \cap \mathcal{S}_i(z_i^{(l)}) = \emptyset$. In continuous representation spaces, Euclidean

distance $d_E = |\epsilon|$ can be arbitrarily small when vectors differ by ϵ in one component, so gradient descent may find solutions satisfying loss tolerances while violating strict support separation.

This motivates discrete encoding as an *architectural enforcement* of A5. In discrete space, the Hamming distance

$$d_H(\mathbf{b}_1, \mathbf{b}_2) = \sum_{i=1}^{n} \mathbf{1}(b_{1,i} \neq b_{2,i}), \quad \mathbf{b}_1, \mathbf{b}_2 \in \{0,1\}^n, \tag{S22}$$

guarantees minimum separation $d_H = 1$ for any distinct codes which is infeasible in continuous space [86]. This ensures different concept values cannot occupy overlapping support regions, directly operationalizing A5. Empirical studies in similarity search [75], metric learning [61], and large-scale retrieval [18] corroborate that discrete constraints yield more robust separation than continuous embeddings.

Hash Center Learning. Given semantic representations $\hat{\mathbf{z}}$, we compute category-level hash centers as the binarized mean of within-category samples [86]: $\bar{\mathbf{b}}_y = \text{sign}\big(\frac{1}{|C_y|}\sum_{i \in C_y} h_{\text{hash}}(\hat{\mathbf{z}}_i)\big)$, where h_{hash} is a learnable projection and C_y denotes samples of category y. Individual samples are mapped to binary codes $\mathbf{b}_i = \text{sign}(h_{\text{hash}}(\hat{\mathbf{z}}_i))$. The hash loss enforces within-category alignment while maintaining cross-category distinctiveness:

$$\mathcal{L}_{\text{hash}} = \frac{1}{|B|}\sum_{i \in B} \ell\big(y_i, \text{sim}(\mathbf{b}_i, \bar{\mathbf{b}})\big), \tag{S23}$$

where $\text{sim}(\mathbf{b}_i, \bar{\mathbf{b}})$ is the vector of cosine similarities between $\mathbf{b}_i$ and all category centers $\{\bar{\mathbf{b}}_y\}_{y \in \mathcal{Y}}$.

E.3. Supervised Contrastive Regularization

While the prototype mechanism (Sec. E.1) induces geometric diversity and hash encoding (Sec. E.2) enforces support separation, neither directly optimizes for the *discriminability* of semantic content. Assumption A3 requires not only that categories have different local geometries, but that these differences arise from discriminative semantics rather than spurious correlations.

We incorporate supervised contrastive learning [33] to explicitly encourage intra-class compactness and inter-class separation:

$$\mathcal{L}_{\text{scl}} = \sum_{i \in B} \frac{-1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \log \frac{\exp\left(\text{sim}(\theta_i, \theta_p)/\tau\right)}{\sum_{a \neq i} \exp\left(\text{sim}(\theta_i, \theta_a)/\tau\right)}, \tag{S24}$$

where B is the batch, $\mathcal{P}(i) = \{p \in B \mid y_p = y_i, p \neq i\}$ indexes same-category samples, θ_i is the representation of sample i, and τ is a temperature parameter. This objective complements prototype and hash losses: $\mathcal{L}_{\text{proto}}$ anchors category-level geometry, $\mathcal{L}_{\text{hash}}$ discretizes concept values, and $\mathcal{L}_{\text{scl}}$ shapes the continuous manifold. They jointly satisfy component-wise identifiability (Theorem 2.2).

E.4. Unified Objective

The complete training objective integrates all components according to their theoretical roles:

$$\mathcal{L}_{\text{all}} = \underbrace{\mathcal{L}_{\text{proto}} + \mathcal{L}_{\text{hash}} + \mathcal{L}_{\text{scl}}}_{\text{supervised comparison (A2, A3, A5)}} + \underbrace{\lambda_1 \mathcal{L}_{\text{flow}} + \lambda_2 \mathcal{L}_s}_{\text{semantic structure (A3, A4)}} + \underbrace{\mathcal{L}_{\text{recon}} + \lambda_3 \mathcal{L}_{\text{ctx}}}_{\text{density \& context (A1, A2)}}. \tag{S25}$$

The supervised comparison terms jointly address multiple conditions: $\mathcal{L}_{\text{proto}}$ induces geometric diversity across categories (A3) while separating semantic from contextual variation (A2); $\mathcal{L}_{\text{hash}}$ enforces discrete support separation (A5); and $\mathcal{L}_{\text{scl}}$ reinforces discriminability of semantic components (A3). For semantic structure, $\mathcal{L}_{\text{flow}}$ provides label-conditioned density estimation satisfying the contrast condition (A3), while $\mathcal{L}_s$ encourages sparse concept activations (A4). The density terms ensure well-posed latent distributions: $\mathcal{L}_{\text{recon}}$ maintains reconstruction fidelity (A1), and $\mathcal{L}_{\text{ctx}}$ regularizes contextual variables toward a label-invariant prior (A2). Assumption A6 (coverage) is a data condition rather than a learning objective.

E.5. Assumption-Component-Intuition

To clarify the motivation of each component in our implemented framework, we present a straightforward assumption-component-intuition specification in Table S2.

Table S2. Mapping from theoretical assumptions (A1–A6) to intuition and concrete components in our implementation.

Assump.	Meaning	Intuition	Components
A1	Well-posed density over $(\hat{\mathbf{z}}, \hat{\mathbf{c}} \mid \mathbf{x})$ and over the semantic subspace.	The latent variables should have a smooth, non-degenerate probability landscape, so that small changes in $\mathbf{z}$ or $\mathbf{c}$ lead to bounded changes in log-probability. This motivates reconstruction objectives with proper regularization.	(i) Reconstruction loss $\mathcal{L}_{\text{recon}}$ ensures a well-defined joint density with reconstruction fidelity. (ii) Conditional normalizing flow with $\mathcal{L}_{\text{flow}}$ yields a smooth, strictly positive density over the semantic subspace.
A2	Context invariance: identifiability of $\mathbf{c}$ without capturing semantic variations.	Contextual variables $\mathbf{c}$ capture shared background information (scene, imaging conditions) that is stable across labels. Separation from semantic information ensures that changing y does not shift the distribution of $\mathbf{c}$.	(i) Context regularization $\mathcal{L}_{\text{ctx}}$ enforces Gaussian prior alignment on $\hat{\mathbf{c}}$ for category-agnostic distributions. (ii) Prototype contrast $\mathcal{L}_{\text{proto}}$ pushes label-discriminative variation into $\hat{\mathbf{z}}$, separated by the MoE gating $M(\mathbf{s})$.
A3	Semantic sufficient contrast: local shape vectors of $\log p(\hat{\mathbf{z}} \mid y)$ are linearly independent across labels.	Each label y induces its own local shape of $\log p(\mathbf{z} \mid y)$ via gradients and curvatures. Linear independence means no class's local shape can be reconstructed as a combination of others, making each concept identifiable.	(i) Prototype loss $\mathcal{L}_{\text{proto}}$ encourages $\hat{\mathbf{z}}$ to span distinct hyperspherical directions across categories. (ii) Supervised contrastive loss $\mathcal{L}_{\text{scl}}$ reinforces discriminability of semantic components. (iii) Conditional flow $\mathcal{L}_{\text{flow}}$ provides explicit label-conditioned parametrization of $\log p(\hat{\mathbf{z}} \mid y)$.
A4	Sparse Markov structure over semantic concepts with modular dependencies.	Semantic concepts form a sparse dependency graph. When the learned graph is sparser than the true one, each learned concept corresponds to a true semantic concept or a small neighborhood around it, preserving interpretable modular structure.	(i) Sparsely-gated MoE learns a sparse adjacency structure $M_{\mathbf{z}}$ over semantic concepts. (ii) ℓ_1 sparsity loss $\mathcal{L}_s = \|\hat{\mathbf{z}}\|_1$ promotes sparse activations. (iii) Flow regularization $\mathcal{L}_{\text{flow}}$ operates on $M_{\mathbf{z}} \odot \hat{\mathbf{z}}$, so only selected edges induce non-zero mixed derivatives.
A5	Support separation: different concept values occupy distinct regions on the semantic manifold.	Different values of a concept (*e.g.*, "striped" vs. "spotted") occupy separated rather than overlapping regions in semantic support; otherwise, the meaning of the same region becomes ambiguous. This guarantees unique decoding of concept values.	(i) Hash loss $\mathcal{L}_{\text{hash}}$ enforces discrete codes with minimum Hamming separation, pushing different semantic values apart. (ii) The discretization bottleneck architecturally guarantees disjoint support regions for distinct concept values.
A6	Coverage: unseen categories share the semantic space with seen categories.	Under sufficient semantic diversity in seen categories, unseen categories can be expressed as new compositions of already learned primitives.	No explicit loss; satisfied when training data exhibit sufficiently diverse semantics across categories.

E.6. Training Algorithm

Algorithm 1 presents the complete training procedure, with loss components organized according to their theoretical roles in Eq. S25.

F. Implementation Details

F.1. Dataset Description

We evaluate our method following established benchmarks in the OCD literature [11, 86]. Specifically, we use four challenging fine-grained subsets from iNaturalist 2017 [71] (Fungi, Arachnida, Animalia, and Mollusca), as well as CUB [74], Stanford Cars [41], Oxford Pets [55], and Food101 [3]. Following the standard protocol [11, 86], the categories are split into two subsets (seen and unseen) at the super-category. By default, 50% of the samples from the seen classes are included in the training set $\mathcal{D}_S$, while the remaining samples are used in the unlabeled set $\mathcal{D}_Q$ for evaluation. During inference, test images

Algorithm 1 C^3 Training Procedure

Training set $\mathcal{D}_S = \{(\mathbf{o}_i, y_i)\}_{i=1}^N$, pretrained encoder f_{enc}, model components Θ Trained parameters Θ each mini-batch $\{(\mathbf{o}, y)\}$ from $\mathcal{D}_S$ $\mathbf{x} \leftarrow f_{\text{enc}}(\mathbf{o})$ Frozen backbone features $[\hat{\mathbf{z}}, \hat{\mathbf{c}}] \leftarrow M_\psi(\hat{g}_\phi^{-1}(\mathbf{x}))$ Semantic-context separation **// Supervised comparison (A2, A3, A5)** $\mathcal{L}_{\text{proto}} \leftarrow$ prototype similarity loss (Eq. S21) $\mathcal{L}_{\text{hash}} \leftarrow$ hash center alignment loss $\mathcal{L}_{\text{scl}} \leftarrow$ supervised contrastive loss (Eq. S24) **// Semantic structure (A3, A4)** $\mathcal{L}_{\text{flow}} \leftarrow$ conditional flow likelihood on $M_z \odot \hat{\mathbf{z}}$ $\mathcal{L}_s \leftarrow \|\hat{\mathbf{z}}\|_1$ **// Density and context (A1, A2)** $\mathcal{L}_{\text{recon}} \leftarrow -\log p_\theta(\mathbf{x}|\hat{\mathbf{z}}, \hat{\mathbf{c}})$ $\mathcal{L}_{\text{ctx}} \leftarrow \text{KL}[q(\hat{\mathbf{c}}|\mathbf{x})\|\mathcal{N}(0, I)]$ **// Combined objective (Eq. S25)** $\mathcal{L}_{\text{all}} \leftarrow (\mathcal{L}_{\text{proto}} + \mathcal{L}_{\text{hash}} + \mathcal{L}_{\text{scl}}) + \lambda_1\mathcal{L}_{\text{flow}} + \lambda_2\mathcal{L}_s + \mathcal{L}_{\text{recon}} + \lambda_3\mathcal{L}_{\text{ctx}}$ $\Theta \leftarrow \Theta - \eta\nabla_\Theta\mathcal{L}_{\text{all}}$

are processed independently as they arrive in a streaming fashion [11].

Table S3. Statistics of datasets.

	CUB	Scars	Pets	Food	Fungi	Arachnida	Animalia	Mollusca
$\lvert Y_S\rvert$	100	196	38	101	121	56	77	93
$\lvert Y_Q\rvert$	200	98	19	51	61	28	39	47
$\lvert\mathcal{D}_S\rvert$	1.5K	2.0K	0.9K	19.1K	1.8K	1.7K	1.5K	2.4K
$\lvert\mathcal{D}_Q\rvert$	4.5K	6.1K	2.7K	56.6K	5.8K	4.3K	5.1K	7.0K

F.2. Experiments on Synthetic Data

Data Generation. We generate synthetic data following the generative process in Eq. 1. The latent space has 4 dimensions: $n_{zc} = 2$ (contextual), $n_z = 2$ (semantic), and $n_d = 2$ (high-level). We sample $\mathbf{z}_c \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\hat{\mathbf{z}} \sim \mathcal{N}(\mu_y, \sigma_y^2\mathbf{I})$, where for each label y (both seen and unseen), $\mu_y \sim \text{Unif}(-4, 4)$ and $\sigma_y^2 \sim \text{Unif}(0.01, 1)$.

Architecture and Training. Since all synthetic data are numerical, observations $\mathbf{x}$ directly correspond to generated outputs without requiring a pretrained backbone. The VAE encoder and decoder are 6-layer MLPs with hidden dimension 32 and Leaky-ReLU activation ($\alpha = 0.2$). The mappings g, f, and m are also MLPs with Leaky-ReLU activations; $\hat{m}$ and $\hat{g}$ are estimated using 2-layer MLPs. For density estimation, we use component-wise spline flows [13] with monotonic linear rational splines (8 bins, bound = 5). To encourage concept discrimination, we apply cross-entropy loss directly to $\hat{\mathbf{z}}$ and $\mathbf{z}$, with an additional sigmoid $\sigma(\cdot)$ on $\hat{\mathbf{z}}$ to promote component-wise separation.

Hyperparameters. We train using AdamW for 200 epochs with learning rate 2×10^{-3}, batch size 128, and weight decay 10^{-4}. The KL loss weight is $\beta = 0.1$.

F.3. Experiments on Real-World Data

Implementation Details. Following the basic OCD setting of [11, 86], we use the DINO-pretrained ViT-B-16 [10] as the backbone. During training, only the final block of ViT-B-16 is fine-tuned. In our approach, the low-level concept learner $\hat{m}^{-1}$ is a single linear layer with an output dimension set to 768, and then use a MLP to map the 3072, the low-level changing concepts $\mathbf{z}$ in our case. Each category has $k = 10$ prototypes. The function m^{-1} consists of three linear layers with an output dimension set to $n_d = 32$. We align all the experiments with setting this dimension for fair comparison. We set $\lambda_1 = 5 \times 10^{-2}$, $\lambda_2 = 1 \times 10^{-2}$, and $\lambda_3 = 1 \times 10^{-3}$.

Comparison with Baselines. Since OCD demands instantaneous inference without access to unlabeled novel-category data during training, traditional NCD and GCD baselines are inapplicable. We compare against two state-of-the-art OCD methods, **SMILE** [11] and **PHE** [86], as well as three alternatives that can operate under the OCD protocol: **Sequential Leader Clustering (SLC)** [22], a classical method for sequential data; **Ranking Statistics (RankStat)** [21], which identifies categories via top-k indices of feature embeddings; and **Winner-Take-All (WTA)** [28], which constructs category descriptors by selecting maximum-value indices within feature groups. All baselines follow configurations established in SMILE [11] for consistency.

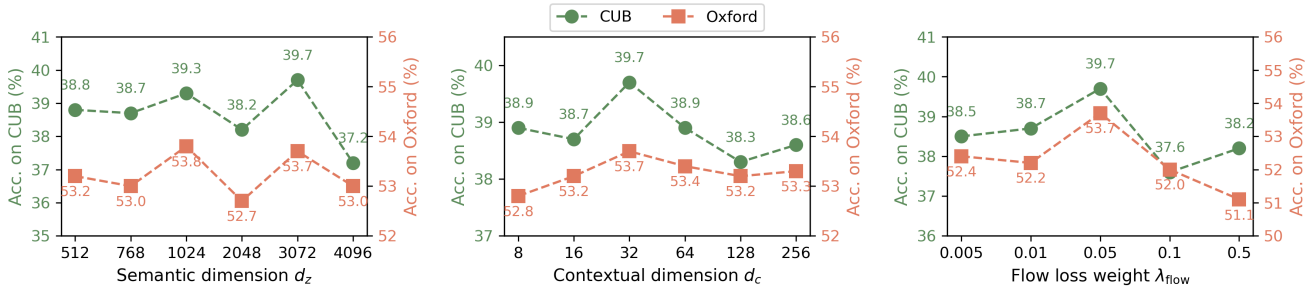

Figure S2. **Hyperparameter sensitivity analysis** on CUB and Oxford-Pet. Performance remains stable within reasonable ranges.

F.4. Statistical Robustness of Main Results

To assess the reliability of our findings, we report the average performance over three independent runs for each experimental setting. Table S4 provides the detailed results for our method, including both the mean and the population standard deviation, thereby quantifying the variability due to stochasticity in training and evaluation.

Table S4. Mean and standard deviation of accuracy across three independent runs for each setting.

Dataset	All	Old	New
CUB	39.7±0.18	60.9±1.43	29.1±0.67
Stanford Cars	33.3±0.24	68.3±1.09	17.2±0.51
Oxford Pets	53.7±0.35	62.5±2.16	49.1±1.42
Food101	30.3±0.27	67.0±0.38	11.2±0.46
Fungi	32.6±0.30	69.2±1.53	16.0±0.49
Arachnida	37.6±0.22	70.1±0.88	13.2±0.54
Animalia	41.5±0.33	58.3±1.24	32.6±1.07
Mollusca	41.6±0.29	69.4±2.01	27.0±1.18

F.5. Analysis on Hyperparameters.

Hyperparameter Analysis. Fig. S2 analyzes three key hyperparameters: semantic dimensionality (d_z), contextual dimensionality (d_c), and structure regularization weight (λ_{flow}). Experiments on CUB and Oxford-Pet show consistent patterns across both datasets. Performance peaks at $d_z = 3072$, indicating sufficient capacity is needed for semantic disentanglement, while excessive dimensionality ($d_z = 4096$) introduces redundancy. For contextual representation, moderate dimensionality ($d_c = 32$) performs best; too small values under-represent context while too large values risk encoding semantic information. The flow weight $\lambda_{\text{flow}} = 0.05$ achieves optimal balance: smaller values provide insufficient structure regularization, while larger values over-constrain the latent space. The low performance oscillation across varying hyperparameter values and datasets demonstrates the robustness of our design choices.

F.6. Effect of Sparsity Regularization

To evaluate the sensitivity of our method to the sparsity regularization coefficient λ_{sparse} on the gated network, which is responsible for identifying distinctive concepts and modular concepts, we conduct experiments under four different settings. Table S5 presents the ALL accuracy (mean ± std) across three independent runs. We observe that an inappropriate choice of sparsity coefficient significantly impairs performance.

Results on Synthetic Data. As shown in Table S6, several results support the validity of our theoretical insights. The Mean Correlation Coefficiency (MCC) quantifies the degree of component-wise identifiability of $\mathbf{z}$, whereas R^2 measures the subspace identifiability of $\hat{\mathbf{c}}$. Higher values for all metrics indicate better identifiability. First, we observe that both R^2 and the MCC of our method increase *monotonically* with the number of known classes within the evaluated range (up to n_s=24), at which both subspace and component-wise latent variables achieve high identifiability scores. This confirms our

Table S5. Effect of λ_{sparse} for concept selection on C^3 performance (mean $\pm$ std over three runs). The setting $\lambda = 0.1$ corresponds to the main results in Table 1.

Dataset	$\lambda = 0.0$	$\lambda = 0.01$	$\lambda = 0.1$	$\lambda = 1.0$
CUB	36.8 ± 0.5	38.9 ± 0.3	$\mathbf{39.7 \pm 0.2}$	35.1 ± 0.6
Stanford Cars	30.7 ± 0.6	32.6 ± 0.4	$\mathbf{33.3 \pm 0.3}$	29.4 ± 0.7
Oxford Pets	49.2 ± 0.4	51.6 ± 0.3	$\mathbf{53.7 \pm 0.2}$	48.1 ± 0.5
Food101	27.9 ± 0.7	29.4 ± 0.4	$\mathbf{30.3 \pm 0.3}$	26.5 ± 0.6
Average	36.1 ± 0.5	38.1 ± 0.3	$\mathbf{39.2 \pm 0.2}$	34.8 ± 0.6

d_z	Metric	n_s=6	n_s=12	n_s=18	n_s=24
2	MCC	0.73	0.89	0.90	0.95
	R^2	62.6	75.3	90.3	95.5
	Acc.	51.3	65.7	73.6	79.8
5	MCC	0.81	0.85	0.91	0.92
	R^2	73.6	77.2	79.3	89.2
	Acc.	43.5	59.9	67.0	72.2
9	MCC	0.31	0.65	0.75	0.74
	R^2	12.6	67.6	82.3	86.1
	Acc.	10.9	35.5	40.4	68.4

Table S6. **Identifiability Results on Synthetic Data**. n_s denotes the number of known categories.

Concept	**Latent Variables**	**Cls Acc. (%)**	**Notes**
Body	z_1, z_2	92.3	discriminative latent concepts, comparison
Tail	z_5	64.1	not discriminative
Wings	z_3, z_4	78.5	not discriminative
Head	z_8	95.0	discriminative latent concepts, comparison

Table S7. Concept-guided classification results by latent variables.

	CUB (%)			**SCars (%)**		
Method	**All**	**Old**	**New**	**All**	**Old**	**New**
CLIP (ViT-B/16)	33.5	58.9	21.8	26.7	52.1	14.9
CLIP (ViT-L/14)	28.9	51.0	14.6	22.3	45.5	10.2
Ours (DINO ViT-B/16)	40.1	62.1	29.5	34.1	69.0	17.8

Table S8. Scaling to larger foundation backbones: comparison on CUB and Stanford Cars using CLIP visual encoders.

theoretical prediction and aligns with the intuition that a sufficient number of known classes is necessary for identifiability. Specifically, when the condition $2d_z + |\mathcal{M}| + 1 \leq n_s$ is not satisfied, both MCC and accuracy drop to impractically low values, further validating the assumption that adequate comparisons are required. Notably, we find that strong performance can still be achieved with a large number of concepts, *e.g.*, $d_z = 9$, highlighting why our method remains effective even in complex fine-grained datasets.

Scaling to larger foundation model backbones. We evaluate our framework with large vision–language backbones by adopting only the *visual encoders* from the CLIP family, since the OCD setting provides no language supervision. The CLIP image embeddings are used as inputs to our network without modifying the training protocol. Results are shown below.

Merge setting	All (%)	Old (%)	New (%)
C^3 (Fungi only)	32.9	69.8	16.2
C^3 (Fungi + Arachnida)	34.5	66.7	19.8
C^3 (Fungi + Arachnida + Animalia)	36.4	63.5	22.9
C^3 (All four merged)	38.7	60.1	25.8

Table S9. **C^3 on merged iNaturalist subsets** (Fungi, Arachnida, Animalia, Mollusca). As the merge size grows, *New* improves and *Old* decreases modestly, indicating stronger open-set generalization with broader semantics.

# Unknown Labels	CUB All (%)	CUB Old (%)	CUB New (%)
150	23.0	32.1	17.5
120	31.5	51.3	22.3
100	40.1	62.1	29.5

Table S10. **Scaling with unknown category count on CUB.** Performance improves as the number of unknown labels decreases (more seen semantics), supporting the scalability of C^3 under increasing semantic coverage.

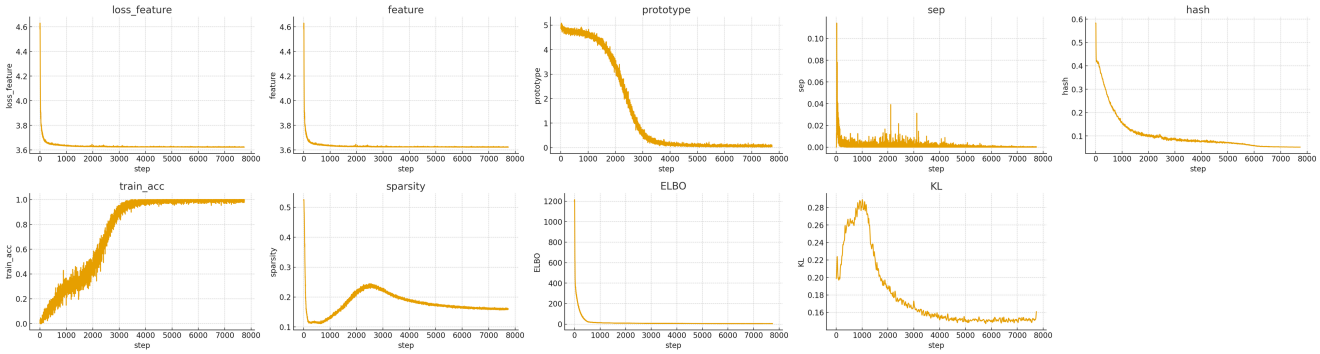

Figure S3. Loss curves in training stage.

Scaling to larger datasets with more semantic categories. We evaluate the scalability of C^3 by merging iNaturalist subsets—*Fungi*, *Arachnida*, *Animalia*, and *Mollusca*—to substantially increase the number of unseen categories. As shown in Table S9, as more subsets are merged, *New* accuracy consistently increases while *Old* decreases moderately; the overall *All* metric also changes accordingly. This confirms that comparison-and-composition benefits scale with available seen semantics rather than overfitting to them.

Varying the number of unknown labels. We follow the same protocol while adapting the known : unknown ratio and report CUB performance as the number of unknown labels increases (Table S10). Results show consistent gains in *All/Old/New* as the unknown set shrinks (i.e., more seen semantics are available).

Ablation on comparison and composition. We perform a leave-one-out ablation to quantify the contribution of each component in C^3, separating *comparison* (e.g., $\mathcal{L}_{\text{flow}}$, $\mathcal{L}_{\text{proto}}$, $\mathcal{L}_{\text{ctx}}$) from *composition* (e.g., sparsity $\mathcal{L}_S$ and residual integration f_{res}), while the reconstruction term $\mathcal{L}_{\text{ELBO}}$ stabilizes training. As reported in Table 2, removing any single module consistently degrades performance; the full model achieves the best overall results, indicating that comparison and composition play complementary roles.

Training Stability: Compatible Regularizers. As shown in Figure S3, we include the training curves of the independent regularizers, including the flow loss, likelihood (ELBO), prototype contrast, and sparsity on the Markov network.

Computational Efficiency and Scalability. We analyze the computational cost and scaling behavior of C^3 on CUB. As summarized in Table S11, the flow and MoE modules introduce only marginal overhead—approximately X% additional parameters and Y% increase in inference latency—while substantially improving model expressiveness and convergence

Model Variant	Params (M)	Training Time (min)	Inference Time (ms)	GPU Memory (GB)
PHE (baseline)	22.4	70.5	1.18	5.2
Base (w/o Flow, MoE)	23.1	72.3	1.21	6.4
+ Flow Module	24.4	75.8	1.25	6.8
+ MoE Module	24.9	76.2	1.27	6.9
Full C^3 (Flow + MoE)	**25.0**	**76.8**	**1.29**	**7.0**

Table S11. **Computational efficiency on CUB dataset.** The flow and MoE modules add minimal overhead while improving convergence stability; both operate in a low-dimensional, sparse regime.

stability. The added costs are limited because both modules operate in a low-dimensional, sparsity-inducing space. Moreover, the training loss exhibits smooth, stable convergence across datasets (Fig. S3), and we observe near-linear scaling with respect to the number of latent variables and input dimensions, indicating practicality for larger-scale applications.

G. Discussion

Compatibility of A2 and A3. We want to clarify that **A2** is a weaker version of **A3** (since isolating subspace $\mathbf{z}$ and $\mathbf{c}$ requires less diversity than disentangling concept), and then, Assumption **A2 / A3** are imposed on the data-generating process and is, in fact, a mild requirement for any successful semantic classification: if the latent concepts themselves lack sufficient diversity, neither humans nor machines can distinguish categories meaningfully. However, standard neural networks often fail to capture such diversity in practice, as their training objectives can converge to degenerate local minima that entangle concepts across categories.

H. Broader Impacts

This work provides theoretical foundations for learning compositional, identifiable representations that generalize to novel categories. By grounding representation learning in identifiability theory, our framework contributes to the broader goal of building AI systems whose learned concepts align with human-interpretable semantics. We anticipate no direct negative societal consequences from this foundational research.