Less Structure is More: Minimal Representations for Supervised Learning

Menghui Zhou Vitaveska Lanfranchi Po Yang
School of Computer Science, University of Sheffield
{menghui.zhou,v.lanfranchi,po.yang}@sheffield.ac.uk

Abstract

Many real-world applications require highly interpretable machine learning systems, particularly in high-stakes domains such as healthcare. A promising recent direction is the paradigm of *Maximal Coding Rate Reduction* (MCR^2), which seeks to characterise and preserve the low-dimensional structure underlying each class. However, we find that exhaustively preserving structural information is often unnecessary and can even be detrimental in supervised learning, especially under noisy and uncertain real-world conditions, as it may overly capture structural details and hence impair generalisation. In contrast, we propose a simple yet effective framework, termed *SiMCoding*, which captures only minimal structural information for each class. Despite relying on substantially less structural information, *SiMCoding* significantly improves the generalisation ability and robustness of deep models. Extensive experiments across diverse benchmark datasets demonstrate the effectiveness of *SiMCoding*. We further validate *SiMCoding* on both a standard Alzheimer's disease classification task using 3D MRI and a challenging real-world application—Parkinson's disease severity assessment from free-living human activity signals—where it consistently demonstrates strong robustness and outperforms competing methods. Moreover, *SiMCoding* can be naturally extended to incremental learning scenarios, where it substantially alleviates catastrophic forgetting and generally surpasses strong baselines. Code is available at `https://github.com/menghui-zhou/SiMCoding`.

1 Introduction

Over the past decade, deep learning has achieved remarkable success across a wide range of applications, including natural language processing [1, 2, 3], healthcare [4, 5], robotics [6, 7], and autonomous systems. Despite these advances, many real-world applications require machine learning systems that are not only accurate, but also interpretable, reliable, and robust. In high-stakes scenarios such as healthcare [8, 9] and autonomous driving [10, 11], understanding the structural information learned by a model is often crucial for practical deployment and user trust.

Motivated by this need, a growing body of work has sought principled approaches to understanding deep representation learning [12, 13], with the goal of developing more reliable and effective tools for real-world applications. Among these efforts, *Maximal Coding Rate Reduction* (MCR^2) [14] provides an appealing geometric perspective. The central assumption of MCR^2 is that, although real-world data are often high-dimensional, samples belonging to the same class typically lie on an underlying low-dimensional submanifold. MCR^2 therefore suggests that a key objective of deep representation learning is to capture and preserve these class-wise low-dimensional structures.

MCR^2 offers clear structural transparency for learned representations and has demonstrated strong robustness, particularly under noisy or uncertain labels. These properties suggest that MCR^2 may be especially valuable for practical high-dimensional data analysis tasks, such as medical imaging [15]

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

and physiological time-series analysis [16], where reliable prediction depends not only on robustness, but also on an interpretable understanding of the underlying data structure. However, we find that MCR^2 often exhibits limited generalisation ability in practical applications, particularly on complex real-world datasets, as shown in Figure 2. We argue that, in many cases, exhaustively preserving class-wise low-dimensional structural information may force the model to capture too much structural detail, thereby limiting its ability to generalise effectively.

To address this issue, we start from a simple intuition: for supervised classification, separating different classes may not require preserving *the maximum amount of structural information* from the data. Instead, preserving only *minimal yet sufficient structural information* may already be enough for distinguishing different objects. For example, in Parkinson's disease assessment, tremor severity alone can often provide strong evidence for distinguishing patients from healthy individuals, without requiring the model to retain all structural information contained in the data [17]. We therefore hypothesise that learning only minimal yet sufficient structural information may already be enough for accurate supervised classification. We argue that this insight aligns with the well- known principle of Occam's razor that unnecessary complexity should be avoided [18]. Under this minimalist setting, the model is still expected to preserve the structural transparency of learned representations, remain sensitive to intrinsic data structures, and retain the robustness of MCR^2 to noisy or uncertain labels.

To realise this idea, we need to explicitly constrain the amount of structural information learned by MCR^2, which technically corresponds to constraining the rank of the representation matrix. However, directly optimising matrix rank is well known to be difficult [19], since the rank function is discrete, non-differentiable, and highly non-convex, making it unsuitable for standard gradient-based optimisation in deep learning. In this work, we identify a simple and effective mechanism that can directly constrain the upper bound of the representation rank *without introducing any additional computational overhead.* This enables explicit control over the complexity of learned representation structures while preserving the efficiency and scalability of standard deep learning training pipelines.

At a high level, MCR^2 is built on the concept of *lossy coding rate* [20], which measures the number of binary bits required to encode a finite set of feature representations under a prescribed distortion level. Intuitively, it characterises the amount of structural information contained in the learned representations. Conventional wisdom generally assumes that higher coding precision is desirable for representing information more accurately. In this work, however, we show that under the MCR^2 (not lossy coding rate itself) framework, coding precision is closely related to the complexity of the learned representation structures, particularly the rank of the learned representations. This relationship enables direct control over structural complexity through the coding precision itself. Consequently, the model can discard less informative structural components while preserving only the minimal information necessary for classification, without introducing additional computational cost.

Our results further suggest that discarding unnecessary structural information can substantially improve the generalisation ability of deep models. Specifically, this minimalist learning paradigm preserves the robustness of MCR^2 to noisy or uncertain labels, while consistently improving generalisation performance under both noisy and clean training settings.

The main contributions of this work are summarised as follows:

- We identify a fundamental limitation of the existing MCR^2 learning paradigm: exhaustively preserving class-wise structural information can substantially restrict the generalisation ability of deep models in practical applications. On complex real-world datasets, maximising structural information may even lead to severe degradation in generalisation performance.
- We show that, by constraining the rank of learned representations, the model can discard less important structural information while preserving only minimal yet sufficient information for classification. This minimalist structural learning paradigm introduces no additional computational overhead, yet substantially improves the generalisation ability of deep models under both clean and noisy settings.
- We extensively validate *SiMCoding* across multiple image benchmarks, standard 3D MRI-based Alzheimer's disease classification using ADNI [21, 9], and a challenging real-world task: free-living Parkinson's disease severity assessment using wearable activity data [17, 16]. Experimental results show that *SiMCoding* can significantly outperform several influential baselines in highly noisy real-world settings, such as free-living activity analysis, or when some training labels are incorrect.

- We further show that *SiMCoding* can be naturally extended to incremental learning scenarios, where it achieves performance comparable to ReduNet on MNIST and higher overall accuracy on ImageNette.

Taken together, our results demonstrate that *SiMCoding* is a highly promising learning framework, particularly under noisy and uncertain conditions. Its strong robustness and practical effectiveness make it especially valuable for challenging real-world applications such as free-living disease assessment.

2 Method

2.1 Preliminary

Given data $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m] \in \mathbb{R}^{D\times m}$ consisting of m samples from K classes, the goal of representation learning is to map high-dimensional observations into low-dimensional features that capture the intrinsic properties of the data, such as its underlying structure and geometry, thereby facilitating downstream tasks such as classification. A widely adopted viewpoint, commonly referred to as the *manifold hypothesis*, suggests that samples from each class lie on a low-dimensional submanifold $\mathcal{M}_i$, and that the full dataset is concentrated near the union $\mathcal{M} = \cup_{i=1}^{K}\mathcal{M}_i$ [22, 23, 24, 25], as illustrated in Figure 1. This perspective motivates the learning of feature representations $\boldsymbol{z}_i \in \mathbb{R}^d$, where $d \ll D$, that preserve essential structural information while discarding redundant variability.

To obtain such features, one typically employs a nonlinear mapping $f_{\boldsymbol{\Theta}} : \mathbb{R}^D \to \mathbb{R}^d$, parameterised by neural network weights $\boldsymbol{\Theta}$: $\boldsymbol{x} \mapsto \boldsymbol{z} = f_{\boldsymbol{\Theta}}(\boldsymbol{x})$, and collects the resulting feature matrix $\boldsymbol{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m] \in \mathbb{R}^{d\times m}$. For simplicity, we further denote $\boldsymbol{Z} = [\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_K]$, where $\boldsymbol{Z}_i \in \mathbb{R}^{d\times m_i}$ represents the feature matrix associated with the i-th class, and m_i denotes the number of samples in that class. Ideally, the learned representations should not only capture the underlying class structure, but also exhibit a compact and structured form [12].

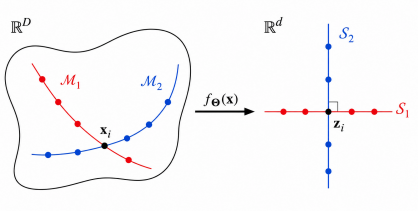


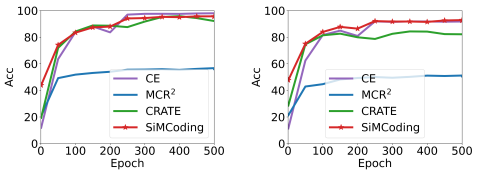
Figure 1: Illustration of MCR^2.

The Maximal Coding Rate Reduction (MCR^2) framework [14, 13] provides an information-theoretic objective for learning compact and discriminative representations. The objective is defined as

$$\max_{\boldsymbol{Z}} \Delta\mathcal{R}(\boldsymbol{Z}) = \underbrace{\frac{1}{2}\log\det\big(\boldsymbol{I} + \alpha\boldsymbol{Z}\boldsymbol{Z}^\top\big)}_{\text{Expansion: } \mathcal{R}(\boldsymbol{Z})} - \underbrace{\sum_{i=1}^{K}\frac{m_i}{2m}\log\det\big(\boldsymbol{I} + \alpha_i\boldsymbol{Z}_i\boldsymbol{Z}_i^\top\big)}_{\text{Compression: } \mathcal{R}_c(\boldsymbol{Z})}, \text{ s.t. } \|\boldsymbol{Z}_i\|_F^2 = m_i, \tag{1}$$

where $\alpha = \frac{d}{m\epsilon^2}$ and $\alpha_i = \frac{d}{m_i\epsilon^2}$. The expansion term $\mathcal{R}(\boldsymbol{Z})$ encourages the overall representation space to occupy a large volume, thereby promoting separation between different classes. In contrast, the compression term $\mathcal{R}_c(\boldsymbol{Z})$ encourages representations within each class to form compact low-dimensional structures. By jointly optimising expansion and compression, MCR^2 learns representations that are both discriminative and geometrically structured.

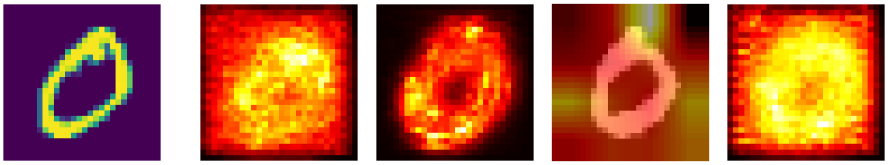
Yu et al. [14] and Chan et al. [13] showed that optimising the MCR^2 objective yields structured representations with two key properties. First, representations within each class tend to concentrate on a low-dimensional linear subspace that reflects the underlying submanifold structure of the data, while subspaces associated with different classes become approximately orthogonal, thereby improving inter-class discriminability. Second, the learned class subspaces collectively expand to occupy the entire feature space $\mathbb{R}^d$. Specifically,



(a) Train (b) Test

Figure 2: Training dynamics on ImageNette.

$$\operatorname{rank}(\boldsymbol{Z}) = \sum_{i=1}^{K}\operatorname{rank}(\boldsymbol{Z}_i) = \sum_{i=1}^{K} d_i = d,$$



(a) Original (b) CE (c) MCR^2 (d) CRATE (e) *SiMCoding*

Figure 3: Comparison of saliency maps for digit 0 from MNIST.

where $d_i = \mathrm{rank}(\boldsymbol{Z}_i)$ denotes the intrinsic dimension of the i-th class subspace. Consequently, each class preserves the maximal possible amount of structural information within its representation matrix $\boldsymbol{Z}_i$. An illustration of this behaviour is provided in Figure 1.

However, we empirically observe that MCR^2 often exhibits both limited fitting performance and poor generalisation ability, as shown in Figure 2. We conjecture that this behaviour arises because the objective encourages the model to preserve excessively rich structural information. As a result, the learned representations may overemphasise fine-grained structural details of the data and even capture noise components, thereby degrading the robustness and generalisation capability of the model. To validate our intuition, we conduct a small experiment on the MNIST dataset and analyse the corresponding saliency maps of the learned representations. The result is shown in Figure 3. We can clearly observe that MCR^2 focuses excessively on fine-grained details and captures substantially more local structural patterns than necessary.

A natural way to address this issue is to explicitly constrain the amount of information preserved in the learned representation. Since MCR^2 tends to drive the class subspaces to collectively fill the entire feature space, one may consider directly restricting the complexity of the learned representation by constraining the rank of $\boldsymbol{Z}$. For example, imposing

$$\mathrm{rank}(\boldsymbol{Z}) \leq r, \qquad r < d,$$

would force the representations to lie in a lower-dimensional subspace, thereby limiting the amount of structural information that can be encoded.

However, directly optimising such a rank-constrained objective is notoriously difficult. The rank function is discrete, non-convex, and non-smooth: $\mathrm{rank}(\boldsymbol{Z}) = \sum_{i=1}^{d} \mathbf{1}(\sigma_i > 0)$, which makes it incompatible with standard gradient-based optimisation for deep neural networks. Classical alternatives, such as nuclear-norm regularisation, $\|\boldsymbol{Z}\|_* = \sum_{i=1}^{d} \sigma_i$, provide convex relaxations of the rank function, but they typically require singular value decomposition or eigenvalue decomposition during training, leading to considerable computational overhead when either the batch size or feature dimension is large. Moreover, nuclear-norm penalties shrink all singular values uniformly, which may suppress informative semantic components together with redundant or noisy ones, potentially leading to underfitting.

Another possible strategy is to explicitly truncate the spectrum of $\boldsymbol{Z}$ by preserving only the top-r singular directions: $\boldsymbol{Z} \approx \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$. However, this approach requires specifying the target rank r in advance, which is highly data-dependent and difficult to tune in practice. Furthermore, spectral truncation introduces unstable gradients around repeated or near-zero singular values, making optimisation sensitive to mini-batch fluctuations. Consequently, the learned representations may vary abruptly across training iterations, leading to unstable training dynamics.

Therefore, although rank control is conceptually appealing, directly imposing rank constraints on deep representations remains computationally expensive, optimisation-unfriendly, and difficult to integrate into large-scale neural network training.

2.2 Definition of Lossy Coding Rate and Geometrical Interpretation

To address this problem, we next revisit the definition of the lossy coding rate and examine its intrinsic relationship with the amount of structural information preserved in the learned representations.

Lossy coding rate. The lossy coding rate $\mathcal{R}(\cdot)$ quantifies the effective volume of a distribution or its finite sample set under a prescribed distortion level $\epsilon > 0$ [26, 27, 20]. Formally, for each $\boldsymbol{z} \in \boldsymbol{Z}$,

let its reconstruction $\widehat{\boldsymbol{z}}$ satisfy $\mathbb{E}[\|\boldsymbol{z}-\widehat{\boldsymbol{z}}\|] \leq \epsilon$. Then, the average number of binary bits required to encode the feature set $\boldsymbol{Z}$ is given by

$$\mathcal{R}(\boldsymbol{Z}) = \tfrac{1}{2}\log\det\big(\boldsymbol{I} + \tfrac{d}{m\epsilon^2}\boldsymbol{Z}\boldsymbol{Z}^\top\big)\,.$$

Beyond its information-theoretic interpretation, Ma et al. [20] further showed that the lossy coding rate admits an equivalent geometric interpretation from the perspective of spherical packing. Specifically, $\mathcal{R}(\boldsymbol{Z})$ characterises how many ϵ-balls (i.e., d-dimensional spheres with radius ϵ) are required to cover the subspace spanned by $\boldsymbol{Z}$. Conventional wisdom typically suggests that higher coding precision is always preferable, since it preserves more information from the original data.

Our initial intuition was that, when the coding precision is too high (i.e., when ϵ is very small), $\mathcal{R}(\boldsymbol{Z})$ is encouraged to capture increasingly fine-grained structural information of the data manifold, including nuisance variations, irrelevant local patterns, and even noise components, rather than only semantically meaningful structures relevant to the downstream task. Conversely, as shown in Figure 4, using a larger distortion tolerance ϵ can be interpreted as covering the underlying structure with larger ϵ-balls, thereby smoothing out insignificant structural details and forcing the representation to retain only the dominant and robust structures of the data. This observation suggests that controlling the coding precision provides a natural mechanism for regulating the amount of structural information preserved in a set of data.

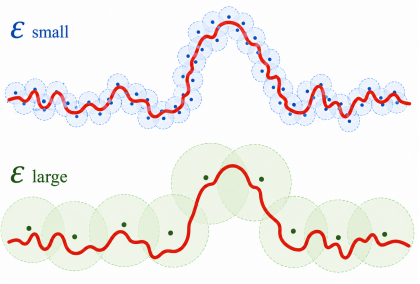


Figure 4: Illustration of the effect of coding precision ϵ on the amount of preserved structural information.

However, up to this point, our discussion has focused only on the lossy coding rate itself. A standalone lossy coding rate $\mathcal{R}(\boldsymbol{Z})$ is insufficient to learn structured representations. Therefore, the above analysis does not directly imply that adjusting the coding precision alone will necessarily produce meaningful effects in the structured representations learned by MCR^2. Although MCR^2 is built upon the lossy coding rate, its optimisation mechanism is fundamentally different from the coding-rate functional itself. Rather than minimising or maximising a single coding-rate term, MCR^2 learns representations through the interaction between global expansion and class-wise compression. Consequently, the final learned representation is determined by a highly coupled optimisation process involving both the global feature matrix $\boldsymbol{Z}$ and the class-wise feature matrices $\{\boldsymbol{Z}_i\}_{i=1}^K$.

Under such a framework, it remains unclear whether the coding precision ϵ establishes a direct connection with the rank of the learned representations. In particular, it is unknown whether varying ϵ affects the structural complexity of each class subspace, the global representation space, or both simultaneously. Nor is it clear how these effects propagate through the expansion–compression dynamics of MCR^2 training.

2.3 Simplest Coding-based Structural Representation

Fortunately, the following theorem provides a precise answer to the above question. Specifically, it shows that the coding precision ϵ explicitly imposes an upper bound on the rank of each learned class representation, thereby establishing a rigorous connection between ϵ-ball covering and the structural complexity of the learned feature subspaces.

Theorem 1. *Let $\boldsymbol{Z}^* = \boldsymbol{Z}_1^* \cup \cdots \cup \boldsymbol{Z}_K^*$ be the optimal solution to equation 1. Define $d_i^\star = \sqrt{\frac{m_i}{m}}\frac{d}{\epsilon^2}$. Then the following properties hold:*

- ***Discriminativeness:*** *Features from different classes reside in mutually orthogonal, low-dimensional linear subspaces; that is, $(\boldsymbol{Z}_i^*)^\top \boldsymbol{Z}_j^* = \boldsymbol{0}$ for all $i \neq j$.*
- ***Bounded dimensionality:*** *Each class-specific subspace satisfies $d_i \leq d_i^\star$. Moreover, for each class, the first $d_i - 1$ singular values of $\boldsymbol{Z}_i^*$ are identical.*

Theorem 1 reveals a fundamental connection between the coding precision ϵ and the structural complexity of the learned representations. Specifically, the intrinsic dimensionality of each class-specific subspace is explicitly upper bounded by $d_i^\star = \sqrt{\frac{m_i}{m}}\frac{d}{\epsilon^2}$. Importantly, among all quantities

in this bound, m_i, m, and d are fixed constants determined by the dataset and feature dimension, whereas ϵ is the only controllable variable. Therefore, the coding precision ϵ directly determines the maximal dimensionality of each learned class subspace.

Moreover, the bound immediately implies that increasing ϵ decreases the upper bound on the dimensionality of the learned representations. In other words, a larger ϵ forces each class representation to concentrate on a lower-dimensional subspace, thereby reducing the amount of preserved structural information and limiting the structural complexity of the learned feature representations.

This result is particularly appealing because it provides an explicit and principled mechanism for controlling representation complexity without introducing additional optimisation objectives or computational overhead. *Unlike conventional rank-regularisation approaches, the proposed mechanism does not require auxiliary penalty terms, spectral truncation, singular value decomposition, or extra hyperparameters for balancing regularisation strength. Instead, the complexity of the learned representations can be directly controlled through the coding precision ϵ itself* ***without any additional computational cost****.*

Notably, Theorem 1 provides an explicit mechanism for controlling the structural complexity of the learned representations through the coding precision ϵ. In principle, one may naturally ask whether there exists an optimal intermediate level of representation complexity, rather than simply minimising the dimensionality of the learned subspaces. However, investigating the optimal trade-off between representation complexity and task performance is beyond the scope of this work. Our primary goal is instead to develop a simple, practical, and effective learning principle that can be applied robustly across different settings without requiring complicated complexity selection strategies.

From this perspective, the most straightforward and stable choice is to directly minimise the structural complexity of each learned class subspace. Concretely, we consider the extreme case where each class representation collapses to the lowest possible non-trivial dimensionality, namely a rank-one subspace. Under such a formulation, the target representation structure becomes explicit and unambiguous, avoiding the need to determine an appropriate intermediate complexity level for different datasets or tasks. Importantly, this design also eliminates the need for additional heuristics or manual tuning of representation complexity, thereby yielding a considerably simpler and more practical learning framework.

Now we can encourage each class to collapse towards its minimal one-dimensional subspace, i.e., $\min_{1\leq i\leq K} d_i \geq 1$. Accordingly, an upper bound on ϵ follows immediately:

$$\epsilon^2 \leq \epsilon_U^2 = \min_{1\leq i\leq K} d\sqrt{\frac{m_i}{m}}. \tag{2}$$

Optimisation flexibility and convergence freedom. Theoretically, the structural complexity of the learned representations can be minimised by setting $\epsilon = \epsilon_U$, thereby constraining each class subspace to the lowest possible non-trivial dimensionality, i.e., we achieve the objective of arguably ***SiM****plest* ***Coding****-based structural representation* (*SiMCoding*). However, such a constraint may be overly restrictive in practical optimisation, especially for highly nonlinear objectives.

Intuitively, effective optimisation still requires sufficient representational flexibility. If the representation complexity is constrained too aggressively from the beginning, the optimisation trajectory may become excessively restricted, leading to unstable training dynamics and poor convergence behaviour. This observation is also supported by prior theoretical results. For example, [28] showed that, for regularised cross-entropy objectives, the global optimality conditions require at least one redundant dimension in the representation space, which enables the optimisation trajectory to escape undesirable local minima. This suggests that excessively restricting representation dimensionality may impair optimisation stability.

Another important consideration follows directly from Theorem 1. The upper bound is guaranteed to be tight only when $d_i^\star = 1$. To ensure that each class subspace satisfies $d_i = 2$, one needs to require at least $d_i^\star > 2$. Note that $d_i^\star$ is real-valued whereas d_i is an integer, any choice satisfying $2 < d_i^\star < 3$ constrains $d_i \leq 2$. Hence, given the common floating-point rounding errors in any computational devices, in practice we recommend simply using $d_i^\star \geq 3$, which provides both stronger practical robustness and greater optimisation flexibility.

Notably, although [29] analysed the global optimisation landscape of a variant of MCR^2, their formulation differs substantially from the objective considered in this work. Moreover, their global

optimality condition still requires the learned discriminative representations to achieve maximal dimensionality, which fundamentally contrasts with our minimal-complexity learning principle.

Overall, these observations suggest that directly enforcing the theoretically minimal representation complexity may overly constrain the optimisation landscape, limiting the model's ability to effectively search for desirable solutions. This issue becomes particularly severe for complex datasets and highly nonlinear learning tasks. Indeed, our experiments confirm this phenomenon. When the representation complexity is minimised too aggressively, *SiMCoding* can still learn discriminative representations on relatively simple datasets such as MNIST, but fails on more challenging datasets including CIFAR-10 and CIFAR-20 (see Figure 7 in Appendix). Therefore, a practically effective choice of ϵ is given by

$$\epsilon^2 \;\leq\; \widetilde{\epsilon_U}^2 = \min_{1\leq i\leq K} \frac{d}{3}\sqrt{\frac{m_i}{m}}.$$

It is worth noting that, given the theoretical characterisation above in Theorem 1, the coding precision ϵ could in principle be selected through empirical trial-and-error. To avoid excessive hyperparameter tuning, however, we instead adopt a theoretically guided choice of ϵ derived from the proposed complexity bounds. In practice, we find that directly using a unified coding precision at the practical upper bound $\widetilde{\epsilon_U}^2$ is already sufficient to achieve the desired effect. Specifically, this choice effectively encourages the learned class representations toward minimal structural complexity while still preserving stable optimisation behaviour and discriminative capability.

3 Experiments

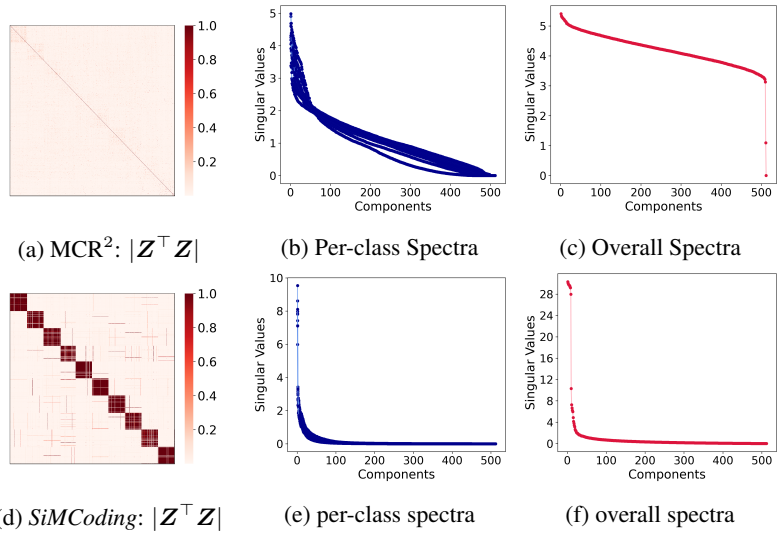
In this section, we experimentally demonstrate that preserving only minimal yet sufficient structural information via *SiMCoding* is already enough to achieve strong representation learning performance. In particular, despite preserving substantially less structural information than MCR^2, *SiMCoding* consistently learns robust and discriminative representations across diverse learning scenarios. Importantly, *SiMCoding* achieves this behaviour without relying on additional heuristics, sophisticated architectural modifications, or complicated engineering strategies. Instead, the framework operates solely through theoretically guided control of the coding precision. Therefore, our goal here is not to pursue state-of-the-art performance through extensive engineering, but rather to validate the effectiveness of the proposed principle under standard and fair experimental settings.

It is worth noting that Cross-entropy (CE), CRATE [30], and MCR^2 represent three highly influential and widely adopted learning paradigms. Remarkably, even under its most basic formulation, *SiMCoding* already demonstrates strong performance and, in several scenarios, substantially outperforms these established approaches. These results highlight the effectiveness of learning only minimal yet sufficient structural information for representation learning.

Experimental setup. We conduct experiments on MNIST [31], CIFAR-10, CIFAR-20 [32], and ImageNette [33]. For MNIST, we use a lightweight convolutional network, while for CIFAR-10, CIFAR-20, and ImageNette, we employ a ResNet-18 backbone for both MCR^2 and *SiMCoding*. For CRATE, we follow the architecture settings in [34]. Since the low-dimensional representations learned by *SiMCoding* make the predefined subspace dimension in NSC classifier [14] ill-defined, we evaluate both *SiMCoding* and MCR^2 using the simple logistic regression classifier and report the resulting classification accuracy. Unless otherwise specified, all methods are trained under identical settings following prior work [14, 13, 34].

Structure analysis of learned features. The results in Figure 5 and Figure 2 provide strong evidence that *SiMCoding* successfully learns minimal-complexity representations. Compared with MCR^2, *SiMCoding* produces much clearer block-diagonal Gram matrices and sharply decaying singular-value spectra, indicating that each class representation is concentrated around an almost rank-one subspace with strong inter-class separation. Despite preserving substantially less structural information, *SiMCoding* simultaneously achieves significantly faster convergence as well as markedly better fitting and generalisation performance than MCR^2, demonstrating that learning only minimal yet sufficient structural information is already enough for effective representation learning.

Discarding structural details. As shown in Figure 3, the qualitative differences in saliency patterns reveal fundamentally different representation mechanisms across methods. CE mainly highlights



(a) MCR^2: $|\boldsymbol{Z}^\top \boldsymbol{Z}|$ (b) Per-class Spectra (c) Overall Spectra

(d) *SiMCoding*: $|\boldsymbol{Z}^\top \boldsymbol{Z}|$ (e) per-class spectra (f) overall spectra

Figure 5: Feature structure analysis on ImageNette under different training objectives.

label-relevant regions without consistently capturing digit geometry, while MCR^2 preserves extensive intra-class details and CRATE produces relatively diffuse attention patterns across digit components. In contrast, *SiMCoding* concentrates almost exclusively on the digit outline—the most discriminative structural component—while suppressing internal noise and irrelevant details. These results provide direct visual evidence that *SiMCoding* effectively filters redundant intra-class variations and learns clean, interpretable discriminative structures, consistent with its formulation as a minimal-complexity, near one-dimensional subspace learning framework.

Varying representation complexity. We now quantify representation complexity on ImageNette to assess Theorem 1, using $\widetilde{\epsilon_U}^2/16$, $\widetilde{\epsilon_U}^2/8$, $\widetilde{\epsilon_U}^2/4$, $\widetilde{\epsilon_U}^2/2$, and $\widetilde{\epsilon_U}^2$. Following the commonly used relative cutoff criterion, we compute effective ranks from PCA singular values using a threshold of 50% for each $\boldsymbol{Z}_i$, with class-wise ranks averaged over ten classes and the resulting mean denoted by $\overline{\mathrm{rank}}_{\mathrm{eff}}(\boldsymbol{Z}_i)$. Here, a threshold of τ means that only singular values greater than τ times the largest singular value of the corresponding matrix are counted. For the whole representation $\boldsymbol{Z}$, minor components from different classes may accumulate and interfere with the identification of its dominant components. We therefore slightly reduce the threshold to 30% when computing $\mathrm{rank}_{\mathrm{eff}}(\boldsymbol{Z})$ to more stably reflect its intrinsic structure. As shown in Table 1, increasing ϵ consistently reduces representation complexity: the global effective rank decreases from 444 to 10, while the average class-wise rank decreases from 28.20 to 1.00. At $\widetilde{\epsilon_U}^2$, the global effective rank equals the number of classes ($K = 10$), and each class has a single dominant PCA component, consistent with the trend predicted by Theorem 1. Training accuracy remains high in the low-rank settings, while test balanced accuracy increases as representation complexity decreases, reaching its highest value at $\widetilde{\epsilon_U}^2$. These results support our central hypothesis that preserving more structural information is not necessarily beneficial; learning a minimal yet sufficient structure can improve generalisation while retaining strong fitting performance.

Table 1: Effect of coding precision on effective representation rank and classification performance on ImageNette. Training and test balanced accuracies are reported in %.

Setting	$\mathrm{rank}_{\mathrm{eff}}(\boldsymbol{Z})$	$\overline{\mathrm{rank}}_{\mathrm{eff}}(\boldsymbol{Z}_i)$	Train	Test
MCR^2	510	53.70	55.95	51.17
$\widetilde{\epsilon_U}^2/16$	444	28.20	92.73	83.38
$\widetilde{\epsilon_U}^2/8$	93	8.00	95.81	90.18
$\widetilde{\epsilon_U}^2/4$	50	3.90	**95.92**	91.41
$\widetilde{\epsilon_U}^2/2$	27	1.60	95.66	92.02
SiMCoding	**10**	**1.00**	95.44	**92.96**

Robustness to label noise. To evaluate robustness under label noise, we follow the protocol of [14] and randomly corrupt a proportion $\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ of CIFAR-20 labels. We also include the clean data where $\alpha = 0$. As shown in Table 2 and Figure 6, *SiMCod-*

Table 3: Balanced accuracy on PD Activity and AD-3DMRI (%, mean ± standard deviation). Gain compares *SiMCoding* with the strongest reported baseline.

Dataset/task	Train → Test	CE	MCR^2	CRATE	*SiMCoding*	Gain	p-value
PD-Binary	–	80.33 ± 5.76	62.00 ± 1.73	77.26 ± 6.40	**87.56 ± 5.35**	↑ 7.23	4.55×10^{-4}
PD-Early	–	86.69 ± 0.60	73.57 ± 1.02	68.58 ± 4.80	**90.38 ± 0.86**	↑ 3.68	4.77×10^{-14}
PD-Severity	–	35.78 ± 0.27	38.54 ± 1.58	36.97 ± 3.96	**40.24 ± 0.73**	↑ 1.70	6.51×10^{-5}
PD-Fine-Grained	–	37.56 ± 0.17	26.49 ± 0.58	40.03 ± 2.27	**41.07 ± 0.19**	↑ 1.04	3.53×10^{-2}
AD-3DMRI (AD vs CN)	ADNI-1 → ADNI-1	80.17 ± 2.44	65.91 ± 2.79	–	**82.19 ± 2.65**	↑ 2.02	7.96×10^{-2}
	ADNI-1 → ADNI-2	81.74 ± 1.41	70.93 ± 3.19	–	**84.57 ± 1.29**	↑ 2.83	8.14×10^{-3}
	ADNI-2 → ADNI-2	84.46 ± 1.96	68.84 ± 3.15	–	**86.61 ± 1.98**	↑ 2.15	1.69×10^{-2}
	ADNI-2 → ADNI-1	77.96 ± 3.12	61.91 ± 4.18	–	**79.32 ± 1.90**	↑ 1.36	5.41×10^{-2}

ing consistently demonstrates the strongest robustness across all corruption levels. Its training accuracy closely follows the ideal behaviour $1 - \alpha$, indicating that the model selectively fits clean samples while suppressing corrupted labels, which leads to the best generalisation performance and consistent improvements of 4.1–10.1 percentage points over competing methods. In contrast, CE severely overfits noisy labels, achieving nearly perfect training accuracy but suffering substantial degradation in test performance as the corruption ratio increases. CRATE exhibits moderate robustness but unstable generalisation under severe corruption, while MCR^2 shows weak fitting ability and poor overall performance despite appearing relatively insensitive to label noise. Overall, these results demonstrate that learning minimal-complexity representations enables *SiMCoding* to effectively suppress noisy supervision while preserving robust discriminative structures.

Table 2: Test accuracy (%) on CIFAR-20 under different label noise ratios.

Ratio α	CE	MCR^2	CRATE	*SiMCoding*	Gain
$\alpha = 0$	80.76	41.89	53.45	**80.87**	↑ 0.11
$\alpha = 0.1$	73.56	30.37	48.00	**77.70**	↑ 4.14
$\alpha = 0.2$	65.94	25.06	42.56	**71.26**	↑ 5.32
$\alpha = 0.3$	58.37	22.12	35.71	**64.80**	↑ 6.43
$\alpha = 0.4$	47.89	16.37	31.21	**57.83**	↑ 9.94
$\alpha = 0.5$	40.57	14.52	25.42	**50.70**	↑ 10.13

3.1 Effectiveness of *SiMCoding* on Real-World Applications

In this subsection, we validate the effectiveness of *SiMCoding* on two real-world applications. **PD Activity** [17] contains wearable-sensor recordings from up to 135 subjects performing up to 16 daily motor activities, represented as 600×384 multivariate time series. Four severity levels (Normal, Mild, Moderate, and Severe) support the *PD-Binary*, *PD-Early*, *PD-Severity*, and *PD-Fine-Grained* classification tasks. **AD-3DMRI** uses volumetric MRI scans from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [35] for binary classification of Alzheimer's disease (AD) versus cognitively normal (CN) subjects. It comprises 374 ADNI-1 and 386 ADNI-2 subjects, with one MRI volume per subject.

Experimental setup. We use a lightweight Transformer for PD Activity and a 3D ResNet-18 for AD-3DMRI, with identical backbones for CE, MCR^2, and *SiMCoding* within each task. Note that CRATE is not suitable for 3D dataset so we omit it for AD-3DMRI classification. As both datasets exhibit class imbalance, we apply minority-class oversampling to balance the class distribution in the training set only, while keeping the validation and test sets unchanged. For PD Activity, we use an 80:20 training/test split. We observed relatively large standard deviations across five independent runs, possibly due to the limited number of subjects in the dataset. Therefore, we increased the number of runs from 5 to 25 to obtain more reliable estimates of the mean and standard deviation. For AD-3DMRI, we conduct five independent runs with within-cohort internal testing and cross-cohort external testing in both training directions. We report balanced accuracy for all experiments.

Table 3 shows that *SiMCoding* achieves the highest mean balanced accuracy across all four PD tasks and all four AD evaluations. On PD Activity, improvements over the strongest baseline range from 1.04 to 7.23 percentage points and are statistically significant ($p < 0.05$), with the largest gain observed on PD-Binary. These results demonstrate the effectiveness of *SiMCoding* in real-world environments characterised by noise and high uncertainty. On AD-3DMRI, improvements over CE range from 1.36 to 2.83 percentage points. The gains are statistically significant for both internal ADNI-2 testing and external ADNI-2 testing after training on ADNI-1, while the two ADNI-1

Table 4: Overall accuracy (%) on all currently seen classes after each incremental stage. Avg. Acc. is the five-stage arithmetic mean; Drop is the relative decrease from Stage 1 to Stage 5, $100(A_1 - A_5)/A_1$. Gain compares SiM-ReduNet with the best baseline in each column, reversing the subtraction for Drop. All derived values use unrounded accuracies. Best values, including ties, are highlighted.

Method	MNIST							ImageNette						
	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5	Avg. Acc.↑	Drop(%) ↓	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5	Avg. Acc.↑	Drop(%) ↓
oEWC	99.95	48.98	31.19	27.17	21.07	45.67	78.92	89.51	37.90	26.98	16.77	13.91	37.02	84.46
SI	99.15	47.49	30.84	24.64	15.64	43.55	84.23	49.49	25.31	17.57	13.36	10.17	23.18	79.46
DER	99.48	97.52	95.77	93.89	89.20	95.17	10.33	87.72	58.23	39.00	31.47	26.17	48.52	70.17
ReduNet	99.86	98.82	98.34	97.39	95.82	98.05	4.04	49.49	31.21	29.04	27.14	29.15	33.20	41.10
SiM-ReduNet	99.86	98.65	98.28	97.39	95.82	98.00	4.04	80.05	55.21	49.10	41.68	39.80	53.17	50.29
Gain	↓ 0.09	↓ 0.17	↓ 0.07	0.00	0.00	↓ 0.05	0.00	↓ 9.46	↓ 3.02	↑ 10.09	↑ 10.20	↑ 10.65	↑ 4.65	↓ 9.18

evaluations do not reach statistical significance at the 0.05 level. These results further suggest that *SiMCoding* is effective at capturing intrinsic and discriminative structures, consistent with Figure 3, where *SiMCoding* produces clearer and more concentrated structural saliency maps. Overall, these findings demonstrate the effectiveness of learning minimal structural representations across different healthcare modalities, including wearable time series and volumetric MRI, as well as their potential for cross-cohort generalisation.

3.2 *SiMCoding* based Incremental Learning

Chan et al. [13] derive ReduNet by unrolling gradient ascent on MCR^2, and Wu et al. [36] extend it to incremental learning. Motivated by the compact class-wise representations encouraged by *SiMCoding*, we similarly construct SiM-ReduNet. We divide each dataset into five tasks with two classes per task and evaluate overall accuracy on all seen classes after each stage. SiM-ReduNet updates $\epsilon_t^2 = d/(3\sqrt{K_t})$, where $K_t = 2t$ and $d = 784$ for MNIST or 3072 for ImageNette. We use 200 layers on MNIST and retain the original 50-layer ImageNette setup. Detailed settings are provided in Appendix C.4.

Table 4 compares the two networks with oEWC [37], SI [38], and DER [39]. On MNIST, ReduNet and SiM-ReduNet achieve comparable five-stage means of 98.05% and 98.00%, respectively; both finish at 95.82% with a Drop of 4.04%. On ImageNette, SiM-ReduNet exceeds ReduNet at every stage and achieves the highest accuracy at Stages 3–5. Its mean accuracy of 53.17% exceeds DER, the strongest baseline by this average, by 4.65 percentage points, and its final accuracy is 39.80% versus 29.15% for ReduNet. However, its relative Drop is larger (50.29% versus 41.10%), reflecting their different starting accuracies. Since the evaluation set expands, Drop measures relative overall accuracy decline rather than forgetting on a fixed task. These single-run results support comparable MNIST performance and stronger overall ImageNette accuracy under the retained settings.

4 Conclusion

This work demonstrates that, in supervised representation learning, preserving only minimal yet highly discriminative structural information can be more effective than retaining full structural complexity. To achieve this, we proposed *SiMCoding*, a simple framework that explicitly controls representation complexity through coding precision, encouraging each class to concentrate around a near one-dimensional discriminative subspace. Across diverse datasets and modalities, *SiMCoding* consistently outperforms CE, MCR^2, and CRATE, particularly under noisy and uncertain settings, while also achieving competitive overall accuracy in incremental learning. Our results suggest that suppressing unnecessary structural complexity leads to cleaner, more robust, and more generalisable representations. However, *SiMCoding* is designed for high-dimensional data with underlying low-dimensional structure, its advantages may diminish when the data are already low-dimensional or lack clear structural organisation, as may be the case for some tabular datasets. Overall, this work highlights minimal-complexity representation learning as a simple and effective paradigm for supervised learning and opens new directions for studying compact structural representations in broader settings.

Acknowledgments and Disclosure of Funding

This work was supported by Innovate UK Project 10081615 and 10180327. We sincerely thank the reviewers for their valuable comments and suggestions, which substantially improved this paper. Due to page limits, we could not include all experiments presented during the rebuttal, but have incorporated the most important ones into the final manuscript. We also refined the experimental setup to better demonstrate the effectiveness of *SiMCoding*. Specifically, the ADNI experiments in the rebuttal used 2D dynamic images derived from 3D MRI scans to reduce computational costs and training time. In the final manuscript, we use full 3D MRI volumes, which better align with the standard AD classification setting in the research community.

References

[1] Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270):20230254, 2024.

[2] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[3] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[4] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pages 1–3, 2024.

[5] Kwanseok Oh, Jee Seok Yoon, and Heung-Il Suk. Learn-explain-reinforce: Counterfactual reasoning and its guidance to reinforce an alzheimer's disease diagnosis model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4843–4857, 2022.

[6] Elia Kaufmann, Leonard Bauersfeld, Antonio Loquercio, Matthias Müller, Vladlen Koltun, and Davide Scaramuzza. Champion-level drone racing using deep reinforcement learning. *Nature*, 620(7976):982–987, 2023.

[7] Guilherme N DeSouza and Avinash C Kak. Vision for mobile robot navigation: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 24(2):237–267, 2002.

[8] Yuzhe Yang, Yuan Yuan, Guo Zhang, Hao Wang, Ying-Cong Chen, Yingcheng Liu, Christopher G Tarolli, Daniel Crepeau, Jan Bukartyk, Mithri R Junna, et al. Artificial intelligence-enabled detection and assessment of parkinson's disease using nocturnal breathing signals. *Nature medicine*, 28(10):2207–2215, 2022.

[9] Menghui Zhou, Xulong Wang, Tong Liu, Yun Yang, and Po Yang. Integrating visualised automatic temporal relation graph into multi-task learning for alzheimer's disease progression prediction. *IEEE Transactions on Knowledge and Data Engineering*, 36(10):5206–5220, 2024.

[10] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10164–10183, 2024.

[11] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17853–17862, 2023.

[12] Yi Ma, Doris Tsao, and Heung-Yeung Shum. On the principles of parsimony and self-consistency for the emergence of intelligence. *arXiv preprint arXiv:2207.04630*, 2022.

[13] Kwan Ho Ryan Chan, Yaodong Yu, Chong You, Haozhi Qi, John Wright, and Yi Ma. Redunet: A white-box deep network from the principle of maximizing rate reduction. *Journal of machine learning research*, 23(114):1–103, 2022.

[14] Yaodong Yu, Kwan Ho Ryan Chan, Chong You, Chaobing Song, and Yi Ma. Learning diverse and discriminative representations via the principle of maximal coding rate reduction. *Advances in Neural Information Processing Systems*, 33, 2020.

[15] Menghui Zhou, Mingxia Wang, Yu Zhang, Zhipeng Yuan, Vitaveska Lanfranchi, and Po Yang. Learning interpretable continuous representation for alzheimer's disease classification. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 6545–6552. IEEE, 2024.

[16] Xulong Wang, Xiyang Peng, Zheyuan Xu, Mingchang Xu, Yun Yang, Menghui Zhou, Zhong Zhao, Peng Yue, and Po Yang. Pdwearml: Leveraging daily activities for fast parkinson's disease severity assessment with wearable machine learning. *IEEE Transactions on Biomedical Engineering*, 2025.

[17] Xiyang Peng, Yuting Zhao, Ziheng Li, Xulong Wang, Fengtao Nan, Zhong Zhao, Yun Yang, and Po Yang. Multi-scale and multi-level feature assessment framework for classification of parkinson's disease state from short-term motor tasks. *IEEE Transactions on Biomedical Engineering*, 2025.

[18] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Occam's razor. *Information processing letters*, 24(6):377–380, 1987.

[19] John Wright and Yi Ma. *High-dimensional data analysis with low-dimensional models: Principles, computation, and applications*. Cambridge University Press, 2022.

[20] Yi Ma, Harm Derksen, Wei Hong, and John Wright. Segmentation of multivariate mixed data via lossy data coding and compression. *PAMI*, 2007.

[21] Menghui Zhou and Po Yang. Automatic temporal relation in multi-task learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3570–3580, 2023.

[22] Matthias Hein and Jean-Yves Audibert. Intrinsic dimensionality estimation of submanifolds in rd. In *Proceedings of the 22nd international conference on Machine learning*, pages 289–296, 2005.

[23] Stefano Spigler, Mario Geiger, and Matthieu Wyart. Asymptotic learning curves of kernel methods: empirical data vs teacher-student paradigm. *arXiv preprint arXiv:1905.10843*, 2019.

[24] Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. *arXiv preprint arXiv:2104.08894*, 2021.

[25] John Wright and Yi Ma. *High-Dimensional Data Analysis with Low-Dimensional Models: Principles, Computation, and Applications*. Cambridge University Press, 2021.

[26] Jorma Rissanen. *Stochastic complexity in statistical inquiry*, volume 15. World scientific, 1998.

[27] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[28] Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. *Advances in Neural Information Processing Systems*, 34:29820–29834, 2021.

[29] Peng Wang, Huikang Liu, Druv Pai, Yaodong Yu, Zhihui Zhu, Qing Qu, and Yi Ma. A global geometric analysis of maximal coding rate reduction. *arXiv preprint arXiv:2406.01909*, 2024.

[30] Yaodong Yu, Sam Buchanan, Druv Pai, Tianzhe Chu, Ziyang Wu, Shengbang Tong, Benjamin Haeffele, and Yi Ma. White-box transformers via sparse rate reduction. *Advances in Neural Information Processing Systems*, 36:9422–9457, 2023.

[31] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[32] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[33] Jeremy Howard. Imagenette: A smaller subset of 10 easily classified classes from imagenet, March 2019.

[34] Yaodong Yu, Sam Buchanan, Druv Pai, Tianzhe Chu, Ziyang Wu, Shengbang Tong, Hao Bai, Yuexiang Zhai, Benjamin D Haeffele, and Yi Ma. White-box transformers via sparse rate reduction: Compression is all there is? *Journal of Machine Learning Research*, 25(300):1–128, 2024.

[35] Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691, 2008.

[36] Ziyang Wu, Christina Baek, Chong You, and Yi Ma. Incremental learning via rate reduction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1125–1133, 2021.

[37] Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In *International conference on machine learning*, pages 4528–4537. PMLR, 2018.

[38] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International conference on machine learning*, pages 3987–3995. PMLR, 2017.

[39] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33:15920–15930, 2020.

[40] Shengbang Tong, Xili Dai, Ziyang Wu, Mingyang Li, Brent Yi, and Yi Ma. Incremental learning of structured memory via closed-loop transcription. *arXiv preprint arXiv:2202.05411*, 2022.

[41] Xili Dai, Shengbang Tong, Mingyang Li, Ziyang Wu, Michael Psenka, Kwan Ho Ryan Chan, Pengyuan Zhai, Yaodong Yu, Xiaojun Yuan, Heung-Yeung Shum, et al. Ctrl: Closed-loop transcription to an ldr via minimaxing rate reduction. *Entropy*, 24(4):456, 2022.

[42] Christopher M Bishop and Hugh Bishop. *Deep learning: Foundations and concepts*. Springer Nature, 2023.

[43] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[44] Shengbang Tong, Xili Dai, Yubei Chen, Mingyang Li, Zengyi Li, Brent Yi, Yann LeCun, and Yi Ma. Unsupervised learning of structured representations via closed-loop transcription. *arXiv preprint arXiv:2210.16782*, 2022.

[45] Tianzhe Chu Chu, Shengbang Tong, Tianjiao Ding, Xili Dai, Benjamin Haeffele, René Vidal, and Yi Ma. Image clustering via the principle of rate reduction in the age of pretrained models. International Conference on Learning Representations (ICLR), 2024.

[46] Ziyang Wu, Jingyuan Zhang, Druv Pai, XuDong Wang, Chandan Singh, Jianwei Yang, Jianfeng Gao, and Yi Ma. Simplifying dino via coding rate regularization. *arXiv preprint arXiv:2502.10385*, 2025.

[47] Menghui Zhou, Gaoshan Bi, Vitaveska Lanfranchi, and Po Yang. On the limits of maximal coding rate reduction for out-of-distribution generalisation. *arXiv preprint arXiv:2609.21001*, 2026.

[48] Jinrui Yang, Xianhang Li, Druv Pai, Yuyin Zhou, Yi Ma, Yaodong Yu, and Cihang Xie. Scaling white-box transformers for vision. *Advances in Neural Information Processing Systems*, 37:36995–37019, 2024.

[49] Ziyang Wu, Tianjiao Ding, Yifu Lu, Druv Pai, Jingyuan Zhang, Weida Wang, Yaodong Yu, Yi Ma, and Benjamin D Haeffele. Token statistics transformer: Linear-time attention via variational rate reduction. *arXiv preprint arXiv:2412.17810*, 2024.

[50] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[51] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Appendix

The appendix is organised as follows:

- **Appendix A** reviews related work on lossy coding rate, maximal coding rate reduction, and white-box transformers.
- **Appendix B** presents the formal statement of Theorem 1, supporting lemmas, and detailed proofs.
- **Appendix C** provides dataset descriptions, model architectures, training and evaluation protocols, and implementation details for the image classification, PD Activity, AD-3DMRI, and incremental learning experiments.
- **Appendix D** discusses potential positive and negative societal impacts.

The NeurIPS Paper Checklist follows these sections.

A Related Work

A.1 Lossy Coding Rate

Information-theoretic principles provide a rigorous foundation for the analysis and design of machine learning methods. Central among these is the notion of *coding rate* [26], which quantifies the efficiency and structural complexity of data representations. Rooted in rate–distortion theory and the minimum description length principle [27], coding-based formulations establish a direct connection between compressibility, geometry, and statistical complexity by seeking compact representations under distortion constraints.

A seminal contribution in this direction is the lossy data coding framework of Ma et al. [20], which formulates data segmentation as minimising the overall coding length of multivariate mixed data. Their analysis derives a closed-form upper bound on the coding length for data drawn from Gaussian distributions supported on subspaces and shows that the expected rate–distortion objective is *concave* over probabilistic assignments. Consequently, the optimum is attained at a vertex of the assignment simplex, implying that deterministic (hard) segmentation is asymptotically optimal. This result provides theoretical justification for bottom–up clustering methods driven by coding-rate reduction and explains their robustness to noise, outliers, and unknown subspace dimensionality.

A.2 Maximal Coding Rate Reduction

However, in the original work of Ma et al. [20], the analysis is conducted primarily on *raw data* and focuses on linear transformations, which inherently limits its applicability to many modern learning scenarios. Nevertheless, the lossy coding rate perspective has had a profound influence on contemporary representation learning. In today's era of large-scale and highly complex data, effective modelling typically requires expressive *nonlinear* mappings, commonly realised through neural networks acting as powerful function approximators. Motivated by this need, Yu et al. [14] assume that each class of complex data lies on a latent low-dimensional submanifold and extend the lossy coding rate formulation to deep neural networks, proposing the principle of Maximal Coding Rate Reduction (MCR^2) as a unifying learning objective. The MCR^2 framework posits that an effective representation should make features within each class as diverse as possible while ensuring that features from different classes remain orthogonal.

The formulation of MCR^2 has subsequently motivated a spectrum of methodological developments, reflecting its flexibility across different representation learning tasks. Building upon MCR^2, Chan et al. [13] introduce ReduNet, a conceptually interpretable white-box multi-layer deep network obtained by unrolling the optimisation of the rate-reduction objective. Wu et al. [36] first explore the potential of MCR^2 in incremental learning, and Tong et al. [40] further develop a closed-loop transcription–based incremental learning framework grounded in MCR^2. Dai et al. [41] extend the MCR^2 objective to provide a unified perspective bridging the benefits of autoencoders [42] and generative adversarial networks [43]. Several other works have adapted MCR^2 for various semi-supervised or unsupervised learning settings [44, 45, 46]. Zhou et al. [47] starts to explore the potential of MCR^2 in out-of-distribution generalisation.

A.3 White-Box Transformer via Sparse Rate Reduction

Yu et al. [30] recently generalise MCR^2 by introducing *sparse rate reduction*, a variant that jointly promotes compression and sparsity in the learned representation. A key contribution of their work is the demonstration that optimising sparse rate reduction via unrolled alternating minimisation naturally yields a *white-box*, fully mathematically interpretable transformer-like architecture, termed CRATE. Following this development, subsequent works [34, 48] further explore the capabilities of CRATE across a broad range of learning paradigms, including supervised classification, masked autoencoding, self-supervised representation learning, and text modelling. In addition, Wu et al. [49] propose a new attention mechanism with linear computational and memory complexity, derived from a variational formulation of the MCR^2 objective, thereby substantially reducing the computational overhead of transformer architectures.

Despite the remarkable success of the MCR^2 family across diverse domains, none of the aforementioned works explicitly addresses a fundamental question: *in supervised learning, is it truly necessary to preserve all structural details of the data distribution?* This question motivates the main contribution of our work. We show that learning only the most discriminative structural component of each class is not merely sufficient but in fact leads to substantial improvements in generalisation performance and robustness against input perturbations and label noise.

B Proof about *SiMCoding*

Notations. We denote by $\mathbb{S}_{++}^d$ the set of $d \times d$ symmetric positive definite matrices, by $\mathbb{R}_{\geq 0}$ the set of non-negative real numbers, and by $\mathbb{Z}_{++}$ the set of positive integers.

B.1 Preliminaries

We begin with a few lemmas that support both the implementation of our algorithm and the subsequent theoretical analysis.

Since $\boldsymbol{Z}\boldsymbol{Z}^\top \in \mathbb{R}^{d\times d}$ and $\boldsymbol{Z}^\top\boldsymbol{Z} \in \mathbb{R}^{m\times m}$ share identical non-zero eigenvalues, the coding rate admits the following commutative form.

Lemma 1 (Commutative Property [20])**.** *For any $\boldsymbol{Z} \in \mathbb{R}^{d\times m}$,*

$$\mathcal{R}(\boldsymbol{Z},\epsilon) \doteq \frac{1}{2}\log\det\big(\boldsymbol{I} + \tfrac{d}{m\epsilon^2}\boldsymbol{Z}\boldsymbol{Z}^\top\big) = \frac{1}{2}\log\det\big(\boldsymbol{I} + \tfrac{d}{m\epsilon^2}\boldsymbol{Z}^\top\boldsymbol{Z}\big)\,.$$

When $d > m$, the second expression is computationally more efficient, which is particularly useful for implementing our *SiMCoding* framework.

In addition, the singular values of $\boldsymbol{Z}$ (equivalently, the eigenvalues of $\boldsymbol{Z}\boldsymbol{Z}^\top$) remain unchanged under orthogonal transformations, yielding the following invariance property.

Lemma 2 (Invariance Property [20])**.** *For any $\boldsymbol{Z} \in \mathbb{R}^{d\times m}$ and any orthogonal matrices $\boldsymbol{U} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{V} \in \mathbb{R}^{m\times m}$,*

$$\mathcal{R}(\boldsymbol{Z},\epsilon) = \mathcal{R}(\boldsymbol{U}\boldsymbol{Z}\boldsymbol{V}^\top,\epsilon).$$

Lemma 3 (Upper Bound of MCR^2 [14])**.** *Let $\boldsymbol{Z} \in \mathbb{R}^{d\times m}$ denote the complete set of representations for K classes, where the i-th class is represented by $\boldsymbol{Z}_i \in \mathbb{R}^{d\times m_i}$, i.e., $\boldsymbol{Z} = \cup_{i=1}^K \boldsymbol{Z}_i$ and $m = \sum_{i=1}^K m_i$. We have*

$$\Delta\mathcal{R}(\boldsymbol{Z},\epsilon) \leq \sum_{i=1}^K \frac{1}{2m}\log\left(\frac{\det^m\left(\boldsymbol{I} + \frac{d}{m\epsilon^2}\boldsymbol{Z}_i\boldsymbol{Z}_i^\top\right)}{\det^{m_i}\left(\boldsymbol{I} + \frac{d}{m_i\epsilon^2}\boldsymbol{Z}_i\boldsymbol{Z}_i^\top\right)}\right), \tag{3}$$

with equality holds if and only if

$$\boldsymbol{Z}_i^\top\boldsymbol{Z}_j \;=\; \boldsymbol{0} \quad \textit{for all} \quad 1 \leq i \neq j \leq K.$$

B.2 Properties of *SiMCoding* Framework

We now present a more formal version of Theorem 1:

Theorem 2. *Consider the following optimisation objective:*

$$\max_{\mathbf{Z}} \Delta\mathcal{R}(\mathbf{Z}, \epsilon) = \frac{1}{2}\log\det\left(\mathbf{I} + \frac{d}{m\epsilon^2}\mathbf{Z}\mathbf{Z}^T\right) - \sum_{i=1}^{K}\frac{m_i}{2m}\log\det\left(\mathbf{I} + \frac{d}{m_i\epsilon^2}\mathbf{Z}_i\mathbf{Z}_i^\top\right), \quad \text{s.t.} \quad \|\mathbf{Z}_i\|_F^2 = m_i \tag{4}$$

Let $\boldsymbol{Z}^* = \boldsymbol{Z}_1^* \cup \cdots \cup \boldsymbol{Z}_K^*$ be the optimal solution to equation 4, and

$$d_i^{\max} = \sqrt{\frac{m_i}{m}} \cdot \frac{d}{\epsilon^2}.$$

Then the following properties hold:

- ***Discriminativeness:*** *Representations from different classes lie in mutually orthogonal low-dimensional linear subspaces:*

$$(\boldsymbol{Z}_i^*)^\top \boldsymbol{Z}_j^* = 0, \ \textit{for all} \ \ i \neq j.$$

- ***Bounded Dimensionality:*** *Each class-specific subspace satisfies $d_i \leqslant d_i^{\max}$. The singular values of $\boldsymbol{Z}_i^*$ fall into one of two patterns:*

  1. *All singular values are equal to $\sqrt{\frac{m_i}{d_i}}$.*
  2. *The largest $d_i - 1$ singular values are equal to σ_H, and the remaining singular value is σ_L, where*

$$\sigma_H \in \left(\sqrt{\tfrac{m_i}{d_i}}, \sqrt{\tfrac{m_i}{d_i - 1}}\right), \quad \textit{and} \quad \sigma_L > 0.$$

B.3 Proof of Main Results

We start with presenting two lemmas that will be useful for the proof for Theorem 2.

Lemma 4. *Consider the following optimisation problem:*

$$\max_{m} \ \ m \cdot f\left(\frac{c}{m}\right) \quad \textit{subject to} \quad m \leqslant \frac{c}{x_p}, \tag{5}$$

where the function $f(x)$ satisfies the following properties:

(i) $f(0) = 0$,

(ii) $f'(0) = 0$ *and* $f'(x) > 0$ *for all* $x > 0$,

(iii) There exists $x_p > 0$ such that $f'(x)$ is strictly increasing on $(0, x_p)$ and strictly decreasing on (x_p, ∞), with $f''(x_p) = 0$.

Let $m^\star$ be the optimal solution to equation 5. Then:

$$m^\star < \frac{c}{x_p}.$$

Proof of Lemma 4. Let $x = \frac{c}{m}$, and define the auxiliary function

$$g(x) = \frac{f(x)}{x}, \quad \text{so that} \quad g'(x) = \frac{xf'(x) - f(x)}{x^2}.$$

Next, define

$$h(x) = xf'(x) - f(x), \quad \text{which implies} \quad h'(x) = xf''(x).$$

By assumption (iii), $f''(x_p) = 0$, and:

- For $x \in (0, x_p)$, we have $f''(x) > 0$, so $h'(x) > 0$;

- For $x > x_p$, $f''(x) < 0$, so $h'(x) < 0$.

Since $h(0) = 0$, this implies that $h(x)$ increases on $(0, x_p)$, decreases on (x_p, ∞), and attains its maximum at $x = x_p$. In particular, $h(x_p) \geqslant 0$, and thus:

$$g'(x_p) = \frac{h(x_p)}{x_p^2} \geqslant 0.$$

Moreover, since $h'(x) < 0$ for $x > x_p$, it follows that $g'(x)$ is strictly decreasing on $[x_p, \infty)$.

Now suppose $g'(x^\star) = 0$ at the optimum. Given that $g'(x_p) \geqslant 0$ and $g'(x)$ is strictly decreasing for $x \geqslant x_p$, we must have $x^\star > x_p$. Recalling that $x^\star = \frac{c}{m^\star}$, this yields:

$$\frac{c}{m^\star} > x_p \quad \Rightarrow \quad m^\star < \frac{c}{x_p}.$$

□

Lemma 5. *Given any twice differentiable function $f : \mathbb{R}_{\geq 0} \to \mathbb{R}$, an integer $r \in \mathbb{Z}_{++}$, and a constant $c \in \mathbb{R}_+$, consider the optimisation problem*

$$\begin{aligned} \max_{\mathbf{x}} \quad & \sum_{i=1}^{r} f(x_i) \\ \textit{s.t.} \quad & x_1 \geqslant x_2 \geqslant \cdots \geqslant x_r \geqslant 0, \\ & \sum_{i=1}^{r} x_i = c, \end{aligned} \tag{6}$$

Let $\mathbf{x}^*$ be an arbitrary global solution to equation 6 and $f(x)$ has the following properties:

(i) $f'(x)$ satisfies $f'(0) = 0$, is positive for $x > 0$;

(ii) There exists some $x_p > 0$ such that $f'(x)$ is strictly increasing on $(0, x_p]$ and strictly decreasing on $[x_p, \infty)$.

Let

$$m \in \mathbb{Z}_{++} \quad \textit{and} \quad m \leq \min\{r, \frac{c}{x_p}\},$$

then $\mathbf{x}^*$ must take one of the following two forms:

- $\mathbf{x}^* = \Big[\underbrace{\frac{c}{m}, \ldots, \frac{c}{m}}_{m}, \underbrace{0, \ldots, 0}_{r-m}\Big]$,
- $\mathbf{x}^* = \Big[\underbrace{x_H, \ldots, x_H, x_L}_{m}, \underbrace{0, \ldots, 0}_{r-m}\Big]$, *for some* $x_H \in \left(\frac{c}{m}, \frac{c}{m-1}\right)$ *and* $x_L > 0$.

Proof. We consider two cases.

Case 1: $r = 1$**.**
In this case the problem reduces to

$$\max_{x_1} \; f(x_1) \quad \text{s.t.} \quad x_1 = c.$$

The unique solution is

$$\mathbf{x}^* = [c],$$

which agrees with the stated result.

Case 2: $r > 1$**.**

Let

$$c_i(x) = -x_i \leqslant 0.$$
$$h(x) = \sum_{i=1}^{r} x_i - c = 0.$$

so the Lagrangian function is

$$\begin{aligned}\mathcal{L}(x_1, \ldots, x_r, \alpha_1, \ldots, \alpha_r, \beta) &= -\sum_{i=1}^{r} f(x_i) + \sum_{i=1}^{r} \alpha_i c_i(x) + \beta h(x) \\ &= -\sum_{i=1}^{r} f(x_i) - \sum_{i=1}^{r} \alpha_i x_i + \beta \left(\sum_{i=1}^{r} x_i - c \right)\end{aligned} \tag{7}$$

Thus, the first-order KKT conditions require that the optimal solution $\boldsymbol{x}^\star$, $\boldsymbol{\alpha}^\star$, and $\beta^\star$ satisfy:

$$\begin{aligned} f'(x_i^\star) &= -\alpha_i^\star + \beta^\star, \quad \forall i \in \{1, \cdots, r\}, \\ \alpha_i^\star x_i^\star &= 0, \quad \forall i \in \{1, \cdots, r\}, \\ x_i^\star &\geqslant 0, \quad \forall i \in \{1, \cdots, r\}, \\ \alpha_i^\star &\geqslant 0, \quad \forall i \in \{1, \cdots, r\}, \\ \sum_{i=1}^{r} x_i^\star &= c. \end{aligned} \tag{8}$$

If $x_i^\star > 0$, then $\alpha_i^\star = 0$, which implies

$$f'(x_i^\star) = \beta^\star.$$

If $x_i^\star = 0$, recalling that $f'(0) = 0$, we have

$$f'(x_i^\star) = -\alpha_i^\star + \beta^\star = 0.$$

So now we assume there are m entries in $\boldsymbol{x}^\star$ are positive, i.e.,

$$\boldsymbol{x}^* = \Big[\underbrace{x_1^*, \ldots, x_m^*}_{m}, \underbrace{0, \ldots, 0}_{r-m}\Big].$$

and

$$f'(x_i^\star) = \beta^\star, \forall i \in \{1, \cdots, m\}.$$

Since there exists some $x_p > 0$ such that $f'(x)$ is strictly increasing on $(0, x_p]$ and strictly decreasing on $[x_p, \infty)$—that is, f' is strictly unimodal—the equation $f'(x) = \beta^\star$ has at most two solutions. Consequently, the optimal vector $\mathbf{x}^\star = [x_1^\star, \ldots, x_m^\star]$ can take at most two distinct values. Two cases arise:

1. *Uniform Case:*

   If $x_1^\star = x_2^\star = \cdots = x_m^\star$, then

   $$m x_i^\star = c \quad \Longrightarrow \quad x_i^\star = \frac{c}{m} \quad \text{for all } i,$$

   and thus

   $$\mathbf{x}^\star = \Big[\underbrace{\frac{c}{m}, \ldots, \frac{c}{m}}_{m}, \underbrace{0, \ldots, 0}_{r-m}\Big].$$

   In this case, it is necessary to note that

   $$\frac{c}{m} \geqslant x_p \Longleftrightarrow m \leqslant \frac{c}{x_p}.$$

We prove it by contradiction. Suppose, on the contrary, that $\frac{c}{m} < x_p$. Then there exists $\sigma > 0$ such that $\frac{c}{m} + \sigma < x_p$. Define

$$\hat{x}_1^\star = \frac{c}{m} - \sigma, \quad \hat{x}_2^\star = \frac{c}{m} + \sigma.$$

Since $f'(x)$ is strictly increasing on $(0, x_p]$, it follows that

$$f'\left(\frac{c}{m} - \sigma\right) < f'\left(\frac{c}{m} + \sigma\right).$$

This inequality implies the slope between $\frac{c}{m} - \sigma$ and $\frac{c}{m}$ is less than the slope between $\frac{c}{m}$ and $\frac{c}{m} + \sigma$, hence

$$\frac{f\left(\frac{c}{m}\right) - f\left(\frac{c}{m} - \sigma\right)}{\sigma} < \frac{f\left(\frac{c}{m} + \sigma\right) - f\left(\frac{c}{m}\right)}{\sigma}.$$

Multiplying by σ and rearranging terms yields

$$f(\hat{x}_1^\star) + f(\hat{x}_2^\star) = f\left(\frac{c}{m} - \sigma\right) + f\left(\frac{c}{m} + \sigma\right) > 2f\left(\frac{c}{m}\right) = f(x_1^\star) + f(x_2^\star),$$

i.e.,

$$f(\hat{x}_1^\star) + f(\hat{x}_2^\star) + \sum_{i=3}^{r} f(x_i^\star) > \sum_{i=1}^{r} f(x_i^\star).$$

This contradiction shows that the assumed inequality $\frac{c}{m} < x_p$ cannot hold. Hence, we conclude

$$\frac{c}{m} \geqslant x_p, \quad \text{equivalently,} \quad m \leqslant \frac{c}{x_p}.$$

We now consider the case where multiple values of m satisfy $m \leqslant \frac{c}{x_p}$. A natural question is whether the optimum lies on the boundary, i.e., $m^\star = \frac{c}{x_p}$. By Lemma 4, it does not: the optimal value strictly satisfies

$$m^\star < \frac{c}{x_p}.$$

2. *Two-Valued Case:*

Otherwise, the set $\{y_1^*, \ldots, y_m^*\}$ consists of two distinct values, say x_H and x_L with $x_H > x_L > 0$. Without loss of generality, assume that, after a suitable reordering,

$$x_1^* = \cdots = x_{m-1}^* = x_H \quad \text{and} \quad x_m^* = x_L.$$

Then the constraint implies

$$(m-1)x_H + x_L = c.$$

Notice that the average c/m must lie strictly between x_L and x_H; hence,

$$0 < x_L < \frac{c}{m} \quad \text{and} \quad \frac{c}{m} < x_H < \frac{c}{m-1}.$$

By the unimodality of f' we have

$$x_L < x_p < x_H.$$

Consequently,

$$f''(x_L) > 0 \quad \text{and} \quad f''(x_H) < 0.$$

In this case, $x_H > x_p$. To make it, we may push a stronger condition, i.e.,

$$x_H > \frac{c}{m} \geqslant x_p \Rightarrow m \leqslant \frac{c}{x_p}.$$

In this case, if we fix x_L, and only consider the rest $(m-1)x_H$, according to Lemma 4, the optimal value satisfies

$$m^\star - 1 < \frac{c - x_L}{x_p} \Rightarrow m^\star < \frac{c}{x_p} + \frac{x_p - x_L}{x_p}.$$

Combining the constraint $m \leqslant \frac{c}{x_p}$ with the optimal $m^\star < \frac{c}{x_p} + \frac{x_p - x_L}{x_p}$, we conclude that

$$m^\star \leqslant \frac{c}{x_p}.$$

Second-Order Analysis:

To ensure that the candidate $\mathbf{y}^*$ is indeed optimal, the second-order necessary conditions must hold (see, e.g., [50, Theorem 12.5]). For any feasible direction $\mathbf{v} = (v_1, \ldots, v_m)$ satisfying

$$\sum_{i=1}^{m} v_i = 0,$$

the Hessian of the Lagrangian must be positive semidefinite on the tangent space, i.e.,

$$\sum_{i=1}^{m} f''(x_i^*)\, v_i^2 \leq 0.$$

Suppose, for the sake of contradiction, that more than one index takes the lower value x_L. For example, if $x_{m-1}^* = x_m^* = x_L$, consider the perturbation direction defined by

$$v_{m-1} = 1, \quad v_m = -1, \quad \text{and } v_i = 0 \text{ for } i \neq m-1,\, m.$$

Clearly, $\sum_{i=1}^{m} v_i = 1 + (-1) = 0$. Then the second-order condition yields

$$\sum_{i=1}^{m} f''(x_i^*)\, v_i^2 = f''(x_{m-1}^*) + f''(x_m^*) = 2f''(x_L) \leq 0.$$

However, since $x_L < x_p$ we have $f''(x_L) > 0$, so that $2f''(x_L) > 0$, which contradicts the second-order necessary condition. Hence, it is necessary that exactly one index takes the lower value x_L, with all remaining $m-1$ indices equal to x_H.

Then the solution in this case is

$$\mathbf{x}^* = \Big[\underbrace{x_H, \ldots, x_H,\, x_L}_{m}, \underbrace{0, \ldots, 0}_{r-m}\Big],$$

where, from the relation $(m-1)x_H + x_L = c$, we obtain

$$(m-1)x_H < c \quad \text{and} \quad m\, x_H > c.$$

Thus,

$$x_H \in \Big(\frac{c}{m}, \frac{c}{m-1}\Big) \quad \text{and} \quad x_L > 0.$$

Step 5: Conclusion.

Combining the results from the two cases, we deduce that given

$$m \leqslant \min\left\{r, \frac{c}{x_p}\right\} \leqslant \frac{c}{x_p},$$

any global solution $\mathbf{x}^*$ to problem equation 6 must have the form

$$\mathbf{x}^* = \Big[\underbrace{x_1^*, \ldots, x_m^*}_{m}, \underbrace{0, \ldots, 0}_{r-m}\Big],$$

where either

$$x_1^* = \cdots = x_m^* = \frac{c}{m},$$

or

$$\mathbf{x}^* = \Big[\underbrace{x_H, \ldots, x_H, x_L}_{m}, \underbrace{0, \ldots, 0}_{r-m}\Big],$$

with

$$x_H \in \Big(\frac{c}{m}, \frac{c}{m-1}\Big) \qquad \text{and} \quad x_L > 0.$$

This completes the proof. □

Proof of Theorem 2. This proof follows a structure similar to that of Theorem A.6 in [14].

Specifically, without loss of generality, let

$$\boldsymbol{Z}^* = [\boldsymbol{Z}_1^*, \ldots, \boldsymbol{Z}_K^*]$$

be an optimal solution. To prove that the matrices $\boldsymbol{Z}_i^*$, for $i \in \{1, \ldots, K\}$, are pairwise orthogonal, assume for the sake of contradiction that

$$(\boldsymbol{Z}_i^*)^\top \boldsymbol{Z}_j^* \neq \boldsymbol{0}$$

for some indices $1 \leq i < j \leq K$.

By applying Lemma 3, the strict inequality in equation 3 holds for the optimal solution $\boldsymbol{Z}^*$. That is,

$$\Delta\mathcal{L}(\boldsymbol{Z}^*, \alpha) < \sum_{i=1}^{K} \frac{1}{2m} \log \left(\frac{\det^m \left(\boldsymbol{I} + \frac{d}{m\epsilon^2} \boldsymbol{Z}_i^* (\boldsymbol{Z}_i^*)^\top \right)}{\det^{m_i} \left(\boldsymbol{I} + \frac{d}{m_i\epsilon^2} \boldsymbol{Z}_i^* (\boldsymbol{Z}_i^*)^\top \right)} \right). \tag{9}$$

On the other hand, let $\operatorname{rank}(\boldsymbol{Z}_i) = d_i$, there exist matrices

$$\{\boldsymbol{U}_i' \in \mathbb{R}^{d \times d_i}\}_{i=1}^K$$

such that the columns of the concatenated matrix $[\boldsymbol{U}_1', \ldots, \boldsymbol{U}_K']$ form an orthonormal set (i.e., for any $i \neq j$, we have $\boldsymbol{U}_i'(\boldsymbol{U}_j')^\top = \boldsymbol{0}$, and each column has unit l_2-norm).

Let the compact singular value decomposition (SVD) of $\boldsymbol{Z}_i^*$ be given by

$$\boldsymbol{Z}_i^* = \boldsymbol{U}_i^* \boldsymbol{\Sigma}_i^* (\boldsymbol{V}_i^*)^\top,$$

and define

$$\boldsymbol{Z}' = [\boldsymbol{Z}_1', \ldots, \boldsymbol{Z}_K'], \quad \text{where} \quad \boldsymbol{Z}_i' = \boldsymbol{U}_i' \boldsymbol{\Sigma}_i^* (\boldsymbol{V}_i^*)^\top.$$

Then, for all $1 \leq i < j \leq K$, we have

$$\begin{aligned} (\boldsymbol{Z}_i')^\top \boldsymbol{Z}_j' &= \boldsymbol{V}_i^* \boldsymbol{\Sigma}_i^* (\boldsymbol{U}_i')^\top \boldsymbol{U}_j' \boldsymbol{\Sigma}_j^* (\boldsymbol{V}_j^*)^\top \\ &= \boldsymbol{V}_i^* \boldsymbol{\Sigma}_i^* \, \boldsymbol{0} \, \boldsymbol{\Sigma}_j^* (\boldsymbol{V}_j^*)^\top \\ &= \boldsymbol{0}. \end{aligned}$$

That is, the matrices $\boldsymbol{Z}_1', \ldots, \boldsymbol{Z}_K'$ are pairwise orthogonal.

Applying Lemma 3 to $\boldsymbol{Z}'$ yields

$$\begin{aligned} \Delta\mathcal{R}(\boldsymbol{Z}', \alpha) &= \sum_{i=1}^{K} \frac{1}{2m} \log \left(\frac{\det^m \left(\boldsymbol{I} + \frac{d}{m\epsilon^2} \boldsymbol{Z}_i' (\boldsymbol{Z}_i')^\top \right)}{\det^{m_i} \left(\boldsymbol{I} + \frac{d}{m_i\epsilon^2} \boldsymbol{Z}_i' (\boldsymbol{Z}_i')^\top \right)} \right) \\ &= \sum_{i=1}^{K} \frac{1}{2m} \log \left(\frac{\det^m \left(\boldsymbol{I} + \frac{d}{m\epsilon^2} \boldsymbol{Z}_i^* (\boldsymbol{Z}_i^*)^\top \right)}{\det^{m_i} \left(\boldsymbol{I} + \frac{d}{m_i\epsilon^2} \boldsymbol{Z}_i^* (\boldsymbol{Z}_i^*)^\top \right)} \right), \end{aligned} \tag{10}$$

where the second equality follows from Lemma 2.

Comparing equation 9 and equation 10 shows that

$$\Delta\mathcal{R}(\boldsymbol{Z}', \alpha) > \Delta\mathcal{L}(\boldsymbol{Z}^*, \alpha),$$

which contradicts the optimality of $\boldsymbol{Z}^*$. Therefore, we must have

$$(\boldsymbol{Z}_i^*)^\top \boldsymbol{Z}_j^* = \boldsymbol{0} \quad \text{for all } 1 \leq i < j \leq K.$$

Therefore, by Lemma 2 we have

$$\Delta\mathcal{R}(\boldsymbol{Z}^*, \alpha) = \sum_{i=1}^{K} \frac{1}{2m} \log \left(\frac{\det^m \left(\boldsymbol{I} + \frac{d}{m\epsilon^2} \boldsymbol{Z}_i^* (\boldsymbol{Z}_i^*)^\top \right)}{\det^{m_i} \left(\boldsymbol{I} + \frac{d}{m_i\epsilon^2} \boldsymbol{Z}_i^* (\boldsymbol{Z}_i^*)^\top \right)} \right). \tag{11}$$

We now prove the result concerning the singular values of $\boldsymbol{Z}_i^*$. To begin with, we claim that the following result holds:

$$\boldsymbol{Z}_i^* \in \arg\max_{\boldsymbol{Z}_i} \log\left(\frac{\det^m\left(\boldsymbol{I} + \frac{d}{m\epsilon^2}\boldsymbol{Z}_i\boldsymbol{Z}_i^\top\right)}{\det^{m_i}\left(\boldsymbol{I} + \frac{d}{m_i\epsilon^2}\boldsymbol{Z}_i\boldsymbol{Z}_i^\top\right)}\right), \quad \text{s.t.} \quad \|\boldsymbol{Z}_i\|_F^2 = m_i. \tag{12}$$

To see why equation 12 holds, suppose that there exists $\widetilde{\boldsymbol{Z}}_i$ satisfying $\|\widetilde{\boldsymbol{Z}}_i\|_F^2 = m_i$ and $\operatorname{rank}(\widetilde{\boldsymbol{Z}}_i) \leq d_i$, such that

$$\log\left(\frac{\det^m\left(\boldsymbol{I} + \frac{d}{m\epsilon^2}\widetilde{\boldsymbol{Z}}_i\widetilde{\boldsymbol{Z}}_i^\top\right)}{\det^{m_i}\left(\boldsymbol{I} + \frac{d}{m_i\epsilon^2}\widetilde{\boldsymbol{Z}}_i\widetilde{\boldsymbol{Z}}_i^\top\right)}\right) > \log\left(\frac{\det^m\left(\boldsymbol{I} + \frac{d}{m\epsilon^2}\boldsymbol{Z}_i^*(\boldsymbol{Z}_i^*)^\top\right)}{\det^{m_i}\left(\boldsymbol{I} + \frac{d}{m_i\epsilon^2}\boldsymbol{Z}_i^*(\boldsymbol{Z}_i^*)^\top\right)}\right). \tag{13}$$

Denote the compact SVD of $\widetilde{\boldsymbol{Z}}_i$ by

$$\widetilde{\boldsymbol{Z}}_i = \widetilde{\boldsymbol{U}}_i\widetilde{\boldsymbol{\Sigma}}_i\widetilde{\boldsymbol{V}}_i^\top,$$

and define

$$\boldsymbol{Z}' = [\boldsymbol{Z}_1^*, \ldots, \boldsymbol{Z}_{i-1}^*, \boldsymbol{Z}_i', \boldsymbol{Z}_{i+1}^*, \ldots, \boldsymbol{Z}_K^*], \quad \text{where} \quad \boldsymbol{Z}_i' := \boldsymbol{U}_i^*\widetilde{\boldsymbol{\Sigma}}_i\widetilde{\boldsymbol{V}}_i^\top.$$

Note that $\|\boldsymbol{Z}_i'\|_F^2 = m_i$, $\operatorname{rank}(\boldsymbol{Z}_i') \leq d_i$, and $(\boldsymbol{Z}_i')^\top\boldsymbol{Z}_j' = \boldsymbol{0}$ for all $i \neq j$. Hence, $\boldsymbol{Z}'$ is a feasible solution to problem equation 4 with its components being pairwise orthogonal.

By invoking Lemma 3, Lemma 2, and using inequality equation 13, we have

$$\begin{aligned}
&\Delta\mathcal{R}(\boldsymbol{Z}', \alpha) \\
&= \frac{1}{2m}\log\left(\frac{\det^m\left(\boldsymbol{I} + \frac{d}{m\epsilon^2}\boldsymbol{Z}_i'(\boldsymbol{Z}_i')^\top\right)}{\det^{m_i}\left(\boldsymbol{I} + \frac{d}{m_i\epsilon^2}\boldsymbol{Z}_i'(\boldsymbol{Z}_i')^\top\right)}\right) + \sum_{j\neq i}\frac{1}{2m}\log\left(\frac{\det^m\left(\boldsymbol{I} + \frac{d}{m\epsilon^2}\boldsymbol{Z}_j^*(\boldsymbol{Z}_j^*)^\top\right)}{\det^{m_j}\left(\boldsymbol{I} + \frac{d}{m_j\epsilon^2}\boldsymbol{Z}_j^*(\boldsymbol{Z}_j^*)^\top\right)}\right) \\
&= \frac{1}{2m}\log\left(\frac{\det^m\left(\boldsymbol{I} + \frac{d}{m\epsilon^2}\widetilde{\boldsymbol{Z}}_i\widetilde{\boldsymbol{Z}}_i^\top\right)}{\det^{m_i}\left(\boldsymbol{I} + \frac{d}{m_i\epsilon^2}\widetilde{\boldsymbol{Z}}_i\widetilde{\boldsymbol{Z}}_i^\top\right)}\right) + \sum_{j\neq i}\frac{1}{2m}\log\left(\frac{\det^m\left(\boldsymbol{I} + \frac{d}{m\epsilon^2}\boldsymbol{Z}_j^*(\boldsymbol{Z}_j^*)^\top\right)}{\det^{m_j}\left(\boldsymbol{I} + \frac{d}{m_j\epsilon^2}\boldsymbol{Z}_j^*(\boldsymbol{Z}_j^*)^\top\right)}\right) \\
&> \frac{1}{2m}\log\left(\frac{\det^m\left(\boldsymbol{I} + \frac{d}{m\epsilon^2}\boldsymbol{Z}_i^*(\boldsymbol{Z}_i^*)^\top\right)}{\det^{m_i}\left(\boldsymbol{I} + \frac{d}{m_i\epsilon^2}\boldsymbol{Z}_i^*(\boldsymbol{Z}_i^*)^\top\right)}\right) + \sum_{j\neq i}\frac{1}{2m}\log\left(\frac{\det^m\left(\boldsymbol{I} + \frac{d}{m\epsilon^2}\boldsymbol{Z}_j^*(\boldsymbol{Z}_j^*)^\top\right)}{\det^{m_j}\left(\boldsymbol{I} + \frac{d}{m_j\epsilon^2}\boldsymbol{Z}_j^*(\boldsymbol{Z}_j^*)^\top\right)}\right) \\
&= \sum_{i=1}^K\frac{1}{2m}\log\left(\frac{\det^m\left(\boldsymbol{I} + \frac{d}{m\epsilon^2}\boldsymbol{Z}_i^*(\boldsymbol{Z}_i^*)^\top\right)}{\det^{m_i}\left(\boldsymbol{I} + \frac{d}{m_i\epsilon^2}\boldsymbol{Z}_i^*(\boldsymbol{Z}_i^*)^\top\right)}\right).
\end{aligned}$$

Combining it with equation 11 shows $\Delta\mathcal{R}(\boldsymbol{Z}', \alpha) > \Delta\mathcal{R}(\boldsymbol{Z}^*, \alpha)$, contradicting the optimality of $\boldsymbol{Z}^*$. Therefore, the result in equation 12 holds.

Observe that the optimisation problem in equation 12 depends on $\boldsymbol{Z}_i$ only through its singular values. More precisely, let

$$\boldsymbol{\sigma}_i := \left[\sigma_{1,i}, \ldots, \sigma_{\min\{m_i,d\},i}\right]$$

denote the singular values of $\boldsymbol{Z}_i$ and

$$\det\left(\boldsymbol{I} + \gamma\,\boldsymbol{Z}\boldsymbol{Z}^\top\right) = \prod_{p=1}^n\left(1 + \gamma\,\sigma_p^2\right).$$

Then, we have

$$\log\left(\frac{\det^m\left(\boldsymbol{I}+\frac{d}{m\epsilon^2}\boldsymbol{Z}_i\boldsymbol{Z}_i^\top\right)}{\det^{m_i}\left(\boldsymbol{I}+\frac{d}{m_i\epsilon^2}\boldsymbol{Z}_i\boldsymbol{Z}_i^\top\right)}\right)=\sum_{p=1}^{\min\{m_i,d\}}\log\left(\frac{\left(1+\frac{d}{m\epsilon^2}\sigma_{p,i}^2\right)^m}{\left(1+\frac{d}{m_i\epsilon^2}\sigma_{p,i}^2\right)^{m_i}}\right).$$

Moreover, note that

$$\|\boldsymbol{Z}_i\|_F^2=\sum_{p=1}^{\min\{m_i,d\}}\sigma_{p,i}^2.$$

Using these relations, the optimisation problem in equation 12 is equivalent to

$$\begin{aligned}\max_{\boldsymbol{\sigma}_i\in\Re_{\geq 0}^{\min\{m_i,d\}}}\quad&\sum_{p=1}^{\min\{m_i,d\}}\log\left(\frac{\left(1+\frac{d}{m\epsilon^2}\sigma_{p,i}^2\right)^m}{\left(1+\frac{d}{m_i\epsilon^2}\sigma_{p,i}^2\right)^{m_i}}\right)\\ \text{s.t.}\quad&\sum_{p=1}^{\min\{m_i,d\}}\sigma_{p,i}^2=m_i.\end{aligned}\tag{14}$$

Let

$$r_i=\min\{m_i,d\}\quad\text{and}\quad\boldsymbol{\sigma}_i^*=\left[\sigma_{1,i}^*,\ldots,\sigma_{r_i,i}^*\right]$$

be an optimal solution to equation 14. Without loss of generality, assume that the entries of $\boldsymbol{\sigma}_i^*$ are sorted in descending order.

Hence,

$$\begin{aligned}\left[\sigma_{1,i}^*,\ldots,\sigma_{r_i,i}^*\right]=\underset{[\sigma_{1,i},\ldots,\sigma_{r,i}]\in\Re_{\geq 0}^{r_i}}{\arg\max}\quad&\sum_{p=1}^{r_i}\log\left(\frac{\left(1+\frac{d}{m\epsilon^2}\sigma_{p,i}^2\right)^m}{\left(1+\frac{d}{m_i\epsilon^2}\sigma_{p,i}^2\right)^{m_i}}\right)\\ \text{s.t.}\quad&\sum_{p=1}^{r_i}\sigma_{p,i}^2=m_i,\quad\sigma_{1,i}\geq\cdots\geq\sigma_{r,i}\geq 0.\end{aligned}\tag{15}$$

To solve the problem equation 15 we define a new function as

$$\begin{aligned}f(x;d,\alpha,m_i,m)&=\log\left(\frac{\left(1+\frac{d}{m\epsilon^2}x\right)^m}{\left(1+\frac{d}{m_i\epsilon^2}x\right)^{m_i}}\right)\\&=m\log\Big(1+\frac{d}{m\epsilon^2}x\Big)-m_i\log\Big(1+\frac{d}{m_i\epsilon^2}x\Big)\\&\text{s.t.}\quad x\geq 0\end{aligned}$$

We compute the first derivative of f with respect to x, which is given by

$$\begin{aligned}f'(x)&=\frac{d\,m}{m\epsilon^2+d\,x}-\frac{d\,m_i}{m_i\epsilon^2+d\,x}\\&=\frac{d^2\,(m-m_i)x}{\Big(m\epsilon^2+d\,x\Big)\Big(m_i\epsilon^2+d\,x\Big)}.\end{aligned}$$

Clearly $f'(0)=0$.

We compute the second derivative of f with respect to x, which is given by

$$f''(x)=\frac{d^2(m-m_i)\left(mm_i\epsilon^4-d^2x^2\right)}{\left(m\epsilon^2+d\,x\right)^2\left(m_i\epsilon^2+d\,x\right)^2}.$$

Let

$$\boldsymbol{x}_p = \frac{\sqrt{mm_i}\epsilon^2}{d}, \;\; f''(\boldsymbol{x}_p) = 0.$$

The optimisation problem equation 15 satisfies

(i) $f'(x)$ satisfies $f'(0) = 0$, and is positive for $x > 0$;

(ii) There exists some $x_p > 0$ such that $f'(x)$ is strictly increasing on $(0, x_p]$ and strictly decreasing on $[x_p, +\infty)$;

Therefore, we may apply Lemma 5 and conclude that let

$$d_i \leq \min\left\{r_i, \frac{m_i}{x_p}\right\} = \min\left\{r_i, \sqrt{\frac{m_i}{m}}\frac{d}{\epsilon^2}\right\}$$

Hence for simplicity, we have

$$d_i \leq \sqrt{\frac{m_i}{m}}\frac{d}{\epsilon^2}.$$

The unique optimal solution to equation 15 is one of the following two forms:

- $\mathbf{x}^* = \Big[\underbrace{\frac{m_i}{d_i}, \ldots, \frac{m_i}{d_i}}_{d_i}, \underbrace{0, \ldots, 0}_{r_i - d_i}\Big]$,
- $\mathbf{x}^* = \Big[\underbrace{x_H, \ldots, x_H, x_L}_{d_i}, \underbrace{0, \ldots, 0}_{r_i - d_i}\Big]$, for some $x_H \in \left(\frac{m_i}{d_i}, \frac{m_i}{d_i - 1}\right)$ and $x_L > 0$.

In fact, if the optimal solution is the first form, according to Lemma 5, $d_i < \sqrt{\frac{m_i}{m}}\frac{d}{\epsilon^2}$.

Equivalently, we have either

- $[\sigma_{1,i}^*, \ldots, \sigma_{r_i,i}^*] = \Big[\underbrace{\sqrt{\frac{m_i}{d_i}}, \ldots, \sqrt{\frac{m_i}{d_i}}}_{d_i}, \underbrace{0, \ldots, 0}_{r_i - d_i}\Big]$, or
- $[\sigma_{1,i}^*, \ldots, \sigma_{r_i,i}^*] = \Big[\underbrace{\sigma_H, \ldots, \sigma_H, \sigma_L}_{d_i}, \underbrace{0, \ldots, 0}_{r_i - d_i}\Big]$, for some $\sigma_H \in \left(\sqrt{\frac{m_i}{d_i}}, \sqrt{\frac{m_i}{d_i - 1}}\right)$ and $\sigma_L > 0$,

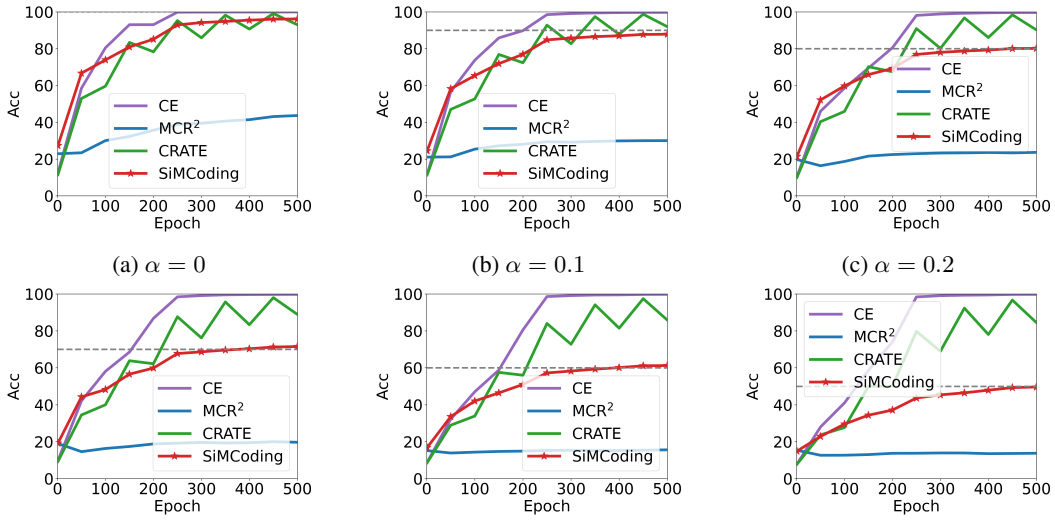
as claimed. □

C Experimental Details

C.1 Image Classification Datasets and Experimental Settings

Performance metric. Traditional MCR^2-based methods typically employ a Nearest Subspace Classifier (NSC), which assigns labels according to the distance between a feature representation and the principal subspace of each class [14, 13]. However, in *SiMCoding*, the dimensionality of each class subspace is intentionally driven toward one, making the predefined subspace dimension required by NSC ill-defined. Therefore, we evaluate both *SiMCoding* and MCR^2 using a simple logistic regression classifier, and report the resulting classification accuracy.

Datasets, networks, and training setup. We conduct experiments on MNIST [31], CIFAR-10, CIFAR-20 [32] (the coarse-label version of CIFAR-100), and ImageNette [33], a 10-class subset of ImageNet. To ensure consistency and comparability, we follow the training hyperparameters used in prior work [14, 13] for MCR^2. For CRATE, we primarily follow the settings in [34]. Unless



(a) $\alpha = 0$ (b) $\alpha = 0.1$ (c) $\alpha = 0.2$

(d) $\alpha = 0.3$ (e) $\alpha = 0.4$ (f) $\alpha = 0.5$

Figure 6: Training accuracy dynamics on CIFAR-20 under different label-noise ratios α. The grey dashed horizontal line indicates the ideal training accuracy $100(1 - \alpha)$.

otherwise specified, all training conditions are identical across methods. CIFAR-20 groups the 100 fine-grained classes of CIFAR-100 into 20 superclasses, each containing 2,500 training images and 500 test images. We apply fivefold data augmentation during training.

- **MNIST.** For MCR^2 and *SiMCoding*, we use a simple convolutional network consisting of two 3×3 convolutional layers (with 32 and 64 channels, each followed by ReLU), a 2×2 max-pooling layer, a fully connected ReLU layer with 1024 units, and a final projection to $d = 64$ dimensions. In all experiments, the learned features are normalised such that $\|\mathbf{Z}_i\|_F^2 = m_i$ for each class. For CE, we add a classification head to the networks used by MCR^2 and *SiMCoding*. For CRATE, we use a patch size of 7×7, model width of 64, depth of 6, and 4 attention heads.
- **CIFAR-10, CIFAR-20 and ImageNette.** For MCR^2 and *SiMCoding*, we employ a ResNet-18 [51] backbone, replacing the final layer with a two-layer MLP with ReLU activation that outputs 512-dimensional representations. For CE, we add a classification head to the networks used by MCR^2 and *SiMCoding*. For CRATE on CIFAR-20 and ImageNette, we use a patch size of 8×8, model width of 512, depth of 8, and 8 attention heads.

C.1.1 Complementary Implementation Details

Architecture. For CRATE on MNIST, we use 28×28 greyscale inputs, 7×7 patches, a model width of 64, six layers, and four attention heads, each with dimension 16. For CIFAR-10 and CIFAR-20, we use 32×32 RGB inputs, 8×8 patches, a model width of 512, eight layers, and eight attention heads, each with dimension 64. ImageNette uses 128×128 RGB inputs with the same patch size, model width, depth, and attention configuration as CIFAR. Model dropout and embedding dropout are both set to zero in all three configurations.

Data Augmentation. We do not apply data augmentation to MNIST. For CIFAR-10 and CIFAR-20, following prior work [14, 13], training images are randomly cropped to 32×32 with eight-pixel padding, randomly flipped horizontally, and converted to tensors.

For ImageNette, training images undergo random resized cropping to 128×128, random horizontal flipping, and conversion to tensors. When Gaussian noise is enabled, zero-mean noise with the specified standard deviation is added before normalisation, with pixel values clipped to $[0, 1]$. RGB channels are normalised using mean $(0.485, 0.456, 0.406)$ and standard deviation $(0.229, 0.224, 0.225)$. At test time, images are resized so that the shorter side is 160 pixels, centre-cropped to 128×128, converted to tensors, and normalised using the same channel statistics.

Training Setting. The following training settings apply to MNIST, CIFAR-10, CIFAR-20, and ImageNette. The task-specific settings for PD Activity and AD-3DMRI are described in Appendix C.2 and C.3, respectively. For CE, MCR^2, and *SiMCoding*, we use the settings below. For MNIST, CIFAR-10, and CIFAR-20, models are trained using stochastic gradient descent (SGD) with an initial learning rate of 0.01, momentum of 0.9, and weight decay of 5×10^{-4}. Training is performed for 500 epochs with a batch size of 1000, and the learning rate is decayed by a factor of 10 every 200 epochs. For ImageNette, due to the larger input resolution (128×128), we reduce the batch size to 400 while keeping the initial learning rate at 0.01.

For CRATE on these four image datasets, we adapt the training strategy described in [34]. The batch size is kept the same as in the other methods for each dataset to ensure a fair comparison. Training is performed using the AdamW optimiser with an initial learning rate of $\eta_0 = 4 \times 10^{-4}$, momentum parameters $\beta = (0.9,\ 0.999)$, and a weight decay of 0.05. We adopt the label smoothing cross-entropy loss with a smoothing parameter of $\alpha = 0.1$. The models are trained for a total of 500 epochs. We use a per-iteration `LambdaLR` scheduler with a linear warmup of 100 steps followed by cosine annealing. For global optimiser step t (starting at 0) and total epochs T, the learning-rate multiplier $\lambda(t)$ is defined as

$$\lambda(t) \;=\; \min\!\left(\frac{t+1}{100+\varepsilon},\ \tfrac{1}{2}\big(1+\cos(\pi\, t/\max(1,T))\big)\right),$$

so that the actual learning rate is $\eta(t) = \eta_0\, \lambda(t)$. The scheduler is updated at every training iteration.

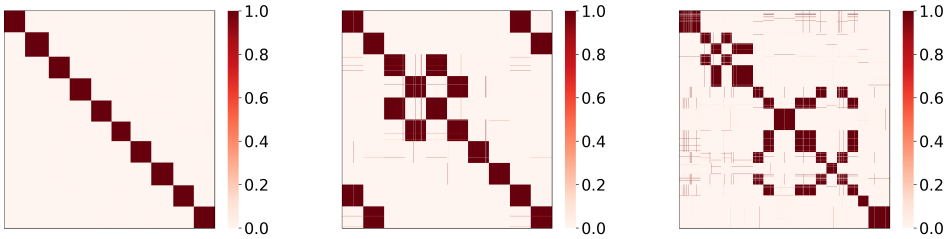
In terms of coding precision ϵ, for MCR^2, we follow [14, 13] and set $\epsilon^2 = 0.5$ uniformly across these four image datasets. For *SiMCoding*, the coding precision is selected according to our theoretical guideline: for MNIST, $\epsilon^2 = 6.4$; for CIFAR-10, $\epsilon^2 = 53.2$; for CIFAR-20, $\epsilon^2 = 38.2$; for ImageNette, $\epsilon^2 = 53.2$.

C.1.2 Saliency and Attention Map Settings

Figure 3 examines digit 0 under the clean MNIST setting. We use the trained CE, MCR^2, *SiMCoding*, and CRATE models described above, with all models in evaluation mode and no data augmentation during visualisation.

For CE, we compute the absolute input gradient of the predicted class logit. Specifically, if $s_c(\boldsymbol{x})$ is the logit for class c and $\widehat{y} = \arg\max_c s_c(\boldsymbol{x})$, the saliency map is $|\nabla_{\boldsymbol{x}} s_{\widehat{y}}(\boldsymbol{x})|$. For MCR^2 and *SiMCoding*, we compute the absolute input gradient of each representation coordinate separately and sum these responses across coordinates: $S(\boldsymbol{x}) = \sum_{j=1}^{d} |\nabla_{\boldsymbol{x}} f_j(\boldsymbol{x})|$. We then average the maps over 50 randomly selected images within each class and min–max normalise the resulting class-level maps to $[0, 1]$ for display. These gradient-based panels use training images.

For CRATE, we extract the final-layer attention from the classification token to the image patches and average it across attention heads. With 7×7 patches on a 28×28 MNIST image, the resulting 4×4 attention grid is min–max normalised, bilinearly upsampled to the input resolution, and overlaid on a representative digit-0 image from the test split. Here, averaging refers to attention heads, rather than multiple images. Attention and input-gradient maps use different mechanisms and are interpreted qualitatively.



(a) MNIST (b) CIFAR-10 (c) CIFAR-20

Figure 7: Heatmaps of $|\mathbf{Z}^\top \mathbf{Z}|$ across datasets, with samples sorted by class. The model is *SiMCoding*-T: *SiMCoding* with $\epsilon = \epsilon_U$

C.2 PD Activity Dataset and Experimental Settings

- **PD Activity [17]:** This dataset comprises up to 135 patients, each performing up to 16 distinct activities, including Finger Tapping, Fist Open–Close, Pronation–Supination, Right/Left Hand Flip, Finger-to-Nose (left/right), Front Arm Raise, Walking, Arising from a Chair, Drinking, Object Pickup, Sitting, Standing, Stand–Swing Arms, and Drawing on Paper. These activities are indexed as Session 1 through Session 16.

  For each activity, Shimmer wearable sensors are placed on both wrists, providing 12-channel signals per hand (accelerometer, gyroscope, and magnetometer). After preprocessing, the 16 activity sequences are concatenated along the feature axis, yielding a unified representation of size 600×384 (16×24 channels) for each patient.

  The dataset includes four classes—Normal, Mild, Moderate, and Severe—corresponding to different levels of Parkinson's disease severity. Following [17], we define four classification tasks: *PD-Binary*, distinguishing Normal from PD; *PD-Early*, distinguishing Normal from Mild; *PD-Severity*, classifying Mild, Moderate, and Severe cases; and *PD-Fine-Grained*, classifying all four severity levels.

Data partitioning and evaluation. We divide the PD Activity dataset into training and test sets in an 80:20 ratio. Minority-class oversampling is applied only to the training set; evaluation uses the original, imbalanced test set.

Networks and training setup for PD Activity dataset.

- **On PD Activity.** We use a lightweight transformer-based backbone. Each input sequence is divided into 25 temporal patches, where each patch consists of 24×384 features and is linearly projected into a $d_{\text{model}} = 128$-dimensional embedding. A learnable positional encoding is added to all patches, which are then processed by a single-layer Transformer encoder with 4 attention heads and a feed-forward dimension of 64. For MCR^2 and *SiMCoding*, the encoded patch representations are averaged to obtain a global feature vector, followed by a small projection head and ℓ_2 normalisation to produce the final embedding. For CE, we use the same transformer backbone but append an additional classification head consisting of a projection layer and a final fully connected layer that outputs the class logits. For CRATE, we set the patch size to 24×384, with a transformer depth of 12, four attention heads, and a hidden dimension of 512.

  We report the mean performance and standard deviation over 25 independent runs. All reported balanced accuracies are computed on the original, imbalanced test sets.

C.3 AD-3DMRI Dataset and Experimental Settings

Dataset and subject partitions. We use one MRI volume per subject for binary AD-versus-CN classification [35]. ADNI-1 contains 374 subjects (204 CN and 170 AD), and ADNI-2 contains 386 subjects (248 CN and 138 AD). For each training direction, the training cohort is split into stratified 80%/20% training and internal test sets at the subject level: ADNI-1 provides 299 training and 75 internal test subjects, while ADNI-2 provides 308 training and 78 internal test subjects. Training

and internal test subjects are disjoint. The other cohort is used in full for external testing. Subject partitions are fixed across the five training repetitions and shared by all methods within each repetition. For class balancing, minority classes are randomly oversampled in the training set to match the largest class; the same rule is applied to training features when fitting the representation classifier. Internal and external test distributions are preserved.

MRI preprocessing and augmentation. Foreground intensities in each volume are clipped at the 1st and 99th percentiles and z-score normalised, with background voxels kept at zero. Volumes are resampled to $96 \times 112 \times 96$ by trilinear interpolation. Training augmentation uses translations of up to three voxels, intensity scaling in $[0.9, 1.1]$, and foreground Gaussian noise with standard deviation 0.02. No anatomical flips are applied.

Architecture and training. All three methods use a randomly initialised 3D ResNet-18, with convolution and pooling operations adapted to volumetric inputs. Global average pooling produces a 512-dimensional feature, followed by a linear $512 \to 128$ projection. MCR^2 and *SiMCoding* normalise the projected representations, while CE jointly trains an additional binary classification head. Models are trained using AdamW for 200 epochs with batch size 32, initial learning rate 3×10^{-4}, weight decay 10^{-4}, cosine annealing to 10^{-6}, and gradient norm clipping at 10. MCR^2 uses $\epsilon^2 = 0.5$, and *SiMCoding* uses $\epsilon^2 = 128/(3\sqrt{2})$.

C.4 Incremental Learning Settings

Tasks and evaluation. The ten classes of MNIST and ImageNette are divided into five sequential tasks, with class-index pairs $(0,1), (2,3), (4,5), (6,7), (8,9)$. At stage t, the number of seen classes is $K_t = 2t$. Evaluation uses all test or validation examples from the seen classes: 2,115, 4,157, 6,031, 8,017, and 10,000 MNIST test examples, and 782, 1,525, 2,328, 3,136, and 3,925 ImageNette validation examples. Stage accuracy A_t is the percentage of these examples classified correctly. All MNIST methods restrict predictions to seen classes. On ImageNette, SI and the ReduNet-based methods restrict predictions to seen classes, while the saved oEWC and DER runs retain their ten-class output heads; only seen-class examples contribute to accuracy. We report

$$\text{Avg. Acc.} = \frac{1}{5}\sum_{t=1}^{5} A_t, \qquad \text{Drop} = 100\frac{A_1 - A_5}{A_1}. \tag{16}$$

Drop is retained as a measure of relative overall accuracy decline. Because both the evaluation set and classification problem expand, it does not isolate forgetting on a fixed task. Gain is SiM-ReduNet's accuracy or mean minus the largest baseline value; for Drop, it is the smallest baseline Drop minus SiM-ReduNet's Drop. Thus, positive Gain always denotes an improvement. All means, drops, and gains are calculated before rounding.

MNIST networks and dynamic coding precision. Both ReduNet-based models use 200 layers, following the network and incremental covariance-update settings of [36]. The step size is $\eta_\ell = 0.5 \times 0.933^\ell$ for $\ell = 0, \ldots, 199$, the confidence parameter is $\lambda = 1$, and evaluation uses a nearest-subspace classifier with 28 principal components per class. Inputs have dimension $d = 28 \times 28 = 784$, without mean subtraction. Training representations use class-wise Frobenius normalisation so that squared norm equals class sample count; inference uses per-example ℓ_2 normalisation. Historical input second moments and class counts are retained for incremental updates, without replaying historical raw training examples. Both runs use float64 and seed 42. ReduNet uses $\epsilon^2 = 0.5$. SiM-ReduNet recomputes

$$\epsilon_t^2 = \frac{d}{3\sqrt{K_t}} \tag{17}$$

at the start of each stage and uses this value throughout all layers at that stage. For MNIST, the five values are 184.79, 130.67, 106.69, 92.40, and 82.64 for $K_t = 2, 4, 6, 8, 10$, respectively.

ImageNette networks and classifier. We retain the original ImageNette implementation: images are resized to $32 \times 32 \times 3$ ($d = 3072$), with no random Gaussian filters or mean subtraction. Both networks use 50 layers, $\lambda = 500$, per-example ℓ_2 normalisation, float32, and seed 42. ReduNet uses $\epsilon^2 = 0.5$ and constant step size $\eta = 0.5$. SiM-ReduNet retains constant $\eta = 10$ and uses the same

dynamic rule, giving $\epsilon_t^2 = 724.08, 512.00, 418.05, 362.04, 323.82$. The original accumulation of per-layer statistics and soft-assignment computation are retained. After each stage, both networks refit an ℓ_2-regularised multinomial logistic regression classifier on the current representations of all seen training examples, including earlier tasks. The classifier uses $C = 1$, L-BFGS, a maximum of 1,000 iterations, and tolerance 10^{-4}. Thus, the ImageNette comparison retains historical training access for classifier fitting and different network step sizes; it evaluates the complete protocols rather than isolating coding precision.

Baselines. We compare with oEWC [37], SI [38], and DER [39]. On MNIST, each is trained for five epochs per task. Online EWC uses a two-block CNN on 28×28 inputs, AdamW with learning rate 10^{-3} and weight decay 10^{-4}, batch size 256, and seed 0. Its empirical Fisher uses 1,024 current-task examples, with penalty coefficient 1,000 and accumulation factor 1. SI uses its 28×28 CNN, Adam with learning rate 10^{-3}, batch size 128, seed 42, $c = 0.1$, and $\xi = 10^{-3}$. DER uses a small CNN with 32×32 inputs, Adam with learning rate 10^{-3}, batch size 128, seed 0, and logit-matching coefficient 1. On ImageNette, oEWC retains its progress-and-compress setup with five epochs in each phase, learning rates 10^{-3} and 10^{-4}, batch size 128, seed 0, distillation temperature 2, and Fisher accumulation factor 0.95. SI uses a small CNN, five epochs per task, Adam with learning rate 10^{-3}, batch size 128, seed 42, $c = 0.1$, and $\xi = 10^{-3}$. DER uses a ResNet-18 trained from scratch for ten epochs per task, AdamW with learning rate 10^{-3} and weight decay 10^{-4}, batch size 128, seed 42, and logit-matching coefficient 0.5. DER retains 200 replay examples on both datasets. Results are from single-seed runs with the stated settings.

Accuracy and Drop across the sequence. On MNIST, SiM-ReduNet attains a five-stage mean of 98.00%, versus 98.05% for ReduNet and 95.17% for DER. Both ReduNet-based models end at 95.82% and have the smallest Drop (4.04%). On ImageNette, SiM-ReduNet has the highest mean (53.17%), followed by DER (48.52%) and ReduNet (33.20%). Its mean Gain is therefore 4.65 percentage points among the included baselines. ReduNet has the smallest relative Drop (41.10%), while SiM-ReduNet's Drop is 50.29%, giving a Drop Gain of -9.18 percentage points calculated from unrounded drops. The higher SiM-ReduNet accuracies at all five stages coexist with a larger relative decline from its substantially higher initial accuracy. We therefore interpret these results as stronger overall ImageNette classification performance, without attributing the difference in Drop solely to forgetting.

D Potential Social Impact

The proposed *SiMCoding* framework may provide positive social impact in applications that require robust and interpretable machine learning systems, particularly in high-stakes domains such as healthcare, autonomous systems, and human activity analysis. By learning compact and discriminative representations with improved robustness to noisy labels and uncertain environments, *SiMCoding* may help improve the reliability and stability of machine learning models deployed in real-world settings. In particular, its strong performance on physiological sensing and neuroimaging tasks suggests potential value for assisting clinical decision-making and health monitoring systems.

At the same time, the proposed method may also introduce negative societal risks if deployed irresponsibly. Although *SiMCoding* improves robustness, it can still inherit biases, annotation errors, or distributional imbalances present in the training data. In healthcare-related applications, incorrect predictions or biased representations may adversely affect underrepresented patient groups or lead to inappropriate clinical decisions if the system is over-relied upon without human oversight. Furthermore, like other representation learning methods, *SiMCoding* could potentially be used in surveillance or behavioural monitoring systems, raising concerns regarding privacy and misuse. Therefore, careful dataset curation, fairness evaluation, privacy protection, and human supervision remain important for responsible deployment.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract and introduction accurately reflect the main contributions and scope of the paper. In particular, the theoretical analysis and experimental results consistently support the proposed minimal-complexity representation learning principle and the effectiveness of *SiMCoding*.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the main limitation of *SiMCoding*, namely that its advantage may diminish when the data itself is ow-dimensional or do not exhibit clear structure.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Please see the Section 2 and the proof in the Appendix.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Code is available at `https://github.com/menghui-zhou/SiMCoding`.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code is available at `https://github.com/menghui-zhou/SiMCoding`.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimiser) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide all necessary experimental details, including those mentioned in this question.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report the mean performance and standard deviation over 25 independent runs for PD Activity and five runs for AD-3DMRI. Table 3 also reports p-values from two-sided tests.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide all necessary details to reproduce the experimental results, along with the implementation code.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in this paper conforms, in all respects, to the NeurIPS Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In Section D, we briefly discuss the potential positive and negative societal impacts of this work.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our paper poses no such risks.

12. **Licences for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the licence and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We only use publicly accessible research datasets and cite the relevant papers.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The released implementation is documented in the GitHub repository's README, which describes installation, dataset preparation, experiment commands, and output files. The code is released under the MIT License.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigour, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in our paper does not involve LLMs as any important, original, or non-standard components.